SAPIENZA
UNIVERSITÀ DI ROMA

Learning of a multilingual bitaxonomy of Wikipedia and its application to semantic predicates

Department of Computer Science

Dottorato di Ricerca in Informatica – XXVII Ciclo

Candidate

Tiziano Flati
ID number 1143472

Thesis Advisor

Prof. Roberto Navigli

A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer science

June 2015

Thesis not yet defended

Learning of a multilingual bitaxonomy of Wikipedia and its application to semantic predicates

Ph.D. thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Version: July 13, 2016

Author's email: flati@di.uniroma1.it

Abstract

The ability to extract hypernymy information on a large scale is becoming increasingly important in natural language processing, an area of the artificial intelligence which deals with the processing and understanding of natural language. While initial studies extracted this type of information from textual corpora by means of lexico-syntactic patterns, over time researchers moved to alternative, more structured sources of knowledge, such as Wikipedia. After the first attempts to extract is-a information from Wikipedia categories, a full line of research gave birth to numerous knowledge bases containing information which, however, is either incomplete or irremediably bound to English.

To this end we put forward MultiWiBi, the first approach to the construction of a multilingual bitaxonomy which exploits the inner connection between Wikipedia pages and Wikipedia categories to induce a wide-coverage and fine-grained integrated taxonomy. A series of experiments show state-of-the-art results against all the available taxonomic resources available in the literature, also with respect to two novel measures of comparison.

Another dimension where existing resources usually fall short is their degree of multilingualism. While knowledge is typically language agnostic, currently resources are able to extract relevant information only in languages providing high-quality tools. In contrast, MultiWiBi does not leave any language behind: we show how to taxonomize Wikipedia in an arbitrary language and in a way that is fully independent of additional resources. At the core of our approach lies, in fact, the idea that the English version of Wikipedia can be linguistically exploited as a pivot to project the taxonomic information extracted from English to any other Wikipedia language in order to have a bitaxonomy in a second, arbitrary language; as a result, not only concepts which have an English equivalent are covered, but also those concepts which are not lexicalized in the source language.

We also present the impact of having the taxonomized encyclopedic knowledge offered by MultiWiBi embedded into a semantic model of predicates (SPred) which crucially leverages Wikipedia to generalize collections of related noun phrases to infer a probability distribution over expected semantic classes. We applied SPred to a word sense disambiguation task and show that, when MultiWiBi is plugged in to replace an internal component, SPred's generalization power increases as well as its precision and recall.

Finally, we also published MultiWiBi as linked data, a paradigm which fosters interoperability and interconnection among resources and tools through the publication of data on the Web, and developed a public interface which lets the users navigate through MultiWiBi's taxonomic structure in a graphical, captivating manner.

Acknowledgments

I would like first to thank my supervisor Roberto Navigli for giving me the opportunity of attending a Ph.D. course. I owe him all the pragmatism I have now and the ability to face apparently insurmountable problems in a matter of an hour. I have also learnt how to manufacture a lot of 'buttons'.

Despite what people say, the Ph.D. has undoubtedly been what I keep calling 'the most demanding and unstable part of my life'. *After having been put through the wringer, I have now learnt what it means to have faith and stay strong. How do they say? What doesn't kill you makes you stronger.*

Thanks God, though, I haven't been alone during this phase of my life.

Thanks to Elena who supported me and bore all my complaints in these last four years. She had the strength to put herself aside and in her silent patience lies the true meaning of friendship.

I think I could never express enough gratitude to Federico. He has been supporting me through all these years with endless patience and he had the strength to stand by my side, despite all the difficulties we have encountered. He rescued me, reminded me who I really am and who I want to be. Thanks to him I managed to focus on what I expect from the rest of my life. I've found in him a precious and faithful companion which knew how to make me feel never alone. He believed in me and in my promises of change; today we are still together, survivors.

My most precious and biggest thanks, though, goes to my family which has continuously supported me and stood by my side by accepting this crazy choice of mine. They suggested me the best way to proceed through this not-so-clear path; I could have never made it without the constant care of my mother, my father, my brothers and sisters. Thanks to all of you. I love you.

Contents

Chapter 1

Introduction

Artificial intelligence (AI) is that branch of computer science which studies and designs intelligent systems. Whereas a general requirement for such systems is the ability to understand and interact with the surrounding environment, of particular interest to AI has always been the interaction between machines and humans. Differently from computers, though, which are instructed with formal languages and whose dialogue is regulated by formal protocols, humans have developed a spontaneous form of interaction called *natural language*, through which they spontaneously express their thoughts, exchange ideas, emotions, information and so on. Machines, however, are not able to *understand* and *produce* natural language autonomously, like humans do, but necessitate the aid of automatically or manually provided rules which embed human mechanisms into their mechanical brain to drive their linguistic choices. Algorithmic approaches are rather required to effectively overcome language modelling problems and minimize human intervention at the same time. Modelling natural language, in fact, is of an overwhelming complexity and easily ends up coping with the most disparate problems, ranging from morphology to phonology, from syntax to semantics. Making even a simple linguistic phenomenon understandable to a machine can at times translate into a very difficult problem, either from a modelling or a computational point of view. To this end, an area of AI arose, called Natural language processing (NLP), with a special focus on the linguistic aspect of the human-computer interaction and the aim of understanding and generating natural language by means of its automatic processing.

One of the simplest, yet useful, human abilities usually developed around the age of 11, during the formal operational stage [Inhelder and Piaget, 1958] – and also reflected in natural language – is that of generalization. Thanks to this cognitive ability, humans are able to reason in terms of abstract concepts, by combining and classifying items in a more sophisticated way, establishing *hypernymy relations* between concrete objects and their generalizations (e.g., thinking about a building in its general sense, be it a church, a house or a school). This type of relation is unconsciously used everyday; for example, when we read "Google and IBM today announced an initiative to promote new software development methods which will help students and researchers", we all know that Google and IBM refer to two big

industry players and, while reading, we silently generalise the two concepts to the same entity (most probably 'company'). Humans are able to effectively and quickly do this thanks not only to a wide, shared background knowledge but also to the wise and complex application of several linguistic mechanisms, including concept generalisation, disambiguation of words based on their context, etc., all tools which machines do not master yet.

The interest in this relation type relies on the fact that it enables hypothetical and deductive reasoning in those who possess it, including computers [Wos et al., 1984, Heit, 2000, Robinson and Voronkov, 2001, Sutcliffe et al., 2010]. For instance, consider the following sentence and question:

> **Fact:** *Renzo Piano* designed the *Shard London Bridge* as a spire-like sculpture emerging from the River Thames.
>
> **Question:** Which *architect* designed the *Shard London Bridge*?

Despite the simplicity of the example, machines are not able to answer this question on their own. Injecting into automatic systems the piece of knowledge that 'Renzo Piano is an architect' would instead enable them to apply a form of deductive reasoning, thanks to which the question above would be correctly answered. For this reason, modelling, discovering and extracting hypernymy relations automatically quickly turned out to be a task of primary interest and importance in NLP and is also the core problem addressed in this thesis.

The goal of the task of hypernymy extraction is to build a semantic taxonomy (from Greek τάξις *taxis*, meaning 'order' or 'arrangement' and νόμος *nomos*, meaning 'law' or 'science') defined as 'a particular system of classifying things',[1] where objects are organized in a structured, tree-like manner. Connections in a taxonomy, also called *is-a* relations, link objects of the taxonomy to their most suitable generalization(s) and an object which is in is-a relation with another is said to be the *hyponym* of the latter (or equivalently, the latter is said to be the *hypernym* of the former). An example of is-a relation is *(singer, person)*, which encodes the fact that a singer *is-a* person.

One of the most fruitful and long-standing human efforts to build a taxonomy in a manual manner so far is certainly WordNet [Fellbaum, 1998], an English lexical database encoding meanings for more than 150 thousand dictionary words, including nouns, verbs, adjectives and adverbs. Besides being a dictionary, WordNet contains semantic relationships between concepts (e.g., meronymy, antonymy, troponymy, etc.), including hypernymy information. Being a resource curated by experts, however, WordNet is difficult to maintain updated over time and its dictionary nature prevents its application to general purpose tasks [Pennacchiotti and Pantel, 2006, Hovy et al., 2009]: for instance, WordNet encodes *Titanic* only as an adjective,

[1] `http://www.oxfordlearnersdictionaries.com/definition/english/taxonomy`

defined as 'of great force or power', while nothing is said about the 1997 movie. Real-world applications, instead, usually need wide-coverage resources which encode and provide hypernyms also for concepts not found in ordinary dictionaries. Another drawback of WordNet is its difficult extension to other languages. There have been recent efforts to build wordnets in other languages [Fišer, 2008, Hitoshi Isahara and Kanzaki, 2008, de Melo and Weikum, 2008a,b, 2009, Montazery and Faili, 2010], including the MultiWordNet project[2] [Pianta et al., 2002], the Open Multilingual WordNets[3] [Bond and Foster, 2013], EuroWordNet[4] [Vossen, 1998] and BabelNet [Navigli and Ponzetto, 2012b]; nonetheless, all these wordnets vary greatly in size and accuracy, have different formats or licenses and the result is often a mere subset of the English version. Moreover, either are they automatically or semi-automatically constructed, possibly achieving suboptimal quality, or are still burdened by the huge cost in terms of human effort required.

The challenge is then to design an automatic intelligent system which is able to either enhance existing taxonomies (e.g., by adding is-a relations to WordNet) or build a taxonomy from scratch (i.e., starting with no previous taxonomic information explicitly available). Furthermore, such an intelligent system should satisfy several *desiderata*: i) first of all, we would like to harvest is-a relations for the biggest number of entities, with the possibility to adapt the algorithm to any domain; ii) we require the information to be correct, i.e., the machine should cover as many entities as possible without committing too many errors; iii) we would like the machine not to be bound to a language in particular, but to work effectively on as many languages as possible; finally iv) the information should be derived in the most automated possible manner, i.e., with no or minimal human intervention during the process.

Of course, over the past decades many researchers tried to fulfill as many requirements as possible while maximising the benefits of each dimension at the same time, and soon this task turned out to be one of the most productive research directions, spawning two main research branches (for an exhaustive survey see [Gómez-Pérez et al., 2003]). A first group includes works which perform taxonomy learning by extracting information from free text. The input of this class of works is simple text, including either textual corpora of usually some millions of words [Caraballo, 1999, Navigli and Velardi, 2004, Roller et al., 2014] or, potentially, the whole Web [Sánchez and Moreno, 2005, Kozareva and Hovy, 2010a, Velardi et al., 2013]; the output is, instead, a taxonomy over words (e.g., *(singer, person), (dog, animal)*, etc). The construction is usually performed in two stages: extraction of terminology and harvesting of the taxonomic relations between the extracted terms. A common assumption made by these approaches, however, is that the set of words is given as input, either by means of seeds to the system [Kozareva and Hovy, 2010a] or by means of an existing taxonomy [Widdows, 2003, Snow et al., 2006], in which case the task thus reduces to expanding the initial taxonomy. This is due to the fact

[2] `http://multiwordnet.fbk.eu/english/home.php`
[3] `http://compling.hss.ntu.edu.sg/omw/`
[4] `http://www.illc.uva.nl/EuroWordNet/`

that very often the task of taxonomy learning originates within applications whose domain is very clear (e.g., healthcare, education, agronomy, etc.), and the taxonomy to be built is not required to encompass the whole human knowledge. On the other hand, however, the problem of how to build a general taxonomy from scratch (with no set of words given in input) remains an open problem. The second phase, instead, that is the extraction of hyponym/hypernym pairs, has been guided by the pioneering works of Hearst [1992], the first to apply lexical patterns (e.g., '* *such as* x *and* y') to retrieve taxonomic information. For example, by applying the previous pattern to the following sentence:

I have traveled to many countries *such as* India *and* Thailand.

it is possible to derive that both India and Thailand are, in fact, instances of *countries* (i.e., *India* is-a *country* and *Thailand* is-a *country*). Methods based on such patterns, though, are affected by coverage problems (due to the sparsity of such patterns throughout text) and strongly depend on the availability, the type (newspaper, gazetteers, email, audio transcriptions, etc.), and the domain (finance, music, animals, etc.) of the textual corpus used as input.

An alternative is offered by a second group of works which, instead of relying on big textual corpora, exploit the biggest collaborative multilingual encyclopedia available, namely Wikipedia. The main advantage of relying on Wikipedia is that it gives the opportunity of achieving unprecedented coverage, being the most comprehensive project currently available encompassing millions of concepts. Not only does Wikipedia provide articles encoding both named entities (e.g., *Barack Obama*, *Apple Inc.*, etc.) and dictionary words (e.g., *person*, *aircraft*, etc.) but extraordinarily offers full-fledged information attached to (almost) each article in the form of textual definitions, images, sounds, hyperlinks to other articles, etc. A critical point is that, in terms of exploitation of human effort, using Wikipedia is not significantly different from relying on hand-crafted lexicons such as WordNet, since Wikipedia has been equally built by human collaborators. However, there exists a remarkable difference between relying on WordNet and relying on Wikipedia. The type of human effort involved in the construction of the two resources is, in fact, very different: in the case of WordNet, the experience of small number of extremely expert lexicographers has been concentrated in building what is mostly a static resource whose aim is *exactly* that of taxonomizing the concepts found in a dictionary; so, deciding to leverage WordNet for building a taxonomy is indeed a crucial design decision which calls into question the real contribution of an approach. The human contribution offered to Wikipedia, instead, is *independent* of the task of building a taxonomy and goes rather in the direction of aggregating the human knowledge into the most comprehensive shared encyclopedia; so the fact that a system manages to build a taxonomy by automatically extracting valuable information from Wikipedia is only another manifestation of the reuse of Wikipedians' human effort, which is offered irrespective of its future potential applications.

The adoption of Wikipedia as the source knowledge repository has triggered a full line of research [Ponzetto and Strube, 2007, Ponzetto and Navigli, 2009, de Melo and Weikum, 2010a, Hoffart et al., 2013, Nastase and Strube, 2013, Kliegr et al., 2014], at the end of which there is MultiWiBi, the main contribution presented in this thesis. These methods often belong to a bigger picture, namely Open Information Extraction (OIE), in which the goal is that of harvesting any type of relationship existing between any two entities [Banko et al., 2007, Fader et al., 2011, Hoffart et al., 2013, Nastase and Strube, 2013]. Sometimes, though, approaches are more taxonomic-centric, meaning that is-a relation represents the only type of relation they focus on [Ponzetto and Strube, 2007, de Melo and Weikum, 2010a, Kliegr et al., 2014, Flati et al., 2014]. MultiWiBi, the most recent effort which automatically structures Wikipedia into a very large-scale multilingual taxonomy, finally brings in several novelties which overcome the limitations affecting the existing alternative approaches. These resources, in fact, either suffer from coverage issues, covering only a fraction of the whole Wikipedia, or exhibit quality issues regarding, mainly, the specificity of the hypernyms returned (e.g., returning PERSON as hypernym of FRANK SINATRA, instead of the more appropriate SINGER): a thorough comparison of these systems according to the above dimensions will be provided in Chapter 2. The advantages produced by MultiWiBi are three-fold:

- First, it taxonomizes both Wikipedia pages and Wikipedia categories for the first time, producing what has been called a *bitaxonomy*; as a result, hypernyms in the two taxonomies are aligned (e.g., the Wikipedia page SINGER is aligned to the Wikipedia category SINGERS). Some alternatives [de Melo and Weikum, 2010a] have taxonomized the two sides of Wikipedia, but some nodes still do not contain any Wikipedia page or category;

- Second, it provides state-of-the-art results when compared against all the existing competitors, also when considering two innovative measures which assess new quality dimensions;

- Third, thanks to its independence from external linguistic tools and resources, MultiWiBi is applicable to any Wikipedia language; innovatively, it is now possible to cover also those concepts present in a certain language but not encoded in the English Wikipedia, a truly ground-breaking added value. For instance, NEBBIONE is a famous Italian wine which MultiWiBi correctly types as VINO (WINE, in English); however, NEBBIONE is encoded in the Italian Wikipedia but not yet in the English Wikipedia.

The availability of intelligent systems able to generalize millions of concepts not only enables a whole palette of tasks in NLP, including question answering ([Moldovan and Novischi, 2002, Cui et al., 2007, Ferrucci et al., 2010]), word sense disambiguation ([Navigli, 2009, 2012]) and entity linking ([Lin et al., 2012]) but also boosts linguistic technologies which need (or have at their core) semantics in the broader sense (e.g., syntactic and semantic parsing). A case study in this

direction is SPred [Flati and Navigli, 2013] (explained in detail in Chapter 3), a system which harvests semantic predicates from textual corpora. The aim of SPred is to learn automatically a distribution of concepts representing the expected classes associated with an arbitrary sequence of text; for example, given the expression '*a cup of*', SPred collects all the possible noun phrases following this expression (e.g., *coffee*, *tea*, *milk*, etc.) and generalizes them to a common, abstract representation in WordNet (e.g., the concept of $beverage_1^n$). To do this, SPred crucially exploits, among others, hypernymy information extracted from Wikipedia, but currently SPred relies on suboptimal mechanisms affected by limited coverage. We will then show that the integration of MultiWiBi into SPred's internal mechanisms will improve the coverage of the semantified arguments, leading potentially to the acquisition of better probability estimates of the expected classes (see Chapter 4 for a detailed investigation).

Finally, in order to foster the interoperability across those tools and systems which might internally reuse the output of MultiWiBi, we published MultiWiBi also to the linguistic linked open data cloud (LLOD), a linguistic database in which resources are provided online, linked on the Web and freely available [Chiarcos et al., 2011]. In Chapter 5 we will describe how we published MultiWiBi as linked data and will also present our graphic Web interface which enables users to navigate through MultiWiBi in a visual, captivating manner. Notably, the English bitaxonomy of MultiWiBi has also been seamlessly integrated into BabelNet 3.0, a large multilingual semantic network: as a result, BabelNet now includes is-a relations which come not only from WordNet, which ensures that dictionary concepts are covered by hand-crafted hypernyms, but also from MultiWiBi, which brings in millions of novel hypernymy information.

1.1 Objectives

The main objectives of this thesis are:

- To analyse, compare and surpass the current status of taxonomic resources available in the literature building on Wikipedia. Given the current limitations affecting such resources, we intend to build a taxonomy which overcomes as many impediments as possible.

- To develop an approach which is independent of the language to which it is applied. While many of the currently available multilingual taxonomic resources simply exploit Wikipedia interlanguage links, we will have special consideration for language-specific concepts which are simply left out for the most part by all the alternative approaches.

- To measure the potential impact that the final taxonomy has on semantically-driven models, when integrated into the latter.

- To make such resource available and exploitable by the research community, in a standard and interoperable manner.

1.2 Contributions

We provided significant contributions to each of the above objectives:

- **A ground-breaking Wikipedia bitaxonomy.** We show how to build an innovative bitaxonomy by leveraging Wikipedia only, with state-of-the-art results against all the available taxonomic resources available in the literature. We also propose an exhaustive wrap-up comparison which groups the resources according to different dimensions;

- **Overcoming language barriers thanks to Wikipedia.** We show how to cope with the language barriers which so far seemingly impeded the construction of a truly multilingual taxonomy on a large scale;

- **Impact on relevant semantic models and their application.** We present the impact of having an encyclopedic taxonomy integrated into SPred, a semantic model which crucially relies on Wikipedia. We will show that overcoming coverage limitations can be as simple as replacing one of the model's internal mechanisms with the English page taxonomy provided by MultiWiBi;

- **Embedding of the bitaxonomies into the linguistic linked data cloud.** We show how to embed the bitaxonomies into the linguistic linked data cloud and discuss the potential impact of such integration on relevant linguistic fields.

1.3 Published material

The part of Chapter 2 presenting the construction of the English bitaxonomy was published in its early stage in [Flati et al., 2014] and part of Chapter 3 was published in [Flati and Navigli, 2013], in the proceedings of the 52nd and the 51st Annual Meeting of the Association for the Computational Linguistics (ACL), respectively. Part of Chapter 5 was published in the proceedings of International Semantic Web Conference (ISWC) conference [Flati and Navigli, 2014a] and as a poster in the SEMANTiCS '14 conference [Flati and Navigli, 2014b] (also winning the best poster award). Finally, part of Chapter 2 has been submitted to the Artificial Intelligence Journal and is currently under review.

The rest of the thesis contains novel, unpublished material which has been added to improve the clarity of the presentation and strengthen the relationships between the parts.

Published material not included in this thesis Other works, which on the one hand did not contribute directly to this thesis and are thus not included but on the other hand represent valuable effort and contribution given during my Ph.D., are, in order of publication:

- The CQC Algorithm: Cycling in Graphs to Semantically Enrich and Enhance a Bilingual Dictionary (winner of an in-house price as 'best 2012 paper' written by a Ph.D. student in the department of Computer Science) [Flati and Navigli, 2013];
- WoSIT: A Word Sense Induction Toolkit for Search Result Clustering and Diversification [Vannella et al., 2014];
- Language Resources and Linked Data: a Practical Perspective [Gracia et al., 2014].

1.4 Individual contributions

I personally contributed to the writing, design and implementation of all the algorithms and the evaluation setup presented in this thesis, with little exceptions which include i) the implementation of SSR module (section 2.4.2) and ii) the design and implementation of the Bitaxonomy algorithm (section 2.5.1) and the Refinement step (section 2.6), all developed by the other authors of the corresponding publications.

1.5 Outline of the thesis

The thesis is organized as follows. Chapter 2 presents MultiWiBi, the idea that contributed to its realisation and an exhaustive comparison of the taxonomic resources currently available. Chapter 3 presents SPred and introduces the notion of Semantic Predicates. Chapter 4 explains the relationship between MultiWiBi and SPred, highlighting the contribution that the former brings into the latter. Chapter 5 shows how MultiWiBi has been converted into linked data, making it reusable by many automatic systems. Finally, Chapter 6 draws conclusions and highlights possible future works.

Chapter 2

The Multilingual Wikipedia Bitaxonomy project

2.1 Introduction

In the last decades knowledge has increasingly become the true oil of our society. As the Web is slowly seeping into our everyday lives, the ability to master knowledge concerns everyone, both the grand mass of users and researchers [Mitchell, 2005, Mirkin et al., 2009, Poon et al., 2010, de Melo and Weikum, 2010b, 2012], and the big industry players [Singhal, 2012, Ferrucci, 2012] which are called to process and serve information in an efficient and accurate manner. Despite the rare cases, such as WordNet [Fellbaum, 1998], in which knowledge has been manually encoded, paving the way for a huge amount of subsequent research, building big repositories of knowledge with limited time and human support is, unfortunately now more than ever, no longer feasible, given the high volume of information, its heterogeneity and the need to have knowledge available in as many languages as possible. Researchers and industrial stakeholders have thus been devoted for decades to design novel mechanisms which were capable to automatically extract valuable information which was both broad and accurate at the same time. This has been accomplished in many different ways during the research lifetime. In the early days (but such methods remain alive as ever) there was the conviction and the desire to extract knowledge from linguistic textual repositories alone. Methods based on distributional word cooccurrence and statistical analysis over linguistic patterns relied on nothing but free text. Given the limited size and source of the textual corpora on which these systems relied on, however, even when proved to be accurate, they trudged to serve as true general domain data providers. As time went by, though, collaborative efforts started to sprout spontaneously, with the aim of developing true encyclopedic stores in which users actively contributed by enhancing the resource with additional information. Wikipedia, started in 2001, is to our knowledge one of the biggest such movements and currently the most active one, with knowledge available in 271 languages at the time of writing. A real added value brought by Wikipedia

was the possibility to decorate text with hyperlinks: this feature, combined with the availability of tabular information, made it possible to extract semi-structured information on a large-scale [Medelyan et al., 2009, Hovy et al., 2013]. Over time, systems have targeted very different types of relations, sometimes very general or open-domain (TextRunner [Banko et al., 2007], ReVerb [Fader et al., 2011]) and sometimes very specific or bound to a particular domain. Semantic relations encode a large number of linguistic aspects, spanning from general relatedness (as is the case for links across Wikipedia articles) up to specific types, such as hypernymy, holonymy, meronymy, and so on. It became increasingly clearer that hypernymy relations represented one of the most important types which could be used to boost current artificial intelligent systems. Starting from the eighties, a whole branch of research had focused on this type of semantic relation, with the pioneering works of Hearst [Hearst, 1992] laying the foundation for the forthcoming literature. Hearst's patterns, however, were designed to be applicable only on free text and did not exploit any peculiar feature of the collaborative machine-readable repositories yet to come. One of the first attempts to extract is-a information from Wikipedia dates back to WikiTaxonomy [Ponzetto and Strube, 2007] which transformed the noisy network of Wikipedia categories into a structured taxonomy of concepts. Subsequently, the example of WikiTaxonomy inspired a full line of research (e.g., YAGO [Hoffart et al., 2013], WikiNet [Nastase and Strube, 2013], MENTA [de Melo and Weikum, 2010a], LHD [Kliegr et al., 2014], etc). On the one hand, the type of knowledge extracted by these resources was either partial (is-a information was provided only for Wikipedia pages or Wikipedia categories), incomplete (lacking full coverage) or heterogeneous (i.e., not drawn from a shared, standard repository). On the other hand, another strong limit was English-centricity: for a long time English has been the only language on which the proposed methods could be applied, due to their dependence on English corpora and tools. The type of knowledge which is usually needed, however, is not constrained to a particular language: an automatic system should therefore have the desirable ability of extracting information in a language-agnostic manner.

Despite more than 50% of the content on the Web being written in English,[1] in fact, almost 90% of the Internet population lives in non-English speaking countries.[2] This consideration should thus urge to foster the production of both content and resources in languages different from English. Here stem multilingual projects such as DBpedia [Bizer et al., 2009b] and BabelNet [Navigli and Ponzetto, 2012b] which provide millions of concepts lexicalized across languages. Although seemingly simple, several factors do prevent us to easily overcome the language barrier. First of all, even though non-English users are actively contributing to populate resources collaboratively, they cannot keep the pace with the number of additions coming from active English users. Second, while having high-quality resources or tools in

[1] http://www.internetworldstats.com/stats7.htm

[2] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

English can already be demanding, expecting such tools to behave with a comparable quality in languages other than English currently seems overambitious. This issue is commonly found in under-resourced languages where no enough information (possibly semantically annotated) is available.

In contrast, MultiWiBi has been designed to address all the above-mentioned issues. First, it does not focus only on Wikipedia pages or categories but taxonomizes the two sides together, showing that they are mutually beneficial for inducing a wide-coverage and fine-grained integrated taxonomy. In particular, hypernyms are returned in a coherent manner, avoiding to intertwine one taxonomy with the other and to adopt a mixed representation depending on inventories of external resources. Second, our method is able to taxonomize Wikipedias in any language, in a way that is fully independent of additional resources. At the core of our approach, in fact, lies the idea that the English version of Wikipedia can be linguistically exploited as a pivot to project the taxonomic information to any other language offered by Wikipedia in order to have a bitaxonomy in a second arbitrary language. Nonetheless, each language is not strongly constrained to the pivot language and, in fact, we prove that our approach overcomes the language barrier and extracts hypernyms also for those concepts which do not have a counterpart in English.

2.2 Background and Contributions

In this section we introduce some background and explain our key idea of a Wikipedia bitaxonomy and clarify our contributions. We also clarify the assumptions our work relies on and introduce notation.

2.2.1 Background

The work presented in this chapter stems from the intuition that the biggest collaborative encyclopedia, namely Wikipedia, can be used for automatically deriving hypernymy information for the entities and concepts described therein. Wikipedians are used to distinguishing between articles and categories. The following paragraphs explain the differences in more detail and present some core terminology.

Wikipedia articles A Wikipedia article provides a thorough description of a single entity or concept; for example the article ALBERT EINSTEIN reports all the known facts about the physicist, while the article PERSON describes the concept of *person*. The text is almost structured, since the information is available in a XML-like language and the information is divided into sections and paragraphs. Whenever possible, articles also contain dates, tables, biographies, citations as well as media files and images. What makes Wikipedia so interesting, though, is the fact that articles are *interlinked*, so that words in an article are associated with articles in Wikipedia. The resulting hypertext can therefore be viewed as a semantic

network of Wikipedia articles. This network, though, is very complicated and links encode not only is-a relations, but also many other types of semantic relations (e.g., *born-in*, *located-in*, etc.), up to, as in the common case, more general relatedness. For example the Wikipedia article ENRICO FERMI contains a link to PHYSICIST (a link which brings the reader to the generalization of ENRICO FERMI) but also to NOBEL PRIZE IN PHYSICS (which, indeed, is strongly related to the physicist, but does not represent an is-a relation).

Besides usual pages, Wikipedia also provides the so-called *redirections*. Redirections are special pages which act as HTML redirections to other Wikipedia pages. For example redirections to the Wikipedia page SINGING include, among others, SINGER and VOCALIST, while redirections to the Wikipedia page HEADPHONES include STEREO HEADPHONES, HEAD PHONES and HEADPHONE, among others. As it should be clear from examples, redirections include misspellings of the final Wikipedia page as well as concepts which are related to the final page but do not necessarily convey the same meaning.

Wikipedia categories Wikipedia categories, instead, represent a categorization of articles into broader classes; for instance THEORETICAL PHYSICISTS is a category of ALBERT EINSTEIN, while PERSON is categorized, among others, into CONCEPTS IN ETHICS. Notably, the two sides are intertwined, as pages are usually associated with multiple categories and a category acts as a bucket for similar pages (we call these page-category associations "cross-links"). However, note that Wikipedia categories do not always represent a proper categorization for that article: for example ALBERT EINSTEIN is associated with THEORETICAL PHYSICISTS, but also to 1879 BIRTHS (which does not characterizes the physicists in a particular manner, if not that of being born in 1879) and also to INSTITUTE FOR ADVANCED STUDY FACULTY which is indeed related to, but does not say much about Albert Einstein as a physicist or, at least, as a person. For this reason, Wikipedia categories can be seen as a noisy graph of categories where nodes are connected by both is-a and relatedness relationships, without explicit distinction between the two.

Cross-links One of the core elements of this chapter is represented by the cross-links. These links are special relations which connect pages to categories. Thanks to this particular type of links, in fact, hypernymy information extracted automatically for the page side of Wikipedia can be transferred to the category side and vice versa. For example, knowing that a lot of articles linked to the category AMERICAN SINGERS have been assigned the page SINGER as hypernym is an important hint to increase the strength of association between the categories AMERICAN SINGERS and SINGERS. Wikipedia articles are usually connected to Wikipedia categories, but this might also not hold: there are, in fact, categories with no pages associated and pages which still needs to be categorised; for example, the Wikipedia article MACQUARRIE has no categories associated with it, while the Wikipedia category TRANSPORT DISASTERS IN YEMEN has no pages associated. In English this

happens about 1.6% and 13.6% of the times for the page and the category sides of the English Wikipedia, respectively.

Sense inventories A sense inventory represents a predefined set of concepts. Two major schools of thoughts emerge in the literature: in the first one all the Wikipedia pages, all the redirections and all the categories form the sense inventory. This is the sense inventory we use in this chapter; in particular hypernyms for pages are drawn from the sets of pages and redirections, while hypernyms for categories are drawn only from the set of categories. The second sense inventory leans on several resources external to Wikipedia (e.g., WordNet or the DBpedia ontology) and this is the case for many alternative approaches, such as MENTA, YAGO, DBpedia, etc. (see Section 2.8).

Our idea Despite the asymmetry between pages and categories, our hunch is that the two sides of Wikipedia can well be exploited mutually and synergistically to extract information about the generalization of both articles and categories. In fact, as a by-product, not only does our system acquire hypernymy information for each article but it also infers generalizations for Wikipedia categories, and vice versa. As the two sides are connected, the output of our system can be seen as a pair of taxonomies, one taxonomy for the Wikipedia articles and one taxonomy for the Wikipedia categories, one linked to the other. We call this pair of taxonomies a *bitaxonomy*.

More formally, a bitaxonomy is a pair $B = (T_P, T_C)$ of taxonomies, where T_P is the taxonomy for the Wikipedia pages and T_C is the taxonomy for the Wikipedia categories. T_P (T_C) is defined as the set of hypernymy edges output by our algorithm for the page (category) side of Wikipedia, that is, $T_P = \{(p, p') \mid p, p' \in P\}$ ($T_C = \{(c, c') \mid c, c' \in C\}$), where P (C) is the set of all Wikipedia articles (categories). These edges represent the hypernymy information found by our algorithm; for instance, if the taxonomy for the Wikipedia pages contains the edge (ALBERT EINSTEIN, PHYSICIST) it means that we have automatically inferred that Albert Einstein is a physicist. Formally, given the edge (p, p'), this characterization is denoted by $p' = \textit{is-a}(p)$,

Figure 2.1 provides a visual grasp of the input and output of our work. The page and the category sides are depicted with full lines and the cross-edges drawn with dashed lines. For instance, consider the Wikipedia page DONALD DUCK which in the Wikipedia page graph points to four pages, among which there are MICKEY MOUSE and CARTOON. Thanks to the application of MultiWiBi, CARTOON is promoted as hypernym of DONALD DUCK and as a result the first edge is discarded. On the other hand, the Wikipedia category DISNEY COMICS CHARACTER which has two super categories (namely, DISNEY CHARACTERS and DISNEY COMICS), is finally associated only with its hypernym category DISNEY CHARACTERS, discarding the other super category.

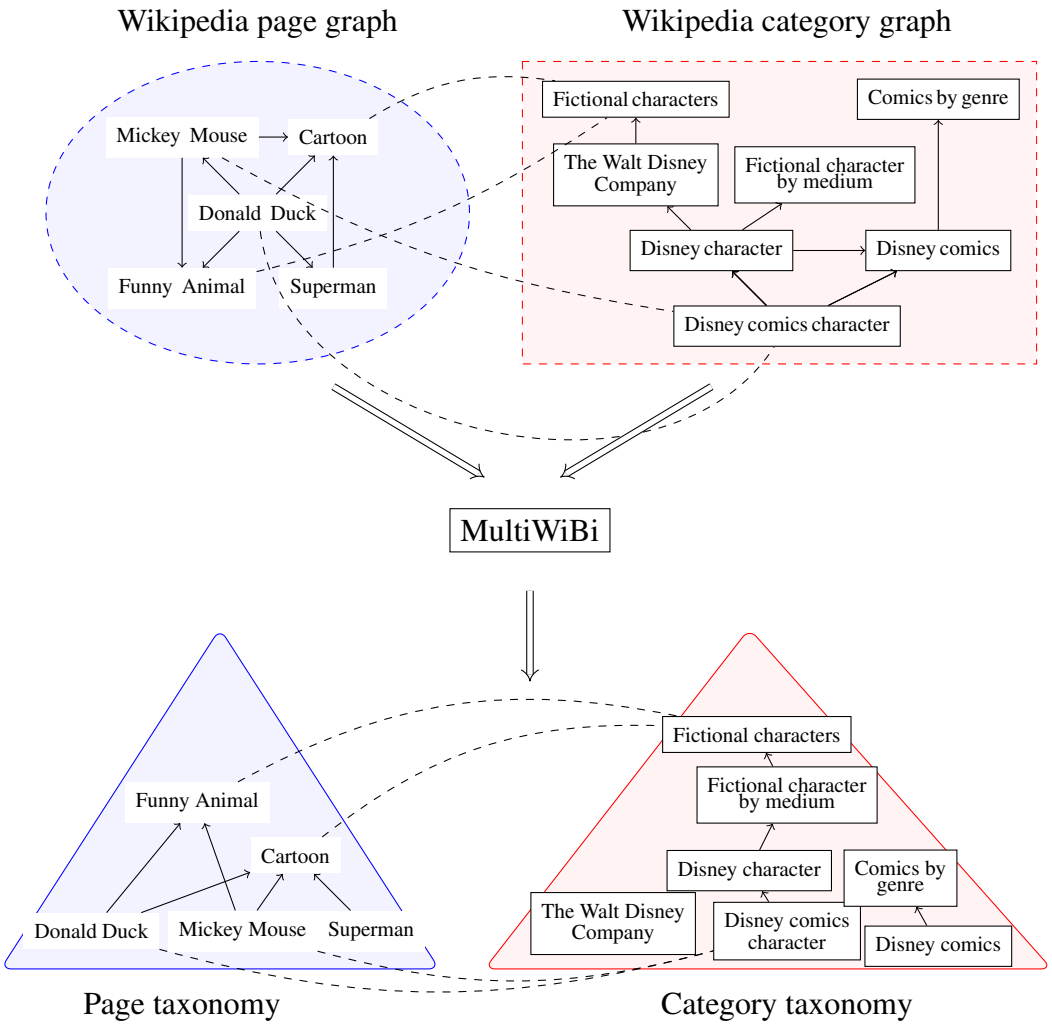


Figure 2.1. Example of input and output of MultiWiBi.

The multilingual case The objective of the work presented in this chapter is not limited to English only, but is applicable to multiple languages. The idea at the core of the extension to the multilingual case is that the English bitaxonomy could effectively be exploited to *project* the information available in English *into* another language. English is seen as a pivot language which allows to infer facts known in English also in other languages. Adding multilinguality, however, is generally easier said than done. Very often, in fact, what has been obtained for English is not as easy to obtain in languages which, differently from English, might be under-resourced or add complexities which are not found in English. With this work we show how we can leverage Wikipedia to overcome the language barriers, making it possible to collect a bitaxonomy in arbitrary languages.

Note that this does not mean that it will be possible to have all the English information transferred to all other languages. For the majority of the languages, in fact, the English Wikipedia contains many more concepts than Wikipedias in

other languages taken individually. On the other hand, note also that the English Wikipedia is far from being the union of the information found in all the other languages individually: each Wikipedia edition contains peculiar information which represents cultural concepts (such as food, dances, people native of a given country) usually not available in English, such as the Italian page PACCHERI, a well-known type of pasta produced in Italy or SAVARIN, a famous French sweet. We study these aspects in detail and we present our conclusions in Section 2.13.

Data used in this chapter In this chapter all the data used for the examples and for the experimental setup is based on the English Wikipedia 2012 (for details, see Section 2.4.3). This was made to ensure a level playing field against alternative approaches, generally leaning on a version of Wikipedia dating back to this year (see Sections 2.8 and 2.9).

2.2.2 Contributions

Our major contributions are the following:

- We provide a fully automatic algorithm for inducing a taxonomy of Wikipedia pages and of Wikipedia categories. Starting from the raw dump of the English Wikipedia, this is performed in three steps. The first step produces a first taxonomy for the page side of Wikipedia; the second step, starting from a noisy category graph, iteratively isolates hypernyms for Wikipedia categories by discriminating is-a relations from general relations thanks to cross-links; the third step refines the category taxonomy, improving the overall coverage, by solving structural flaws in the category graph.

- The two taxonomies are aligned, meaning that concepts and entities in the page taxonomy are linked to categories in the category taxonomy, and vice versa.

- For the English case, the procedure that leads to the bitaxonomy relies on English-specific tools, such as a syntactic parser for analyzing Wikipedia definitions. This choice is justified by the fact that we preferred to exploit only the most reliable language, where not only are tools and resources more studied and established among researchers, but they are also generally guaranteed to have high performance.

- This chapter is an extension of the conference paper "Two Is Bigger (and Better) Than One: The Wikipedia Bitaxonomy Project" [Flati et al., 2014] and provides a method for the automatic extension to the multilingual case and we provide mechanisms aimed at ensuring the same quality of data as of English. In strong contrast to the English case, though, the procedure does not rely on any existing resource or tool external to Wikipedia, making MultiWiBi

virtually independent and replicable on any new version of Wikipedia, in any language.

- We analyzed the behaviour of our algorithm when changing the temporal edition of Wikipedia, from 2012 to 2014; experiments show that the update has impact on the number of items covered and the quality itself is generally improved, both in terms of precision and recall.

Chapter organization The chapter is organized as follows. Section 2.2 introduces the problem to the reader, with insights and examples. Sections 2.3-2.7 present the construction of a multilingual bitaxonomy. Section 2.8 presents the related work and introduces the main competitors we compare against, while the comparative evaluation is reported in Section 2.9. The extension to the multilingual case is explained in Section 2.10 and the corresponding multilingual evaluation is presented in Section 2.11. Finally, given that the whole work relies on the Wikipedia dump dating back to 2012, Section 2.12 discusses the impact of having the underlying data updated to 2014 with regard to the potential increase in quality. Section 2.13 finally draws conclusions.

2.3 A Wikipedia Bitaxonomy for English

In order to induce the English Wikipedia bitaxonomy, i.e., a taxonomy of pages and categories, we proceed in 3 phases:

1. **Creation of the initial page taxonomy:** we first create a taxonomy for the Wikipedia pages by i) parsing the textual definitions of each page and extracting the hypernym lemma(s) and ii) by disambiguating each hypernym lemma according to the Wikipedia sense inventory.

2. **Creation of the bitaxonomy:** we leverage the hypernyms in the page taxonomy, together with their links to the corresponding categories, to induce a taxonomy over Wikipedia categories in an iterative way. At each iteration, the links in the page taxonomy are used to identify category hypernyms and, conversely, the new category hypernyms are used to identify more page hypernyms.

3. **Refinement of the bitaxonomy:** finally we employ structural heuristics to overcome inherent problems affecting certain classes of both category and page hypernyms.

The output of our three-phase approach is a bitaxonomy of millions of pages and hundreds of thousands of categories for the English Wikipedia.

2.4 Phase 1: Inducing the Page Taxonomy

The goal of the first phase is to induce a taxonomy of Wikipedia pages. Let P be the set of all Wikipedia pages and let $T_P = (P, E)$ be the directed graph of the page taxonomy whose nodes are pages and whose edge set E is initially empty ($E := \emptyset$). For each $p \in P$ our aim is to identify the most suitable generalization $p_h \in P$ so that we can create the edge (p, p_h) and add it to E. For instance, given the page APPLE, which represents the fruit meaning of *apple*, we want to determine that its hypernym is FRUIT and add the hypernym edge connecting the two pages (i.e., $E := E \cup \{(\text{APPLE}, \text{FRUIT})\}$). To do this, we proceed in two steps: i) a syntactic step, which extracts from a page's textual definition the lemma which best represents the hypernym for the page and ii) a semantic step, which identifies the most suitable sense for the lemma extracted in the syntactic step, according to our Wikipedia sense inventory.

2.4.1 Syntactic step: hypernym extraction

Given a page's textual definition, the aim of the syntactic step is to identify the lemma which best generalises the page's concept. To do this, for each page $p \in P$, we extract zero, one or more hypernym lemmas from the gloss of p, that is, we output potentially ambiguous hypernyms for the page. The first assumption, which follows the Wikipedia guidelines[3] and is validated in the literature [Navigli and Velardi, 2010, Navigli and Ponzetto, 2012a], is that the first sentence of each Wikipedia page p provides a textual definition for the concept represented by p. The second assumption we build upon is the idea that a lexical taxonomy can be obtained by extracting hypernyms from textual definitions. This idea dates back to the early 1970s [Calzolari et al., 1973], with later developments in the 1980s [Amsler, 1981, Calzolari, 1982] and the 1990s [Ide and Véronis, 1993].

To extract hypernym lemmas, we draw on the notion of copula, that is, "the relation between the complement of a copular verb and the copular verb itself".[4] Therefore, we apply the Stanford parser [Klein and Manning, 2003] to the definition of a page in order to extract all the dependency relations of the sentence. For example, given the definition of the page NOAM CHOMSKY, i.e., "Avram Noam Chomsky is an American linguist, philosopher, cognitive scientist, logician, historian, political critic, and activist", the Stanford parser outputs the set of dependencies shown in Figure 2.2. The noun involved in the copula relation is *linguist* and thus it is taken as the page's hypernym lemma.

Finally, to capture multiple hypernyms, we iteratively follow the *conj_and* and *conj_or* relations starting from the initially extracted hypernym. For example, consider the definition of NOAM CHOMSKY given above. Initially, the *linguist* hypernym is selected thanks to the copula relation; then, following the conjunction

[3] See `http://en.wikipedia.org/wiki/Wikipedia:Writing_better_articles`

[4] Cf. `http://nlp.stanford.edu/software/dependencies_manual.pdf`

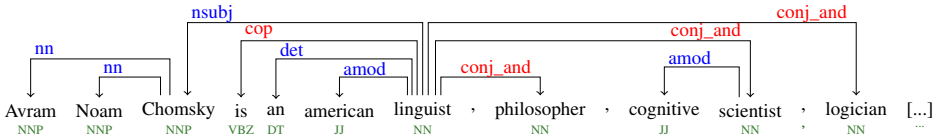


Figure 2.2. The dependency tree for the Wikipedia definition of NOAM CHOMSKY.

relations, also *philosopher*, *scientist*, *logician*, etc., are extracted as hypernyms. To understand the relevance of this step, consider that MultiWiBi succeeded to extract more than one hypernym lemma for about 12% of all the English Wikipedia pages. We acknowledge that more sophisticated approaches like [Navigli and Velardi, 2010] or [Saggion, 2004] could be applied, especially if we consider that this is a more light-weight solution than ours which, instead, leverages a syntactic parser to extract the hypernym lemmas. Obtaining high coverage, though, is critical in our case and we found that, in practice, our hypernym extraction approach is able to cover significantly more pages.

Handling special cases Words such as *one*, *kind*, *type*, etc., are often selected as hypernym lemmas. However, these are not always desirable lemmas, because they represent a class of objects. Consider, for instance, the definition of the page TRESSETTE, "Tressette or Tresette is one of Italy's major national trick-taking card games, together with Scopa and Briscola"; the only copula relation extracted is between *is* and *one*, so the hypernym lemma which is extracted is *one*. Despite being correct, the latter should be rejected in favor of *game*. Thus, to cope with this problem we use an especially designed class of nouns.[5] To avoid discarding valuable hypernyms, though, we handle only those cases in which the class term is followed by the preposition *of* (e.g., *"one of"*, *"a type of"*, etc). Hence, when this occurs we replace the class term x with the noun n involved in the dependency relation $prep\_of(x, n)$. In the previous example, since the latter is involved in the dependency relation $prep\_of(\text{one}, \text{games})$, the lemma *one* is replaced with the more concrete and informative hypernym lemma *game*.

Filling the gaps of the syntactic parser: the sister approach Analyzing the coverage of the lemmas extracted thanks to the syntactic parser, we found that for 400,286 of the English pages (about 10% of the total) no hypernym lemma could be extracted. We considered a sample of 100 pages for which the syntactic parser could not extract the hypernym. Out of the corresponding 100 glosses, we found that only 4 glosses contained the hypernym lemma in the copula relation, representing cases for which the syntactic parser failed to parse correctly, 8 were unrecognized disambiguation pages which we were not able to remove from the total list of pages, 18 contained the hypernym lemma expressed through relations other than copula

[5] *species*, *genus*, *one*, *list*, *term*, *part*, *form*, *type*, *collection*, *group*, *set*, *branch*, *order*, *class*, *family*, *series*, *name*, *style*, *variety*, *kind* and *pair*

(e.g., in the gloss "Arthur Walworth is most noted as a biographer of Woodrow Wilson." the word *biographer* is only involved in a *prep_as* dependency relation and not in a *copula* relation), and 70 were ill-formed glosses which do not clearly define the concept represented by the Wikipedia page and briefly describe its history, its role in the world or leave the generalization implicit. The latter class of ill-formed glosses include for example AUDI which is defined as "AUDI Aktiengesellschaft and its subsidiaries design, engineer, manufacture and distribute automobiles and motorcycles under the Audi, Ducati and Lamborghini brands".[6] In order to cover the pages affected by these problems, we applied an algorithm which is able to assign a hypernym lemma by inducing the information from other pages. Given a page p, the algorithm considers the so-called *sister pages* of p, i.e., pages which share with p at least one category, for which the syntactic parser has been able to provide a hypernym lemma. The algorithm then builds a distribution of such hypernym lemmas and selects the one which overlaps the most with the lemmas of p's Wikipedia categories. For the above page, for instance, the selected hypernym lemma is *manufacturer* which overlaps with the AUDI categories MOTOR VEHICLE MANUFACTURERS OF GERMANY and CAR MANUFACTURERS OF GERMANY, among others. Thanks to the sister approach we are able to recover a hypernym lemma for about 70% of the pages which could not be covered by the syntactic parsing approach.

To visually grasp the impact of the application of the two above approaches, we report in Figure 2.3 the coverage of Wikipedia pages. The bar on top reports the number of pages which have at least one hypernym lemma extracted thanks to the syntactic parsing and the sister approaches; as can be seen, 3,712,201 pages are covered, that is, approximately 97% of the total number of Wikipedia pages. The second bar reports, instead, the overall number of hypernym lemmas extracted with the two approaches. Remember that our hypernym extraction procedure possibly extracts multiple hypernyms from a single definition, so that the total number of hypernym lemmas extracted can be much higher than the number of all the Wikipedia pages (vertical red line in the figure); in fact, for the 3,712,201 Wikipedia pages covered, 4,288,709 hypernym lemmas have been extracted in total.

2.4.2 Semantic step: hypernym disambiguation

Since our aim is to connect pairs of pages via hypernym relations, our second step consists of disambiguating the obtained hypernym lemmas of page p with their most suitable senses. For instance, given *fruit* as the hypernym for APPLE we would like to link APPLE to the page FRUIT as opposed to, e.g., FRUIT (BAND) or FRUIT (ALBUM). As explained in Section 2.2.1, going beyond previous work

[6] Note that the definition for this page has been improved in 2014 into "Audi AG is a German automobile manufacturer that designs, engineers, produces, markets and distributes luxury automobiles.", so that our syntax-based approach would have been effective.

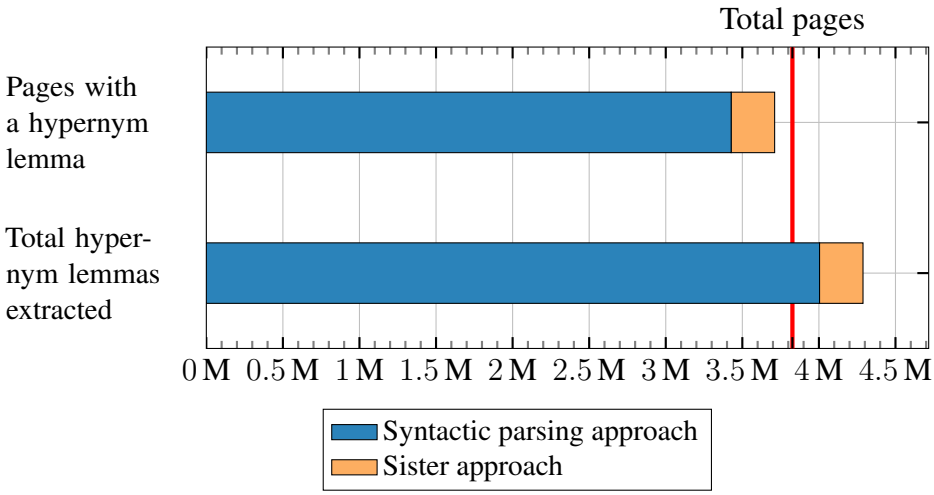


Figure 2.3. Coverage of Wikipedia pages during the hypernym lemma extraction step.

[Navigli and Ponzetto, 2012a, Ruiz-Casado et al., 2005], as inventory for a given lemma we consider the set of pages and redirections whose main title is the lemma itself, except for the sense specification in parentheses. It is very important to do so because frequent concepts, such as SINGER, PHILOSOPHER, and VOLLEYBALL PLAYER, lack their own pages in Wikipedia. If on the one hand, though, Wikipedians are continuously mitigating this issue over time (e.g., PHILOSOPHER has its own Wikipedia page in 2014), on the other hand this kind of problem is likely to persist in the future (e.g., SINGER does not exist as an independent page yet).

Another design option regards how to represent the hypernyms of a given concept. For example Avram Noam Chomsky is defined as "an American linguist, philosopher, cognitive scientist, logician, political commentator, social justice activist, and anarcho-syndicalist advocate", and it can be argued that these represent the different roles of the entity in the world. Roles in AI can be represented as properties, individuals or classes. The property-based representation is out of scope and does not serve our goals, since we do not have (and are not focusing on) objects. As regards individuals and classes, Wikipedia respectively define them as "Individuals (instances) are the basic, "ground level" components of an ontology. The individuals in an ontology may include concrete objects such as people, animals, tables, automobiles, molecules, and planets, as well as abstract individuals such as numbers and words " and "Classes – concepts that are also called type, sort, category, and kind – can be defined as [...] abstract groups, sets, or collections of objects.". For convenience, we decided to represent all the nodes in our taxonomy to be classes. This was done because it is not easy to determine automatically which nodes in the taxonomy represent individuals and which represent classes. Future work might investigate an automatic procedure to distinguish between the two.

Finally, it is worthy to point out that, historically speaking, researchers have been focusing on extracting hypernymy relations only at the lexical level (i.e., between words); for example to extract that *fruit* is the hypernym word of the word

apple, but the two words are not embedded in a semantic network, i.e., they are not disambiguated. This work, in contrast, belongs to a class of works which relies on a semantic representation of the hypernymy relation, so that the previous hypernym relation would be encoded as (APPLE, is-a, FRUIT), where the subject and the object of the triple are Wikipedia senses and not just words.

In order to disambiguate hypernym lemmas extracted in the previous step, we apply a battery of hypernym linkers, which output the most suitable sense for a given lemma, combined with two procedures which limit sense-drifts during the application of the linkers.

Hypernym linkers

To disambiguate hypernym lemmas, we exploit the structural features of Wikipedia through a pipeline of hypernym linkers $\mathcal{L} = \{L_i\}$, applied in cascade order. We start with the set of page-hypernym pairs $H = \{(p, h)\}$ as obtained from the syntactic step. The successful application of a linker to a pair $(p, h) \in H$ yields a page p_h as the most suitable sense of h, resulting in setting $isa(p, h) = p_h$. At step i, the i-th linker $L_i \in \mathcal{L}$ is applied to H and all the hypernyms which the linker could disambiguate are removed from H. This prevents lower-precision linkers from overriding decisions taken by more accurate ones (cf. Section 2.4.3). Hypernym linkers are applied in the same order with which they are presented (for details, see Section 2.4.3).

In what follows we denote with $p \xrightarrow{h} p_h$ the fact that the definition of a Wikipedia page p contains an occurrence of h linked to page p_h. Note that we do not constrain p_h to be necessarily a sense of h and let it represent an arbitrary Wikipedia page; for instance, we let the hypernym lemma *person* to be linked to the Wikipedia page INDIVIDUAL which is not a sense of *person* in Wikipedia.

Category linker Given the set $W \subset P$ of Wikipedia pages which have at least one category in common with p, we select the majority sense of h, if there is one, as hyperlinked across all the definitions of pages in W:

$$isa(p, h) = \arg\max_{p_h} \sum_{p' \in W} 1(p' \xrightarrow{h} p_h)$$

where $1(p' \xrightarrow{h} p_h)$ is the characteristic function which equals 1 if h is linked to p_h in page p', 0 otherwise. For example, the linker sets $isa(\text{EGGPLANT}, \textit{plant})$ = PLANT because most of the pages associated with TROPICAL FRUIT, a category of EGGPLANT, contain in their definitions the term *plant* linked to the PLANT page.

Crowdsourced linker If $p \xrightarrow{h} p_h$, i.e., the hypernym h is found to have been manually linked to p_h in p by Wikipedians, we assign $isa(p, h) = p_h$. For example, because *capital* was linked in the BRUSSELS page definition to CAPITAL CITY, we set $isa(\text{BRUSSELS}, \textit{capital})$ = CAPITAL CITY.

Distributional linker This linker provides a distributional approach to hypernym disambiguation. We represent the textual definition of page p as a distributional vector $\vec{v}_p$ whose components are all the English lemmas in Wikipedia (we consider nouns, adjectives, adverbs and verbs). The value of each component is the occurrence count of the corresponding content word in the definition of p. We perform no compounding, discard lemmas whose length is equal to 1 and discard the verb *to be* because contained in almost all Wikipedia definitions. The goal of this approach is to find the best link for hypernym h of p among the pages h is linked to, across the whole set of definitions in Wikipedia. Formally, for each p_h such that h is linked to p_h in some definition, we define the set of pages $P(p_h)$ whose definitions contain a link to p_h, i.e., $P(p_h) = \{p' \in P | p' \xrightarrow{h} p_h\}$. We then build a distributional vector $\vec{v}_{p'}$ for each $p' \in P(p_h)$ as explained above and create an aggregate vector $\vec{v}_{p_h} = \sum_{p'} \vec{v}_{p'}$. For discriminating among vectors, we also remove the target lemma from $\vec{v}_{p_h}$. Finally, we determine the similarity of p to each p_h by calculating the dot product between the two vectors $sim(p, p_h) = \vec{v}_p \cdot \vec{v}_{p_h}$. If $sim(p, p_h) > 0$ for any p_h we perform the following association:

$$isa(p, h) = \arg\max_{p_h} sim(p, p_h)$$

For example, consider the Wikipedia page ARISTOTLE and its hypernym lemma *teacher*. Among all Wikipedia textual definitions in which it occurs, the latter has been linked to several senses, among which there are TEACHER and PIANO TEACHER. The vectors for the starting page ARISTOTLE and these two senses are shown below:

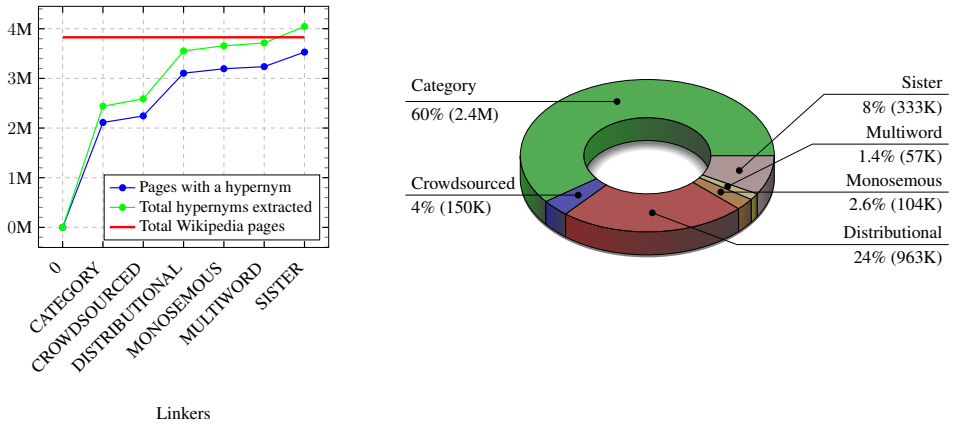
$$\vec{v}_{\text{ARISTOTLE}} = (\text{polymath:1}, \text{philosopher:1}, \text{student:1}, \dots)$$

$$\vec{v}_{\text{TEACHER}} = (\text{student:30}, \text{philosopher:14}, \text{polymath:1}, \dots)$$
$$\vec{v}_{\text{PIANO TEACHER}} = (\text{pianist:1}, \text{virtuoso:1}, \text{composer:1}, \dots)$$

The similarities between the vector for the starting page and the vectors of the two senses are thus

$$similarity(\text{ARISTOTLE}, \text{TEACHER}) = 1 \times 30 + 1 \times 14 = 45.0$$
$$similarity(\text{ARISTOTLE}, \text{PIANO TEACHER}) = 0$$

In the first case the two vectors share lemmas such as *student* and *philosopher*, so their similarity is greater than zero, while in the second case the two vectors have no word in common. Hence, since TEACHER is the sense of *teacher* which maximises the similarity with ARISTOTLE, this linker sets $isa(\text{ARISTOTLE}, \textit{teacher})$ = TEACHER.



(a) Coverage of pages as linkers are applied.

(b) Distribution of disambiguated hypernyms by linker (displayed from top-left, counter-clockwise).

Figure 2.4. Absolute number and distribution of hypernyms disambiguated by our hypernym linkers.

Monosemous linker If h is monosemous in Wikipedia (i.e., there is only a single sense p_h for that lemma), link it to its only sense by setting $isa(p, h) = p_h$. For example, the syntactic step extracted the hypernym lemma *businessperson* from the definition of MERCHANT and, being unambiguous, we link it to BUSINESSPERSON.

Multiword linker If $p \xrightarrow{m} p_h$ and m is a multiword expression containing the lemma h as one of its words, set $isa(p, h) = p_h$. For example, we set $isa(\text{AREA 51}, \textit{base})$ = MILITARY BASE, because the multiword expression *military base* is linked to MILITARY BASE in the definition of AREA 51.

Sister linker Finally, given the set $W \subset P$ of Wikipedia pages which have at least one category in common with p and share the hypernym lemma with it, we select the most frequent hypernym across these. For example we determine *is-a*(GUITARIST, *person*) = PERSON, thanks to the fact that seven pages (e.g., COMPOSER and DISC JOCKEY) have PERSON as common hypernyms and share the category OCCUPATIONS IN MUSIC with the starting page GUITARIST.

Figure 2.4a plots the coverage of the Wikipedia pages as hypernym linkers are applied in the presented order. Two lines are shown: the blue line plots the number of pages with at least one hypernym, the green line shows the number of total hypernyms found up to a certain phase. Again, since MultiWiBi extracts more than one hypernym lemma for any given page, the total number of hypernyms is higher than the total number of Wikipedia pages. Figure 2.4b shows also the absolute number of the hypernym links found and the corresponding relative ratios. As can be seen, the first two heuristics provide, alone, about two thirds of the total hypernyms contained in the Wikipedia page taxonomy, while the others increasingly

disambiguate hypernym lemmas, until 4,046,411 total hypernyms are found for 3,529,647 Wikipedia pages, covering more than 92% of the total Wikipedia pages.

In order to limit the potential noise introduced by the linkers, after the application of each of them, we apply two special modules whose aim is to preserve quality during the linking pipeline and detect possible shifts in meaning.

Preserving meaning between hypernym lemmas and hypernym senses

As a result of the application of the entire linking pipeline we obtain a large number of disambiguated hypernym lemmas. However, a non-negligible number of disambiguated hypernyms suffer from the problem of *semantic shift*. This phenomenon occurs when a page's hypernym lemma is linked to another page which is very related to it but is not a sense for the hypernym at hand. Consider for example the gloss "Heinrich von Tenner was an Austrian fencer$_{\text{FENCING}}$." in which the hypernym term *fencer* is linked to the page FENCING. This is not something inappropriate *per se* but rather reflects a very common phenomenon which consists in annotating text with the domain rather than the word sense (i.e., FENCING can be considered as the topic or domain usually associated with *fencer* but not a sense of it).[7] Furthermore, this phenomenon involves different kinds of linguistic aspects, such as gender differentiation (e.g., *actress/actor*), distinction between an activity and the associated role (e.g., *singing/singer*, *painting/painter*), etc. In addition, it is important to point out that links in Wikipedia can be pages as well as redirections. As such, redirections include mispellings of the final Wikipedia page as well as concepts which are related to the final page but do not necessarily convey the same meaning. Note that redirections do not have any text associated with them so that it becomes hard to define solid linguistic rules which measure the relationship between a redirection and the target page.

Lemma preserver (LP) As a first simple attempt to cope with the semantic shift phenomenon we apply a procedure that we called *Lemma preserver*. Whenever any of the linkers presented in Section 2.4.2 outputs a Wikipedia page p as the disambiguation of the hypernym lemma l this routine tries to preserve the meaning of l by looking at the possible redirections of p. For example, the Category linker disambiguated the hypernym lemma *linguist* of the Wikipedia page NOAM CHOMSKY with LINGUISTICS. Of course, as explained above, this is a very related page, but should not be considered a valid disambiguation for the hypernym lemma extracted. As a result of the LP procedure, instead, LINGUISTICS is replaced by LINGUIST, which is a redirection to the former. This is a very frequent and important action to take; consider that about 17% of the links output by the first (i.e., category) linker are replaced by the lemma preserver.

[7] In fact, these edges bear very important semantics and could in principle be left in the taxonomy with an opaque RELATED-TO label. As for now, we decided to discard them to provide a cleaner and more coherent taxonomy for the Wikipedia page side.

Input: Strings s_1, s_2
Output: true if (s_1, s_2) is a semantic shift, false otherwise

1: $s_1 \leftarrow$ normalize(s_1);
2: $s_2 \leftarrow$ normalize(s_2);
3: $h_1 \leftarrow$ get_head(s_1);
4: $h_2 \leftarrow$ get_head(s_2);
5: r $\leftarrow$ is_shift(h_1, h_2);
6: **if** r == true OR r == false **then return** r;
7: $t_1 \leftarrow$ get_last_token(s_1);
8: $t_2 \leftarrow$ get_last_token(s_2);
9: r $\leftarrow$ is_shift(t_1, t_2);
10: **if** r == true OR r == false **then return** r;
11: **return** false;

(a) The Semantic Shift Recognizer (SSR) Algorithm.

Freq.	s_1	s_2
47299	footballer	ASSOCIATION FOOTBALL
15872	football player	ASSOCIATION FOOTBALL
3671	fencer	FENCING
1980	manager	COACH (SPORT)
1648	peer	PEERAGE
1352	sprinter	SPRINT (RACE)
1177	wrestler	AMATEUR WRESTLING
1130	title	BRITISH NOBILITY
1087	volleyball player	VOLLEYBALL

(b) Excerpt of the most frequent semantic shifts recognized by the SSR module.

Figure 2.5. The SSR algorithm (a) and an excerpt of the most frequent shifts returned by the algorithm (b).

Semantic Shift Recognizer (SSR) A second, more general and linguistic-bound attempt is represented by a specific module called *Semantic Shift Recognizer* (SSR) which, on the basis of English hand-crafted rules, automatically discriminates is-a relations from semantic shifts.

We now describe in detail the mechanism behind the SSR module, whose pseudocode is reported in Figure 2.5a. To recognize if there is a semantic shift between the two concepts represented by two strings s_1 and s_2 we first normalize them (lines 1–2) so that i) all words within parentheses are removed (e.g. *Person (sport)* is cut to *Person*), ii) s_1 and s_2 are lowercased, iii) acronyms are normalized (e.g., s_1=*ep* and s_2=*extended play* get normalized into *ep*), and iv) several separators are normalized with a space (e.g., *business_man* and *business-man* get both normalized to *business man*).

The core of the SSR module consists of isolating the heads of the two strings (lines 3–4) and subsequently applying the following string matching rules (line 5 of the algorithm in Figure 2.5a):

> **identity test**: return false if s_1 is identical to s_2 (e.g., s_1=*plant* and s_2=*plant*);
>
> **suffix test**: strip off the stems of the two heads and compare the suffixes of s_1 and s_2; consider as shifts any person/profession or subject/domain shift between the two suffixes (e.g., *singer/singing* and *novelist/novel* are considered shifts).

Note that the two tests above might not give any answer; the first test identifies

negative cases, while the second identifies positive cases. In case the previous tests do not detect any semantic shift (i.e., the variable r at line 6 is undefined), the analysis is repeated on the last two respective tokens of s_1 and s_2 (lines 7 and 8) and finally applies the two string matching rules described above. If no semantic shifts have been detected, the SSR applies the simple baseline of not detecting a shift.

Figure 2.5b reports the most frequent semantic shifts detected by the Semantic Shift Recognizer module; as can be seen, most of them consist of topic drifts. Since this is an automatic procedure, of course this list might well include errors not detected by our module (e.g., *manager* and COACH (SPORT)).

2.4.3 Page Taxonomy Evaluation

We now describe the setup of our experiments and discuss the results. As explained above, all our experiments are based on the 2012 edition of Wikipedia in 4 different languages.[8] This was almost a forced choice, since nearly all the available taxonomic resources refer to a version of Wikipedia which dates back to 2012.[9]

Dataset

To evaluate the quality of our page taxonomy we randomly sampled 1,000 Wikipedia pages. For each page we provided: i) a list of suitable hypernym lemmas for the page, mainly selected from its definition; ii) for each lemma the correct hypernym page(s).

Evaluation measures

Unfortunately, measuring the quality of a taxonomy is not a trivial task. Currently there is still no agreement on how to perform this [Velardi et al., 2013]. On the one hand, performing a complete validation of all the edges contained in a taxonomy is unattainable, on the other hand, even when a smaller sample of the edges is validated, it is not clear which measures to use for a correct and fair evaluation. For these reasons, we defined three measures that take values between 0 and 1 and try to characterise three different dimensions of quality: precision, recall and coverage.

We used macro precision, i.e. the average ratio of correct items to the total number of items returned. This measure is intended to count the average correctness of the information provided for each single node covered by the taxonomy. Note that if the taxonomy contains only one correct edge, its precision is 1; this means that this measure alone cannot truly grasp the overall quality of the taxonomy.

Given the wide range of possible answers that could be considered to be correct, standard recall across resources could not be calculated. We thus calculated recall as the ratio of the items for which the system outputs at least one correct answer.

[8] The exact dates for the different languages are: 2012/10/01 for English, 2012/10/07 for French, 2012/10/12 for Italian and 2012/09/27 for Spanish.

[9] All but MENTA, which instead extracts data from Wikipedia 2010. See Table 2.1 for details.

	P	R*	C	# items
Lemma (syntactic)	93.80	89.80	94.50	1,000
Lemma (syntactic + sisters)	**93.17**	**93.30**	**98.90**	
Sense (simple)	83.20	81.80	96.00	1,000
Sense (only MP)	85.89	84.20	96.00	
Sense (only SSR)	89.68	85.90	94.20	
Sense (LP + SSR)	**90.89**	**87.40**	**94.60**	

Figure 2.6. Page taxonomy performance at lemma- and sense- level. Performance related to the chosen configurations are shown in bold.

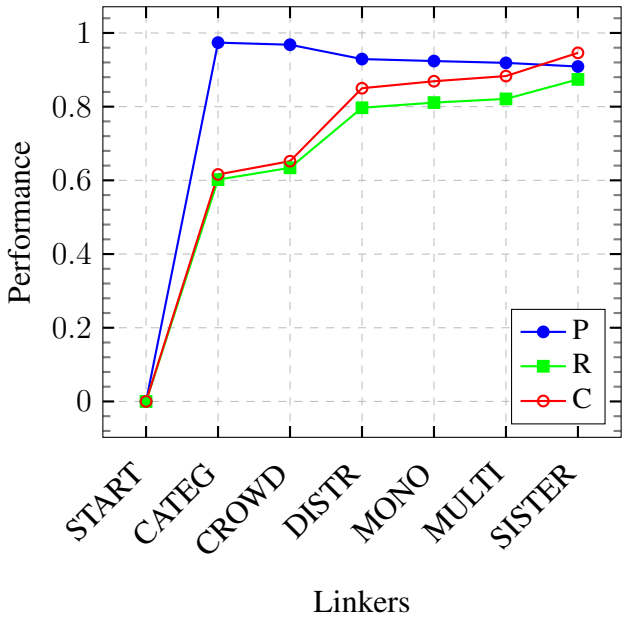


Figure 2.7. Page taxonomy performance as linkers are applied.

For example, FRANK SINATRA can be considered to be a singer, a person, an artist or even an actor. Furthermore these concepts correspond to several lexicalizations (singer vs. vocalist or actor vs. performer), so that it is quite difficult to identify a shared, well-posed inventory of expected answers. In order to calculate precision and recall, for each resource we therefore manually marked each hypernym returned as correct or not.

Another useful measure which acts as the upper bound to precision and recall is coverage, defined as the fraction of items for which at least one answer is returned, independently of their correctness; the rationale behind this measure is to have a rough idea of the amount of information provided by the taxonomy by considering the number of covered items.

Hypernym linker order

The optimal order of application of the above linkers is the same as that presented in Section 2.4.2. It was established by selecting the combination, among all possible permutations, which maximized macro precision on a tuning set of 100 randomly sampled Wikipedia pages, disjoint from our page dataset. The Sister linker, instead, is employed as the last one, since it exploits hypernym links found by previous linkers.

Results

Results, both at lemma- and sense-level, are reported in Figure 2.6. The first two lines show performance when considering the quality of the extraction of the ambiguous hypernyms. As can be seen, at lemma-level, the configuration that exploits the sister pages in combination with the simple syntactic extraction phase produces a modest increment in both coverage and recall, to little detriment of precision. The final configuration is shown in bold (syntactic + sisters). The following lines in the table show results after i) the disambiguation step (vanilla), ii) when the LP module is used after the application of the linkers (only LP), iii) when only the SSR module is applied after the application of the linkers, and finally iv) when both modules are applied (LP + SSR). As can be seen, applying the LP module does not alter coverage, because this module does not filter out any linker's answer. In constrast, both precision and recall are boosted modestly. When the SSR module is applied, instead, coverage lowers to 94.20, but precision and recall receive an important increase. Finally, when the two modules are applied, the climax is reached for precision and recall, while coverage attests somewhat in between the vanilla setting and the more restrictive one when using the LP or the SSR individually. In bold we highlighted the final, chosen configuration, that is, the combination of the linkers, the LP and the SSR procedures. Figure 2.7 shows the performance in terms of precision, recall and coverage as the hypernym linkers are applied (cf. Section 2.4.3). Precision, generally very high, has a positive spike after the application of the first linker and then decreases slowly as subsequent linkers are chained, attesting around 90%. Recall and coverage consistently increase when more linkers are considered, going on par with precision.

2.5 Phase 2: Inducing the Bitaxonomy

The Wikipedia page taxonomy built in Section 2.4 will now serve as a stable, pivotal input to the second phase, the aim of which is to build our bitaxonomy, that is, a taxonomy of pages aligned to a taxonomy of categories. Our key idea is that the generalization information available in each of the two partial taxonomies is mutually beneficial. We implement this idea by exploiting one taxonomy to add new hypernymy relations to the other, and vice versa, in an iterative way, until a fixed point is reached. The final output of this phase is, on the one hand, a page taxonomy

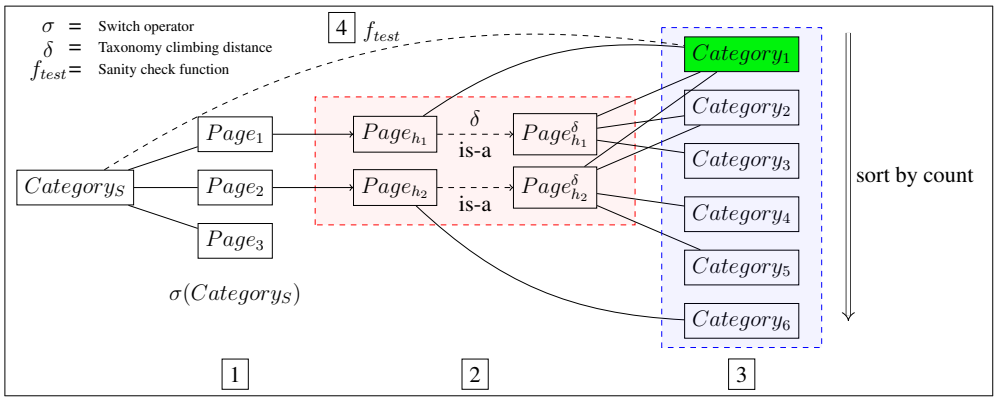


Figure 2.8. Example of the application of the MultiWiBi iterative algorithm on the category side of Wikipedia. $Category_S$ and $Category_i$ denote the starting and the candidate categories, respectively.

augmented with additional hypernymy relations and, on the other hand, a category taxonomy which is built starting from the noisy category graph (see Section 2.2).

2.5.1 The Bitaxonomy Algorithm

We now describe in detail the bitaxonomy algorithm. To help the reader throughout the explanation, we will support the presentation with Figure 2.8, which shows the steps in which the algorithm is divided. As can be seen, we identified four steps (each step is represented by a number enclosed in a square) called as follows: Item switch (step 1), Taxonomy climbing (step 2), Candidate discovery (step 3) and Sanity check (step 4). Before going into the details of each single step, let us explain how the data structures are initialised.

2.5.2 Initialization

Our initial bitaxonomy $B = (T_P, T_C)$ is a pair consisting of the page taxonomy $T_P = (P, E)$, as obtained in Section 2.4, and the category taxonomy $T_C = (C, \mathbb{1}_{super})$, where C contains all the Wikipedia categories and $\mathbb{1}_{super} := \{e = (u, v) \in E(CG) \mid deg^+(u) = 1\}$, where CG is the Wikipedia category graph; in simpler words, the initialization of the category taxonomy considers all those nodes which have outdegree equal to 1 (i.e., which have only one super category in the noisy category graph) and adds these edges to the set $E(T_C)$. The algorithm is started on the category taxonomy with the (partial) page taxonomy as input (line 1).

In the algorithm we denote with T the taxonomy being refined and with T' the taxonomy that the algorithm draws on to update T. Initially $T = T_C$ and $T' = T_P$ (see line 1).

Algorithm 1 The Bitaxonomy Algorithm

Input: T_P, T_C

1: $T := T_C, T' := T_P, \xi \leftarrow 1000, \lambda \leftarrow 1, \delta \leftarrow 1, \delta_{max} \leftarrow 3, \lambda_{max} \leftarrow 6$
2: **repeat**
3: $sizeT \leftarrow |E(T)|$
4: $convergence \leftarrow false$
5: **for all** $t \in V(T)$ s.t. $\nexists\, t_h \in T, (t, t_h) \in E(T)$ **do**
6: **reset** $candidate\_count$
7: $\Sigma \leftarrow \sigma(t)$ ▷ step 1
8: $H \leftarrow get\_hypernyms(\Sigma, \delta, T')$ ▷ step 2
9: **for all** $t'_h \in H$ **do**
10: **for all** $t_h \in \sigma(t'_h)$ **do** ▷ step 3
11: $candidate\_count(t_h)$++
12: **end for**
13: **end for**
14: **for all** $t_h \in sort(candidate\_count)$ **do**
15: **if** $sanity\_check(t, t_h, T)$ **then** ▷ step 4
16: $E(T) := E(T) \cup (t, t_h)$
17: **break**
18: **end if**
19: **end for**
20: **end for**
21: **if** $T == T_C$ AND $(|E(T)| - sizeT < \xi)$ **then** ▷ Parameter update and stop condition
22: $\lambda \leftarrow \lambda + 1$
23: **if** $\lambda >= \lambda_{max}$ **then**
24: $\lambda \leftarrow 1$
25: $\delta \leftarrow \delta + 1$
26: **end if**
27: **if** $\delta >= \delta_{max}$ **then** $convergence \leftarrow$ **true**
28: **end if**
29: **end if**
30: **swap** T and T'
31: **until** convergence
32: **return** $\{T, T'\}$

Figure 2.9. The Bitaxonomy Algorithm

2.5.3 The four steps

We now describe the core algorithm of our approach, which iteratively populates and refines the edge sets $E(T_C)$ and $E(T_P)$.

Item switch (step 1) In the first step we start by considering an uncovered node $t \in T$. Depending on the current iteration, t can be either a page or a category (line 5). We then apply an operator σ, that we call *switch operator*, which takes as input a Wikipedia item (either a page or a category) and returns the set of its Wikipedia counterpart elements, i.e., those items which belong to the other side of Wikipedia and are connected to it by means of a cross-link (see Section 2.2). In a few words, σ expresses the mutual membership relation existing between pages and categories. More formally, given $c \in C$, $\sigma(c)$ is the set of pages categorized with c, while given $p \in P$, $\sigma(p)$ is the set of categories associated with page p in Wikipedia. In this step, the algorithm starts from t and uses $\sigma(t)$ to switch from one taxonomy to the other (line 7 and Figure 2.8, 1).

Example Consider the uncovered Wikipedia category $t =$ OLYMPICS (line 5). By applying the switch operator to OLYMPICS, we reach the following set of pages σ(OLYMPICS)={ PARALYMPIC GAMES, OLYMPIC GAMES, OLYMPIC CUP, ... } ($|\sigma(\text{OLYMPICS})| = 26$).

Taxonomy climbing (step 2) Given the dual Wikipedia items $\sigma(t) = \{t'_1, \ldots, t'_{|\sigma(t)|}\}$, the goal of this step is to harvest hypernyms of the dual nodes in $\sigma(t)$ which will then be switched back to the starting taxonomy. To do this, we build a set $H(\sigma(t))$ by "climbing" the taxonomy T', reaching all the hypernyms at distance less than or equal to the hypernymy distance parameter δ starting from each item $t'_i \in \sigma(t)$ (line 8). The maximum climbing distance changes during the iterations, so as to constrain the algorithm to favor closer hypernyms over the first iterations and allow it to reach farther hypernyms as it proceeds (line 21 and Figure 2.8, 2).

Example (cont'd) Out of the total 26 pages contained in σ(OLYMPICS), 23 pages come with a hypernym, discovered during the construction of the page taxonomy (line 8); for example, PARALYMPIC GAMES is a MULTI-SPORT EVENT. All the hypernyms at distance 1 are added to $H(\sigma(\text{OLYMPICS}))$, which the set of the pages to project back to the category taxonomy; for example, MULTI-SPORT EVENT is contained in this set.

Candidate discovery (step 3) The goal of this step is to identify a set of candidate hypernyms for the starting node t. To this end, having $H(\sigma(t))$ as input to this step, we apply the *switch operator* to each t'_h in $H(\sigma(t))$ (lines 9–10) and we count the number of times we reach a node in T (line 11). As Wikipedia items in one taxonomy are usually associated with multiple items in the other taxonomy, items will be counted multiple times, so as to generate a distribution. The result of this step is thus a distribution over candidate nodes which notably belong to the same taxonomy given as input to the algorithm (cf. Figure 2.8, 3). This is the core of the bitaxonomy algorithm, in which hypernymy knowledge is transferred from one taxonomy to the other.

Example (cont'd) For each hypernym page in $H(\sigma(t))$ we apply the *switch operator*, obtain the candidate categories and sum 1 for each of them. As a result we obtain the following distribution: {MULTI-SPORT EVENTS:4, ..., AWARDS:1, SWIMSUITS:1}, meaning that we end up counting the category MULTI-SPORT EVENTS four times and other categories, such as AWARDS and SWIMSUITS, only once.

Sanity check (step 4) The input for this step is the same in either of the two sides of the bitaxonomy, i.e., a starting node $t \in T$ and a candidate hypernym $t_h \in T$, belonging to the same taxonomy. The goal of this step is to select, whenever possible, the best hypernym amongst the candidate list found in the previous step. Such promotion is performed only if the candidate hypernym t_h passes a sanity check which guarantees the compatibility with the starting node t. Given the different nature of the two sides of Wikipedia, we devised specialised conditions; this step is thus the only one which depends on the current taxonomy being updated. As regards the page taxonomy, given the page $p \in T_P$ and the candidate hypernym page $p_h \in T_P$, the sanity check verifies whether p_h is a sense for some of the hypernym lemmas extracted for p (see Section 2.4.1). As for the category taxonomy, given the category $c \in T_C$ and the candidate hypernym category $c_h \in T_C$, the sanity check verifies whether c and c_h are connected by a path of length $\leq \lambda$ (see Section 2.5.4). If this holds, we then select the direct super-category of c lying on the shortest path between c and c_h. The rationale behind this asymmetry lies in the fact that only the category side of Wikipedia is backed with an underlying noisy graph and connectivity techniques cannot be generalised easily also to the page side.

This fourth step considers the items contained in the distribution of step 3 in decreasing order, promotes the node t_h^* with the highest count which passes a sanity check, if any (line 15), and a new edge $e = (t, t_h^*)$ is finally added to the taxonomy (line 16). Note that as soon as a candidate node passes the sanity check a new edge is added to the taxonomy and all the remaining candidates discarded 17. The sanity check has the aim of discriminating among the hypernym candidates contained in the set $H(\sigma(t))$, by checking whether it is safe to add an edge between the starting node and the candidate.

Example (cont'd) We proceed in decreasing order of vote and verify whether the sanity check for categories holds. As MULTI-SPORT EVENTS has the highest count and is connected to the starting category OLYMPICS by a path in the Wikipedia category network (in fact, the former is a direct super-category of the latter), we finally add the hypernym edge (OLYMPICS, MULTI-SPORT EVENTS) to T_C (line 16) and exit step 4 (line 17).

2.5.4 Parameter update and stop condition

At the end of each iteration the role played by the two taxonomies is swapped and the (partial) taxonomy becomes the new input for a new iteration (line 30). The four steps are repeated until a stop condition is satisfied (line 27). The algorithm is governed by two parameters, the maximum path length parameter λ and the maximum hypernymy distance parameter δ. The former controls the maximum length of the path in the category sanity check; the latter regulates the maximum hypernymy distance in the taxonomy climbing step (step 2). We voluntarily let δ take smaller values than λ in order not to assign over-generalised hypernyms. At the end of a given iteration, whenever less than ξ edges have been added, λ is incremented. When a maximum value λ_{max} is reached, λ is reset to 1, in order to prefer closer categories, and δ increased by one. As a safety stop condition, we constrain also the hypernymy distance parameter δ to a maximum value δ_{max}. By starting from a page (or category) and climbing a taxonomy without such care, in fact, we would easily risk to reach the top of the taxonomy and assign hypernyms which are too general (such as ENTITY or BEING) and which would likely contribute to generate errors. When eventually the hypernymy distance parameter δ reaches the maximum value allowed, the algorithm is stopped and the two taxonomies returned. Note that the parameters are modified only when a temporary convergence with these parameters is reached: the fact that the algorithm assigns a small number of edges during a certain iteration, in fact, means that the path length parameter is not high enough to let the algorithm generalize sufficiently. Hence, the need to increase the path parameter and spin the algorithm through an additional iteration. Note also that, since λ depends on ξ, it is not possible to know *a priori* the number of iterations that the algorithm has to perform.

2.6 Phase 3: Bitaxonomy refinement

Despite the successful application of the Bitaxonomy Algorithm to the two taxonomies, the latter still suffer from structural shortcomings we will now focus on.

As regards the page taxonomy, the algorithm crucially leverages two important features to discover the right hypernym to promote: first, a Wikipedia page needs to have categories associated with it, and, second, it also needs to provide at least one hypernym lemma. This means that the algorithm cannot be applied to a (small) class of Wikipedia articles which are, in fact, left out. This class mainly contains redirections promoted to hypernym which, by construction, have neither categories nor definitions associated. For this reason we introduced a final refinement for the page taxonomy which addresses the problem of finding a proper generalization for this set of redirections.

As regards categories, the problem is very similar. Since the bitaxonomy algorithm crucially exploits the *switch operator* to harvest the pages associated with a

certain category, it fails whenever the latter has no pages categorised under it. To this end we designed an ad-hoc procedure that overcomes this structural shortcoming.

2.6.1 Page taxonomy refinement

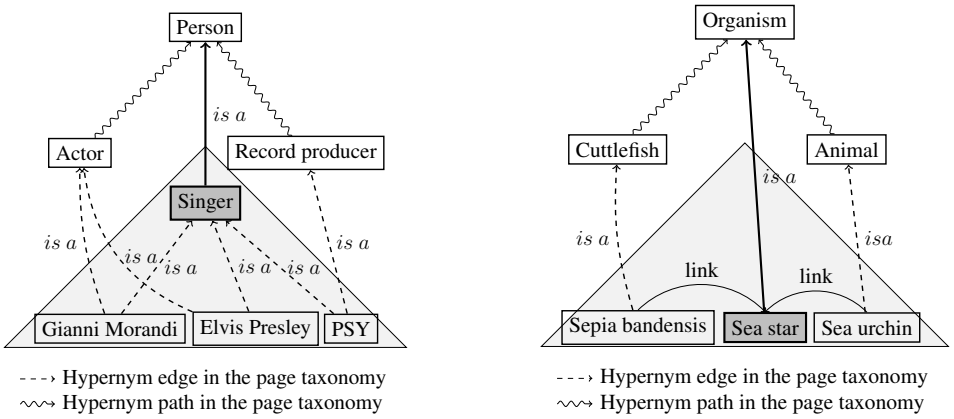
At the core of the refinement of the page taxonomy there are two simple ideas which, applied in cascade order, both use a trivial taxonomical property: if a node in a taxonomy has two hypernyms, then these must reconcile somewhere up in the hierarchy, i.e., they must have a common ancestor. For example, in an ideal taxonomy the two hypernyms of ELVIS PRESLEY, namely SINGER and ACTOR, should both have PERSON or ARTIST as their lowest common ancestor. The two ideas, called IYA (I am if You Are) and ILY (I am Like You), both exploit this principle. The former, IYA (Figure 2.10a), exploits the ancestors of the hyponyms of a given redirection. For example, in order to discover the hypernym for the redirection SINGER, we first consider, among others, the pages GIANNI MORANDI and PSY and consider in turn their alternative hypernyms, ACTOR and RECORD PRODUCER, respectively. We then climb the taxonomy until a common ancestor is encountered, i.e., PERSON, which is finally promoted as hypernym of the initial redirection SINGER. The latter, ILY (Figure 2.10b), contrarily to IYA, leverages the ancestors of those pages which have an outgoing link to the redirection considered. For example, in order to find the hypernym for the redirection SEA STAR, we consider all the pages pointing to the redirection, among which there are SEPIA BANDENSIS and SEA URCHIN. Similarly to the previous procedure, ILY determines the common ancestor ORGANISM and sets is-a(SEA STAR, ORGANISM).

Note that the two procedures differ only in the set of starting Wikipedia pages considered. In the first case preference is given to pages which have the redirection as direct hypernym; in the second one, instead, the condition is relaxed and all the pages that contain an outgoing link to the redirection are considered.

2.6.2 Category taxonomy refinement

The refinement of the category taxonomy aims to address a structural weaknesses, represented by the fact that for a given Wikipedia category the cross-links are missing or limited in number. For this reason it is very difficult to provide hypernyms for this type of categories on the basis of the cross-links which are thus not sufficient to infer all the hypernymy information required. For example, note that the English categories that are associated with 5 or less pages represent the 40% of the total number of the categories in Wikipedia.

We thus designed a simple enrichment heuristic which, applied iteratively until convergence, adds hypernyms to those categories c for which no hypernym could be found in phase 2, i.e., $\nexists c'$ s.t. $(c, c') \in E(T_C)$ (see Figure 2.11). Note that this heuristic does not leverage the cross-links but only the information learned during the application of the algorithm. Given an uncovered category c, we consider its direct Wikipedia super-categories and let each of them vote for their direct hypernym



(a) *IYA* heuristic. **(b)** *ILY* heuristic.

Figure 2.10. Patterns for the coverage refinement of the page taxonomy. Edges in bold represent inferred hypernymy relations.

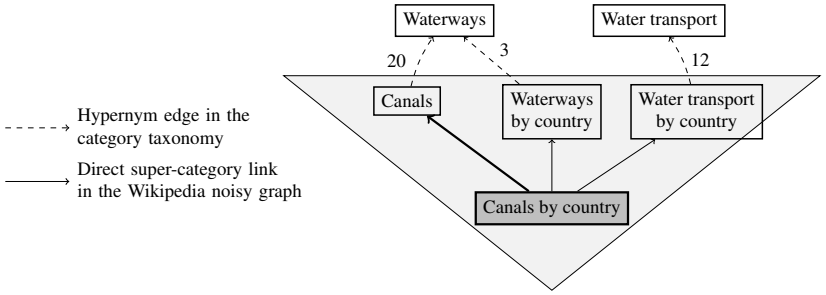


Figure 2.11. Pattern for the category taxonomy refinement.

categories. Then we proceed in decreasing order of vote and select the highest-ranking category c' which is connected to c in T_C. We finally pick up the direct super-category c'' of c which lies in the path from c to c' and add the edge (c, c'') to $E(T_C)$. In the case of ties categories which contributed the most to the score of c'' are favored. For example, as shown in Figure 2.11, given the category CANALS BY COUNTRY, we take all its super-categories (namely CANALS, WATERWAYS BY COUNTRY and WATER TRANSPORT BY COUNTRY) and let each of them vote according to their hypernym categories in T_C. For example WATERWAYS accumulates a score of 23 because during the bitaxonomy algorithm 20 pages contributed to the insertion of the edge (CANALS, WATERWAYS) and 3 pages contributed to the insertion of the edge (WATERWAYS BY COUNTRY , WATERWAYS). Given that WATERWAYS is the most voted hypernym, the algorithm chooses CANALS as hypernym because it is the category which contributes the most to the score of WATERWAYS and therefore adds the edge (CANALS BY COUNTRY, CANALS) to T_C.

2.7 English Bitaxonomy Evaluation

2.7.1 Page taxonomy improvement

After the application of the first phase, in the Wikipedia page taxonomy 359,925 items out of 3,889,572 were still uncovered, i.e., had no hypernym(s) associated (cf. Section 2.4.2). After phases 2 and 3, however, 59,303 total edges are added to the page taxonomy, covering 58,113 nodes, about 15% of the total uncovered pages after the first phase.

2.7.2 Category taxonomy statistics

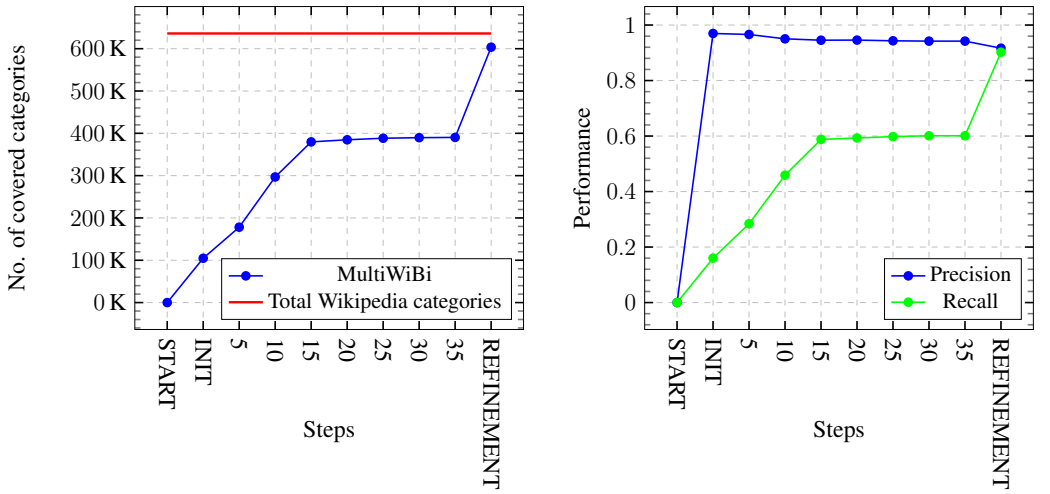
We applied phases 2 and 3 to the output of phase 1, which was evaluated in Section 2.4.3. In Figure 2.12a we show the increase in category coverage at each iteration as well as after phase 3. The final outcome is a category taxonomy which includes 603,590 hypernymy links between categories, covering about 95% of the 635,972 categories in the 2012 English Wikipedia dump. The graph shows the steepest slope in the first iterations of phase 2, which converges around 400k categories at iteration 30, and a significant boost of another 213k hypernymy edges as the result of the refinement (phase 3).

2.7.3 Category taxonomy quality

To estimate the quality of the category taxonomy, we randomly sampled 1,000 categories and, for each of them, we manually associated the super-categories which were deemed to be appropriate hypernyms. Figure 2.12b shows the performance trend as the algorithm iteratively covers more and more categories. Phase 2 is particularly robust across iterations, as it leads to increased recall while retaining very high precision. As regards phase 3, the refinement leads to only a slight precision decrease, while improving recall considerably. Overall, the final taxonomy T_C achieves 91.67% precision, 90.20% recall and 98.40% coverage on our dataset.

2.8 Related Work

Although the extraction of taxonomies from machine-readable dictionaries was already studied in the early 1970s [Calzolari et al., 1973], pioneering work on large amounts of data only appeared in the early 1990s [Hearst, 1992, Ide and Véronis, 1993]. More recently, approaches based on hand-crafted patterns and pattern matching techniques have been developed to provide a supertype for the extracted terms [Navigli and Velardi, 2010, Etzioni et al., 2004, Blohm, 2007, Kozareva and Hovy, 2010b, Velardi et al., 2013, inter alia]. However, these methods do not link terms to existing taxonomies, whereas those that explicitly link do so



(a) Category coverage trend. **(b)** Category precision and recall trend.

Figure 2.12. Category taxonomy evaluation.

by adding new leaves to the existing taxonomy instead of acquiring wide-coverage taxonomies from scratch [Pantel and Ravichandran, 2004, Snow et al., 2006].

The recent upsurge of interest in collaborative knowledge curation has enabled several approaches to large-scale taxonomy acquisition [Hovy et al., 2013]. Most approaches initially focused on the Wikipedia category network, an entangled set of generalization-containment relations between Wikipedia categories, to extract the hypernymy taxonomy as a subset of the network. The first approach of this kind was WikiTaxonomy [Ponzetto and Strube, 2007, 2011], based on simple, yet effective lightweight heuristics, totaling more than 100k is-a relations. Another approach of this type is YAGO [Hoffart et al., 2013, Suchanek et al., 2008] which yields a taxonomical backbone by linking Wikipedia leaf categories to the first (i.e., most frequent) sense of their category heads in WordNet.

Interest in taxonomizing Wikipedia pages, instead, developed with DBpedia [Auer et al., 2007], which pioneered the current stream of work aimed at extracting semi-structured information from Wikipedia templates and infoboxes. In DBpedia, entities are mapped to a coarse-grained ontology which is collaboratively maintained and contains only about 270 classes corresponding to popular named entity types. Freebase [Bollacker et al., 2008] is a later development and a merger of other resources, such as MusicBrainz and ChefMoz. While being also based on infoboxes, it is a set of more than four million topics loosely organized and relies on the collaborative editing of what is now a mixture of both strict and informal relations. The Linked Hypernym Dataset (LHD) [Kliegr et al., 2014] is the most recent effort which tries to taxonomize the Wikipedia encyclopedia by associating Wikipedia articles with a DBpedia entity or a DBpedia ontology concept as their type. The types are hypernyms mined from articles' free text using hand-crafted lexico-syntactic patterns. Furthermore, LHD has been released in two versions:

LHD 1.0 provides hypernyms which can be either DBpedia entities or concepts drawn from the DBpedia ontology (version 3.9), while LHD 2.0 contains concepts drawn from the smaller DBpedia ontology only. To our knowledge LHD is also the only other approach which, in addition to providing hypernyms for Wikipedia pages, also attaches the corresponding ambiguous hypernym lemmas. A few notable efforts to reconcile the two sides of Wikipedia, i.e., pages and categories, have been put forward only very recently: WikiNet [Nastase and Strube, 2013, Nastase et al., 2010] is a project which heuristically exploits different aspects of Wikipedia to obtain a concept network by deriving not only is-a relations, but also other types of relations. A second project, MENTA [de Melo and Weikum, 2010a], creates one of the largest multilingual lexical knowledge bases by interconnecting more than 13M articles in 271 languages. Hypernym extraction, though, is supervised in that decisions are made on the basis of labelled training examples and requires a reconciliation step owing to the heterogeneous nature of the hypernyms.

Finally, our work differs substantially from the others in many respects: first, in marked contrast to most other resources, but similarly to WikiNet and WikiTaxonomy, our resource is self-contained and does not depend on other resources such as WordNet; second, similarly to MENTA and differently from all others, we address the taxonomization task on both sides, i.e., pages and categories, by providing an algorithm which mutually and iteratively transfers knowledge from one side of the bitaxonomy to the other; third, we provide a wide coverage bitaxonomy closer in structure and granularity to a manual WordNet-like taxonomy, in contrast, for example, to DBpedia's flat type-oriented hierarchy.

The following section presents evidence of these facts by comparing statistics about the structure of the taxonomies and by presenting experiments that assess their quality across all other resources.

2.9 Comparative Evaluation

In this section we provide a thorough comparison of MultiWiBi's English bitaxonomy against all the major alternatives in the literature. Section 2.9.1 presents and discusses several dimensions characterizing the different resources. In Section 2.9.2 we report on different measures concerning the taxonomical structure. In Section 2.9.3 we describe the selection and construction of the datasets used, while in Section 2.9.4 we report and discuss the results obtained.

2.9.1 Features of taxonomic resources

In order to examine the differences between MultiWiBi and all other resources analyzed in this chapter, we first present several features in Table 2.1. For each resource we report i) the data timestamp, i.e., the most accurate date of the Wikipedia dumps from which data has been derived, ii) whether the resource provides hypernyms for Wikipedia pages, iii) whether the resource provides hypernyms for Wikipedia

Resource	Timestamp (dd/mm/yy)	Pag.	Categ.	Page hyp. inventory	Categ. hyp. inventory	Type of add. sources	# Languages
MultiWiBi	01/10/12	✓	✓	P+R	C	H,Syntax	271
WikiNet	04/01/12	✓	✓	P+C	P+C	H,Syntax	1 (EN)
DBpedia	01/06/12	✓	✗	D	-	D	125
LHD	10/12	✓	✗	D	-	H,PoS tagger,D	3 (EN, DE, NL)
WikiTaxonomy	01/10/12	✗	✓	-	C	H,Syntax	1 (EN)
YAGO	01/12/12	✗	✓	-	W	H,Syntax,W	1 (EN)
MENTA	10/04/10	✓	✓	P+C+W	P+C+W	H,Syntax,W,WKT	271

Table 2.1. Features of the main taxonomic resources. *W* stands for WordNet, *P* for Wikipedia pages, *R* for Wikipedia redirections, *C* for Wikipedia categories, *D* for DBpedia ontology, *H* for Human and *WKT* for Wiktionary.

categories, iv) the inventory from which the hypernyms are drawn, v) the type of dependencies to language, tools and other resources, and finally vi) the degree of multilingualism, measured as the number of languages covered by the resource at the time of writing.

Timestamp First, as can be seen from Table 2.1 (second column), all resources but MENTA are isochronous: apart from small differences in the reference month, they all come from 2012. This makes comparison much easier.

Wikipedia sides covered and the hypernym inventories For years, resources have been covering only one of the two sides and approaches could mainly be divided into two groups: those which provide hypernyms only for pages and those which provide hypernyms only for categories. Within this classification we can further discriminate on the basis of the inventories from which hypernyms are drawn: i) DBpedia and LHD are consistent and return hypernyms drawn from the DBpedia upper ontology, ii) WikiTaxonomy is also consistent by returning Wikipedia categories and, finally, iii) YAGO uses WordNet synsets as inventory for the categories. In contrast with the four resources above, instead, the other systems presented in Table 2.1 try to cover both two sides of Wikipedia; WikiNet, MENTA and MultiWiBi are, in fact, the only resources which provide hypernyms for both pages and categories. As regards the hypernym inventory, though, they differ substantially. On the one hand WikiNet mixes the two sides of Wikipedia up together and MENTA returns hypernyms in which Wikipedia pages, Wikipedia categories and WordNet synsets are all amassed together. MultiWiBi, in contrast, by returning two separate (but aligned) taxonomies with two disjoint hypernym sets, associates hypernyms in a coherent and consistent manner; as a result, Wikipedia pages have pages and categories have categories as hypernyms.

Dependency from additional sources Another feature, which discriminates the systems in different classes, is the need for any sort of human intervention, additional resource or sense-tagged corpora. The second to last column of Table 2.1 for each resource shows the type of such dependency. The degree with which each resource

is tied to human effort is in turn linked to the ease of converting such resource into another language and, of course, the lesser the better. As regards human intervention, MultiWiBi depends only on the list of stopwords introduced in the syntactic step and does not need any additional human effort. LHD, WikiNet and WikiTaxonomy, instead, heavily rely on lexico-syntactic patterns (e.g., *X by Y* or *X [VBN IN] Y*): LHD learns the patterns using 600 manually annotated training examples for each language; in WikiNet and WikiTaxonomy, instead, patterns are defined by hand, making such pattern-based models at least laborious to generalize across languages. YAGO involves human effort because the category-to-WordNet mappings were corrected by hand, also making it difficult to generalize to many languages automatically. Finally, in the case of MENTA: i) Wikipedia-to-WordNet mappings are established by a supervised linker, trained on 200 manually labelled training examples, ii) the Category-WordNet subclass relationship is learnt thanks to a supervised learning model, trained on 1,539 labelled training mappings. This, however, is done only once for all the languages.

As regards the amount of dependence on external tools, we note that MultiWiBi needs a syntactic parser only for extracting hypernym lemmas in English and for extracting heads from the strings passed to the SSR module. LHD requires only a PoS-tagger in order to train the transducer which learns lexical-syntactic patterns. WikiNet, WikiTaxonomy, YAGO and MENTA all need a syntactic parser to extract heads from categories. Needless to say, the dependency from tools which are language-specific limits the applicability of a system only to languages having such tools. Even though MultiWiBi relies on syntax in the English case, in Section 2.10 we will introduce a new mechanism for extracting hypernym lemmas in other languages which requires nothing but the raw Wikipedia dump in the desired language.

As regards instead the dependence from external resources, we can distinguish between two types of dependency: in some cases approaches use an external resource only as hypernym inventory. This is the case for DBpedia, LHD and YAGO. DBpedia and LHD use the DBpedia Ontology as hypernym inventory (letter D in Table 2.1), while YAGO links Wikipedia categories to WordNet synsets. These systems, however, do not exploit the external resource any further. In contrast, MENTA heavily uses external resources for different purposes: not only countability information about category heads is based both on WordNet and Wiktionary, but its is-a classifier takes as input also the hypernymy information already contained in WordNet. In strong contrast, WikiNet, WikiTaxonomy and MultiWiBi do not lean on any additional information: all these are self-contained taxonomies which exploit data coming solely from Wikipedia itself.

Degree of multilingualism Last, another dimension we considered, strictly intertwined with the previous one, is multilingualism: first of all, note that all the resources are (or could be made) multilingual, thanks to the interlanguage links which connect the different editions of Wikipedia (see Section 2.10 for details). By

merely analysing the resources as they have been publicly released, instead, we note that three resources rely only on the English Wikipedia, namely WikiNet, WikiTaxonomy and YAGO. The possibility of an extension to other languages for these three resources is at the very least arguable for the language limitations explained above. LHD has been applied to 3 languages, and separate data repositories are released independently; DBpedia has been released in 125 languages, which correspond to the separate, isolated chapters administrating each language; MultiWiBi and MENTA are the only two resources which are applicable to every Wikipedia language, making them the only truly language-indepedent approaches (even though the latter relies on WordNet, Wiktionary and labelled training examples).

2.9.2 Structural analysis of the taxonomic resources

Going beyond what can be considered only a qualitative analysis of the resources, we would like to present some comparative evaluation of the structure of the taxonomies. As already mentioned in Section 2.4.3, it is not easy to find valid measures capable of evaluating a taxonomy in a comprehensive manner. Before reporting on precision, recall and coverage we wish to present and discuss several structural measures for all the competitors, both on the page and the category side of Wikipedia (Tables 2.2 and 2.3, respectively). To this end we considered several indicators, among which the number of nodes, edges and leaves contained in the taxonomies, as well as the average height, the number of nodes, the in-degree, etc., as well as a new measure called *granularity*.

Structural features of taxonomies

Page resources Table 2.2 reports the statistics concerning Wikipedia pages. In order to have a reference point, we report the same measures also for WordNet. Note, however, that WordNet contains far fewer concepts, so the number of nodes and edges is not much informative. The first measure that we reckon important is the number of nodes (first row), along with the coverage with respect to the whole Wikipedia (second row). As can be seen, MultiWiBi is the best resource in terms of coverage, with 92.33% of nodes covered by a hypernym.

The second dimension considered is the number of edges (row 3) which express how much hypernymy information a resource contains. As can be seen, MultiWiBi provides the largest number of edges, surpassed only by WikiNet which, however, has a very high average in-degree, meaning that all the leaves point to a few internal nodes. This is due to a phenomenon which we will study in Section 2.9.2, which assesses how many distinct hypernyms each resource offers. The high number of edges returned by WikiNet, in fact, reflects in the fact that a lot of nodes have a lot of hypernyms, not all of which are correct, thus making WikiNet quite a noisy resource.

The most important feature is probably the height of a taxonomy (row 5), measured as the length of the longest hypernymy path linking a leaf to a root. This

Feature	Resources						
	MultiWiBi	**WikiNet**	**DBpedia**	**MENTA**	**LHD 1.0**	**LHD 2.0**	**WordNet**
# nodes	3,600,781	4,071,094	4,000,823	2,980,273	3,038,604	2,960,780	82,115
coverage	92.33%	75.47%	49.18%	66.16%	67.14%	65.94%	-
# edges	4,098,772	14,279,134	2,112,715	2,958,215	3,013,581	2,960,508	84,427
avg. in degree	36.91	1,259.40	6,269.18	24.43	111.34	10,884.22	4.92
avg. height	4.99	1.91	4.23	1.76	1.02	1	8.07

Table 2.2. Structural analysis of the page taxonomies in the literature.

feature gives an idea of the generalization power of a taxonomy, since the longer a path, the more fine grained the generalization. Among all the resources LHD has the lowest height, namely 1, which means that each node in the taxonomy has exactly one hypernym, approximately all distinct from the others. MENTA and WikiNet provide slightly higher taxonomy, followed by DBpedia, with a height of 4.23. MultiWiBi, with an average height of 5, surpasses all other approaches, making it the resource structurally closest to WordNet, which has height 8.07. The DBpedia Ontology, instead, with its 270 ontology classes, makes DBpedia and LHD 2.0 the resources with the highest average in-degree, equal to 6,269 and 10,884, respectively. This is due, of course, to the extremely small size of the hypernym inventories adopted by these resources.

Category resources As regards categories (see Table 2.3), MultiWiBi is the resource with the best coverage. While WikiNet has the highest number of hypernyms, MultiWiBi exhibits the highest height, more than five times as high as any other approach. However, the higher amount of hypernymy information does now correlate with quality (see Section 2.9.3), since the around double number of hypernyms returned by WikiNet is often noisy.

We also note that the average height of the category taxonomy T_C is much greater than that of the page taxonomy T_P, due to the fact that the category taxonomy distinguishes between very subtle classes (such as ALBUMS BY ARTISTS vs. ALBUMS BY RECORDING LOCATION, etc), which get all merged into the same concept ALBUM.

Note that YAGO is the resource with the lowest coverage possible, due to the fact that attention has been paid only to leaf categories. Also here YAGO and MENTA, the only two resources adopting WordNet as hypernym inventory, show the highest average in-degree. This is due to the small size of the hypernym inventory used which reflects in a lot of Wikipedia categories pointing to a small number of hypernyms (WordNet synsets for YAGO, aggregated synsets for MENTA).

Taxonomy granularity

A second important aspect that we analyzed was the granularity of each taxonomy, determined by drawing each resource on a bi-dimensional plane with the number of distinct hypernyms (i.e., non-leaf nodes) on the x axis and the total number of hypernyms (i.e., edges) in the taxonomy on the y axis. Figures 2.13a and 2.13b show

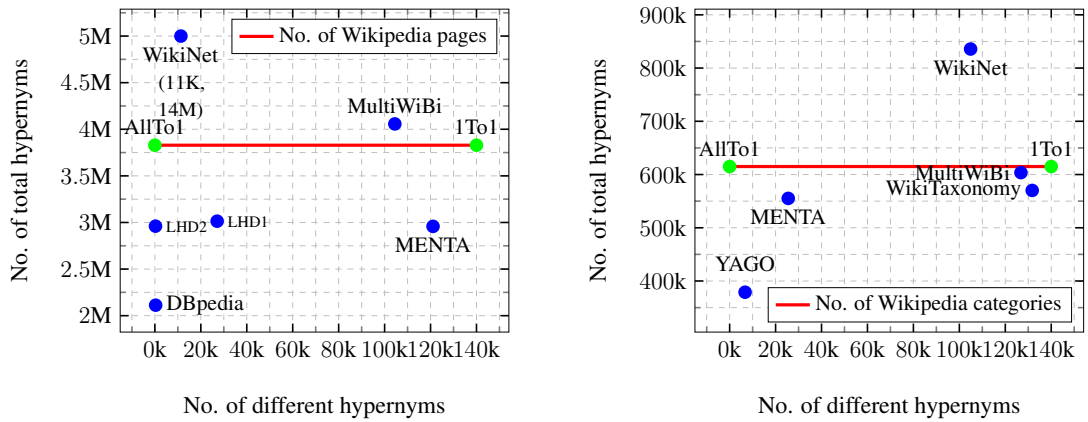
Feature	Resources					
	MultiWiBi	WikiNet	WikiTax	YAGO	MENTA	WordNet
# nodes	605,887	620,870	528,668	588,450	569,643	82,115
# edges	603,557	835,765	570,116	378,942	555,173	84,427
coverage	94.91%	63.44%	56.02%	52.10%	56.18%	-
avg. in degree	4.59	7.97	4.33	56.29	21.76	4.92
avg. height	16.69	3.29	3.46	1	1.33	8.07

Table 2.3. Structural analysis of the category taxonomies in the literature.

the position of each resource for the page and the category taxonomies, respectively. The figures show also two baselines, considered bad systems and essentials to grasp the difference among the position of the different systems in the two-dimensional plane. These two baselines are called *AllTo1* and *1To1* and display two opposite decisions: the former represents the baseline system which always assigns the same hypernym to all the Wikipedia items, while the latter represents the system which assigns a different fictitious hypernym to each Wikipedia item. As can be seen, MultiWiBi, as well as the page taxonomy of MENTA, is the resource with the best granularity, as not only does it attain high coverage, but it also provides a larger variety of classes as generalizations of pages and categories. Specifically, MultiWiBi provides hypernyms for over 4M hypernym pages, chosen from a range of 104k distinct hypernyms, while others exhibit a considerably smaller range of distinct hypernyms (e.g., DBpedia by design, but also WikiNet, with around 11k distinct page hypernyms). The large variety of classes provided by MENTA, however, is due to providing more than 100k Wikipedia categories as page hypernyms (among which, categories about *deaths* and *births* alone represent more than 3% of the distinct hypernyms). As regards categories, while the number of distinct hypernyms of MultiWiBi and WikiTaxonomy is approximately the same (around 130k), the total number of hypernyms returned by WikiTaxonomy (around 580k for both taxonomies) refers to half of the categories covered by MultiWiBi. We remind that the dimension on the y-axis in the Figure represents the number of edges contained in the taxonomy, so the higher number of hypernyms returned by WikiNet overall reflects in a lower coverage. WikiTaxonomy, in fact, has an average out-degree around 1.59, so the number of covered categories is significantly lower (cf. Table 2.3). As regards WikiNet, its large number and variety of category hypernyms is instead counterbalanced by an extremely low precision and recall, as we will show in the experimental results (Section 2.9.3).

2.9.3 Experimental Setup

We compared MultiWiBi against the Wikipedia taxonomies of the major knowledge resources in the literature providing hypernym links, namely DBpedia, WikiNet, MENTA, WikiTaxonomy and YAGO (see Section 2.8). As datasets, we used our gold standards of 1,000 randomly-sampled pages (see Section 2.4.3) and categories (see Section 2.7.3). For all the experiments we used the original outputs from the four resources, but, given the heterogeneity in the release date of the resources, in



(a) Page taxonomies **(b)** Category taxonomies

Figure 2.13. Hypernym granularity for the resources.

order to ensure a level playing field, we detected those pages (categories) which do not exist in any of the above resources and removed them to ensure (potential) full coverage of the dataset across all resources. As already explained in Section 2.9.1, in fact, MENTA is the only resource based on a dump dating back to 2010, a bit far from the others. However, if on the one hand we acknowledge its performance might be relatively higher on a 2012 dump, on the other hand the software for generating MENTA over a different Wikipedia dump is not available and a direct comparison is not possible.[10] WikiTaxonomy, originally based on a 2009 dump, was instead re-implemented in order to align it to same dump used by MultiWiBi. The last column of Table 2.5 reports the size of the leveled datasets after the item deletion.[11]

2.9.4 Results

Wikipedia pages We first report the results of the knowledge resources which provide page hypernyms, i.e., we compare against WikiNet, DBpedia, LHD and MENTA. We show the results on our page hypernym dataset in Table 2.5 (top) using the same measures as defined in Section 2.4.3. As can be seen, all systems but WikiNet exhibit very good precision. WikiNet on one side and DBpedia and LHD 1.0 on the other side stick to the two opposite poles of the precision-recall trend: the former achieves high recall (around 71%) at the cost of a much lower precision (around 57%) due to the high number of hypernyms provided, many of which are incorrect, whereas the latter are characterised by high precision, but low recall. In the case of DBpedia this is also accompanied by low coverage, due to the dependency on the presence of infoboxes in Wikipedia pages. Despite LHD 2.0 being the system with the highest precision, it shows only modest coverage

[10] We contacted the authors, but they stated to be unable to provide us with updated data.

[11] We wish to make very clear that these datasets are the same as those presented in Section 2.4.3 and Section 2.7.3, but are different just in size, because of the item deletion.

	P	R*	C	# items
Lemma	94.83	90.20	98.50	1,000
Lemma LHD	92.78	81.00†	87.30	

Table 2.4. Taxonomy comparison at lemma level. † denotes statistically significant difference, using χ^2 test, $p < 0.01$ between MultiWiBi and LHD.

Dataset	System	P	R*	C	# items
Pages	MultiWiBi	90.76	87.48	94.78	767
	WikiNet	56.86†	71.32†	82.01	
	DBpedia	87.06†	51.50†	55.93	
	MENTA	81.52†	72.49†	88.92	
	LHD 1.0	76.20†	53.85†	70.66	
	LHD 2.0	91.57	63.75†	69.62	
Categories	MultiWiBi	90.65	89.06	98.26	631
	WikiNet	64.05†	49.92†	71.16	
	WikiTax	89.68	55.15†	59.43	
	YAGO	93.58	53.09†	56.74	
	MENTA	87.11	84.63	97.15	
	MENTA$^{-\text{ENT}}$	85.18	71.95†	84.47	

Table 2.5. Page and category taxonomy evaluation. † denotes statistically significant difference, using χ^2 test, $p < 0.01$ between MultiWiBi and the daggered resource.

and recall and an inspection of the answers returned revealed that 32% and 11% of the answers were `http://dbpedia.org/ontology/Agent` and `http://dbpedia.org/ontology/Place`, respectively, which, despite being correct, are very general. MENTA is the closest resource to ours; however, we remark that the hypernyms output by MENTA are very heterogeneous: in fact, 48% of the hypernyms are represented by a WordNet synset, 37% by Wikipedia categories and only 15% by Wikipedia pages. In contrast to all other resources, MultiWiBi outputs hypernyms in a coherent manner, by associating pages with hypernym pages, while at the same time achieving the highest performance of 90.76% precision, 87.48% recall and 94.78% coverage.

Wikipedia categories We then compared all the knowledge resources which deal with categories, i.e., WikiNet, WikiTaxonomy, YAGO and MENTA.

We show the results on our category dataset in Table 2.5 (bottom). MultiWiBi is definitely the best resource when compared against the others: it achieves almost the highest precision (90.65%), the highest recall (89.06%) and the highest coverage (98.26%). WikiNet is the worst system, characterised by the lowest precision and recall. The lowest coverage, between 56% and 59%) is reached by WikiTaxonomy and YAGO: in the former case this is likely due to the inadequacy of lexical-syntactic patterns which do not succeed in capturing all category variants, whereas in the latter

Category	Resources				
	MultiWiBi	**WikiNet**	**WikiTax**	**YAGO**	**MENTA**
Nigerian culture	Culture by nationality	Cultures of Africa	African culture by nationality	-	entity^1_n
Racism in Russia	Racism by country	Black-on-black racism	-	-	entity^1_n
Orellana Province	Provinces of Ecuador	Province capitals of Ecuador	Provinces of Ecuador	-	entity^1_n
Salvadoran cuisine	Latin American cuisine	Latin American culture	Cuisine by nationality	politician^1_n	entity^1_n
Turkish artists	Artists by nationality	-	Artists by nationality	artist^1_n	person^1_n
People from Campania	People by region in Italy	People by region in Italy	People by region in Italy	person^1_n	person^1_n

Table 2.6. Excerpt of the answers given by the different systems on the category dataset.

Dataset	System (X)	MultiWiBi =X	MultiWiBi >X	MultiWiBi <X
Pages	WikiNet	21.51	75.36	3.13
	DBpedia	29.99	59.19	10.82
	MENTA	20.47	54.24	25.29
	LHD 1.0	46.68	46.41	6.91
	LHD 2.0	25.68	64.41	9.91
Categories	WikiNet	42.00	45.32	12.68
	WikiTax	42.31	40.25	17.43
	YAGO	9.51	86.05	4.44
	MENTA	10.62	78.61	10.78

Table 2.7. Specificity comparison.

case this is due to the fact that only leaf categories are considered. MENTA is, again, the closest system to ours, obtaining comparable performance overall. Notably, however, MENTA outputs the first WordNet sense of *entity* for 13% of all the given answers, which, despite being correct and accounted in precision and recall, is uninformative. Since a system which always outputs *entity* would maximise all the three measures, we also calculated the performance for MENTA when discarding *entity* as an answer; as Table 2.5 shows (bottom, $\text{MENTA}^{-\text{ENT}}$), recall drops to 71.95%.

Table 2.6 shows for example the different answers given by the systems for some items in the category dataset. As can be seen, MENTA's answers are quite general and much less specific than those returned by other systems. Further analysis, presented below, shows that the specificity of other systems' hypernyms is considerably lower than that of MultiWiBi.

2.9.5 Taxonomy specificity

To get further insight into our results we performed also an additional analysis by means of a last quality measure. We estimated the level of specialization of the hypernyms in the different resources on our two datasets. The idea is that a

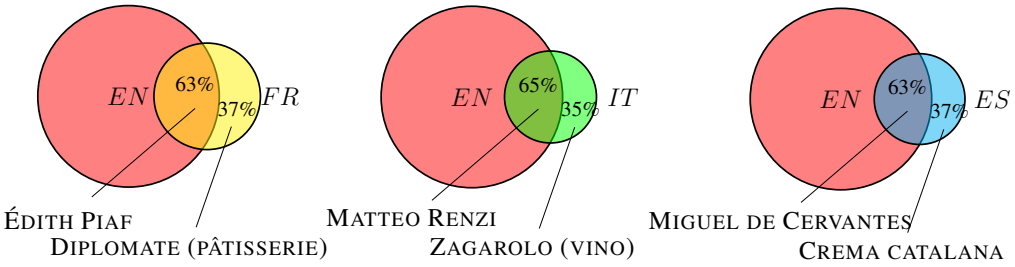


Figure 2.14. Relationship between the English Wikipedia and the versions in other languages.

hypernym should be valid while at the same time being as specific as possible (e.g., SINGER should be preferred over PERSON). We therefore calculated a measure, which we called specificity, that computes the percentage of times a system outputs a more specific answer than another system. To do this, we manually annotated each hypernym returned by a system S with a score, as follows:

$$score(a_S) = \begin{cases} -1 & \text{if } a_S \text{ is wrong} \\ 0 & \text{if } a_S \text{ is missing} \\ x > 0 & \text{if } a_S \text{ is correct and has specificity degree equal to } x \end{cases}$$

Note that higher scores correspond to more specific answers. Since a system S is allowed to return more than one hypernym for each item, for each system we considered only the most specific answer a_S. When comparing two systems S_1 and S_2, we say that S_1 is more specific than S_2 whenever $score(a_{S_1}) > score(a_{S_2})$. We then calculated three types of configurations, depending on $score(a_{S_1})$ was equal, greater or lesser than $score(a_{S_2})$ and denote these with $S_1 = S_2$, $S_1 > S_2$ and $S_1 < S_2$, respectively. More formally:

$$S_1 \odot S_2 := |\{a \in D : score(a_{S_1}) \odot score(a_{S_2})\}| \,/\, |D|$$

where $\odot \in \{=, >, <\}$. Table 2.7 shows the results for all the resources and for both the page and category taxonomies: MultiWiBi consistently provides considerably more specific hypernyms than any other resource (middle column).

2.10 Projecting the Bitaxonomy

We now will explain how to obtain a bitaxonomy in any other language different from English. Before describing the details, it is very important to introduce a new element characterizing Wikipedia, namely the Interlanguage links. This type of connection plays, in fact, a very important piece of information which allows MultiWiBi, as well all other systems, to make the taxonomic information available across many languages. We reveal in advance, though, that MultiWiBi goes beyond

the simple exploitation of such Interlanguage links, by an innovative approach that will be able to cover also Wikipedia items which do not have an English counterpart.

Interlanguage links and the projection rule We begin by showing the definition of interlanguage link extracted of Wikipedia: "*Interlanguage links are links from a page in one Wikipedia language to an equivalent page in another language. [...] For example, the Irish Wikipedia has a page on Ireland titled "Éire", so the English Wikipedia article on Ireland will link to the Irish one, and vice versa*".[12] Thanks to the interlanguage links it is possible to align pages contained in the English Wikipedia to pages in another language, preserving the original meaning during the procedure. Notably, interlanguage links are present also between the categories of a Wikipedia language and the categories of Wikipedia in another language. This kind of links paves the way for a simple, yet effective mechanism which allows to project the hypernymy information coming from a language onto another language. We called this mechanism *Projection rule*. The Projection rule is a technique which exploits the interlanguage links to copy information from a language to another and we will exploit this rule to project the bitaxonomies across languages. Simply put, by means of the interlanguage links, this rule checks whether a given Wikipedia item in a source language (a page or category) and its hypernym exist also in the target language. More formally, the *Projection rule* is defined as follows:

$$X_E \text{ is-a } Y_E \wedge X_E \,||\, X_F \wedge Y_E \,||\, Y_F \;\Rightarrow\; X_F \text{ is-a } Y_F \qquad \forall X_E, Y_E \in T_E, X_F, Y_F \in T_F$$

(Projection rule)

According to this rule, given the English language E and another arbitrary language $F \neq E$, if we know that i) an English page has a hypernym (X_E is-a Y_E), that ii) the English page has an equivalent in language F ($X_E \,||\, X_F$) and that iii) the English hypernym has an equivalent in language F ($Y_E \,||\, Y_F$), then we can safely infer that the latter is a valid hypernym for the foreign page (X_F is-a Y_F). This idea is also depicted in Figure 2.15 with an example. As can be seen, the English page SUBAPICAL CONSONANT has CONSONANT as hypernym and, furthermore, we know that the corresponding Italian equivalents are CONSONANTE SUBAPICALE and CONSONANTE, respectively; under these facts, then, we can derive the fact that the same is-a relation holds between the two corresponding Italian Wikipedia pages (i.e., CONSONANTE SUBAPICALE is-a CONSONANTE).[13] We will draw upon the

[12] `http://en.wikipedia.org/wiki/Help:Interlanguage_links`

[13] The projection rule, though, despite correct in principle, might not hold in the real world. It might happen, in fact, that interlanguage links do not preserve meaning across two languages because they do not align exactly the same concept (e.g., very specific types of snow might not be represented in all the Wikipedia or they might be translated as general snow, without fine-grained distinction). Thus, in order to evaluate, in general in the whole Wikipedia, whether quality was preserved across languages by means of the Projection Rule, we sampled 500 random Wikipedia English pages which presented an interlanguage link in Italian and evaluated the correctness of the links. We found, in practice, only 2 times out of the total (i.e., the 0.004% of the total) the equivalence between aligned concepts did not hold. Note that wrong cases include also alignments which are not completely

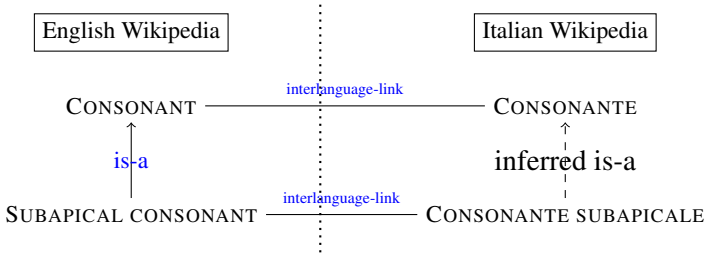


Figure 2.15. Example of the application of the Projection rule. The dashed edge in the Italian Wikipedia (right) represents new is-a information that can be inferred by English (left).

Projection rule basically in two moments, for projecting the English bitaxonomy (see Section 2.10.2) and for building a multilingual gold standard which will enable fair comparison across languages (see Section 2.11).

A limitation of the Interlanguage links As can be seen in Figure 2.14, though, the English Wikipedia overlaps only to a certain extent to Wikipedia in other languages. For example, only 65% of the set of Italian pages has an equivalent in English. With regard to English, Wikipedias in other languages contain additional concepts which either are typical of that very culture and often existing only in that language (such as the Italian page CASTAGNOLE (DOLCE), a typical Italian sweet) or, despite not being culture-specific, represent some other culture's concept (e.g., the French page TEATRO DEL GIGLIO about a famous Italian theatre, which could exist also in English but has not been encoded yet). From here on we will call this set of pages *WEE pages* (pages Without English Equivalent). Therefore, any procedure which relies only on interlanguage links for producing a multilingual taxonomy will have the strong drawback that its application will be limited only to those pages which have an equivalent in English.

Going beyond the interlanguage links We will then present an innovative approach which will overcome this limitation and will provide hypernyms also for those pages which do not have a corresponding page in the English Wikipedia. Our method is general and can be applied to any version and any language of Wikipedia, having as system input only the respective XML Wikipedia dump: none of the algorithmic procedures presented from here on is bound to any language whatsoever. However, for our convenience we present, discuss and evaluate the bitaxonomies in three languages: French (FR), Italian (IT) and Spanish (ES) (see details in Section 2.11).

Note that our approach is completely resource independent because it does not use any other additional resource or tool which is language-specific. Even though the

incorrect: for example the English page MYTHOLOGICAL HYBRID is linked to the Italian page CECAELIA which is a particular mythological hybrid.

syntactic step draws on an English syntactic parser (which exploits English-specific information), we do so only in English, only because these tools are nowadays very performing and fast at the same time. Given that it is a language where researchers and communities have been devoted for decades, we wanted to benefit from it to the maximum extent possible and only then project all the possible information to languages where resources or tools can well be either less powerful or missing at all (e.g., languages such as Tagalog or Latvian).

In order to obtain a full bitaxonomy in a language different from English we will proceed in three steps:

1. **Construction of a Translation Table (TT)**: we will provide a mechanism to build a translation table for a large number of lemmas contained in Wikipedia;
2. **Construction of a Lemma Taxonomy**: we will show how to exploit the translation table built in the previous step to build a lemma taxonomy for another language;
3. **Application of WiBi**: we will apply exactly the same procedure presented for the English case (cf. Sections 2.4-2.6).

We will provide a mechanism which compensates for the lack of a syntactic parser in another language (used in the syntactic step, cf Section 2.4.1), by means of a TT built thanks to Wikipedia itself.

2.10.1 Construction of Translation Tables

In this phase we will show, starting from the English Wikipedia, how to build a Translation Table (TT) for an arbitrary Wikipedia language. The TT will be shortly exploited to translate the hypernym lemmas offered by the English page taxonomy and associate them as hypernym lemmas of Wikipedia pages in the second language (see Section 2.10.2). A translation table can be seen as a sort of bilingual dictionary in which lemmas of the source language are translated into lemmas of a target language. In contrast with standard bilingual dictionaries, though, translation tables generated within this step contain explicit probabilities associated with the translations of a given lemma. An excerpt of the Italian translation table (obtained thanks to this step) is shown in Table 2.8. Here, the (ambiguous) English lemma *plane* is translated in Italian as *piano cartesiano* (the x-y plane) with probability 0.20, as *piano* (the metaphoric sense of *plane*) with probability 0.15, as *pialla* (the carpenter's plane) with probability 0.04, as *aeroplano* (airplane) with probability 0.03, and so on.

We will now present the method through which it is possible to build a TT for an arbitrary Wikipedia language. We will denote with $TT_{E \to F}$ the translation table which has language E and F as source and target language, respectively. The input of the procedure is a lemma l_E in the source language E and the output is a probability distribution $P(\cdot \mid l_E)$ over lemmas in the target language F.

English lemma	Translations
plane	piano_cartesiano:0.20 piano:0.15 pialla:0.04 aeroplano:0.03 aereo:0.023 piano_astrale:0.02 ...
car	automobile:0.33 autovettura:0.11 automobili:0.05 auto:0.02 autovetture:0.01 vettura:0.01 ...
key	chiave:0.37 chiavi:0.03 chiave_crittografica:0.001 chiave_segreta:0.0005 ...

Table 2.8. Excerpt of the English-Italian translation table (numbers indicate translation confidence).

We set up the problem of finding suitable translations for a given lemma by exploiting on the one hand the interlanguage links provided by Wikipedia and on the other hand the association between Wikipedia pages and the associated text anchors occurring in the whole Wikipedia. Figure 2.16 shows this by means of an example. The data on the left side of the figure belong to the English Wikipedia, while data on the right side belong to the Italian Wikipedia. Edges between a lemma and a page represent the fact that that lemma has been linked to that sense and numbers report the number of times the link occurs in Wikipedia. The English lemma *plant* on the left, for example, is linked 10,634 times to the PLANT Wikipedia page, 30 times to FACTORY and so on. The pages linked by an anchor represent the senses that the given lemma can have in different contexts. A similar configuration can be seen on the right side of the figure, where Italian lemmas are linked to the corresponding senses. Note, however, that in general an anchor can link to different senses (*plant* pointing to PLANT, FACTORY, etc.) and a given sense is linked by many anchors (the Italian page FABBRICA is pointed by both *fabbrica* and *stabilimento*). Finally, interlanguage links are shown as undirected edges linking the two sides of the figure; for example the English page PLANT is aligned to the Italian PLANTAE and FACTORY to FABBRICA.

Our hunch is that this network can be exploited to derive translation probabilities. Starting from a given English lemma, in fact, it is possible to reach all its translations in the other language by following the paths which join the two sides. For example, in order to infer that *pianta* is a valid translation for *plant*, it is sufficient to follow the path *plant* $\rightarrow$ PLANT $\rightarrow$ PLANTAE $\rightarrow$ *pianta*. Each path is made of exactly three edges which represent, respectively, i) the association between the source lemma and one of its senses, ii) the interlanguage link between the source and the target senses, and iii) the association between the target sense and the target lemma. By calculating the paths between any pair of lemmas (i.e., if we do this for all the source and target lemmas) we can then easily obtain all the possible translations. Note that in general, however, there might be more than a single path between any two lemmas (for instance when two ambiguous lemmas share the same senses across the two languages) and it is thus necessary to take into account all the paths which link the source lemma in a language to another lemma in the other language.

We are now then ready to define formally the probabilities provided by a translation table. Given a lemma l_E in the source language E, we define the probability

that the lemma l_F in the target language F is a translation for l_E as:

$$P(l_F|l_E) := \frac{1}{Z} \cdot \sum_{\substack{p_E \in O(l_E) \\ p_F \in O(l_F)}} P(p_E|l_E) \cdot P(p_F|p_E) \cdot P(l_F|p_F) \tag{2.1}$$

with:

$$P(p_E|l_E) := \frac{c(l_E \rightarrow p_E)}{\sum_{p'_E \in O(l_E)} c(l_E \rightarrow p'_E)} \tag{2.2}$$

$$P(p_F|p_E) := \begin{cases} 1 & \text{if there exists an interlanguage link between } p_E \text{ and } p_F \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

$$P(l_F|p_F) := \frac{c(l_F \rightarrow p_F)}{\sum_{l'_F \in I(p_F)} c(l'_F \rightarrow p_F)} \tag{2.4}$$

where Z is a normalization constant, $O(l_X)$ denotes the set of senses linked by l_X in language X and $c(l_X \rightarrow p_X)$ is the count of how many times l_X points to p_X in language X.

The Equation 2.1 is a probability over all the paths going from l_E and ending into l_F. Each of the terms in the sum represents the probability of a single path and is made of three terms: i) the probability of having l_E linking to p_E (first term $P(p_E|l_E)$, Equation 2.2), ii) the probability of having p_E aligned to p_F (second term $P(p_F|p_E)$, Equation 2.3), and iii) the probability of having p_F linked by a lemma l_F in that language (third term $P(l_F|p_F)$, Equation 2.4).

For example, given the term $l_E = plant$, the probability $P(p_E = \text{PLANT} \mid l_E = plant)$ is .99, the probability $P(p_E = \text{FACTORY} \mid l_E = plant)$ is .001, while the probability $P(p_E = \text{PLANT (PERSON)} \mid l_E = plant)$ is .00023.

By adapting Equation 2.1 to the case of $l_E = plant$ and $l_F = pianta$ (i.e., the probability that *plant* translates into the Italian *pianta*) is:

$$P(pianta \mid plant) = \sum_{\substack{p_E \in O(plant) \\ p_F \in O(pianta)}} P(p_E \mid plant) \times P(p_I|p_E) \times P(pianta \mid p_I)$$

The set $O(plant)$ includes for example PLANT, FACTORY and FLOWERING PLANT, among others. The only path with three non-zero terms is *plant* $\rightarrow$ PLANT $\rightarrow$ PLANTAE. Since $P(\text{PLANT} \mid plant) = .99$, $P(\text{PLANTAE}|\text{PLANT}) = 1$ and $P(pianta \mid \text{PLANTAE}) = .38$, the overall product $P(pianta \mid plant) = .99 \times 1 \times .38 = .37$.

In conclusion, as a result of the application of the technique described above, we have obtained one TT for each Wikipedia language. Each TT will then be used to associate hypernym lemmas with Wikipedia pages in the particular language ($TT_{E \rightarrow F}$ will be used for French, $TT_{E \rightarrow I}$ will be used for Italian and $TT_{E \rightarrow S}$ will be used for Spanish).

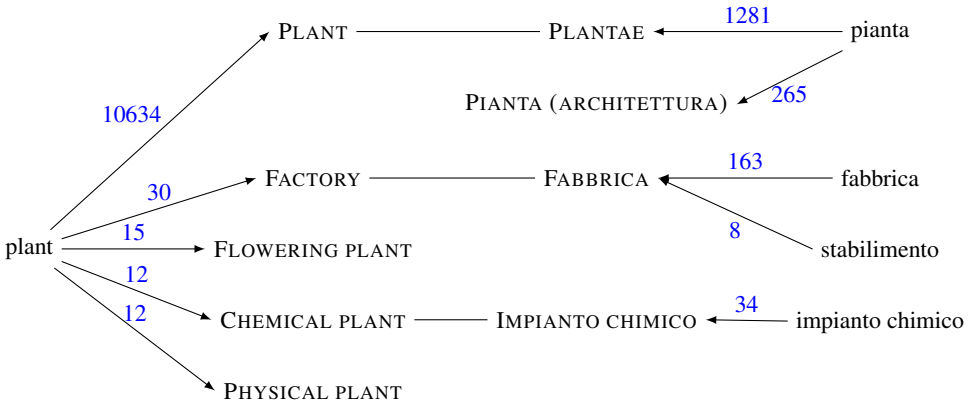


Figure 2.16. Paths connecting the surface anchor *pianta* in Italian to the surface anchor *plant* in English. Numbers report the number of times the link $a \rightarrow p$ between the anchor a and the page p occurs in Wikipedia.

2.10.2 Extraction of Multilingual Lemmas

At this point we have shown a mechanism for obtaining automatically translation tables which provide translations of lemmas from a source language into a target language, notably by using Wikipedia only.

The aim of this Section is to show that the Translation Tables, in conjunction with the Interlanguage links, can be exploited to provide hypernym lemmas for any page in an arbitrary Wikipedia language. The assignment of hypernym lemmas is based on heuristics which exploit four different sources of information: i) the interlanguage links between two Wikipedia languages; ii) the TT and the local context of a given page (such as its gloss, its categories, etc.); iii) the context provided by the sisters of a given page (basically, the distribution of hypernym lemmas of the sister pages); iv) global features of both Wikipedia and the hypernym lemmas discovered up to this point across all the Wikipedia pages. These heuristics are applied in the same order of presentation, in cascade order.

Exploiting the interlanguage links The first heuristic exploits the interlanguage links by means of the application of the Projection rule. The hypernym lemma assigned to the page in the second language is the title lemma of the projected hypernym. Thanks to this heuristic, for example, the Wikipedia page MADRID is assigned the hypernym lemma *ciudad*. This is, in fact, the title lemma of the hypernym CIUDAD which is aligned to CITY, the hypernym for the English concept corresponding to MADRID. However, note that this heuristic could not cover concepts which are not covered in English (which are covered, instead, by subsequent heuristics).

Exploiting the Translation Tables and the local context This heuristic draws on the TT presented in Section 2.10.1 and thus represents the first effort to translate

a hypernym lemma associated with an English page to its equivalent in a given language. At this step only local features are exploited, such as the page's gloss and the titles of its category. Starting from the English hypernym lemma l_E, this heuristic considers in decreasing order of probability all the translations of l_E and checks whether one of these is contained within the gloss of p or in some of the category titles of p. For instance, the hypernym lemma for the Italian page KARL POPPER is *filosofo*; in fact, since in English KARL POPPER has *philosopher* as hypernym lemma, the heuristic considers all its translations, including *filosofo*, *filosofia*, *filosofi*, *filosofa*, etc. Since the Italian gloss for KARL POPPER *"Popper è anche considerato un filosofo politico di statura considerevole, difensore della democrazia e del liberalismo [...]"* contains the translation *filosofo*, the latter is promoted to hypernym lemma of this page.

Exploiting context provided by sister pages In order to cover also those pages of a language which do not have an equivalent in English, we designed another heuristic which draws on the sister pages of a given page and exploits the distribution of hypernym lemmas already discovered for these.

The heuristic considers in decreasing importance the distribution of hypernym lemmas of p's sisters and assigns to p the first hypernym lemma which is contained in the gloss of p_F or in some of the categories of p. For example, with this heuristic the Wikipedia French page YAHOO! MESSENGER is assigned the hypernym lemma *logiciel*, because contained in the following categories: LOGICIEL PROPRIÉTAIRE, LOGICIEL DE MESSAGERIE INSTANTANÉE, LOGICIEL POUR MAC OS and LOGICIEL POUR UNIX.

This heuristic provides two nice added values: first of all, since it exploits the neighborhood of a page, pages which do not have an equivalent in English have the opportunity to be covered. For example this heuristic succeeds to identify *actrice* as hypernym lemma for the page French STÉPHANIE REYNAUD even though this page is available in French only. The other added value is that, given it exploits features which go beyond the mere gloss of a page, it is able to extract suitable hypernyms even when the pages's gloss does not contain the hypernym lemma or contains a lemma which is less specific than expected (e.g., the gloss for the page PLATINUM TOWER is "El Platinum Tower es una lujosa edificación [...]" and the lemma contained therein is *edificación*, which is less specific than the expected *rascacielos*).

Exploiting global features There is still a non-negligible fraction of Wikipedia WEE pages which, however, are still in their early stage and thus suffer from problems of content. For example more than 30% of Italian WEE pages lack a Wikipedia category, and 25% of these do not even have a definition. Note that the first group of heuristics above cannot be applied to this class of pages, since there is no English equivalent. The second group of heuristics, instead, might work here, but only on those pages which have categories whose meaning is related to the

hypernym lemma to be discovered, i.e., such that they bring in sister pages with a valid hypernym lemma. For example, the second group cannot be useful on the Spanish page HEMISFERIO NORTE (i.e., NORTHERN HEMISPHERE in English), because its only category is GEOGRAFÍA (i.e., GEOGRAPHY in English).

To overcome this limitation, we propose an approach called *Mimic* that overcomes this problem by considering global features. This heuristic takes into account all the possible content available for a given page, by considering i) the page's gloss (when present), ii) its categories (when present) and iii) the words of the title appearing between parentheses (e.g., the word *fiume* in the title TICINO (FIUME)). Given this context, all the possible n-grams are then collected (with $n \leq 5$), and the n-gram that maximises the following formula is promoted to hypernym lemma.

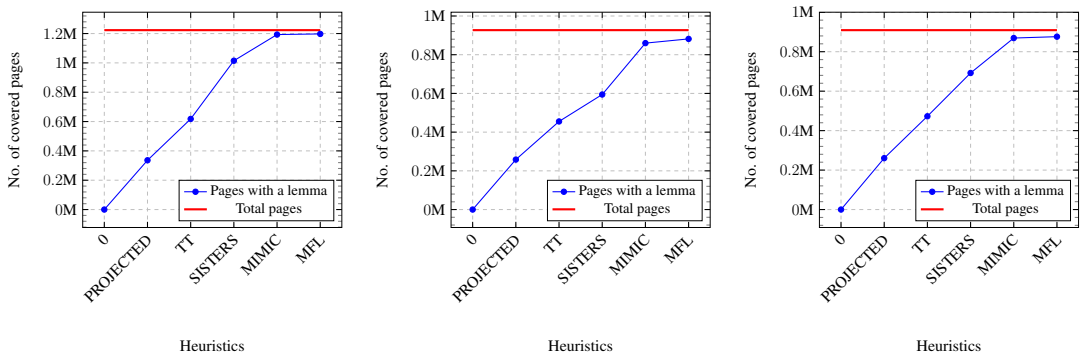
$$score(w) = f_w \cdot igf_w$$

where f_w is the frequency of the word gram w as hypernym lemma (as obtained after the previous heuristics) and igf_w (inverse gloss frequency) is the inverse frequency of w across all Wikipedia's glosses. The former prefers word grams which are common hypernym lemmas across the whole Wikipedia lemma taxonomy built so far, the latter favours specific word grams (such as the Italian *brano musicale*, whose igf is $\frac{1}{2761}$, vs. *brano*, whose igf is $\frac{1}{4298}$). For instance, consider the Italian page CERCAMI (RENATO ZERO), about a famous song of an Italian singer, whose gloss is "Cercami è un famoso brano di Renato Zero, secondo singolo estratto dall'album Amore dopo amore del 1998". The word grams extracted for this page include, among others, *Renato Zero*, *Cercami*, *brano*, *famoso brano*, *secondo*, etc. Among these, only four have also been assigned as hypernyms in the Italian taxonomy, namely *brano* (98 times, i.e., $f_{brano} = 98$), *singolo* ($f_{singolo} = 85$), *secondo* ($f_{secondo} = 16$) and *zero* ($f_{zero} = 7$). The corresponding igf are $\frac{1}{4,298}$ for *brano*, $\frac{1}{9501}$ for *singolo*, $\frac{1}{15667}$ for *secondo* and $\frac{1}{660}$ for *zero*; *brano* is thus the n-gram which is finally preferred, with a score equal to $98 \cdot \frac{1}{4,298} = 0.023$ and thus the candidate which is promoted as hypernym lemma of CERCAMI (RENATO ZERO).

Finally, for all those pages having an equivalent in English for which none of the above heuristics succeeded in assigning a hypernym lemma, we backoff to the MFT, the Most Frequent Translation l_F of l_E. For instance, the lemma assigned to the Spanish page FREDERICK LUGARD is *explorador*, since the latter is the translation with the highest probability for the English hypernym lemma *explorer*.

2.10.3 Statistics for the Multilingual Hypernym Lemmas

Figure 2.17 reports the coverage during the application of the heuristics for each language. As can be seen, coverage increases consistently as more heuristics are applied and about the 80% of the total pages are covered at lemma level in each language.



(a) French **(b)** Italian **(c)** Spanish

Figure 2.17. Multilingual lemma taxonomy coverage with the different heuristics.

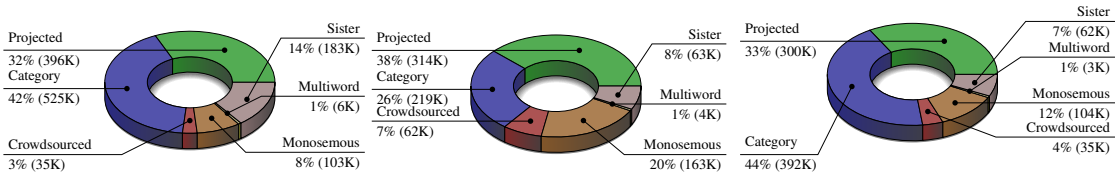
Note that the trend seems very similar across the different languages and we attribute this phenomenon to a similar distribution of hypernym lemmas across the editions of Wikipedia.

2.10.4 Construction of the multilingual Page Taxonomies

Now that we have obtained a lemma taxonomy for a foreign language, our aim is to build a taxonomy also at the sense level. Note that we are in the same situation as we were in the English case, right before the application of the semantic step (see Section 2.4.2). We can then re-apply exactly the same hypernym linkers adopted for the English page taxonomy (cf. Section 2.4.2), with the exception of the Distributional linker which assumes the availability of a PoS-tagger in the language.

Statistics for the Multilingual Page Taxonomies

Figure 2.18 shows the distribution of the disambiguated hypernym lemmas across the languages. Differently from the case of English, the pies include on the one hand the hypernyms coming from the application of the Projection rule and on the other hand do not display information regarding the Distributional linker. As can be seen, apart from the projected edges, the distribution of the linkers varies according to the language considered. This does not hold for the projected is-a information, since we have seen that the overlap between English and the three languages does not change significantly and we expect the same amount of information being transferred across languages. As regards the other linkers, the is-a edges returned by the Category linker represent a substantial fraction of the total. In terms of number of links, the impact of the Monosemous linker is comparable to that of the Distributional linker in the English case, while the Multiword linker proves to be marginally contributing. Overall, we extracted 1,246,524 is-a relations for French, 825,465 for Italian and 895,301 for Spanish, providing hypernyms respectively for 1,116,330 pages out of 1,221,845 (91%), 742,796 pages out of 926,129 (80%) and 809,410 pages out of



(a) French **(b)** Italian **(c)** Spanish

Figure 2.18. Distribution of multilingual disambiguated hypernyms.

908,820 (89%).

2.10.5 Running the Bitaxonomy Algorithm on the multilingual taxonomies

As done in the English case (cf. Section 2.5), in order to obtain a bitaxonomy in the foreign language (i.e., a taxonomy for the page side and a taxonomy for the category side of Wikipedia) we start the iterative algorithm i) over the partial page taxonomy built so far and ii) over a raw category taxonomy initialised exactly in the same manner as explained in Section 2.5.2.

2.10.6 Refinement of the multilingual taxonomies

Exactly as done in Section 2.6, we apply the refinement step also on the bitaxonomy obtained in the other language after the application of the Bitaxonomy algorithm.

2.10.7 Statistics for the Multilingual Category Taxonomies

We made two types of experiments, reported in Figure 2.19: the first one (blue, lower line), in which we start the category taxonomy as explained in Section 2.5.2; a second one (green, upper line), in which, before initialising the category taxonomy, we apply the Projection rule to each vertex of the English category taxonomy (see Projection rule), in order to project as many hypernym edges as possible from English to the language of interest.

Figure 2.19 reports the coverage trend for categories when applying the iterative algorithm. We note that, similarly to the case of English, coverage progressively increases until iteration 25, where it finally reaches a plateau. Thanks to the application of the category refinement step, the gap with respect to the total is greatly reduced, reaching approximately full coverage of all Wikipedia categories. As can be seen in the figure (green line, x-label START), the initialisation which exploits the Projection rule yields a taxonomy which is already covered for about one third for all languages. Interestingly, though, after the Bitaxonomy algorithm and the category refinement step have been applied, the two lines reconcile approximately

at the same point, meaning that starting with a projected category taxonomy does not necessarily reflect into a significantly greater coverage.

2.10.8 Analysis of the Page taxonomies across language

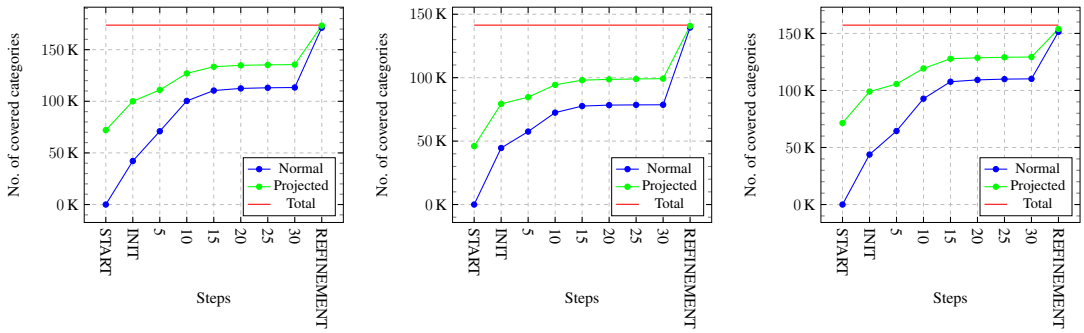
WEE pages After running MultiWiBi on a different language of Wikipedia, in fact, what we obtain is an augmented bitaxonomy which overlaps to a certain extent with the concepts contained in the English Wikipedia but which also differs significantly from it. We thus distinguished among four types of taxonomised pages:

- **Pages with projected hypernyms**. Pages which *do* have a corresponding page in English and whose hypernyms all coincide with those found in English (i.e., have the same hypernym aligned across languages). For example, the Italian page SCIROPPO D'ACERO, aligned to the English MAPLE SYRUP, has the Italian page SCIROPPO as hypernym, aligned in turn to the English SYRUP.

- **Pages with different hypernyms**. Pages which *do* have a corresponding page in English but the concept expressed by their hypernyms differs with respect to those represented by the equivalent concept in English. This happens when the hypernym in English has a different specificity: in the case of French, RICHARD II D'ANGLETERRE has MONARQUE as hypernym but KING OF ENGLAND in English (note that KING OF ENGLAND has no equivalent in French, but MONARQUE aligns to MONARCH in English).

- **Pages with hypernyms not aligned.** Pages which *do* have a corresponding page in English but whose hypernyms are not aligned to those returned in English. This is the case, for example, for ARNOLFO DI CAMBIO which has the hypernym SCULPTEUR in French and the hypernym SCULPTOR in English, but the two are not aligned across the two languages because both redirections. This class represents, thus,false negatives because the hypernyms returned in the languages, while expressing exactly the same concept, are only considered different at the surface level only because no automatic equivalence has been established so far in Wikipedia.

- **WEE pages with a hypernym**. Pages which *do not* have a corresponding page in English and for which one or more hypernyms has been found. This class is of extremely importance to us, because hypernyms belonging to this class are unique, in the sense this information could not be derived from English by means of the simple Projection rule. The pages included in this class are usually (even though not always, see discussion in Section 2.10) culture-specific: here we find the Italian Wikipedia page SAN GIMIGNANO (VINO) about a famous Italian wine which has no English counterpart and which has been taxonomised as VINO (WINE).



(a) French **(b)** Italian **(c)** Spanish

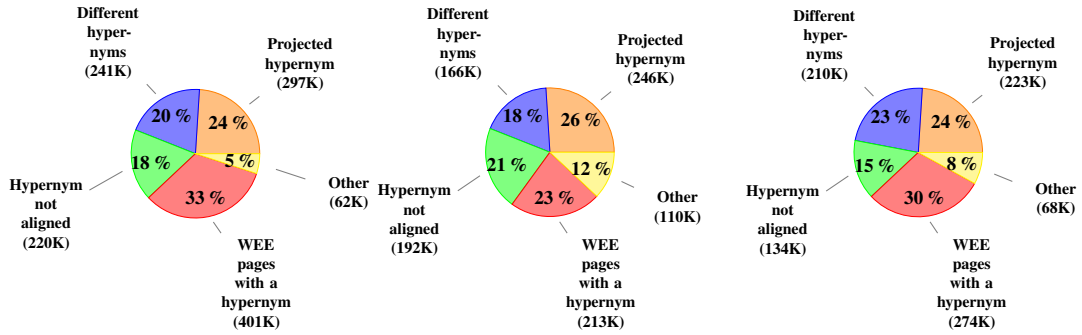
Figure 2.19. Multilingual category taxonomy coverage over the iterations.

Finally, there is a fraction of pages which have not been taxonomised and thus do not have hypernyms in the final foreign taxonomy. The latter, however, often includes WEE pages: for example, the Italian E PENSO A TE (ALBUM) has no English counterpart and MultiWiBi did not succeed to taxonomise it (even though the hypernym lemma extraction step managed to associate the lemma *album*). This class, however, also includes a small fraction of pages which are lexicalized also in English but which have either not been taxonomised in one of the two languages or have not been taxonomised in neither of the two.

In Figure 2.20 we plotted the distribution of the four types of pages in the different languages. Each pie reports the percentage as well as the absolute number of pages included in each type with regard to the total. As can be seen, roughly 60% of the taxonomised pages are also lexicalized in English and, within this set i) about one third has hypernyms which have been projected and thus share their hypernyms with the English bitaxonomy, ii) one third has hypernyms with different granularity, and iii) one third has hypernyms which are not aligned. As regards WEE pages, instead, MultiWiBi managed to disambiguate 86%, 66% and 80% of hypernym lemmas for French, Italian and Spanish, respectively. This is a very important piece of information that states clearly the amount of potentially language-specific information that MultiWiBi is able to extract. Just to grasp the ground-breaking effect derived by covering this type of concepts, in Table 2.9 we report a selected list of WEE pages for French, Italian and Spanish MultiWiBi managed to find suitable hypernyms for. Even though these might also be contained in other Wikipedia languages, they have the characteristic of not being lexicalized in English and thus represent additional concepts which MultiWiBi succeeded to taxonomize.

2.11 Multilingual evaluation

We will now present the experimental setup in a multilingual setting and show the results of the multilingual bitaxonomies. We will show the creation of the gold standards in Section 2.11.1, present results at lemma level in Section 2.11.2 and



(a) French **(b)** Italian **(c)** Spanish

Figure 2.20. Characterization of pages and their hypernyms in other languages.

French	Italian	Spanish
Langhe Chardonnay	Trenette al pesto	Crema catalana
Diplomate (pâtisserie)	Lasagne (gastronomia)	Baldomero Fernández Moreno
Gâche de Vendée	Paccheri	Luis García Montero
Savarin	Maltagliati	El Rey (canción)
Fromages en Provinois	Nebbione	Cantares (canción de Joan Manuel Serrat)
Chécy (fromage)	Piemonte Brachetto	No soy de aquí ni soy de allá
Le Brebiou	Zagarolo (vino)	Mediterráneo (canción)

Table 2.9. Example of culture-specific concepts.

at sense level in Section 2.11.3 and 2.11.3 for the page and the category sides, respectively.

2.11.1 Experimental setup

(Automatically) projected gold standards Our aim is to estimate the quality of the multilingual bitaxonomies obtained and compare these with the English bitaxonomy. We could have sampled new articles in the new language from scratch, but not only this would have forced us to annotate multiple datasets in vain but, above all, it would also have not guaranteed comparability of the obtained results across languages. For these reasons, we decided to exploit the datasets presented in Section 2.4.3 by projecting them automatically. The obtained datasets can then be considered the multilingual versions of the English datasets. Of course full projection coverage cannot be obtained because not all the articles are aligned across all language editions of Wikipedia (especially if we consider that interlanguage links are added manually). After the projection we obtained 256 pages for French, 218 for Italian and 205 for Spanish. We report results on i) the subset of the English pages which we managed to translate in another language and then ii) on a new set of pages which we used to fill out translated datasets. After having identified the set

of pages which compose the multilingual datasets, due to the lack of interlanguage links between lemmas, we manually provided hypernym lemmas for each of these pages, similarly as done in English. Thus, while datasets for pages and categories were obtained automatically thanks to the Projection rule, datasets for lemmas still required human intervention.

(Manual) WEE gold standards In order to evaluate the overall quality of the approach also in the other languages we manually created additional datasets which contain a certain number of language-specific articles. From now on, these datasets will be called D_F^{WEE} (dataset of WEE in language F). The criterion followed to decide how many WEE pages to add was that of preserving the balance between pages which exist also in English and pages which do not. Consider for example the relation between the French and the English Wikipedia, shown in Figure 2.14; since the ratio $r_F = \frac{|F \cap E|}{|F|}$ is 0.62, in the final dataset we should have that the number P_F of articles which we succeeded to project automatically from the English dataset should be the 63% of the total. Since, in general, P_F is different than expected, the size of D_F^{WEE} should take into account both P_F and r_F, such that the proportion is preserved. Thus, to preserve the balance between pages belonging to the intersection and language-specific pages, we sampled $|D_F^{WEE}| = P_F \times \frac{1-r_F}{r_F}$ random pages which exist in the language of interest but not in English.[14] The first factor takes into account the number of pages which we succeed to project; the second factor represents the coefficient that preserves the proportion.

2.11.2 Results for Multilingual hypernym lemmas

In this section we provide experimental evidence about the quality of the taxonomies at lemma level. In particular, we will also i) compare the quality of the automatic translation procedure (described in Section 2.10.2) against a tool-based syntactic lemma extraction (similarly as done in the English setting, see Section 2.4.1) and ii) analyse and discuss the results when considering WEE pages.

Results for automatically translated hypernym lemmas As can be seen in Table 2.10 (rows 'Projection'), the automatic assignment of hypernym lemmas provides very good results. As regards the projected datasets, we can see that more than 99% of pages have at least one lemma, which means that practically all the pages in the other languages are covered at lemma-level. Also quality is very high: while in Italian precision and recall are around 70%, in the other two languages they are both very high, between 80% and 85%. This means that the heuristics presented in Section 2.10.2 managed either to translate (whenever possible) or identify meaningful hypernym lemmas in the context provided by the Wikipedia pages.

[14] This is obtained by setting up the proportion $P_F : |D_F^{WEE}| = r_F : (1 - r_F)$, which leads to $|D_F^{WEE}| = P_F \times \frac{1-r_F}{r_F}$

Language	Setting	P	R*	C	# items
FR	Projection	81.76	82.42	99.61	256
	Syntactic	82.41	69.53	84.38	
	Syntactic + sisters	78.69	75.00	95.31	
	WEE	76.67	74.19	96.77	155
IT	Projection	71.68	73.39	99.08	218
	Syntactic	-	-	-	-
	Syntactic + sisters	-	-	-	-
	WEE	58.77	57.76	98.28	116
ES	Projection	83.95	84.39	99.51	205
	Syntactic	70.00	65.37	90.24	
	Syntactic + sisters	67.57	68.78	98.54	
	WEE	74.55	66.67	89.43	123

Table 2.10. Multilingual lemma taxonomy quality.

Comparison with language-specific hypernym lemma extractors What if, as we did for the English version, we used a language-specific syntactic parser to extract hypernym lemmas for the Wikipedia pages in another language? To test whether the lemma taxonomy actually benefits from using a syntactic parser designed ad-hoc for that language, on each specific version of Wikipedia we syntactically extracted all the terms involved in a dependency relation corresponding to the English *cop*. Since the names of the dependency relations change across the languages, labels were provided manually: in French, for instance, in French we identified the *ast* relation of the Malt syntactic parser[15]; in Spanish we used the *att* relation output by the FreeLing syntactic parser.[16]. As done in English, whenever possible, we also employed the list of stop hypernym lemmas, also manually translated from English (e.g., *variety* was translated into *variedad* in Spanish and *variété* in French), and exploited the relations corresponding to the English *conj_and* and *conj_or* relations as well (*coord* in French and *co-n* in Spanish). Unfortunately, it was not possible to find a syntactic parser for Italian.[17] In Table 2.10 (rows 'Syntactic') we show the performance on the lemma extraction, compared to those obtained by automatically translating the lemmas from the English taxonomy.

Similarly to Section 2.4.1, in Table 2.10 (rows 'Syntactic' and 'Syntactic + sisters'), we report the different performance when the *vanilla syntactic setting* is used and when the local context is also considered. As can be seen, coverage in the vanilla syntactic setting is not very high, in complete analogy with the English case. When also sister pages are considered, instead, coverage increases consistently,

[15] `http://www.maltparser.org`

[16] `http://nlp.lsi.upc.edu/freeling/`

[17] We have submitted a request on the Web site `http://ai-nlp.info.uniroma2.it/external/chaosproject/` but we never received any reply with a link where to download the software.

almost totally covering the page inventory and going on par with the one reached in the automatic setting. Also precision and recall, initially far, converge at around 75% in French and 69% in Spanish, but even in this setting a significant gap in performance is observed when comparing against the lemma taxonomy obtained thanks to the automatic lemma translation procedure. The higher performance of the latter setting is also due to the broader context made available to the lemma extraction heuristics which make extensive use not only of translation information about the English hypernym lemma, but also of local context such as the gloss of the page and its categories.

In conclusion we can say that the automatic projection of hypernym lemmas provides better hypernym translations overall, with a significant gap in precision and recall when compared with the setting in which a syntactic parser is exploited. This phenomenon is likely due to two factors: on the one hand the heuristics used to project the taxonomical information exploit more context than that made available to the syntactic parsers, on the other hand the latters might not be as mature as the English Stanford counterpart and still need more extensive training data as input.

Lemma extraction for WEE pages Table 2.10 (rows 'WEE'), reports also on the performance on the D_F^{WEE} gold standards. Also when considering WEE pages, the same trend of the pages with an English equivalent are observed: Italian ranks last both in precision and recall, with only around 57% performance, while the other two languages show performance around 75%. Note that coverage is very high also in this setting: this shows that, even though it is not possible to achieve the same results obtained when an English counterpart is available, MultiWiBi somewhat compensates for this by drawing on the context provided by the sister pages and the global features.

2.11.3 Results for Multilingual Page taxonomies

Results for the multilingual page taxonomies are presented in Table 2.11 for the three languages, also compared with the possible alternative approaches, namely DBpedia and MENTA. LHD was excluded from the comparison because not available in any of the three languages, while WikiNet was excluded because the average coverage on the normal datasets and the D_F^{WEE} datasets was around 49.51% and 5.46%, respectively.

As can be seen, performance are very high for all the languages when the projected datasets are considered and, to a certain extent, they are comparable with the results obtained in English. For all the three languages we observe around 85% precision, 80% recall and 93% coverage. This results might be expected by considering that these datasets are somewhat related to their English equivalents, even though hypernyms found by MultiWiBi in the two settings is different (see Section 2.10.8): because of the application of the hypernym linkers, hypernyms found in different languages might change considerably, causing relatively small

Language	Resource	P	R*	C	# items
FR	MultiWiBi	84.51	80.86	94.14	256
	MENTA	81.37	48.83†	59.77	
	DBpedia	79.61	25.00†	29.69	
IT	MultiWiBi	84.33	78.44	92.20	218
	MENTA	82.72	51.83†	62.39	
	DBpedia	91.98	55.50†	60.09	
ES	MultiWiBi	86.98	81.95	93.66	205
	MENTA	81.02	42.93†	52.68	
	DBpedia	65.15†	32.20†	48.29	

Table 2.11. Multilingual page taxonomy evaluation. † denotes statistically significant difference, using χ^2 test, $p < 0.01$ between MultiWiBi and the daggered resource.

differences in the evaluation results.

We also show results for the D_F^{WEE} datasets in Table 2.12. We observe that it was not possible to obtain performance comparable to those of English. First of all we point out that results greatly vary according to the reference language: for example, French exhibits very good results with regard to the three measures, achieving 76% precision and around 71% recall. Italian and Spanish have a lower coverage than French and also lower performance in general. Spanish, however, despite affected by low recall stands up with a quite high 69.47% precision.

These results are, however, very promising because we are the true first resource which provides this kind of information: no other resource is able to provide a comparable coverage or quality in hypernym extraction on WEE pages. As can be seen in the table, all other alternative approaches suffer from critical coverage problems. In addition to this, their answers, despite being correct, are either very generic (DBpedia returning `http://dbpedia.org/ontology/Monument` instead of ÉGLISE (ÉDIFICE)) or uninformative (MENTA returning 1970 BIRTHS instead of ÉCRIVAIN FRANÇAIS).

MultiWiBi represents thus the first approach which manages to extract specific-language information automatically from Wikipedia with performance that are, however, depending on the language and not on the methodology. We believe that the results correlate with the quality of the individual Wikipedia languages, being very disparate at the time of writing.

2.11.4 Results for Multilingual Category taxonomies

In Table 2.13 we report the results of the application of the Bitaxonomy Algorithm on the category taxonomy. The only competitor we compared against is MENTA, because this proved to be the closest competitor to our approach in terms of performance in the English experimental setup. Note also that YAGO and WikiTaxonomy could not be compared anyway, because it does not exists a multilingual version of

Language	Setting	P	R*	C	# items
FR	MultiWiBi	76.39	70.97	92.90	155
	MENTA	85.71	34.84†	40.65	
	DBpedia	100.00	16.77†	16.77	
IT	MultiWiBi	55.56	43.10	77.59	116
	MENTA	91.67	37.93	41.38	
	DBpedia	100.00	43.97	43.97	
ES	MultiWiBi	69.47	53.66	77.24	123
	MENTA	76.19	26.02†	34.15	
	DBpedia	66.13	33.33†	50.41	

Table 2.12. Multilingual WEE page taxonomy evaluation. † denotes statistically significant difference, using χ^2 test, $p < 0.01$ between MultiWiBi and the daggered resource.

Language	Resource	P	R*	C	# items
FR	MultiWiBi (Not projected)	78.07	74.17	95.00	140
	MultiWiBi (Projected)	80.71	80.71	100.0	
	MENTA	82.61	55.00†	65.71	
IT	MultiWiBi (Not projected)	86.67	82.28	94.94	91
	MultiWiBi (Projected)	83.54	83.54	100.0	
	MENTA	75.86	49.45†	63.74	
ES	MultiWiBi (Not projected)	84.75	81.97	96.72	131
	MultiWiBi (Projected)	84.43	84.43	100.0	
	MENTA	80.46	54.20†	66.41	

Table 2.13. Multilingual category taxonomy evaluation. † denotes statistically significant difference, using χ^2 test, $p < 0.01$ between MultiWiBi and the daggered resource.

these systems (see Section 2.9.1).

The table shows the difference in quality between the two category taxonomies obtained by projecting the English category taxonomy or not (cf. Section 2.13). Results without and with the application of the Projection rule are presented in rows labeled with 'Not projected' and 'Projected', respectively. We can see that, in all the languages, projecting the category taxonomy from English greatly benefits the category taxonomies in the other languages. First of all we point out that, by means of the automatic projection, we achieved full coverage on the category dataset. Second, apart precisions in Italian and Spanish (however, the 0.32 decrement in Spanish can be considered negligible), all other measures exhibit a remarkable increase, ranging between 1.27 and 5.83.

Language	P	R*	C	# items
EN	42.29	52.35	90.71	2,316,053
FR	32.03	23.59	68.50	732,687
IT	47.50	42.17	77.19	544,590
ES	33.32	35.47	86.57	621,125

Table 2.14. Multilingual page taxonomy evaluation when considering Wikidata as the gold standard (multilingual setting).

2.11.5 Automatic multilingual evaluation using Wikidata as gold standard

In Table 2.14 we report the results when evaluating the page taxonomy against Wikidata. This is an automatic evaluation, so it might well be the case that valid answers are not accounted as such, just because not contained in the gold standard. The size of the gold standard prevented us to manually edit the gold standard by including MultiWiBi correct answers. For example, FREDDIE MERCURY is annotated as a HUMAN in Wikidata, but MultiWiBi outputs SINGER-SONGWRITER as its hypernym: the two, despite being different, are in fact one the specialization of the other and should be accounted as valid (actually, MultiWiBi outputs a hypernym which is even more specific than Wikidata). Thus, to understand the impact of the automatic vs. the manual evaluation, we inspected and manually corrected the top 50 most frequent wrong answers and calculated performance again. Results showed that both precision and recall increased considerably, to 48.67 and 67.35 respectively.

2.12 The impact of 2014

1,047,476 pages have seen their glosses change between 2012 and 2014: this is a number which represents more than the 27% of the total glosses available in 2012. Inspired by this observation, we thus decided to repeat the most important experiments presented in this chapter by simply updating MultiWiBi's input data to a more recent Wikipedia version, in order to assess the potential increase in performance. It makes no senses to compare against other competitors, because their resources are outdated to 2012, so results are reported only for MultiWiBi in Table 2.15.

Lemma extraction For assessing the impact of the update to 2014 on lemmas, we use the same evaluation dataset used in Section 2.9 containing information at the lemma level for 1,000 pages. Out of the total, 56 pages have been removed (48 have become redirections, 8 have been removed), so the gold standard in 2014 contains 944 pages in total.

	P	R*	C	# items
Lemma	92.27	92.58	98.94	944
Lemma '14	92.05 (-0.22)	92.80 (+0.22)	98.73 (-0.21)	
Sense	81.47	77.33	94.92	944
Sense '14	88.78 (+7.30)	82.94 (+5.61)	93.43 (-1.48)	
Category	91.33	90.76	99.37	952
Category '14	90.39 (-0.94)	88.97 (-1.79)	98.42 (-0.95)	

Table 2.15. Overall English Bitaxonomy performance when using 2014 dumps.

Page taxonomy As can be seen in Table 2.15 (rows Sense vs. Sense '14) the quality of the page taxonomy has benefited greatly from the update to 2014. Despite a little decrease in coverage, precision and recall have received a boost of more than 5%.

Category taxonomy As regards categories, 48 categories were removed in Wikipedia between 2012 and 2014, so the final dataset contains 952 categories. In this case, however, categories do not exhibit the same trend observed for the page taxonomy and show a little decrease in performance. This phenomenon might be due to Wikipedians being more intent on enriching Wikipedia articles than categories.

After the update of the data to the 2014, the overall performance of the English bitaxonomy has greatly increased. Not only this is an interesting behaviour per se, but we also think that the significant change is very promising and reflects a significant change in the quality of the underlying Wikipedia data, continuously improved by its collaborators.

2.13 Conclusions

In this chapter we have presented MultiWiBi, a language independent approach for constructing bitaxonomies of Wikipedia in any language, where each bitaxonomy is made of two taxonomies which establish is-a relations between Wikipedia articles and categories, respectively. For each langauge, the approach is mainly divided in three phases. The first phase aims at building a taxonomy for the page side of Wikipedia; the second phase triggers an iterative algorithm that incrementally populates a taxonomy for the category side of Wikipedia by exploiting the interlanguage links existing between the two sides; the third phase is aimed at solving some structural problem affecting the structure of Wikipedia categories so as to output a polished category taxonomy.

Our contribution is multiple. First, the two taxonomies of each bitaxonomy are aligned (pages are aligned to categories) and the bitaxonomies in the different languages are aligned (concepts in English are aligned to the corresponding concepts in all other languages). Second, in strong contrast to others, our work crucially

pivots on the English edition of Wikipedia for inducing taxonomies in the other languages, without relying on any external resource, parallel corpus or tool at all. Third, experiments show that our bitaxonomies are characterized by higher accuracy and specificity than all past and current competitors, making MultiWiBi the best set of taxonomies in the literature at the moment.

In the future work we might experiment further on directions remained unexplored, such as the integration of all the bitaxonomies into a single, unified multi-bitaxonomy, or the integration of MultiWiBi into well-known knowledge bases.

Chapter 3

SPred

3.1 Introduction

Acquiring semantic knowledge from text automatically is a long-standing issue in Computational Linguistics and Artificial Intelligence. Over the last decade or so the enormous abundance of information and data that has become available has made it possible to extract huge amounts of patterns and named entities [Etzioni et al., 2005], semantic lexicons for categories of interest [Thelen and Riloff, 2002, Igo and Riloff, 2009], large domain glossaries [De Benedictis et al., 2013] and lists of concepts [Katz et al., 2003]. Recently, the availability of Wikipedia and other collaborative resources has considerably boosted research on several aspects of knowledge acquisition [Hovy et al., 2013], leading to the creation of several large-scale knowledge resources, such as DBPedia [Bizer et al., 2009b], BabelNet [Navigli and Ponzetto, 2012a], YAGO [Hoffart et al., 2013], MENTA [de Melo and Weikum, 2010a], to name but a few. This wealth of acquired knowledge is known to have a positive impact on important fields such as Information Retrieval [Chu-Carroll and Prager, 2007], Information Extraction [Krause et al., 2012], Question Answering [Ferrucci et al., 2010] and Textual Entailment [Berant et al., 2012, Stern and Dagan, 2012].

Not only are these knowledge resources obtained by acquiring concepts and named entities, but they also provide semantic relations between them. These relations are extracted from unstructured or semi-structured text using ontology learning from scratch [Velardi et al., 2013] and Open Information Extraction techniques [Etzioni et al., 2005, Yates et al., 2007, Wu and Weld, 2010, Fader et al., 2011, Moro and Navigli, 2013] which mainly stem from seminal work on *is-a* relation acquisition [Hearst, 1992] and subsequent developments [Girju et al., 2003, Pasca, 2004, Snow et al., 2004, among others].

However, these knowledge resources still lack semantic information about language units such as phrases and collocations. For instance, which semantic classes are expected as a direct object of the verb *break*? What kinds of noun does the adjective *amazing* collocate with? Recognition of the need for systems that are aware

a	full [[bottle]]	of milk
a	nice hot [[cup]]	of milk
a	cold [[glass]]	of milk
a	very big bottle	of milk
a	brand	of milk
a	constituent	of milk

Table 3.1. An excerpt of the token sequences which match the lexical predicate *a * of milk* in Wikipedia (filling argument shown in the second column; following the Wikipedia convention we provide links in double square brackets).

of the selectional restrictions of verbs and, more in general, of textual expressions, dates back to several decades [Wilks, 1975], but today it is more relevant than ever, as is testified by the current interest in semantic class learning [Kozareva et al., 2008] and supertype acquisition [Kozareva and Hovy, 2010a]. These approaches leverage lexico-syntactic patterns and input seeds to recursively learn the semantic classes of relation arguments. However, they require the manual selection of one or more seeds for each pattern of interest, and this selection influences the amount and kind of semantic classes to be learned. Furthermore, the learned classes are not directly linked to existing resources such as WordNet [Fellbaum, 1998] or Wikipedia.

The goal of our research is to create a large-scale repository of semantic predicates whose lexical arguments are replaced by their semantic classes. For example, given the textual expression *break a toe* we want to create the corresponding semantic predicate *break a* BODY PART, where BODY PART is a class comprising several lexical realizations, such as *leg*, *arm*, *foot*, etc.

In this chapter we provide three main contributions:

- We propose SPred, a novel approach which harvests predicates from Wikipedia and generalizes them by leveraging core concepts from WordNet.
- We create a large-scale resource made up of semantic predicates.
- We demonstrate the high quality of our semantic predicates, as well as the generality of our approach, also in comparison with our closest competitor.

3.2 Preliminaries

We introduce two preliminary definitions which we use in our approach.

Definition 1 (lexical predicate). A lexical predicate

$$w_1\ w_2\ \ldots\ w_i\ *\ w_{i+1}\ \ldots\ w_n$$

is a regular expression, where w_j are tokens ($j = 1, \ldots, n$), $*$ matches any sequence of one or more tokens, and $i \in \{0, \ldots, n\}$. We call the token sequence which matches $*$ the filling argument of the predicate.

For example, *a * of milk* matches occurrences such as *a full bottle of milk*, *a glass of milk*, *a carton of milk*, etc. While in principle * could match any sequence of words, since we aim at generalizing nouns, in what follows we allow $*$ to match only noun phrases (e.g., *glass*, *hot cup*, *very big bottle*, etc.).

Definition 2 (semantic predicate). A semantic predicate is a sequence

$$w_1\ w_2\ \ldots\ w_i\ c\ w_{i+1} \ldots\ w_n$$

where w_j are tokens ($j = 1, \ldots, n$), $c \in C$ is a semantic class selected from a fixed set C of classes, and $i \in \{0, \ldots, n\}$.

As an example, consider the semantic predicate *cup of* BEVERAGE,[1] where BEVERAGE is a semantic class representing beverages. This predicate matches phrases like *cup of coffee*, *cup of tea*, etc., but not *cup of sky*. Other examples include: MUSICAL INSTRUMENT *is played by*, *a* CONTAINER *of milk*, *break* AGREEMENT, etc.

Semantic predicates mix the lexical information of a given lexical predicate with the explicit semantic modeling of its argument. Importantly, the same lexical predicate can have different classes as its argument, like *cup of* FOOD vs. *cup of* BEVERAGE. Note, however, that different classes might convey different semantics for the same lexical predicate, such as *cup of* COUNTRY, referring to cup as a prize instead of cup as a container.

3.3 Large-Scale Harvesting of Semantic Predicates

The goal of the work presented in this chapter is to provide a fully automatic approach for the creation of a large repository of semantic predicates in three phases. For each lexical predicate of interest (e.g., *break* $*$):

1. We extract all its possible filling arguments from Wikipedia, e.g., *lease, contract, leg, arm*, etc. (Section 3.3.1).

2. We disambiguate as many filling arguments as possible using Wikipedia, obtaining a set of corresponding Wikipedia pages, e.g., *Lease*, *Contract*, etc. (Section 3.3.2).

3. We create the semantic predicates by generalizing the Wikipedia pages to their most suitable semantic classes, e.g., *break* AGREEMENT, *break* LIMB, etc. (Section 3.3.3).

[1] In what follows we denote the SEMANTIC CLASS in small capitals and the *lexical predicate* in italics.

Figure 3.1. Argument cloud for the lexical predicate *cup of* *. Bigger words correspond to arguments with higher frequency.

We can then exploit the learned semantic predicates to assign the most suitable semantic class to new filling arguments for the given lexical predicate (Section 3.3.4).

3.3.1 Extraction of Filling Arguments

Let π be an input lexical predicate (e.g., *break* $*$). We search the English Wikipedia for all the token sequences which match π, resulting in a list of noun phrases filling the $*$ argument. We show an excerpt of the output obtained when searching Wikipedia for the arguments of the lexical predicate *a * of milk* in Table 3.1. As can be seen, a wide range of noun phrases are extracted, from quantities such as *glass* and *cup* to other aspects, such as *brand* and *constituent*.

The output of this first step is a set L_π of triples (a, s, l) of filling arguments a matching the lexical predicate π in a sentence s of the Wikipedia corpus, with a potentially linked to a page l (e.g., see the top 3 rows in Table 3.1; $l = \epsilon$ if no link is provided, see bottom rows of the Table).[2] Note that Wikipedia is the only possible corpus that can be used here for at least two reasons: first, in order to extract relevant arguments, we need a large corpus of a definitional nature; second, we need wide-coverage semantic annotations of filling arguments.

Figure 3.1 shows the cloud of arguments harvested for the lexical predicate *cup of* *. Bigger words correspond to arguments with higher frequency. As can be seen, expected (ambiguous) arguments, such as *coffee*, *tea*, *water*, etc. are well represented in the Figure. Also less expected arguments, though, are collected as a result of the matching procedure, such as *powder*, *rice*, *wine*, etc. The aim of the next phase will be to disambiguate these arguments, whereas the last phase will

[2]We will also refer to l as the *sense* of a in sentence s.

generalize the disambiguated arguments up to semantic classes.

Extraction optimization In order to handle an arbitrary number of lexical predicates, we store the set of lexical predicates L in a trie and then perform the extraction step in an optimized way which uses this data structure. A trie is a special n-ary tree, where each node contains a piece of information, such as a letter or a word. Figure 3.2 shows the trie used to store all the lexical predicates starting with the token '*to*', such as *to break * down*, *to break into **, *to * cup of*, etc. The root node (i.e., that containing the token "*to*") subsumes other sub-tries representing the corresponding subset of lexical predicates. For example, the sub-trie rooted in *break* represents all those lexical predicates which start with "*to break*". Leaf nodes represent the final lexical predicates; for example the leaf node labelled with "*of*" is solely associated with the lexical predicate *to * cup of*. This data structure is particularly useful when performing a sort of generalized binary search. In fact, assume we have the following sentence:

You need to be able to break *problems* down *into smaller parts.*

and assume also that we are on the node labelled with *break* in the trie, after having analyzed the word *break* in the sentence. Since the word coming after *break* in the sentence is "*problems*", we are interested only in those lexical predicates which go ahead with the $*$ character or with the word *problems*. Thanks to the trie we can quickly discard all the subtries (children of *break*) whose token does not match the next word; in the example, we can discard the subtrie anchored in *into*, but we let the match algorithm go onto the node anchored with $*$.

Note that the introduced mechanism based on a trie has the nice added value to be much faster than a simple matching procedure: thanks to the trie-based search algorithm, in fact, only predicates whose tokens match the current word being analyzed are retrieved, whereas as encountered words do not match some sub-trie's token, the latter gets subsequently discarded. In contrast, the simple routine that tries to match each predicate $\pi \in L$ against each sentence in the corpus has a remarkable overhead due to the useless iteration over predicates which have empty intersection in terms of tokens with the text being analyzed.

3.3.2 Disambiguation of Filling Arguments

The objective of the second step is to disambiguate as many arguments in L_π as possible for the lexical predicate π. We denote $D_\pi = \{(a, s, l) : l \neq \epsilon\} \subseteq L_\pi$ as the set of those arguments originally linked to the corresponding Wikipedia page (like the top three linked arguments in Table 3.1). Therefore, in the rest of this section we will focus only on the remaining triples $(a, s, \epsilon) \in U_\pi$, where $U_\pi = L_\pi \setminus D_\pi$, i.e., those triples whose arguments are not semantically annotated. Our goal is to replace ϵ with an appropriate sense, i.e., page, for a. For each such triple $(a, s, \epsilon) \in U_\pi$, we apply the following disambiguation heuristics:

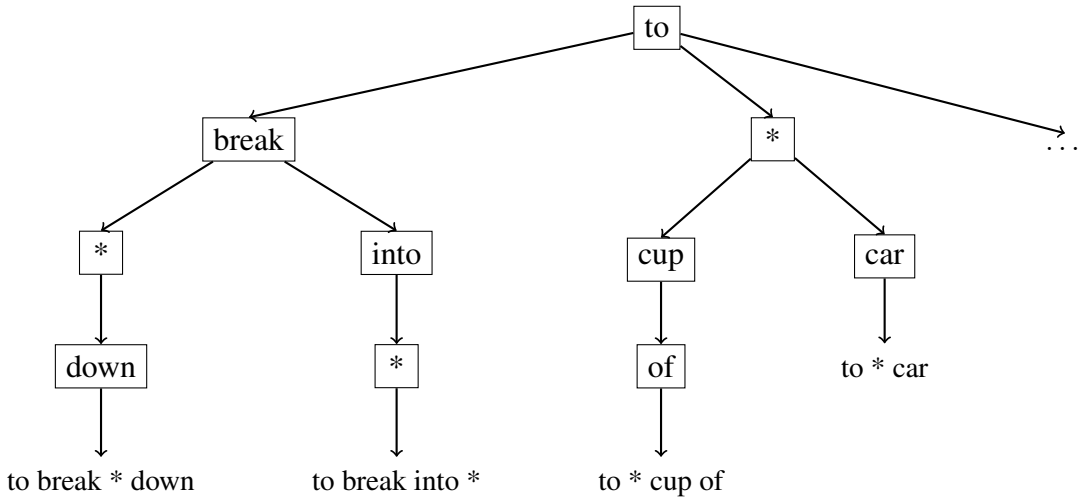


Figure 3.2. Trie-based representation of a set of lexical predicates starting with the token "*to*" (e.g., *to break * down*, *to * cup of*, etc).

- **One sense per page:** if another occurrence of a in the same Wikipedia page (independent of the lexical predicate) is linked to a page l, then remove (a, s, ϵ) from U_π and add (a, s, l) to D_π. In other words, we propagate an existing annotation of a in the same Wikipedia page and apply it to our ambiguous item. For instance, *cup of coffee* appears in the Wikipedia page *Energy drink* in the sentence "[...] energy drinks contain more caffeine than a strong *cup of coffee*", but this occurrence of *coffee* is not linked. However the second paragraph contains the sentence "[[Coffee]], tea and other naturally caffeinated beverages are usually not considered energy drinks", where *coffee* is linked to the *Coffee* page. This heuristic naturally reflects the broadly known assumption about lexical ambiguity presented in [Yarowsky, 1995], namely the one-sense-per-discourse heuristic.

- **One sense per lexical predicate:** if $\exists(a, s', l) \in D_\pi$, then remove (a, s, ϵ) from U_π and add (a, s, l) to D_π. If multiple senses of a are available, choose the most frequent one in D_π. For example, in the page *Singaporean cuisine* the occurrence of *coffee* in the sentence "[...] combined with a *cup of coffee* and a half-boiled egg" is not linked, but we have collected many other occurrences, all linked to the *Coffee* page, so this link gets propagated to our ambiguous item as well. This heuristic mimes the one-sense-per-collocation heuristic presented in [Yarowsky, 1995].

- **Trust the inventory:** if Wikipedia provides only one sense for a, i.e., only one page title whose lemma is a, link a to that page. Consider the instance "At that point, Smith threw down a *cup of Gatorade*" in page *Jimmy Clausen*; there is only one sense for *Gatorade* in Wikipedia, so we link the unannotated occurrence to it.

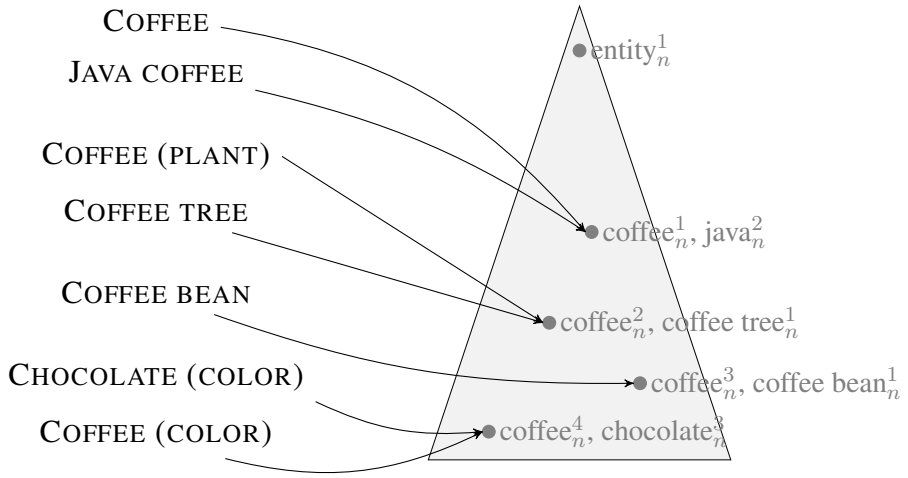


Figure 3.3. Wikipedia-WordNet mapping for the senses of *coffee* in WordNet.

As a result, the initial set of disambiguated arguments in D_π is augmented with all those triples for which any of the above three heuristics apply. Note that D_π might contain the same argument several times, occurring in different sentences and linked many times to the same page or to different pages. Notably, the discovery of new links is made through one scan of Wikipedia per heuristic. The three disambiguation strategies, applied in the same order as presented above, contribute to promoting the most relevant sense for a given word.

Finally, let A be the set of arguments in D_π, i.e., $A := \{a : \exists(a, s, l) \in D_\pi\}$. For each argument $a \in A$ we select the majority sense $sense(a)$ of a and collect the corresponding set of sentences $sent(a)$ marked with that sense. Formally, $sense(a) := \arg\max_l |\{(x, y, z) \in D_\pi : x = a \wedge z = l\}|$ and $sent(a) := \{s : (a, s, sense(a)) \in D_\pi\}$.

3.3.3 Generalization to Semantic Classes

Our final objective is to generalize the annotated arguments to semantic classes picked out from a fixed set C of classes. As explained below, we assume the set C to be made up of representative synsets from WordNet. We perform this in two substeps: we first link all our disambiguated arguments to WordNet (Section 3.3.3) and then leverage the WordNet taxonomy to populate the semantic classes in C (Section 3.3.3).

Linking to WordNet

So far the arguments in D_π have been semantically annotated with the Wikipedia pages they refer to. However, using Wikipedia as our sense inventory is not desirable; in fact, contrarily to other commonly used lexical-semantic networks such as WordNet, Wikipedia is not formally organized in a structured, taxonomic

hierarchy. While it is true that attached to each Wikipedia page there are one or more categories, these categories just provide shallow information about the class the page belongs to. Indeed, categories are not ideal for representing the semantic classes of a Wikipedia page for at least three reasons: i) many categories do not express taxonomic information (e.g., the English page *Albert Einstein* provides categories such as DEATHS FROM ABDOMINAL AORTIC ANEURYSM and INSTITUTE FOR ADVANCED STUDY FACULTY); ii) categories are mostly structured in a directed acyclic graph with multiple parents per category (even worse, cycles are possible in principle); iii) there is no clear way of identifying core semantic classes from the large set of available categories. Although efforts towards the automatic taxonomization of Wikipedia categories do exist in the literature [Ponzetto and Strube, 2011, Nastase and Strube, 2013], the results are of a lower quality than a hand-built lexical resource. Therefore, as was done in previous work [Mihalcea and Moldovan, 2001, Ciaramita and Altun, 2006, Izquierdo et al., 2009, Erk and McCarthy, 2009, Huang and Riloff, 2010], we pick out our semantic classes C from WordNet and leverage its manually-curated taxonomy to associate our arguments with the most suitable class. This way we avoid building a new taxonomy and shift the problem to that of projecting the Wikipedia pages – associated with annotated filling arguments – to synsets in WordNet. We address this problem in two steps:

Wikipedia-WordNet mapping. We exploit an existing mapping implemented in BabelNet [Navigli and Ponzetto, 2012a], a wide-coverage multilingual semantic network that integrates Wikipedia and WordNet.[3] Based on a disambiguation algorithm, BabelNet establishes a mapping $\mu : Wikipages \rightarrow Synsets$ which links about 50,000 pages to their most suitable WordNet senses.[4] Figure 3.3 shows the mapping from Wikipedia to WordNet for the four senses of *coffee* in WordNet. As can be seen, COFFEE and JAVA COFFEE (and a lot of redirections to this two pages, not shown in the figure) get mapped to $coffee_n^1$ (the coffee sense); COFFEE (PLANT) and COFFEE TREE are mapped to $coffee_n^2$, and so on. Note that the mapping might well regard redirections: for example COFFEE TREE is a redirection to COFFEA and it is mapped to the second sense of coffee in WordNet (whereas the Wikipedia page COFFEA is mapped to the WordNet synset $\{Coffea_n^1, genus\ Coffea_n^1\}$ representing the genus of the coffee trees).

Mapping extension. Nevertheless, BabelNet is able to solve the problem only partially, because it still leaves the vast majority of the 4 million English Wikipedia pages unmapped. This is mainly due to the encyclopedic nature of most pages, which do not have a counterpart in the WordNet dictionary. To address this issue, for each unmapped Wikipedia page p we obtain its textual definition as the first

[3] `http://babelnet.org`

[4] We follow [Navigli, 2009] and denote with w_p^i the i-th sense of w in WordNet with part of speech p.

$P_{class}(c\|\pi)$	c	$support(c)$
0.189	wine^1_n	wine, sack, white wine, red wine, wine in china, madeira wine, claret, kosher wine
0.180	coffee^1_n	turkish coffee, drip coffee, espresso, coffee, cappucino, caffè latte, decaffeinated coffee, latte
0.114	herb^2_n	green tea, indian tea, black tea, orange pekoe tea, tea
0.110	water^1_n	water, seawater
0.053	beverage^1_n	chinese tea, 3.2% beer, orange soda, boiled water, hot chocolate, hot cocoa, tejuino, cider, beverage, cocoa, coffee milk, lemonade, orange juice
0.040	milk^1_n	skim milk, milk, cultured buttermilk, whole milk
0.035	beer^1_n	3.2% beer, beer
0.027	alcohol^1_n	mead, umeshu, kava, rice wine, jägermeister, kvass, sake, gin, rum
0.018	poison^1_n	poison

Table 3.2. Highest-probability semantic classes for the lexical predicate π = *cup of* *, according to our set C of semantic classes.

sentence of the page.[5] Next, we extract the hypernym from the textual definition of p by applying Word-Class Lattices [Navigli and Velardi, 2010, WCL[6]], a domain-independent hypernym extraction system successfully applied to taxonomy learning from scratch [Velardi et al., 2013] and freely available online [Faralli and Navigli, 2013]. If a hypernym h is successfully extracted and h is linked to a Wikipedia page p' for which $\mu(p')$ is defined, then we extend the mapping by setting $\mu(p) := \mu(p')$. For instance, the mapping provided by BabelNet does not provide any link for the page *Peter Spence*; thanks to WCL, though, we are able to set the page *Journalist* as its hypernym, and link it to the WordNet synset *journalist*^1_n.

This way our mapping extension now covers 539,954 pages, i.e., more than an order of magnitude greater than the number of pages originally covered by the BabelNet mapping.

Populating the Semantic Classes

We now proceed to populating the semantic classes in C with the annotated arguments obtained for the lexical predicate π.

Definition 3 (semantic class of a synset). The semantic class for a WordNet synset S is the class c among those in C which is the most specific hypernym of S according to the WordNet taxonomy.

For instance, given the synset *tap water*^1_n, its semantic class is *water*^1_n (while the other more general subsumers in C are not considered, e.g., *compound*^2_n, *chemical*^1_n, *liquid*^3_n, etc).

For each argument $a \in A$ for which a Wikipedia-to-WordNet mapping $\mu(sense(a))$ could be established as a result of the linking procedure described above, we associate a with the semantic class of $\mu(sense(a))$. For example, consider the case in

[5] According to the Wikipedia guidelines, "The article should begin with a short declarative sentence, answering two questions for the nonspecialist reader: *What (or who) is the subject?* and *Why is this subject notable?*", extracted from `http://en.wikipedia.org/wiki/Wikipedia:Writing_better_articles`.

[6] `http://lcl.uniroma1.it/wcl`

which a is equal to *tap water* and $sense(a)$ is equal to the Wikipedia page *Tap water*, in turn mapped to *tap water*^1_n via μ; we thus associate *tap water* with its semantic class *water*^1_n. If more than one class can be found we add a to each of them.[7]

Ultimately, for each class $c \in C$, we obtain a set *support*(c) made up of all the arguments $a \in A$ associated with c. For instance, *support*(*beverage*^1_n) = { *chinese tea*, *3.2% beer*, *hot cocoa*, *cider*, ..., *orange juice* }. Note that, thanks to our extended mapping (cf. Section 3.3.3), the support of a class can also contain arguments not covered in WordNet (e.g., *hot cocoa* and *tejuino*).

Since not all classes are equally relevant to the lexical predicate π, we estimate the conditional probability of each class $c \in C$ given π on the basis of the number of sentences which contain an argument in that class. Formally:

$$P_{class}(c|\pi) = \frac{\sum_{a \in support(c)} |sent(a)|}{Z}, \tag{3.1}$$

where Z is a normalization factor calculated as $Z = \sum_{c' \in C} \sum_{a \in support(c')} |sent(a)|$. As an example, in Table 3.2 we show the highest-probability classes for the lexical predicate *cup of* $*$.

As a result of the probabilistic association of each semantic class c with a target lexical predicate $w_1\ w_2\ \ldots\ w_i\ *\ w_{i+1}\ \ldots\ w_n$, we obtain a semantic predicate $w_1\ w_2\ \ldots\ w_i\ c\ w_{i+1}\ \ldots\ w_n$.

Figure 3.4 shows graphically the steps that SPred goes through for inducing the semantic classes associated with the lexical predicate. Starting with a lexical predicate (*cup of* * in the Figure), SPred iterates through Wikipedia sentences, trying to match the star character. After such a match occurs (Figure 3.4, step ①), the three heuristics presented in Section 3.3.2 are applied and, as a result, the filling argument *earl grey tea* is linked to the Wikipedia page EARL GREY TEA by the Trust-the-inventory heuristic (Figure 3.4, step ②). Since no mapping is provided for the latter, WCL manages to generalize the Wikipedia page to TEA 3.4, step ③). Now, thanks to the Wikipedia-to-WordNet mapping presented above, SPred manages to map the Wikipedia page to the corresponding WordNet concepts *tea*^1_n (Figure 3.4, step ④). Finally, thanks to the set of semantic classes explained above, SPred climbs up the WordNet hierarchy until it finds the semantic class *beverage*^1_n (Figure 3.4, step ⑤). At this point the procedure stops and sums one to the score of the semantic class involved. The procedure is applied to each and every filling argument of the lexical predicate (including words different than *earl grey tea*, but also including all other occurrences of the matched argument *earl grey tea*, which might be linked to some other Wikipedia page). At the end of the procedure, SPred ends up with a probability distribution over semantic classes, and an example output is reported in Table 3.2.

[7] This can rarely happen due to multiple hypernyms available in WordNet for the same synset.

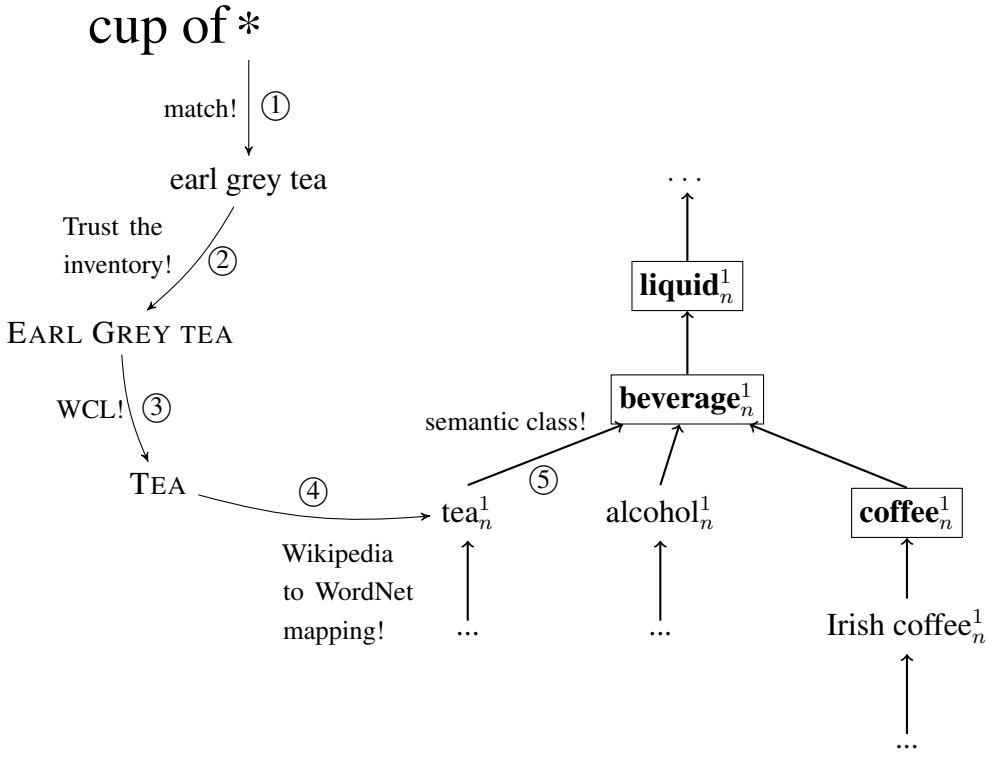


Figure 3.4. Example of a semantic class for the lexical predicate *cup of **. Circled numbers indicate the steps of SPred used to generalize the seen argument.

3.3.4 Classification of new arguments

Once the semantic predicates for the input lexical predicate π have been learned, we can classify a new filling argument a of π. However, the class probabilities calculated with Formula 3.1 might not provide reliable scores for several classes, including unseen ones whose probability would be 0.

To enable wide coverage we estimate a second conditional probability based on the distributional semantic profile of each class. To do this, we perform three steps:

1. For each WordNet synset S we create a distributional vector $\vec{S}$ summing the noun occurrences within all the Wikipedia pages p such that $\mu(p) = S$. Next, we create a distributional vector for each class $c \in C$ as follows:

   $$\vec{c} = \sum_{S \in desc(c)} \vec{S},$$

   where $desc(c)$ is the set of all synsets which are descendants of the semantic class c in WordNet. As a result we obtain a predicate-independent distributional description for each semantic class in C.

2. Now, given an argument a of a lexical predicate π, we create a distributional vector $\vec{a}$ by summing the noun occurrences of all the sentences s such that $(a, s, l) \in L_\pi$ (cf. Section 3.3.1).

3. Let C_a be the set of candidate semantic classes for argument a, i.e., C_a contains the semantic classes for the WordNet synsets of a as well as the semantic classes associated with $\mu(p)$ for all Wikipedia pages p whose lemma is a. For each candidate class $c \in C_a$, we determine the cosine similarity between the distributional vectors $\vec{c}$ and $\vec{a}$ as follows:

$$sim(\vec{c}, \vec{a}) = \frac{\vec{c} \cdot \vec{a}}{||\vec{c}||\,||\vec{a}||}.$$

Then, we determine the most suitable semantic class $c \in C_a$ of argument a as the class with the highest distributional probability, estimated as:

$$P_{distr}(c|\pi, a) = \frac{sim(\vec{c}, \vec{a})}{\sum_{c' \in C_a} sim(\vec{c}', \vec{a})}. \tag{3.2}$$

We can now choose the most suitable class $c \in C_a$ for argument a which maximizes the probability mixture of the distributional probability in Formula 3.2 and the class probability in Formula 3.1:

$$P(c|\pi, a) = \alpha P_{distr}(c|\pi, a) + (1 - \alpha) P_{class}(c|\pi), \tag{3.3}$$

where $\alpha \in [0, 1]$ is an interpolation factor.

We now illustrate the entire process of our algorithm on a real example. Given a textual expression such as *virus replicate*, we: (i) extract all the filling arguments of the lexical predicate * *replicate*; (ii) link and disambiguate the extracted filling arguments; (iii) query our system for the available *virus* semantic classes (i.e., {*virus*^1_n, *virus*^3_n}); (iv) build the distributional vectors for the candidate semantic classes and the given input argument; (v) calculate the probability mixture. As a result we obtain the following ranking, *virus*^1_n:0.250, *virus*^3_n:0.000894, so that the first sense of *virus* in WordNet 3.0 is preferred, being an "ultramicroscopic infectious agent that replicates itself only within cells of living hosts".

3.4 Experiment 1: Oxford Lexical Predicates

We evaluate on the two forms of output produced by SPred: (i) the top-ranking semantic classes of a lexical predicate, as obtained with Formula 3.1, and (ii) the classification of a lexical predicate's argument with the most suitable semantic class, as produced using Formula 3.3. For both evaluations, we use a lexical predicate dataset built from the Oxford Advanced Learner's Dictionary [Crowther, 1998].

3.4.1 Set of Semantic Classes

The selection of which semantic classes to include in the set C is of great importance. In fact, having too many classes will end up in an overly fine-grained inventory

of meanings, whereas an excessively small number of classes will provide little discriminatory power. As our set C of semantic classes we selected the standard set of 3,299 core nominal synsets available in WordNet.[8] However, our approach is flexible and can be used with classes of an arbitrary level of granularity.

3.4.2 Datasets

The Oxford Advanced Learner's Dictionary provides usage notes that contain typical predicates in various semantic domains in English, e.g., Traveling.[9] Each predicate is made up of a fixed part (e.g., a verb) and a generalizable part which contains one or more nouns.

Examples include *fix an election/the vote*, *bacteria/microbes/viruses spread*, *spend money/savings/a fortune*. In the case that more than one noun was provided, we split the textual expression into as many items as the number of nouns. For instance, from *spend money/savings/a fortune* we created three items in our dataset, i.e., *spend money*, *spend savings*, *spend a fortune*. The splitting procedure generated 6,220 instantiated lexical predicate items overall.

3.4.3 Evaluating the Semantic Class Ranking

Dataset. Given the above dataset, we generalized each item by pairing its fixed verb part with * (i.e., we keep "verb predicates" only, since they are more informative). For instance, the three items *bacteria/microbes/viruses spread* were generalized into the lexical predicate * *spread*. The total number of different lexical predicates obtained was 1,446, totaling 1,429 distinct verbs (note that the dataset might contain the lexical predicate * *spread* as well as *spread* *).[10]

Methodology. For each lexical predicate we calculated the conditional probability of each semantic class using Formula 3.1, resulting in a ranking of semantic classes. To evaluate the top ranking classes, we calculated precision@k, with k ranging from 1 to 20, by counting all applicable classes as correct, e.g., $location^1_n$ is a valid semantic class for *travel to* * while $emotion^1_n$ is not.

Results. We show in Table 3.3 the precision@k calculated over a random sample of 50 lexical predicates.[11] As can be seen, while the classes quality is pretty high with low values of k, performance gradually degrades as we let k increase. This is mostly due to the highly polysemous nature of the predicates selected (e.g., *occupy* *, *leave* *, *help* *, *attain* *, *live* *, etc.). We note that high performance, attaining

[8] `http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt`

[9] `http://oald8.oxfordlearnersdictionaries.com/usage_notes/unbox_colloc/`

[10] The low number of items per predicate is due to the original Oxford resource.

[11] One lexical predicate did not have any semantic class ranking.

k	Prec@k	Correct	Total
1	0.94	46	49
2	0.87	85	98
3	0.86	124	145
4	0.83	160	192
5	0.82	194	237
6	0.81	228	282
7	0.80	261	326
8	0.78	288	370
9	0.77	318	414
10	0.76	349	458
11	0.75	379	502
12	0.75	411	546
13	0.75	445	590
14	0.76	479	634
15	0.75	510	678
16	0.75	544	721
17	0.76	577	763
18	0.76	612	806
19	0.76	643	849
20	0.75	671	892

Table 3.3. Precision@k for ranking the semantic classes of lexical predicates.

above 80%, can be achieved by focusing up to the first 7 classes output by our system, with a 94% precision@1.

3.4.4 Evaluating Classification Performance

Dataset. Starting from the lexical predicate items obtained as described in Section 3.4.2, we selected those items belonging to a random sample of 20 usage notes among those provided by the Oxford dictionary, totaling 3,245 items. We then manually tagged each item's argument (e.g., *virus* in *viruses spread*) with the most suitable semantic class (e.g., $virus^1_n$), obtaining a gold standard dataset for the evaluation of our argument classification algorithm (cf. Section 3.3.4).

Methodology. In this second evaluation we measure the accuracy of our method at assigning the most suitable semantic class to the argument of a lexical predicate item in our gold standard. We use three customary measures to determine the quality of the acquired semantic classes, i.e., precision, recall and F1. Precision is the number of items which are assigned the correct class (as evaluated by a human) over the number of items which are assigned a class by the system. Recall is the number of items which are assigned the correct class over the number of items to be classified. F1 is the harmonic mean of precision and recall.

Method	Precision	Recall	F1
SPred	85.61	68.01	75.80
Random	40.96	40.96	40.96

Table 3.4. Performance on semantic class assignment.

Tuning. The only parameter to be tuned is the factor α that we use to mix the two probabilities in Formula 3.3 (cf. Section 3.3.4). For tuning α we used a held-out set of 8 verbs, randomly sampled from the lexical predicates not used in the dataset. We created a tuning set using the annotated arguments in Wikipedia for these verbs: we trained the model on 80% of the annotated lexical predicate arguments (i.e., the class probability estimates in Formula 3.1) and then applied the probability mixture (i.e., Formula 3.3) for classifying the remaining 20% of arguments. Finally, we calculated the performance in terms of precision, recall and F1 with 11 different values of $\alpha \in \{0, 0.1, \ldots, 1.0\}$, achieving optimal performance with $\alpha = 0.2$.

Results. Table 3.4 shows the results on the semantic class assignments. Our system shows very high precision, above 85%, while at the same time attaining an adequate 68% recall. We also compared against a random baseline that randomly selects one out of all the candidate semantic classes for each item, achieving only moderate results. A subsequent error analysis revealed the common types of error produced by our system: terms for which we could not provide (1) any WordNet concept (e.g., *political corruption*) or (2) any candidate semantic class (e.g., *immune system*).

3.4.5 Disambiguation heuristics impact

Total triples	Linked in Wikipedia	One sense per page	One sense per lexical predicate	Trust the inventory	Not linked
73,843,415	1,795,608	1,433,634	533,946	1,716,813	68,363,414

Table 3.5. Statistics on argument triple linking for all the lexical predicates in the Oxford dataset.

As a follow-up analysis, for each dataset we considered the impact of each disambiguation heuristic described in Section 3.3.2 according to how many times it was triggered. Starting from the entire set of 1,446 lexical predicates from the Oxford dictionary (see Section 3.4.3), we counted the number of argument triples (a, s, l) already disambiguated in Wikipedia (i.e., $l \neq \epsilon$) and those disambiguated thanks to our disambiguation strategies. Table 3.5 shows the statistics. We note that, while the amount of originally linked arguments is very low (about 2.5% of total), our strategies are able to considerably increase the size of the initial set of linked instances. The most effective strategies appear to be the *One sense per page*

and the *Trust the inventory*, which contribute 26.16% and 31.33% of the total links, respectively.

Even though most of the triples (i.e., 68 out of almost 74 million) remain unlinked, the ratio of distinct arguments which we linked to WordNet is considerably higher, calculated as 3,723,979 linked arguments over 12,431,564 distinct arguments, i.e., about 30%.

3.5 Experiment 2: Comparison with Kozareva & Hovy (2010)

Due to the novelty of the task carried out by SPred, the resulting output may be compared with only a limited number of existing approaches. The most similar approach is that of [Kozareva and Hovy, 2010a, K&H] who assign supertypes to the arguments of arbitrary relations, a task which resembles our semantic predicate ranking. We therefore performed a comparison on the quality of the most highly-ranked supertypes (i.e., semantic classes) using their dataset of 24 relation patterns (i.e., lexical predicates).

Dataset. The dataset contained 14 lexical predicates (e.g., *work for* * or * *fly to*), 10 of which were expanded in order to semantify their left- and right-side arguments (e.g., * *work for* and *work for* *); for the remaining 4 predicates just a single side was generalized (e.g., * *dress*). While most of the relations apply to persons as a supertype, our method could find arguments for each of them.

Methodology. We carried out the same evaluation as in Section 3.4.3. We calculated precision@k of the semantic classes obtained for each relation in the dataset of K&H. Because the set of applicable classes was potentially unbounded, we were not able to report recall directly.

Results. K&H reported an overall accuracy of the top-20 supertypes of 92%. As can be seen in Table 3.6 we exhibit very good performance with increasing values of k. A comparison of Table 3.3 with Table 3.6 shows considerable differences in performance between the two datasets. We attribute this difference to the higher average WordNet polysemy of the verbal component of the Oxford predicates (on average 2.64 senses for K&H against 6.52 for the Oxford dataset).

Although we cannot report recall, we list the number of Wikipedia arguments and associated classes in Table 3.7, which provides an estimate of the extraction capability of SPred. The large number of classes found for the arguments demonstrates the ability of our method to generalize to a variety of semantic classes.

k	Prec@k	Correct	Total
1	0.88	21	24
2	0.90	43	48
3	0.88	63	72
4	0.89	85	96
5	0.91	109	120
6	0.91	131	144
7	0.92	154	168
8	0.91	175	192
9	0.92	198	216
10	0.92	221	240
11	0.92	242	264
12	0.92	264	288
13	0.91	284	312
14	0.90	304	336
15	0.91	327	360
16	0.91	348	384
17	0.90	367	408
18	0.89	386	432
19	0.89	407	456
20	0.89	429	480

Table 3.6. Precision@k for the semantic classes of the relations of [Kozareva and Hovy, 2010a].

Predicate	Number of args	Number of classes
cause *	181,401	1,339
live in *	143,628	600
go to *	134,712	867
* cause	92,160	1,244
work in *	79,444	770
* go to	71,794	746
* live in	61,074	541
work on *	58,760	840
work for *	58,332	681
work at *	31,904	511
* work in	24,933	528
* celebrate	23,333	408

Table 3.7. Number of arguments and associated classes for the 12 most frequent lexical predicates of [Kozareva and Hovy, 2010a] extracted by SPred from Wikipedia.

3.6 Related work

The availability of Web-scale corpora has led to the production of large resources of relations [Etzioni et al., 2005, Yates et al., 2007, Wu and Weld, 2010, Carlson

et al., 2010, Fader et al., 2011]. These systems take the Web in input and apply lexico-syntactic patterns to extract relations between noun phrases. For example, TextRunner [Yates et al., 2007] extracts facts under the form of triples, such as *(company, purchased by, Google)* or *(Rome, capital of, Italy)* thanks to a Naive Bayes model with shallow linguistic features. These systems differ in the set of tools used to analyze the data and in the amount of human intervention. One of the requirements for these extractors of raw knowledge is speed: a common assumption of the underlying model is that data will be scanned once and at such a speed that complex models which parse the incoming text syntactically on the fly are not feasible. More complex models based on linear-chain CRF [Banko et al., 2008], Markov Logic Network [Zhu et al., 2009] or dependency parse features [Wu and Weld, 2010] have been also applied and indeed lead to improved extraction over shallow linguistic features, but at the cost of increased processing time. However, a common drawback of all these resources is that they operate purely at the lexical level, that is, they do not provide any information on the semantics of their arguments or relations. Several studies have examined adding semantics through grouping relations into sets [Yates and Etzioni, 2009], ontologizing the arguments [Pantel and Ravichandran, 2004], or ontologizing the relations themselves [Moro and Navigli, 2013]. However, analysis has largely been either limited to ontologizing a small number of relation types with a fixed inventory [Mohamed et al., 2011, Moro and Navigli, 2013], which potentially limits coverage, or has been focusing on implicit definitions of semantic categories (e.g., clusters of arguments, as in [Pantel and Ravichandran, 2004], or clusters of relations, as in [Yates and Etzioni, 2009]), which limits interpretability. As regards the first group, for example, [Mohamed et al., 2011] use the semantic categories of the NELL system [Carlson et al., 2010] to learn roughly 400 valid ontologized relations from over 200M web pages, whereas WiSeNet [Moro and Navigli, 2012] leverages Wikipedia to acquire relation synsets for an open set of relations. Despite these efforts, though, no large-scale resource has existed to date that contains ontologized lexical predicates. As regards the second group, [Pantel and Ravichandran, 2004], building upon [Pantel and Lin, 2002], represent semantic classes as clusters of similar words (e.g., the two semantic classes for *plant* are {plant, factory, facility, refinery} and {shrub, ground cover, perennial, bulb}), while relations in [Yates and Etzioni, 2009] and [Moro and Navigli, 2012] are represented as sets of strings (for example, [Moro and Navigli, 2012] represent the relational phrase *is a field of* as a relation synset *{is a field of, is an area of, is studied in}*). This type of implicit representation, though, makes comparison complicated and it also might turn out to be hard for a machine to interpret automatically. In contrast, the present work, thanks to its WordNet-based notion of semantic classes, provides a high-coverage method for learning argument supertypes from a broad-coverage ontology and can in turn potentially be leveraged in relation extraction to ontologize relation arguments.

Our method for identifying the different semantic classes of predicate arguments is closely related to the task of identifying selectional preferences, an area pioneered by [Resnik, 1996]. Selectional preferences model the strength of association be-

tween an argument and a predicate; for example, given the verb *shoot* and the two arguments *deer* and *pen*, the former is said to have a stronger semantic association strength (i.e., selectional preference) than the latter. The strength of associations of two arguments for the same predicate is so important because it is exploited as a hint for clustering arguments with similar selectional preferences. For example, the two arguments *apple* and *pear* should be considered similar with respect to the verb *eat*, since their selectional preferences are similar. Several methods have been proposed which differ in the way the arguments are collected or the similarity between arguments calculated. A whole branch of works, the most similar to ours, are the taxonomy-based ones [Resnik, 1996, Li and Abe, 1998, Clark and Weir, 2002, Pennacchiotti and Pantel, 2006, Rooth et al., 1999, Agirre and Martinez, 2001, Pantel et al., 2007, Bergsma et al., 2008, Ritter et al., 2010, Séaghdha, 2010, Bouma, 2010, Jang and Mostow, 2012] which leverage the semantic types of the relations arguments to infer the similarity between arguments, generating probability distributions over the arguments. These methods rely either on existing taxonomies (such as WordNet) or on sense tagged corpora [Agirre and Martinez, 2001] to generalize over observed arguments, but, despite their high quality sense-tagged data, they often suffer from lack of coverage. As a result, alternative, non hierarchy–based approaches have been proposed that eschew taxonomies and rely on distributional similarity between arguments in order to obtain higher coverage of preferences [Erk, 2007, Erk et al., 2010, Chambers and Jurafsky, 2010]. However, what makes these latter approaches hard to appreciate is the implicit representation of preference, whose quality is historically evaluated via pseudo-words disambiguation that revealed to be problematic due to the correct methodology to adopt (i.e., distribution of seen vs. unseen arguments, the strategy to extract the pseudo-word confounder, etc).

In contrast, we overcome the data sparsity of class-based models by leveraging the large quantity of collaboratively-annotated Wikipedia text in order to connect predicate arguments with their semantic class in WordNet using BabelNet [Navigli and Ponzetto, 2012a]; since we map directly to WordNet synsets, we provide a more readily-interpretable collocation preference model than most similarity-based or probabilistic models. In fact, each cluster of generalized arguments is labelled with a semantic class, that is a WordNet concept, so that the meaning is not implicit in the representation.

There is also another very productive line of research which, stemmed with the creation of a resource called FrameNet [Baker et al., 1998], is also related to our work, represented by approaches which generalize the notion of lexical predicate in order to account for multiple generalization slots [Green et al., 2004, Surdeanu et al., 2003, Yakushiji et al., 2006]. In this paradigm, called Frame Semantics [Fillmore, 1976], the knowledge is encoded by means of *frames*, descriptions of prototypical situations in which a variable number of semantic roles, called *frame elements*, are involved. For example, the *Travel* frame is defined as: "In this frame a TRAVELER goes on a journey, an activity, generally planned in advance, in which the TRAVELER moves from a SOURCE location to a GOAL along a PATH or within an AREA. [...] The DURATION or DISTANCE of the journey, both generally long,

may also be described as may be the MODE OF TRANSPORTATION" and TRAVELER, DISTANCE, DURATION, etc., are the expected frame elements involved in the frame. Manually sense-tagged lemmas, called *lexical units*, are said to *evoke* a certain frame; lexical units are mainly verbs, but they might include also nouns, adjectives and adverbs (e.g., both *tour.v* and *safari.n* are lexical units evoking the *Travel* frame). The aim of such an area is very similar in spirit to ours, since the arguments are generalized to the above roles (for example in the sentence *John flew for more than two hours*, *John* would fill the TRAVELER role and *more than two hours* would fill the DURATION role), but it also differs for a number of aspects: first, the generality of frame-based approaches goes beyond our intentions, as we focus on semantic predicates, which is much simpler and free from syntactic parsing; second, the number and the type of roles involved in a frame is sometimes arbitrary and not widely accepted; third, roles accomodate arguments of any length (for instance, in the sentence "a book was written describing a voyage, by balloon, to the newly discovered planet Uranus", the GOAL role is filled by the whole text fragment '*to the newly discovered planet Uranus*'), while arguments in SPred are simple noun-phrases (e.g., *Scotland*, *holy land*, etc). Finally, FrameNet has been shown to suffer from serious coverage problems [Burchardt et al., 2005, Shen and Lapata, 2007, López de Lacalle et al., 2014]. Due to its ongoing growth, FrameNet is still well below the size of established lexicons such as WordNet: in fact, not only does it lack lexical entries (about 10,000 lemmas in FrameNet 1.5, against more than 150,000 lemmas in WordNet) but it also lacks word senses for many word, due to the fact that it is increased one frame at a time rather than one lemma at a time. In contrast, one of the strong points of SPred is that it is applicable to, potentially, all the verbs and lemmas found in English and is not bound to a fixed set of frames or lexical units.

Another closely related work is that of [Hanks, 2013] concerning the Theory of Norms and Exploitations, where norms (exploitations) represent expected (unexpected) classes for a given lexical predicate. For example "an army of *mercenaries*" or "a swarm of *bees*" are examples of norms, because *mercenaries* and *bees* are expected arguments for the corresponding predicates *an army of* and *a swarm of*. Examples of exploitations are, instead, "an army of *lawyers*" or "a swarm of *teenagers*", since the two arguments somewhat do not satisfy the expected classes of the two predicates (being $soldier_n^1$ and $insect_n^1$, most probably). Exploitations are to be seen by grounding it into norms, so lawyers that collectively form an army-like group are behaving like fierce soldiers, while teenagers that collectively form a swarm-like crowd are behaving like insects. Although our semantified predicates do, indeed, provide explicit evidence of norms obtained from collective intelligence and would provide support for this theory, exploitations present a more difficult task, different from the one addressed here, due to its focus on identifying property transfer between the semantic class and the exploited instance.

The closest technical approach to ours, however, is that of [Kozareva and Hovy, 2010a], who use recursive patterns to induce semantic classes for the arguments of relational patterns. For example, given the seed noun *John* and the lexical predicate

fly to, [Kozareva and Hovy, 2010a] build the recursive pattern "* and <seed> fly to *", where initially <seed> is equal to *John*. In the following, they apply the lexical predicate on the Web and harvest left-side arguments (e.g., Brian, Kate, bees, Delta, Alaska, etc.) and right-side arguments (e.g., flowers, trees, party, New York, Italy, France, etc). The procedure then proceeds by replacing the initial seed with the newly learned arguments, and as a result new arguments are learned. The procedure continues harvesting new arguments in a breadth-first fashion, until no more arguments are encountered. However, the approach of [Kozareva and Hovy, 2010a] requires both a relation pattern and one or more seeds which though have the potential to bias the types of semantic classes that are learned. In contrast, our proposed method requires only the pattern itself, and as a result it is capable of learning an unbounded number of different semantic classes.

3.7 Conclusions

In this chapter we presented SPred,[12] a novel approach to large-scale harvesting of semantic predicates. In order to semantify lexical predicates we exploit the wide coverage of Wikipedia to extract and disambiguate lexical predicate occurrences, and leverage WordNet to populate the semantic classes with suitable predicate arguments. As a result, we are able to ontologize lexical predicate instances like those available in existing dictionaries (e.g., *break a toe*) into semantic predicates (such as *break* a BODY PART). For each lexical predicate (such as *break* $*$), our method produces a probability distribution over the set of semantic classes (thus covering the different expected meanings for the filling arguments) and is able to classify new instances with the most suitable class. Our experiments show generally high performance, also in comparison with previous work on argument supertyping.

We hope that our semantic predicates will enable progress in different Natural Language Processing tasks such as Word Sense Disambiguation [Navigli, 2009], Semantic Role Labeling [Fürstenau and Lapata, 2012] or even Textual Entailment [Stern and Dagan, 2012], all needing reliable supertyping power. While we focused on semantifying lexical predicates, as future work we will apply our method to the ontologization of large amounts of sequences of words, such as phrases or textual relations (e.g., considering Google n-grams appearing in Wikipedia). Our method should, in principle, also generalize not only to any semantically-annotated corpus (e.g., Wikipedias in other languages) but also to large textual corpora, such as Gigaword, provided efficient argument disambiguation techniques will be available. However, we do acknowledge the need for further testing to quantify the relationship between semantic predicate quality and the corpus size.

One of the main problems left open is represented by the number of disambiguated arguments. The majority of predicate arguments has not been disambiguated yet and a more mature mechanism is needed to perform full-fledged WSD.

[12] `http://lcl.uniroma1.it/spred`

A current limitation is also represented by the fact that the set of lexical predicates processed by SPred is given in input. A real-world application might well not have this set in advance, since the latter might potentially include all the possible word sequences occurring in natural language. We are currently experimenting on applying SPred on more than 10M English lexical predicates, extracted by combining Wikipedia and the Google n-gram corpus.

Chapter 4

Impact of hypernymy information on SPred

4.1 Introduction

The goal of this chapter is to show that the integration of MultiWiBi (see Chapter 2) into SPred's internal mechanisms (see Chapter 3) is beneficial for the latter in terms of quality, both in terms of arguments linked to WordNet and in terms of precision and recall.

In particular, we will apply SPred to a word sense disambiguation task and will show the impact of is-a information when integrated into the semantic system. We underline that this chapter presents only preliminary results on the application of SPred on a real word task; the aim of this chapter, in fact, is not that of comparing SPred against alternative approaches in the same task, but rather study to which extent the quality and quantity of hypernymy information injected into SPred impacts on the final overall results. We acknowledge that substantial work needs still to be done in order for SPred to ever provide state-of-the-art results in this task and also that other tasks might be more suitable for better revealing SPred's power.

4.2 Experimental setup

To evaluate SPred we selected the WSD Task 12 of Semeval2013 [Navigli et al., 2013]. This was done because it is one of the most recent tasks which requires to disambiguate nouns only and provides keys in WordNet 3.0 sense inventory (while tasks in previous or subsequent SemEval were open to all the word classes including verbs, adjectives and adverbs and used different versions of WordNet, so that it would have been necessary to use automatic mappings across WordNet versions). The dataset contains 306 sentences with 1,931 target nouns of which only 1,644 have a sense-annotation. A target noun might be not annotated in the gold standard because WordNet does not contain the sense intended in that context (for example the target noun *emerging economy* appearing in the sentence "[...] how developed

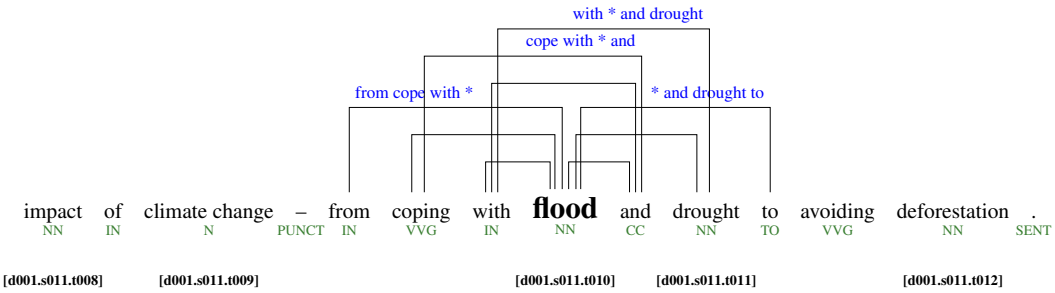


Figure 4.1. The lexical predicates of length $l \leq 4$ for the noun with target ID *d001.s011.t010*, extracted from the 11th sentence of the Task 12 of Semeval 2013.

and major emerging economies would cut their carbon output [...]" has no senses in WordNet). The total number of annotations contained in the gold standard is 1,656, meaning that a given noun might have more than one annotation (for example in the sentence "[...] to avoiding deforestation.", the target noun *deforestation* has been annotated with both deforestation^1_n and deforestation^2_n).

4.3 Statistics

We extracted the set Π of all the possible lexical predicates involving a target noun of length $2 \leq l \leq 5$. Figure 4.1 shows the predicates extracted for a particular given sentence and target noun in the gold standard. The procedure generated 17,030 distinct lexical predicates.

We then matched Π against Wikipedia, resulting in 9,572 lexical predicates with at least one argument. On the total, 730 have length two, 2,507 have length three, 3,490 have length four and 2,845 have length five. We then applied the same procedure explained in Chapter 3, Section 3.3.2. Table 4.1 shows the statistics for the resulting set of lexical and semantic predicates found. As can be seen, more than 251 million triples have been extracted after matching the lexical predicates against Wikipedia. The three heuristics presented in Section 3.3.2 produced about 152 millions linked arguments overall, with the One-sense-per-predicate heuristic being the most contributing one. When analysing the triples at the word-level (i.e., not considering multiple occurrences of the same argument word), we can see that the heuristic which contributes the most is the Trust-the-inventory, and the three heuristics produce more than 16 millions linked words overall (the 43.3% of the total arguments words). This table hightlights that a considerable number of both argument words and occurrences still need to be disambiguated and future work will have to tackle this limitation in an effective way.

We then applied the procedure explained in Section 3.3.3 for generalizing linked arguments to WordNet. We remind that this step crucially draws upon a taxonomy defined over Wikipedia pages used for extending the mapping to WordNet also to those arguments for which a direct mapping was not provided. In SPred (see

	Arguments	Linked arguments	One sense per page	One sense per predicate	Trust the inventory
Occurrences	251,635,404	152,321,081	11,506,271	117,923,034	22,891,776
Words	37,370,475	16,201,876	2,567,704	1,746,952	11,887,220

Table 4.1. Statistics of the impact of the amount of hypernymy information on the lexical predicates extracted from Semeval 2013.

Taxonomy used	Generalized	Linked to WordNet	Avg. semantic classes
None	0	94,780,954	121.32
WCL	3,065,328	97,846,282	128.53
MultiWiBi	13,495,048	108,276,002	140.69

Table 4.2. Distribution of arguments for the lexical predicates extracted from Semeval 2013.

Chapter 3) we used WCL to generalize the disambiguated arguments. In order to test how the underlying Wikipedia taxonomy impacts on SPred's generalization power, we experimented the usage of i) an empty taxonomy (a taxonomy providing no is-a relations), ii) the taxonomy obtained thanks to WCL, and iii) the English Wikipedia page taxonomy offered by MultiWiBi (see Chapter 2). Table 4.2 shows the statistics when using the three different taxonomies. The first column reports the number of arguments linked to Wikipedia for which the taxonomy could provide a generalization which, in turn, could be mapped to WordNet. As can be seen, by using WCL, it is already possible to map more than 3 millions of arguments to WordNet; MultiWiBi, however, manages to map more than 13 millions arguments, more than four times the number of arguments mapped by WCL. This augmented generalization power, of course, is reflected also by the average number of semantic classes per predicate: while being 121 when no taxonomy is used, it increases to 128 when using WCL and finally jumps up to 140 when MultiWiBi is employed. This is interesting because the higher number of semantic classes per predicate translates directly into a more fine-grained expressivity associated with the semantic predicates.

4.4 Methodology

We decided to design a new disambiguation framework in which all the possible predicates extracted from the sentence are exploited and all the associated semantic classes exploited accordingly.

In order to disambiguate each target noun n we apply the following formula:

$$s^* = \underset{s_i \in Senses(n)}{\arg\max}\ score(s_i) = \underset{s_i \in Senses(n)}{\arg\max} \sum_{\pi \in \Pi} \sum_{c \in \pi} \omega(s_i, c) \cdot P_{class}(c|\pi) \qquad (4.1)$$

$$\omega = \begin{cases} e^{-(1+d(s_i,c))} & \text{if } s_i \text{ is-a}^* \ c \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

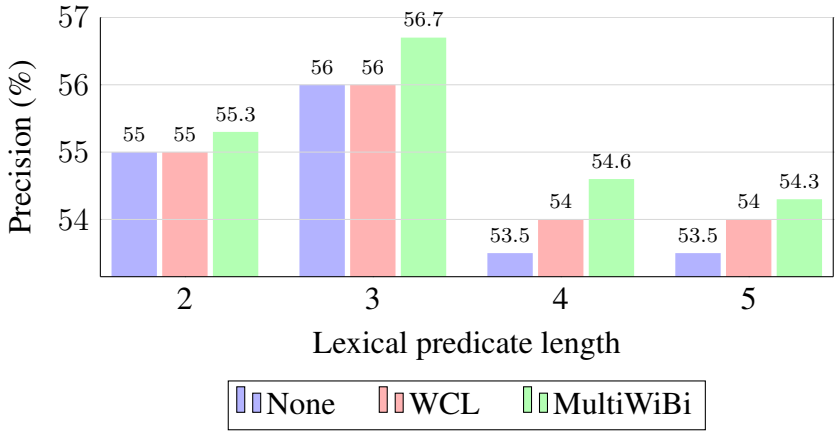
ω is a weighting parameter which is equal to 0 if there is no is-a path between the sense and the semantic class and is inversely proportional to the distance between the two, otherwise. The role of ω is to favor semantic classes closer to the sense being scored and discard those which instead are not compatible. The formula considers all the possible semantic predicates, centered on the target noun, extracted from the given sentence. For each of the predicates, all the semantic classes are exploited and whenever the predicate's semantic class is compatible with the current sense being scored, the corresponding score is incremented. In this way semantic classes which are incompatible with the word's sense are not accounted whereas closer classes are favored.

4.5 Results

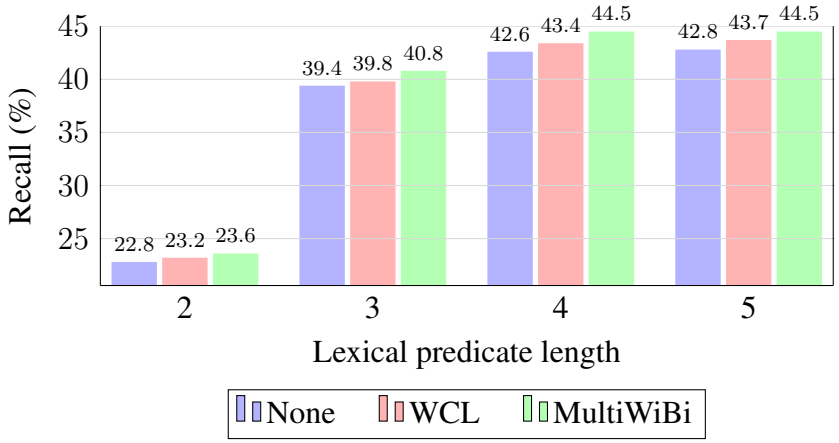
Our experiments focused on two parameters: i) the taxonomy being used when generalizing the lexical predicate arguments (namely, no taxonomy, the WCL taxonomy and the English page taxonomy of MultiWiBi), and ii) the maximum lexical predicate length used to disambiguate the target noun (with values between 2 and 5). Figure 4.2 displays the results by reporting precision in Figure 4.2a, recall in Figure 4.2b and attempt ratio in Figure 4.2c. The first two are standard precision and recall, while the third measure counts the number of items for which the system provided an answer, irrespective of its correctness.

As regards precision, we can see that, generally, all systems behave similarly, with a precision around 55%. Notwithstanding, while WCL provides a little improvement when lexical predicates of length four and five are exploited, MultiWiBi provides consistently the best precision. A general comment is that, except for length 3, as longer lexical predicates are considered, precision decreases considerably.

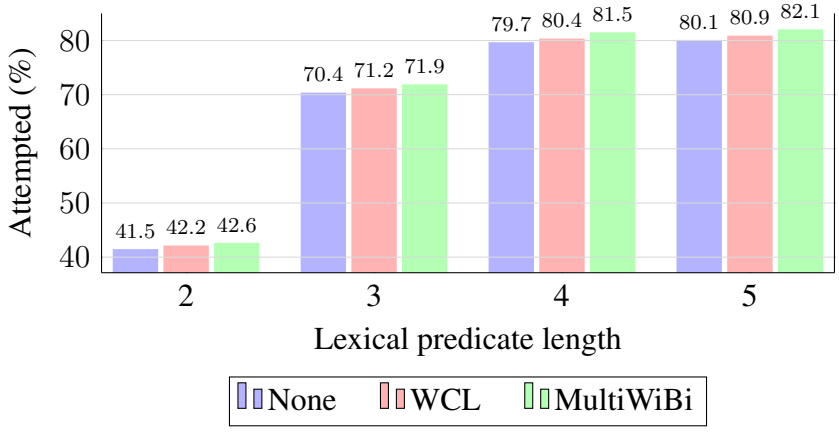
As regards recall and attemped ratio, instead, it can be clearly see that when only predicates of length 2 are taken into account, both the two measures exhibit a big drop, poorly attesting around 23% and 42% respectively, for all the three systems. This is probably due to the low number of predicates of this length being used during the disambiguation (amounting to 7.6% of the total predicates found in Wikipedia) which, therefore, contribute little to the system's disambiguation ability. As longer predicates are considered, though, performance witness an appreciable increase and exploiting longer predicates always leads the better results. This suggests that the higher amount of information, obtained as longer and more predicates are considered, always correlates with an increase in performance. WCL provides performance about one point higher than using no taxonomy at all. When MultiWiBi is used, instead, the two measures receive another point boost, totaling almost two points above the setting in which no taxonomy is being used, with 44.5% recall and 82.1%



(a) Precision



(b) Recall



(c) Attempt ratio

Figure 4.2. Performance on the Semeval 2013 Task 12.

of attempted items, respectively.

4.6 Conclusions

We have seen that the amount of hypernymy information provided by different taxonomies plays an important role on the final performance of a semantic system, such as SPred. The latter, in fact, crucially leverages the is-a relations to map Wikipedia pages to WordNet and finally generalize WordNet synsets to semantic classes in WordNet. Thanks to MultiWiBi, we demonstrated that the generalization power embedded into SPred's internal mechanisms truly corresponds to higher quality in a real world application. We have shown, in fact, that greater hypernymy information directly corresponds not only to an increased number of arguments linked to WordNet, but also to an increase in overall performances, when evaluating SPred on a word sense disambiguation task.

Future work will tackle the problem of ambiguous arguments left out from the disambiguation phase and test the possible further impact on final performance.

Another likely direction for future work will be to extend the current work in order to account for multilingual lexical predicates or even predicates aligned across languages.

Chapter 5

MultiWiBi and the Linguistic Linked Data

In this chapter we will introduce the paradigm of linked data, according to which resources are converted in a graph-based representation where facts are encoded as triples involving a subject, an object and a predicate linking the two. We will introduce the basics of the linked data and the principles regulating the publishing of resources to the linked data cloud, a shared space in which resources are linked to each other. We will present the case study of the conversion of MultiWiBi into linked data and its integration into BabelNet 3.0 as well as into a visual interface, available at `wibitaxonomy.org`.

5.1 Data & Linked Data

In what has been called 'the digital era' *data* plays a crucial role. Our society owes its wealth partly to data and a lot of companies and markets exist thanks to the production, consumption and reuse of data. From the last year's report of the European Commission expert group on taxation of the digital economy:

> "ICT continuously drives down the cost of collecting, storing and analysing data following Moore's law. This has helped to reduce transaction costs, many of which are information related, making many more market transactions possible than previously. [...] Furthermore, Big Data has helped digital firms develop innovative goods and services, with lower costs associated with innovation, in terms of measurement, experimentation, sharing and replication than in the pre-digital age. It is now possible to measure and analyse phenomena to an extent never imaginable before, thus making it easier to run controlled experiments and measure their success with great precision."[1]

[1] `http://ec.europa.eu/taxation_customs/resources/documents/taxation/gen_info/good_governance_matters/digital/report_digital_`

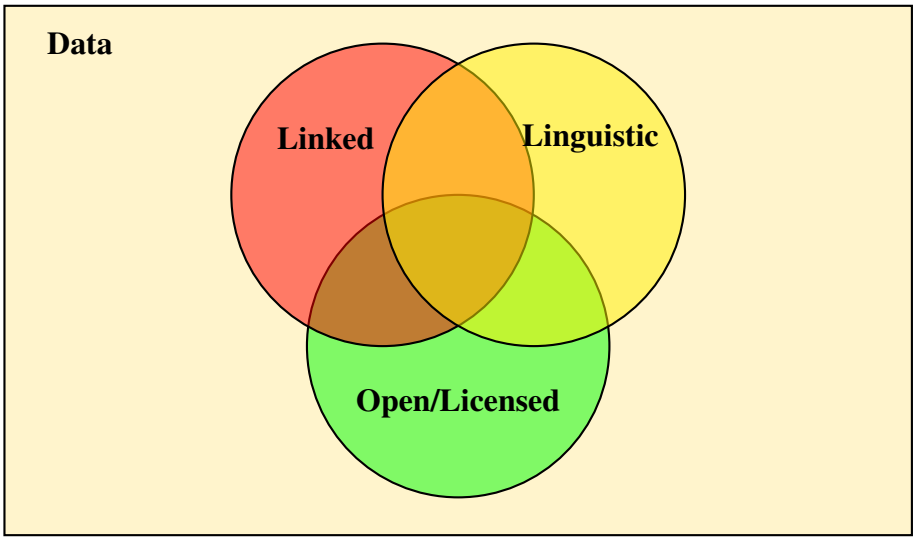


Figure 5.1. Relationship among linked, licensed and linguistic data.

It thus becomes apparent that data is a precious source not only for all those who work with data, but also for all those who, not even aware of its importance, get benefits indirectly.

L*D: the different dimensions of 'L' Besides what can be considered just simple data, researchers and industry stakeholders have witnessed the rise of additional 'L' dimensions which have been outlining the world of data over the last few decades: the **l**inking (linked data), the **l**inguistic (linguistic data) and the **l**icensing aspect (licensed data).

- **Linked data** is a publishing paradigm which allows to have a global data space based on open standards. This class comprehends many types of data, including, but not limited to, linguistic, geographic, media, government, up to cross-domain data. For a comprehensive conceptual and technical survey on linked data, we suggest [Heath and Bizer, 2011].

- **Linguistic data** deals with data which is linguistic, i.e., related to any form of textual or multimedia data regarding language. This category notably includes linguistic datasets containing text usually harvested from different sources (gazetteers, newspapers, e-mail, transcripts, etc.), such as BNC[2] or Gigaword [Graff et al., 2003], and machine-readable dictionaries and lexicons, such as WordNet [Fellbaum, 1998] or BabelNet [Navigli and Ponzetto, 2012b].

economy.pdf

[2]The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

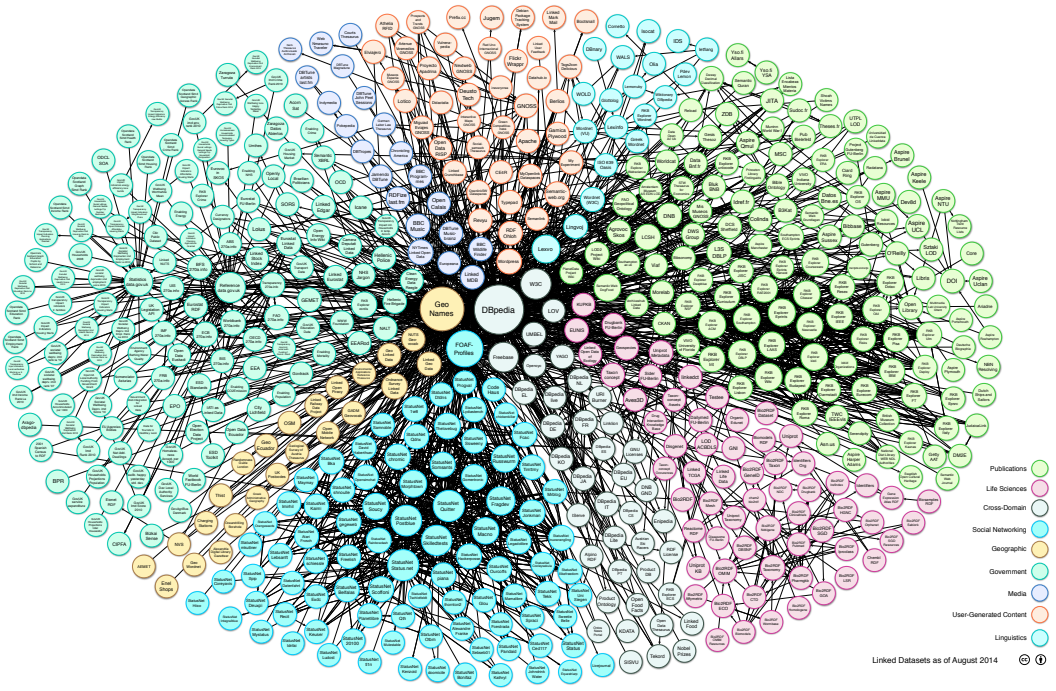


Figure 5.2. Current status of the LOD data cloud, Linking Open Data cloud diagram Aug. 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. `http://lod-cloud.net/`.

- **Open/Licensed data** is data which is licensed, i.e., subject to limitations in the reproduction, redistribution and copy [Rodríguez Doncel et al., 2013, Rodríguez-Doncel et al., 2013]. The most wide-spread licensing system is represented by *Creative Commons (CC)*, 'a global nonprofit organization that enables sharing and reuse of creativity and knowledge through the provision of free legal tools',[3] which issued six main types of licenses based on four attributes regulating the rights: Attribution (BY), Noncommercial (NC), NoDerivatives (ND) and ShareAlike (SA). For example a resource released under a BY-NC-SA license can be freely copied and redistributed i) provided the original author is attributed, ii) only for non-commercial purposes and iii) under licenses identical to those governing the original work. A case in point of (un)licensed data is the so-called *open data* (OD), whose spirit is that of publishing data without restrictions from copyright or other controlling mechanisms, in line with the idea of *open* promoted by the *open source* movement.

Figure 5.1 shows the relationship existing among these three areas. By combining the linguistic circle with the linked circle (but excluding the licensed circle) we obtain the so-called LLOD (Linguistic Linked Open Data), that is, data which is both linked, linguistic and open (i.e., not subject to licenses), while combining

[3] `https://wiki.creativecommons.org/FAQ`

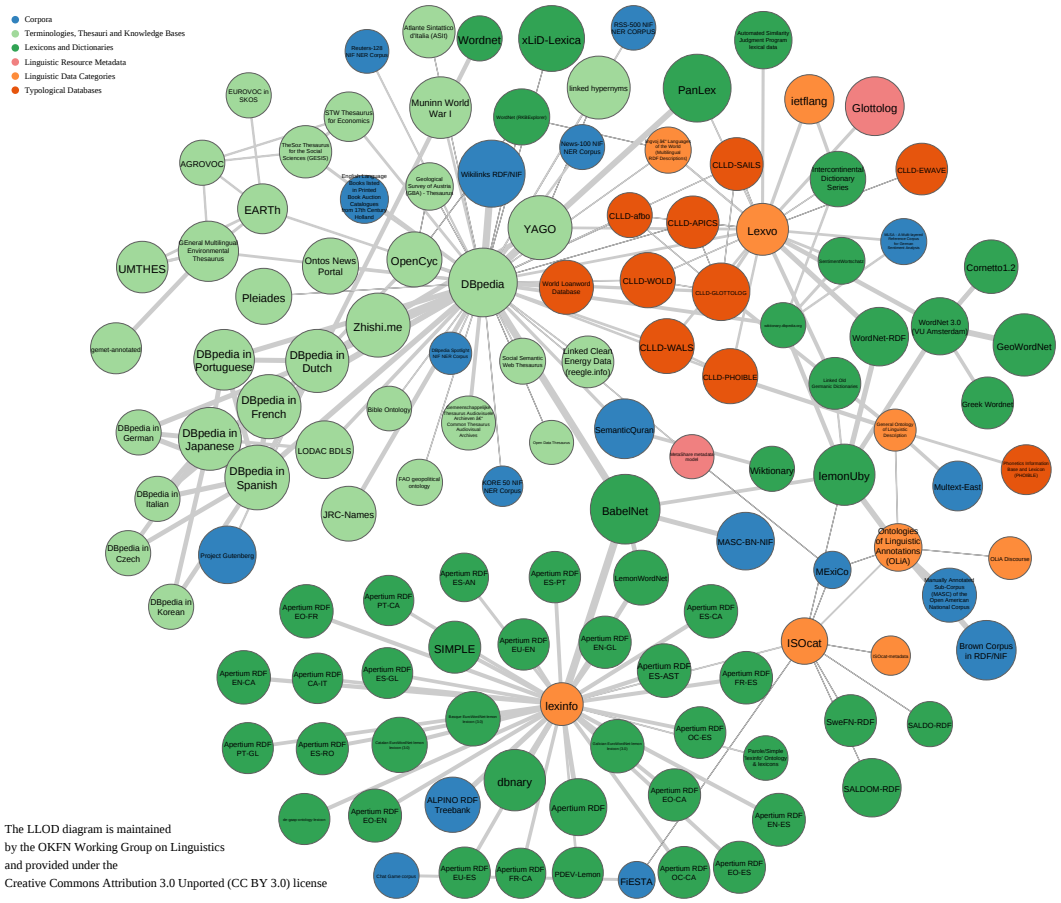


Figure 5.3. Current status of the LLOD data cloud, divided by resource type (as of Apr. 2015).

the three circles together we obtain the Linguistic Linked Licensed Data (LLLD or 3LD), that is, data published along and regulated by a machine-readable license. In the real world, however, the three dimensions are at times blurred, according to the openness of the data being published and the fact that it is hard sometimes to label data as being fully linguistic or not (e.g., transcriptions of videos might be considered linguistic or not).

The collection of all the datasets, resources, etc. which are also linked (and, in particular, linked to each other) is called *linked data cloud*. Figure 5.2 shows the snapshot of the linked open data cloud (LOD cloud) as taken in August 2014. As can be seen, a lot of resources have been migrated as linked data, including linguistic (in cyan), media (in violet), geographic (in yellow), etc. In this representation each linked resource is displayed as a node in a directed graph and there is a weighted edge $e = (R_A, R_B)$ if the resource R_A contains a non-negligible number of links to the resource R_B (where the weight of the edge is proportional to the number of links in between).

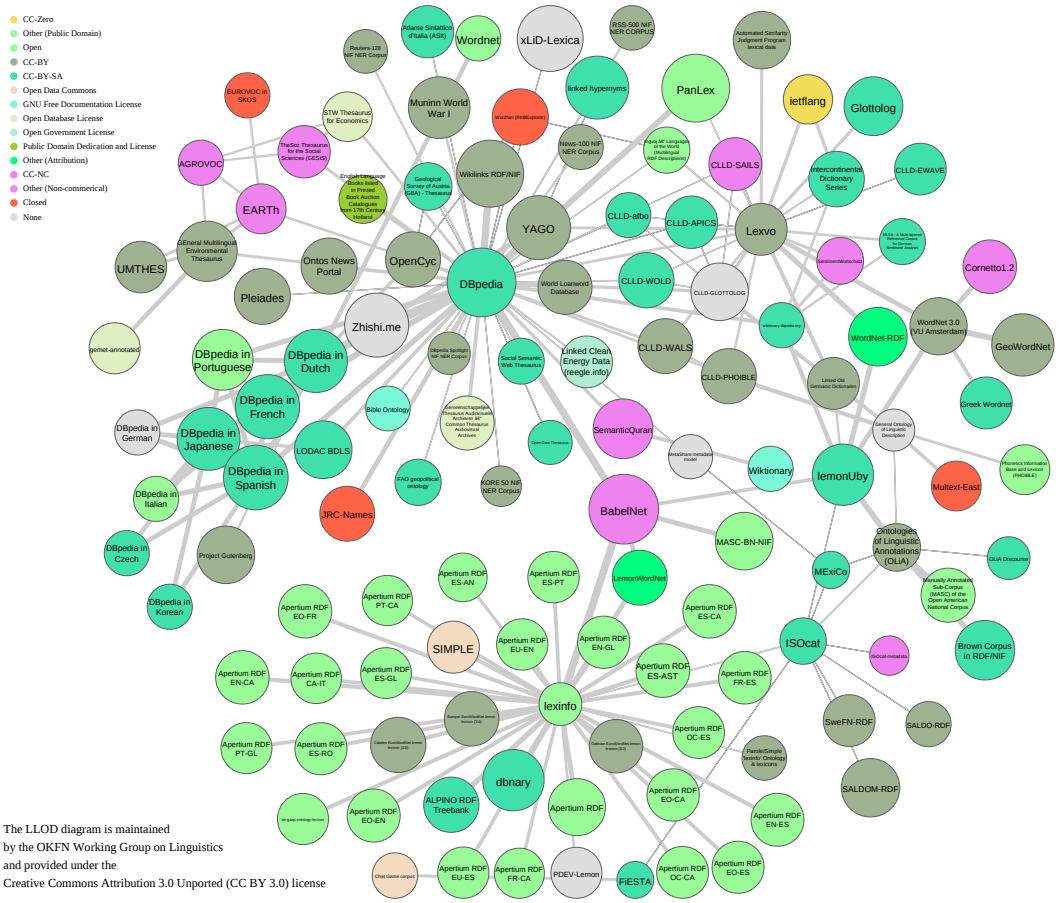


Figure 5.4. Current status of the LLOD cloud, divided by license (as of Apr. 2015).

Figure 5.3, instead, allows to grasp the type of resources contained in the LLOD cloud in a graphic manner. Here resources are associated with different colors, according to the different linguistic category they belong to. For example, nodes in light green are thesauri and knowledge bases, such as DBpedia and YAGO, while nodes in dark green are lexicons and dictionaries (such as BabelNet, lemonUby and WordNet). As can be seen, the cloud includes also textual corpora (in blue) and vocabularies used to encode metadata (such as *LexInfo*[4], used for attaching lexical information to ontologies [Buitelaar et al., 2009], and OLiA[5], used to encode morphological, morphosyntactic and syntactic annotations [Chiarcos, 2010]). As more authorities and companies publish their resources as linked data, the linguistic linked data cloud will get increasingly interconnected, towards an easier integration across datasets.

Finally, Figure 5.4 shows the same cloud shown in Figure 5.3, but according to the type of license released along with the linguistic resource. As can be seen,

[4] `http://lexinfo.net/ontology/2.0/lexinfo.owl`

[5] `http://www.acoli.informatik.uni-frankfurt.de/resources/olia/html/`

almost all the interconnected resources are open, with the notable exception of BabelNet and a few others, which are either released for non-commercial purposes (but still open for research purposes) or closed (such as JRC-Names, a multilingual named entity resource for person and organisation names). An important remark is that the cloud now contains many more resources than it used to do in the past but is currently characterized by a few hub nodes (such as BabelNet, DBpedia, lemonUby and lexinfo) and several satellite nodes linking to the hubs. The hope in the linked data ecosystem is to see the cloud grow in the next few years, as user communities get involved and new European projects are funded to foster the publishing and the exploitation of the linguistic linked open data cloud. In this line we find, for instance, European projects such as LIDER[6], born with the mission of establishing a roadmap [Sasaki, 2014, Klimek and Sasaki, 2014, Sasaki, 2015] and a community [Lewis et al., 2014] around the linked data ecosystem, by fostering public dialogue and developing a reference architecture [Koidl et al., 2014] and best practices [Cimiano et al., 2014].

5.1.1 Linked data principles and the LLD cloud

Publishing resources to the Web is not all about putting data available online. The material being published should preferably satisfy a set of principles outlined by Tim Berners Lee[7] [Bizer et al., 2009a] and further fostered by the Linking Open Data community,[8]:

1. *Things should be described by URIs*. Things described as linked data should have resolvable http:// URIs (unique resource identifiers) as identifiers;

2. *Avoid shallow URI*. The URIs should not just identify objects, but also refer to a concrete resource which should be possible to see, download and reuse later on (in this case they are called dereferenceable URI);

3. *Reuse existing standards*. Linked data should be accessible: this is performed by providing access to a RDF dump or to a SPARQL endpoint in order to enable automatic systems to query and consume linked data. Besides, data should be represented in one of the established RDF formats (RDFa, RDF/XML, Turtle, N-Triples);

4. *Foster the interconnection*. In order to have an increasingly interconnected linked data cloud, resources published as linked data should be connected to other linked data (i.e., by including links to or being linked by resources already in the LOD cloud diagram). The lower bound to see your own resource

[6] `http://www.lider-project.eu/`

[7] Tim Berners-Lee (2006-07-27). "Linked Data—Design Issues". `http://www.w3.org/DesignIssues/LinkedData.html`

[8] `http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`

added to the cloud is to contain at least 1,000 triples and be connected to other resources in the cloud with at least 50 links.

The first two rules promote data consistency, the third is intended for fostering reuse and dynamic access to data, while the last rule has been added for filtering out resources too small for deserving attention or resources providing only some spurious links.

5.1.2 RDF, URIs and vocabularies

RDF model When data needs to be published as linked data, we represent the information using the Resource Description Framework (RDF). This model allows to describe objects and the relations occurring among these. For example, the information that

Colin Firth is-a Person

would be published as linked data in the following manner:

http://en.wikipedia.org/wiki/Colin_Firth	(Colin Firth)
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	(is-a)
http://xmlns.com/foaf/0.1/Person	(person)

As can be seen, each piece of information, according to the RDF model, is represented as a *triple*, with a *subject*, a *predicate* and an *object*. In the previous statement, *Colin Firth* is the subject (encoded with *http://en.wikipedia.org/wiki/Colin_Firth*), *is-a* is the predicate (encoded with a standard URI as *http://www.w3.org/1999/02/22-rdf-syntax-ns#type*, see below) and *person* is the object (encoded as *http://xmlns.com/foaf/0.1/Person*).

Objects can either be URI (unique identifiers, see below) or simple labels. For example, for representing the first name of a person, the property *foaf:firstName* can be used, which has *rdfs:domain foaf:Person* and *rdfs:range rdfs:Literal*, meaning that any string can be provided as object of the triple. Literals are used to identify values such as strings, numbers, dates. Usage examples of literals are *foaf:name "Jack"*, *foaf:age "28"^^xsd:integer* or *foaf:birthday "2010-03-23T13:40:22.489+00:00"^^xsd:dateTime*.[9]

The RDF model is as simple as it is powerful. Considering that every common fact can be expressed by means of a triple of the kind subject-predicate-object, it stands to reason that all data, under all its forms, can be easily published as linked data. The model, in fact, has been kept simple deliberately. Note that RDF objects can in turn be also subjects of other RDF triples (e.g., Person is-a Agent). The mathematical model behind the scenes is that of directed graph, in which RDF

[9] See `http://www.w3.org/TR/rdf11-concepts/#section-Graph-Literal` for further details.

subjects and objects are encoded as nodes and RDF predicates are represented by labeled edges between nodes.

URI Each (non literal) component of RDF triples is encoded as a *Unique Resource Identifier* (URI) which uniquely identifies a resource on the Web. Examples of URIs are: *http://en.wikipedia.org/wiki/Colin_Firth*, *http://brown.nlp2rdf.org/lod/a01.ttl* or *http://xmlns.com/foaf/0.1/Document*.

In order to ease the readability and reduce the size of the RDF files, the base URI of resources is often shortened to a reference string, called *prefix*. For example, the URI *http://xmlns.com/foaf/0.1/Person* is usually contracted to *foaf:Person*, where the prefix *foaf:* is a shortcut for the more verbose string *http://xmlns.com/foaf/0.1/*. The triple shown above would then be encoded as follows:

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix wiki: <http://en.wikipedia.org/wiki/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

wiki:Colin_Firth rdf:type foaf:Person .

Some predicates are so common that they get shortened even more. The prototypical example is that of *rdf:type*, more than ever replaced by the single letter '*a*', meaning 'is-a'.

Vocabularies Relationships between RDF entities are regulated by ontologies, known also as *vocabularies* [Gómez-Pérez, 1999, Fernández-López and Gómez-Pérez, 2002]. Ontologies contain the definition of classes and properties between classes. The primitives of all ontologies are the *rdfs:Class* and *rdf:Property*.[10] A RDF resource is considered a class if it is declared to be an instance of rdfs:Class (by means of the predicate rdf:type). For example, assume we want to define a new ontology for the representation of entities in cinema, having *http://cinemaontology.org/* as base uri, encoded with the *cinema:* prefix. Adding a new class which represents an actor is then as simple as defining "*cinema:Actor* a rdfs:Class". Also new predicates can be defined in order to create a new type of relationship between two RDF resources, by means of the rdf:Property.[11] For example one could define a new property called *cinema:plays* which represents the act of playing as an actor in a movie. In order to constrain the types of classes which can be compatible subjects and objects of a property, two additional predicates have been introduced, respectively: rdfs:domain and rdfs:range (subclasses of rdf:Property). For example, in order to enforce the subjects of the property *cinema:plays* to be actors, the cinema ontology

[10] The *rdfs:* and the *rdf:* prefixes are shortcuts for `http://www.w3.org/2000/01/rdf-schema#` and `http://www.w3.org/1999/02/22-rdf-syntax-ns#`, respectively.

[11] Note, however, that rdf:Property is in turn defined as an instance of rdfs:Class.

should add the following constrain to the definition of *cinema:plays*: "cinema:plays rdf:domain cinema:Actor".

Different semantic web communities have developed several ontologies over time, with the aim of modelling different common aspects of knowledge. Notable examples include:

- **Friend-of-a-Friend** (FOAF),[12] a vocabulary for describing people, their attributes and relations. For example, the property *foaf:mbox* models the fact that a person has an e-mail address.
- **Dublin Core** (DC)[13] defines general metadata attributes.
- **Simple Knowledge Organization System** (SKOS),[14] a vocabulary for representing taxonomies and loosely structured knowledge. For example, the property owl:sameAs is used to link two individuals in an ontology, and indicate that they are the same individual.
- **Creative Commons** (CC),[15] a vocabulary for describing license terms.

5.1.3 Problems

One of the most common problems still existing in the linked data world is that of publishing a new resource (e.g., a linguistic dataset, a lexicon, etc.). A lot of technical issues arise when trying to do so; fortunately, though, there exist many guidelines and best practises which explain how to best perform this task [Cimiano et al., 2014].[16] In general, when publishing a new resource as linked data, one should keep into account the following aspects:

- **URI design**: URI, at the very least, should be persistent and abstract away from all implementation details (such as extensions about the particular scripting language used)[17];
- **Correctly modelling our resource**: choose the right representation for the objects to publish (e.g., which classes, which properties to use) and study how to correctly link our resource to others (e.g., understand the difference between skos:sameAs vs. skos:exactMatch, etc.);
- **Reuse of vocabularies**: whenever possible, existing terms and vocabularies should be reused in order to limit redundancy [Fernández-López et al., 2013].

[12] `http://xmlns.com/foaf/0.1/`

[13] `http://purl.org/dc/elements/1.1/`

[14] `http://www.w3.org/2004/02/skos/core#`

[15] `http://creativecommons.org/ns#`

[16] For a detailed discussion we point the interested reader to `https://www.w3.org/community/bpmlod/`, one of the emerging communities for best practises, or to [Heath and Bizer, 2011].

[17] See `http://www.w3.org/Provider/Style/URI.html`.

5.2 Converting MultiWiBi to RDF

As a case in point, we present the conversion of MultiWiBi data into linked data. The next section will instead present a Web-based visual explorer which easily integrates search facilities with customization tools to personalize the user's experience, as well as a single-click facility for exporting the displayed data as RDF.

The first problem we had to solve was how to represent Wikipedia pages and Wikipedia categories. DBpedia[18] is the most famous project which represents the world in the same way Wikipedia does, by using Wikipedia objects (pages, categories, infoboxes, etc). We decided to convert the MultiWiBi dataset by using a proprietary ontology, linked to the DBpedia entries. Currently the conversion has been carried out only in English, but the conversion of MultiWiBi will be extended to the multilingual case soon [Vila-Suero et al., 2014].

We thus find ourselves in the situation to model three types of objects:

- Wikipedia articles: these include both pages and categories;
- Cross-links: the links existing between a page and a category;
- Hypernymy relations: relationships linking a Wikipedia article to its generalization(s).

Modeling Wikipedia articles In order to represent WiBi's objects, we created the prefix *http://wibitaxonomy.org/* that, once attached to the article's title, uniquely identifies the URI in our domain. The encoding of a Wikipedia article is then just the juxtaposition of the prefix and the article's title.

For example, the URI of the Wikipedia page INTERNATIONAL SEMANTIC WEB CONFERENCE is encoded as follows:

```
wibi:International_Semantic_Web_Conference a skos:Concept;
```

while the Wikipedia category CONFERENCES is encoded as:

```
<http://wibitaxonomy.org/Category:Conferences> a skos:Concept;
```

Note that URIs enconding Wikipedia pages and articles are both defined as subclass of *skos:Concept* (by means of the *a* predicate), under the assumption that each article represents a concept. It might well be, though, that a Wikipedia page and a Wikipedia category encode the same meaning (such as in the case of the page CHARLIE HEBDO and the category CHARLIE HEBDO), but we are not able to establish equivalence links yet.

[18] dbpedia.org

```
@prefix wibi: <http://wibitaxonomy.org/> .
@prefix wibi-model: <http://wibitaxonomy.org/model/wibi#> .
@prefix skos:  <http://www.w3.org/2004/02/skos/core#> .

wibi:International_Semantic_Web_Conference a skos:Concept;
  wibi-model:hasWikipediaCategory  <http://wibitaxonomy.org/Category:Web-related_conferences> ;
  skos:broader               wibi:Academic_conference .

<http://wibitaxonomy.org/Category:Conferences> a skos:Concept ;
  wibi-model:hasWikipediaPage  wibi:Academic_conference ;
  skos:narrower      <http://wibitaxonomy.org/Category:Technology_conferences> .
```

Figure 5.5. RDF excerpt of the taxonomy view for the ISWC Wikipedia page.

Modeling hypernymy relations We decided to leverage the SKOS predicates *skos:broader* and *skos:narrower* which associate a *skos:Concept* with its generalization or its specification, respectively. The is-a relation between the Wikipedia page INTERNATIONAL SEMANTIC WEB CONFERENCE and ACADEMIC CONFERENCE is for example encoded as:

```
wibi:International_Semantic_Web_Conference a skos:Concept ;
skos:broader               wibi:Academic_conference .
```

Modeling cross-links Cross-links encode relations between Wikipedia pages and categories (cf. Section 2.2 in Chapter 2). For example, the page INTERNATIONAL SEMANTIC WEB CONFERENCE is associated with many categories, among which WEB-RELATED CONFERENCES. We decided thus to introduce two new predicates in our model: *hasWikipediaCategory* and *hasWikipediaPage*. The first encodes the fact that a Wikipedia page has a Wikipedia category associated, while the second encodes the fact that a Wikipedia category is associated to a Wikipedia page. Both the predicates have *domain* and *range* equal to *skos:Concept* and are defined in the model *http://wibitaxonomy.org/model/wibi#*.

5.3 The Web interface

We will now describe the Web interface of MultiWiBi. The interface, available at `wibitaxonomy.org`, welcomes the user with a home page which enables a dynamic search of Wikipedia articles and displays an excerpt of the bitaxonomy centered on the queried item (result page). It is possible to customize different visualization aspects and automatically export displayed data into RDF.

The home page. An excerpt of the interface's home page is shown in Fig. 5.6. As can be seen, this page has been kept very clean with as few elements as possible. On the top of the page a navigation bar contains links to i) the *about* page, which contains release information about the website content, ii) a *download* area, where it is possible to obtain the data underlying the interface and iii) the *search* page, which represents the core contribution of this work.

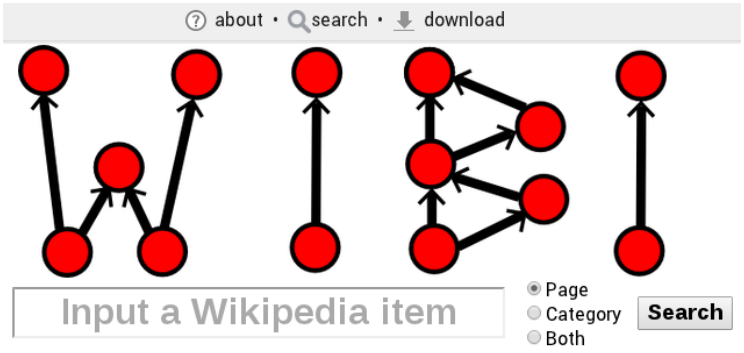


Figure 5.6. The Wikipedia Bitaxonomy Explorer home page.

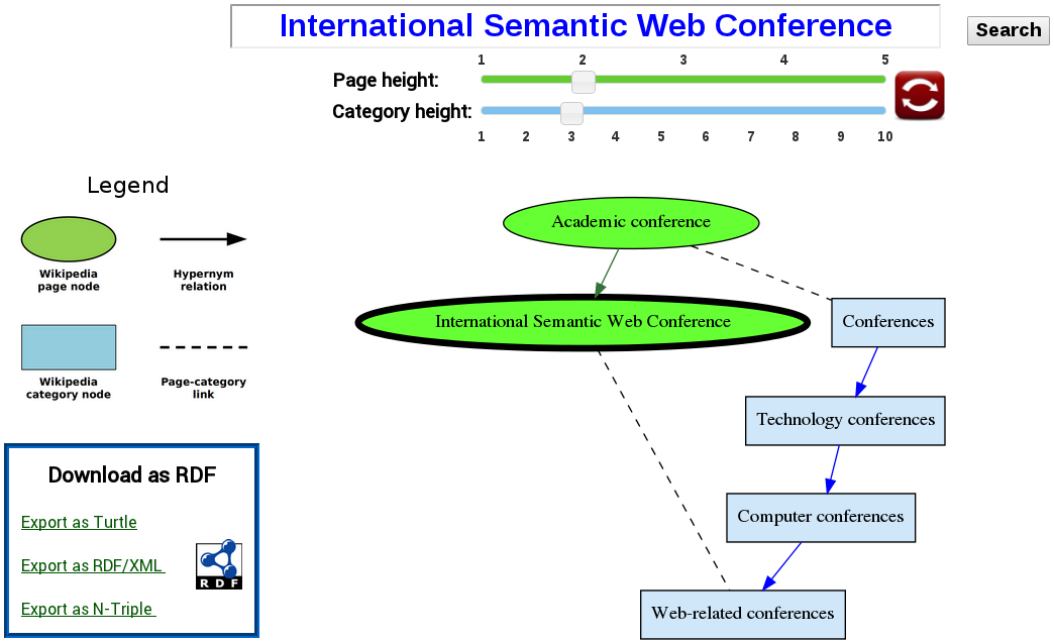


Figure 5.7. Result for the ISWC Wikipedia page.

The search page mainly contains a text area in which the user is requested to input her query of interest, additionally opting for searching through either the page inventory, the category inventory or both, thanks to dedicated radio buttons. After the query is sent, the search engine tries to match the input text against the whole database of Wikipedia pages (or categories) and, if a match is found, the engine displays the final result to the user. Otherwise, the query is interpreted as a lemma and the user is returned with the (possible) list of all Wikipedia pages/categories whose lemma matches against the query.

The result page. Starting from the Wikipedia element provided by the user, the objective of the result page is to show a relevant excerpt of the bitaxonomy, that is, the nearest (or more relevant) nodes connected to it, drawn from both of the two

taxonomies. To do this, MultiWiBi Explorer performs a series of steps:

1. Start a DFS of maximum length δ_1 from the given element p of a taxonomy. As a result, a subgraph $ST_1 = (SV_1, SE_1)$ is obtained;
2. Collect all the nodes $\sigma(p)$ belonging to the other taxonomy (i.e, those whose cross-edges are incident to p). Start a DFS of maximum length δ_2 from each element in $\sigma(p)$. As a result, a subgraph $ST_2 = (SV_2, SE_2)$ is obtained;
3. Display ST_1 and ST_2, as well as all the possible cross-edges linking nodes of the two subgraphs. Prune out low-connected nodes from the displayed bitaxonomy.

As a result, the interface displays a meaningful excerpt of the two taxonomies, centered on the issued query. The result for the Wikipedia page INTERNATIONAL SEMANTIC WEB CONFERENCE is shown in Fig. 5.7.

Customization of the view Since a user might be interested in a more general view of the bitaxonomy, two additional sliders are provided to the user in order to manually adjust the two maximum depths δ_1 and δ_2 (see Fig. 5.7 on top). Moreover, the interface provides the user with the capability to click on nodes and interactively explore different parts of the taxonomy. The application thus acts as a dynamic explorer that enables users to navigate through the structure of the bitaxonomy and discover new relations as the visit proceeds.

RDF export facility of the interface Interestingly, data can also be exported in RDF format, in line with recent work on (linguistic) linked open data and the Semantic Web [Ehrmann et al., 2014]. To this end, the explorer is backed by the Apache Jena framework[19] and thus also integrates a single-click functionality that seamlessly converts the displayed data into RDF format. The user can opt for Turtle, RDF/XML or N-Triples format (see blue box in Fig. 5.7, bottom left). An excerpt of a view of the bitaxonomy converted into RDF for the query ISWC is shown in Fig. 5.5.

Integration into BabelNet The hypernym relations contained in MultiWiBi have notably been integrated also into BabelNet 3.0.[20] BabelNet [Navigli and Ponzetto, 2010] is an encyclopedic semantic network which provides knowledge for millions of concepts. At the core of BabelNet lies the integration of the dictionary knowledge offered by WordNet and the encyclopedic information offered by Wikipedia, aligned thanks to a high-precision mapping established between the two resources. As a result, BabelNet encodes not only dictionary words (e.g., *house*, *balloon*, etc.), but also encyclopedic entries (e.g., HOUSE (2008 FILM), HOT AIR BALLOON, etc). Notably, there is a small, but important fraction of synsets which represents the

[19] `https://jena.apache.org/`
[20] `babelnet.org`

information encoded in the two sources and merged thanks to the mapping: for instance, BALLOON (AIRCRAFT) and $balloon_n^1$) encode the same meaning and, in fact, appear in the same synset. Thanks to MultiWiBi, BabelNet now provides full hypernymy coverage of both Wikipedia pages and categories: in fact, while all the synsets containing a WordNet sense were already covered since version 1.0, now nearly all the synsets are covered, thanks to the extensive is-a knowledge offered by MultiWiBi.

The integration of MultiWiBi into BabelNet is part of a bigger picture which tries to reconcile the hypernymy information coming from different sources. The process is automatic but, however, does not draws on linked data mechanisms yet. For example, in order to provide the hypernym for the synset *bn:03370424n*, containing the Wikipedia page FLY ME TO THE MOON, MultiWiBi is exploited and the hypernym SINGLE (MUSIC) is finally extracted. In this case, the synset does not integrate information coming from WordNet, so MultiWiBi is the only source of hypernymy information available. When multiple sources of information are available, though, the hypernymy information can be combined. For example, the synset *bn:00018323n* (containing the Wikipedia page PRESIDENT OF THE UNITED STATES as well as the WordNet synset with offset *10467395* and the Wikidata entry *chief_executive*, among others) has four hypernyms in total (three coming from WordNet and one coming from Wikidata). In case hypernyms coming from more accurate resources are available, the latters are preferred over hypernyms coming from MultiWiBi.

5.4 Conclusions

In this chapter we have introduced the world of linked data, its fundamentals and the principles at the base of the linked data cloud, a virtual space where resources are published and linked to each other. We presented the case of the conversion of MultiWiBi into RDF, showing how to represent hypernymy information extracted from Wikipedia. We introduced also the Wikipedia Bitaxonomy Explorer, an extensible Web interface that allows the navigation of the recently created Wikipedia Bitaxonomy [Flati et al., 2014] and MultiWiBi, its extension to the multilingual setting. In the interface, in addition to the default settings, several parameters concerning the general appearance of the results can also be customized according to the user's preferences. The demo is available at `wibitaxonomy.org` and seamlessly integrated into the BabelNet interface (`http://babelnet.org/`) while the data is freely downloadable under a CC BY-NC-SA 3.0 license. A future work might be that of providing access to the RDF data of MultiWiBi also through a SPARQL endpoint to foster further reuse of the linked data.

Chapter 6

Conclusions and Future Work

Knowledge is growing in size at an unprecedented speed. Now, more than ever, we need an automatic systematization of such knowledge which, otherwise, remains hardly navigable. One of the primary efforts in this direction is that of taxonomizing the entities and the concepts involved, by means of automatic systems which extract as many is-a relations as possible and structure them into a full-fledged hierarchy of concepts. At the same time, though, we also require systematization to preserve the quality of the original data being structured so as not to commit errors during subsequent automatic processing of derived information. Due to the difficulty of the task, the amount of human involvement plays a major role; if on the one hand human effort generally leads to higher performance over automatic systems, on the other hand constructing or validating data by hand is a time-consuming process and makes the reproducibility or the extension of the approach to different settings much harder, due to the need to repeat the whole process, often from scratch (for example, by providing seeds in another domain or in another language). The incentive to design automatic systems which attain performance comparable to that achieved by humans has thus led to a very fertile area which, over the last decades, has generated several taxonomic resources, obtained automatically [Navigli and Velardi, 2004, Auer et al., 2007, Bollacker et al., 2008, Ponzetto and Navigli, 2009, de Melo and Weikum, 2010a, Ponzetto and Strube, 2011, Hoffart et al., 2013, Nastase and Strube, 2013, Velardi et al., 2013, Kliegr et al., 2014].

Along this line, this thesis presents MultiWiBi, an innovative, state-of-the-art multilingual Wikipedia bitaxonomy which arranges Wikipedia articles into a well-defined structure. Historically speaking, previous efforts either focused only on Wikipedia categories and only more recently they started to taxonomize the Wikipedia pages or both the sides of Wikipedia. MultiWiBi is new for many different aspects. First of all, thanks to a novel iterative algorithm, we show that the information contained in Wikipedia pages and Wikipedia categories is strictly intertwined and mutually beneficial in order to build what has been called a *bitaxonomy*, a pair of taxonomies, each structuring either pages or categories. Second, similarly to WikiNet, DBpedia and MENTA, it is applicable to all the languages in which Wikipedia has been written, but, differently from these, the bitaxonomy in each

language is obtained in a linguistically agnostic manner, without relying on any additional resource except for Wikipedia itself. An outstanding feature which puts MultiWiBi one step further is its power to cover concepts which are present in a specific language but do not have an equivalent in English (such as ZAGAROLO, a renown Italian wine, not yet described in the English Wikipedia).

We also presented the various taxonomic resources currently available in the literature and we analysed and compared these resources against MultiWiBi, according to several dimensions. We performed a thorough evaluation that assesses the quality of each resource and demonstrated that MultiWiBi consistently provides state-of-the-art results, with a granularity and specificity superior to all other alternative approaches. We also introduced two novel measures which ease comparison across resources. The first global measure is called granularity and measures the number of distinct hypernyms returned by a system (thus differentiating between resources with only hundreds hypernyms from those providing considerable discriminative power). The second measure, called specificity, aims at assessing the degree of specialization of the hypernyms returned by a system (so as to favor systems returning SINGER rather than PERSON).

At the core of MultiWiBi lies a mixture of content and structural features which, combined, lead to a high quality bitaxonomy covering millions of concepts encoded by the pages of Wikipedia and hundreds of thousands categories. The process that leads to the construction of the page taxonomy is divided in two steps: the first step aims at identifying (ambiguous) hypernym lemmas for each Wikipedia page, i.e., lemmas which represent the hypernym at lexical level; a second step is then needed to disambiguate the hypernym lemmas extracted in the first step. The disambiguation step combines linkers which rely on content heuristics (e.g., by exploiting sense-tagged hypernyms added by Wikipedians) and linkers which disambiguate the hypernym lemmas according to a distributional similarity. The iterative algorithm and the refinement of the bitaxonomy, instead, are crucially based on heuristics which exploit the structure of Wikipedia (e.g., by analyzing the neighbourhood of pages and categories).

We later show the seamless extension to the multilingual case which not only preserves quality but has the nice added value of covering concepts which do not have an English equivalent. We developed and analyzed a novel mechanism, namely the Translation Tables, to generate multilingual translations for hundreds of thousands English lemmas which are subsequently exploited to obtain hypernym lemmas for Wikipedia pages in other languages. Remarkably, our approach for generating multilingual bitaxonomies is language agnostic: the identical algorithmic procedures employed in the English case are, in fact, applied exactly as they are to any other arbitrary Wikipedia language.

Aware that releasing bitaxonomies for many languages is however not sufficient for a complete interoperability and reuse in many other tasks and resource, we also published MultiWiBi as linked data, a paradigm which fosters interoperability and interconnection among resources on the Web. We also developed a public interface which lets users navigate through the MultiWiBi's taxonomic structure in a graphical,

appealing way. The interface also enables users to download hypernymy information represented as linked data, in the most established formats, namely Turtle, N-triple and RDF/XML.

As a final study, in Chapter 4 we have shown the impact that MultiWiBi can have on semantic systems, by means of the case study of SPred, an approach to generalize arguments of given lexical predicates. Given an arbitrary sequence of words, called lexical predicate, SPred matches the latter against a textual corpus and, by means of linking and disambiguation heuristics, infers the probability distribution associated with the lexical predicate (e.g., $beverage^1_n$ as the expected class for the lexical predicate *cup of* *). In order to perform the generalization step, SPred leverages on WCL, a system for the automatic extraction of hypernym from textual definitions, for building a taxonomy over Wikipedia pages. However, due to the limited coverage of WCL, many disambiguated arguments are not generalized effectively. Thus, we have shown that replacing WCL with MultiWiBi not only improves SPred's generalization power but also its performance when evaluated in the task of word sense disambiguation.

However, many questions remain open to future work, linked either to idiosyncrasies of Wikipedia or to the limitations of MultiWiBi and SPred:

- Wikipedia lacks independent articles which encode certain concepts. Many common concepts, including more concrete ones (such as SINGER or VOLLEYBALL PLAYER), as well as more abstract ones (such as the terms *extraction* or *member* are still missing). The Wikipedia meanings of the word *member*, for example, include only a subset of the senses contained in WordNet, such as LIMB (ANATOMY) or MEMBER (LOCAL CHURCH), but the sense of 'member belonging to a club' is definitely still needed. Of course, due to their level of abstraction, some concepts might be well harder to define than others and might be inappropriate for an encyclopedia. Nonetheless, many concepts are currently still out of the inventory: as explained in Chapter 2, in fact, some of these concepts are encoded, but only as *shallow* redirections. On the other hand, it is also true that Wikipedians seem to be conscious of this lack and continuously try to remedy, as witnessed by the gradual creation of new independent articles, absent before (e.g., PHILOSOPHER, a redirection between 2006 and 2012, when it was promoted to an independent article). Future work might for example investigate more in depth new methods to compensate this lack (as done by BabelNet [Navigli and Ponzetto, 2010], which defines a mapping from arbitrary articles – including redirections – to WordNet synsets).

- Our choice to integrate redirections as possible senses for hypernym lemmas brought in a major drawback not present in the previous work, represented by the fact that, since redirections lack an explicit definition, it is not possible to extract hypernym lemmas for these. This poses a problem when finding hypernym senses for the latter. Currently we adopt two heuristics which

overcome this limitation at the structure level, but future work might further investigate how to outplay this more effectively.

- One of the major problems we encountered was represented by the phenomenon we called *semantic shift*, consisting in a drift of meaning between two highly correlated concepts, such as SINGER and SINGING or FOOTBALL PLAYER and FOOTBALL. We developed a module which, based on handcrafted rules, detects the main shifts. An ad-hoc algorithm is definitely needed here, since tackling this problem effectively could lead to a major boost in performance. Not only could future work lead to the development of an improved version of the SSR module, but could also investigate how to extend the latter to the multilingual case, where currently we are not able to reproduce the detection automatically (if not at the cost of duplicating the human effort by means of ad-hoc patterns, similarly to those implemented for English).

- Our hypernym lemma extraction builds on the assumptions that i) an article is well defined by a high-quality gloss, ii) the hypernym is explicitly contained in the gloss and iii) the hypernym lemma can be harvested by means of the simple syntactic relation *copula*. However, data showed (see Section 2.4.1 in Chapter 2) that definitions (more in other languages than English) are often ill-formed, meaning that the hypernym lemma is either missing or not in an is-a relation. Our multilingual hypernym lemma extraction module crucially exploits hypernym lemmas found in English to assign hypernym lemmas to articles in other languages, by means of Translation Tables; however, more sophisticated approaches, based on more complex lexico-syntactic as well as distributional features, could lead to overall higher quality. In addition to this, we acknowledge that low-quality extraction of hypernym lemmas in languages other than English is often correlated to the low quality of the underlying data; concepts in other languages are, in fact, often defined only in an indirect way (e.g., for celebrities, by citing the successes achieved in life or by recalling co-authors and collaborators, without explicitly stating his/her profession or role in the world). As said above, it is also true that, as time passes and more contributors get involved, the quality of Wikipedia articles is constantly getting better; as a result, also Wikipedia definitions slowly see their quality improving over time.

- A major phenomenon affecting MultiWiBi, as well as all other taxonomic systems relying on interlanguage links, is the low number of interlanguage links between two editions of Wikipedia in different languages. This type of information is unfortunately hard to maintain aligned, especially when new articles are created and removed in an asynchronous manner. Strictly intertwined with the problem of sense inventory, this Wikipedia idiosyncrasy considerably hinders the alignment of common concepts such as SINGER, SURGEON and NOVELIST, to name a few. There have been efforts to automatically align equivalent knowledge in wiki-like knowledge bases written in

different languages [Lefever et al., 2012, Wang et al., 2012, Sorg and Cimiano, 2008]. However, not only are these based on supervised classification, so that a new model should then be learned for each language pair, but they are also applied only to articles, while it is not clear how to extend their applicability to redirections. Future work could try to integrate these into MultiWiBi's workflow.

- Another research direction would be to explore a different approach which builds bitaxonomies by combining all the Wikipedia languages together, by integrating the information coming from many different sources all at once, and not only pair-wise, like we do in Chapter 2.

- As regards SPred, it is currently limited in that it generalizes only nouns; it remains thus unclear how to generalize SPred over arbitrary parts of speech, other than nouns only. This is mainly due to its dependence on an external taxonomy defined over nouns, whereas no taxonomy is yet available for adjectives and adverbs (WordNet provides some support for verb generalization, but this is limited, since verbs usually have only one or two ancestors up in the hierarchy).

- Another direction left to future work is certainly the application of SPred to large textual corpora, such as Gigaword or ClueWeb [Pomikálek et al., 2012] (a collection of 1 billion Web pages in ten languages, estimated to be more than 1% of the whole Web), in order to get more variegated sets of lexical arguments, as well as more lexical predicates. The integration of resources external to the textual corpus might be very useful; an example of this is WikiLinks[1] which contains over 10M of Web pages with more than 40M of entity mentions under the forms of links to Wikipedia. The dataset comes along with mention contexts so that it might be also used as an additional source corpus. In addition, it would be desirable to automatically identify semantic predicate equivalences; at the current stage, in fact, 'write $book^1_n$' and '$book^1_n$ is written by', while being two equivalent expressions of the same fact, are considered different at the surface level.

- A direction which has finally completely been left out is the extension of SPred to the multilingual setting. Furthermore, while MultiWiBi has been integrated into SPred only in English, it might instead play a crucial role also when dealing with other languages. In particular, it would be interesting to explore how to align lexical predicates expressed in different languages, automatically [Lewis and Steedman, 2013]. For example, knowing that 'sing $song^1_n$' translates into 'cantare $song^1_n$' in Italian or 'chanter $song^1_n$' in French, would dramatically help tasks such as question answering [Poon and Domingos, 2009] or textual entailment [Mehdad et al., 2010].

[1] `http://www.iesl.cs.umass.edu/data/wiki-links`

Bibliography

Eneko Agirre and David Martinez. Learning class-to-class selectional preferences. In *Proceedings of the Fifth Workshop on Computational Language Learning (CoNLL '01)*, pages 15–22, 2001.

Robert A. Amsler. A Taxonomy for English Nouns and Verbs. In *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics (ACL '81)*, pages 133–138, Stanford, California, USA, 1981.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ive. DBpedia: A nucleus for a web of open data. In *Proceedings of 6th International Semantic Web Conference joint with 2nd Asian Semantic Web Conference (ISWC+ASWC 2007)*, pages 722–735, Busan, Korea, 2007.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL '98)*, pages 86–90, Montreal, Canada, 1998.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI '07), Hyderabad, India, 6–12 January 2007*, pages 2670–2676, 2007.

Michele Banko, Oren Etzioni, and Turing Center. The Tradeoffs Between Open and Traditional Relation Extraction. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL '08)*, volume 8, pages 28–36, 2008.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. Learning entailment relations by global graph structure optimization. *Computational Linguistics*, 38(1):73–111, 2012.

Shane Bergsma, Dekang Lin, and Randy Goebel. Discriminative learning of selectional preference from unlabeled text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, pages 59–68, Stroudsburg, PA, USA, 2008. URL `http://dl.acm.org/citation.cfm?id=1613715.1613725`.

Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data – the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009a.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - a crystallization point for the Web of Data. *Web Semantics*, 7(3):154–165, 2009b. ISSN 1570-8268. doi: 10.1016/j.websem.2009.07.002. URL http://dx.doi.org/10.1016/j.websem.2009.07.002.

Sebastian Blohm. Using the web to reduce data sparseness in pattern-based information extraction. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 18–29, Warsaw, Poland, 2007. Springer.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the International Conference on Management of Data (SIGMOD '08)*, SIGMOD '08, pages 1247–1250, New York, NY, USA, 2008. ISBN 978-1-60558-102-6. doi: 10.1145/1376616.1376746. URL http://doi.acm.org/10.1145/1376616.1376746.

Francis Bond and Ryan Foster. Linking and extending an open multilingual wordnet. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL '13)*, pages 1352–1362, 2013.

Gerlof Bouma. Collocation Extraction beyond the Independence Assumption. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10), Short Papers*, pages 109–114, Uppsala, Sweden, 2010. URL http://www.aclweb.org/anthology/P10-2020.

Paul Buitelaar, Philipp Cimiano, Peter Haase, and Michael Sintek. Towards linguistically grounded ontologies. In *The semantic web: research and applications*, pages 111–125. Springer, 2009.

Aljoscha Burchardt, Katrin Erk, and Anette Frank. A WordNet Detour to FrameNet. In Bernhard Fisseni, Hans C. Schmitz, and Bernhard Schröder, editors, *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, volume 8 of *Computer Studies in Language and Speech*, Frankfurt am Main, 2005. Peter Lang.

Nicoletta Calzolari. Towards The Organization Of Lexical Definitions On A Database Structure. In *Proceedings of the 9th Conference on Computational Linguistics (COLING '82)*, pages 61–64, Prague, Czechoslovakia, 1982.

Nicoletta Calzolari, Laura Pecchia, and Antonio Zampolli. Working on the Italian Machine Dictionary: a Semantic Approach. In *Proceedings of the 5th Conference on Computational Linguistics (COLING '73)*, pages 49–70, Pisa, Italy, 1973.

Sharon A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Md., 20–26 June 1999*, pages 120–126, 1999.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the 24th Annual Meeting of the Association for the Advancement of Artificial Intelligence (AAAI '10)*, pages 1306–1313, Atlanta, Georgia, 2010.

Nathanael Chambers and Dan Jurafsky. Improving the use of pseudo-words for evaluating selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pages 445–453, Stroudsburg, PA, USA, 2010. URL `http://dl.acm.org/citation.cfm?id=1858681.1858727`.

Christian Chiarcos. Grounding an ontology of linguistic annotations in the data category registry. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC '10), Workshop on Language Resource and Language Technology Standards (LT&LTS), Valetta, Malta*, pages 37–40, 2010.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. Towards a Linguistic Linked Open Data cloud: The Open Linguistics Working Group. *TAL*, 52 (3):245–275, 2011.

Jennifer Chu-Carroll and John Prager. An experimental study of the impact of information extraction accuracy on semantic search performance. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM '07)*, pages 505–514, Lisbon, Portugal, 2007.

Massimiliano Ciaramita and Yasemin Altun. Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*, pages 594–602, Sydney, Australia, 2006.

Philipp Cimiano, John McCrae, Jorge Gracia, Bettina Klimek, Martin Brümmer, Ciro Baroni, Dave Lewis, and Víctor Rodríguez-Doncel. D2.1.1 - guidelines and best practices for linguistic linked data-based content analytics - phase i - v2.0, October 2014. URL `http://www.lider-project.eu/sites/default/files/D2.1.1.Phase-I-v2.0.pdf`.

Stephen Clark and David Weir. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206, 2002.

Jonathan Crowther, editor. *Oxford Advanced Learner's Dictionary*. Cornelsen & Oxford, 5th edition, 1998.

Hang Cui, Min-Yen Kan, and Tat-Seng Chua. Soft Pattern Matching Models for Definitional Question Answering. *ACM Transactions on Information Systems*, 25 (2), 2007. ISSN 1046-8188. doi: 10.1145/1229179.1229182.

Flavio De Benedictis, Stefano Faralli, and Roberto Navigli. GlossBoot: Bootstrapping multilingual domain glossaries from the Web. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL '13)*, pages 528–538, Sofia, Bulgaria, 2013.

Gerard de Melo and Gerhard Weikum. On the utility of automatically generated wordnets. In *Proceedings of the 4th Global WordNet Conference (GWC '08)*. Citeseer, 2008a.

Gerard de Melo and Gerhard Weikum. A machine learning approach to building aligned wordnets. In *Proceedings of the First International Conference on Global Interoperability for Language Resources*, pages 163–170, 2008b.

Gerard de Melo and Gerhard Weikum. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the Eighteenth ACM Conference on Information and Knowledge Management, Hong Kong, China, 2009*, pages 513–522, 2009.

Gerard de Melo and Gerhard Weikum. MENTA: Inducing Multilingual Taxonomies from Wikipedia. In *Proceedings of the 19th ACM international Conference on Information and Knowledge Management (CIKM '10)*, pages 1099–1108, New York, NY, USA, 2010a. ISBN 978-1-4503-0099-5. doi: 10.1145/1871437.1871577. URL `http://doi.acm.org/10.1145/1871437.1871577`.

Gerard de Melo and Gerhard Weikum. Towards universal multilingual knowledge bases. In *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the Fifth Global WordNet Conference (GWC 2010)*, pages 149–156, 2010b.

Gerard de Melo and Gerhard Weikum. UWN: A Large Multilingual Lexical Knowledge Base. In *Proceedings of the ACL 2012 System Demonstrations*, pages 151–156. Association for Computational Linguistics, 2012.

Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John Philip Mccrae, Philipp Cimiano, and Roberto Navigli. Representing multilingual data as linked data: the case of babelnet 2.0. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Katrin Erk. A Simple, Similarity-based Model for Selectional Preferences. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL '07)*, pages 216–223, Prague, Czech Republic, 2007.

Katrin Erk and Diana McCarthy. Graded word sense assignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, pages 440–449, Stroudsburg, PA, USA, 2009. ISBN 978-1-932432-59-6. URL http://dl.acm.org/citation.cfm?id=1699510.1699568.

Katrin Erk, Sebastian Padó, and Ulrike Padó. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4): 723–763, 2010.

Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Web-scale information extraction in knowItAll: (preliminary results). In *Proceedings of the 13th International Conference on World Wide Web (WWW '04)*, pages 100–110, New York, NY, USA, 2004. ACM. ISBN 1-58113-844-X. doi: 10.1145/988672.988687. URL http://doi.acm.org/10.1145/988672.988687.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence*, 165 (1):91–134, 2005.

Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying Relations for Open Information Extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pages 1535–1545, Edinburgh, UK, 2011.

Stefano Faralli and Roberto Navigli. A Java framework for multilingual definition and hypernym extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL '13)*, pages 103–108, Sofia, Bulgaria, 2013.

Christiane Fellbaum, editor. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA, 1998.

Mariano Fernández-López and Asunción Gómez-Pérez. Overview and analysis of methodologies for building ontologies. *The Knowledge Engineering Review*, 17 (02):129–156, 2002.

Mariano Fernández-López, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. Methodological guidelines for reusing general ontologies. *Data & Knowledge Engineering*, 86:242–275, 2013.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. Building watson: An overview of the deepqa project. *AI magazine*, 31(3): 59–79, 2010.

David A. Ferrucci. Introduction to "This is Watson". *IBM Journal of Research and Development*, 56(3):1, 2012.

Charles J Fillmore. Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences*, 280(1):20–32, 1976.

Darja Fišer. Using multilingual resources for building SloWNet faster. In *Proceedings of the 4th Global WordNet Conference (GWC '08)*, 2008.

Tiziano Flati and Roberto Navigli. SPred: Large-scale Harvesting of Semantic Predicates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1222–1232, Sofia, Bulgaria, 2013.

Tiziano Flati and Roberto Navigli. The Wikipedia Bitaxonomy Explorer. In *Proceedings of the 13th International Semantic Web Conference (ISWC '14) (Posters & Demonstrations)*, 2014a.

Tiziano Flati and Roberto Navigli. Three birds (in the llod cloud) with one stone: Babelnet, babelfy and the wikipedia bitaxonomy. In *Proceedings of the 10th International Conference on Semantic Systems (SEMANTiCS 2014)*, 2014b.

Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 945–955, Baltimore, Maryland, 2014. Association for Computational Linguistics.

Hagen Fürstenau and Mirella Lapata. Semi-supervised semantic role labeling via structural alignment. *Computational Linguistics*, 38(1):135–171, 2012.

Roxana Girju, Adriana Badulescu, and Dan Moldovan. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL '03)*, pages 1–8, Edmonton, Canada, 2003.

Asunción Gómez-Pérez. Ontological engineering: A state of the art. *Expert Update: Knowledge Based Systems and Applied Artificial Intelligence*, 2(3):33–43, 1999.

Asunción Gómez-Pérez, David Manzano-Macho, et al. A survey of ontology learning methods and techniques. *OntoWeb Deliverable D*, 1(5), 2003.

Jorge Gracia, Daniel Vila-Suero, John McCrae, Tiziano Flati, Ciro Baron, and Milan Dojchinovski. Language Resources and Linked Data: a Practical Perspective. In Patrick Lambrix, Eero Hyvönen, Eva Blomqvist, Valentina Presutti, Guilin Qi, Uli Sattler, Ying Ding, and Chiara Ghidini, editors, *Proceedings of the 19th International Conference on Knowledge Engineering and Knowledge Management (EKAW '14) Satellite Events, Linköping (Sweden)*, November 2014.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English Gigaword. *Linguistic Data Consortium, Philadelphia*, 2003.

Rebecca Green, Bonnie J. Dorr, and Philip Resnik. Inducing Frame Semantic Verb Classes from WordNet and LDOCE. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL '04)*, pages 375–382, Barcelona, Spain, 2004.

Patrick Hanks. *Lexical Analysis: Norms and Exploitations*. University Press Group Limited, 2013. ISBN 9780262018579. URL `http://books.google.it/books?id=MpymkHkC0WYC`.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING '92)*, pages 539–545, Nantes, France, 1992.

Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.

Evan Heit. Properties of inductive reasoning. *Psychonomic Bulletin & Review*, 7(4): 569–592, 2000.

Kiyotaka Uchimoto Masao Utiyama Hitoshi Isahara, Fransis Bond and Kyoko Kanzaki. Development of the Japanese WordNet. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2420–2423, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. http://www.lrec-conf.org/proceedings/lrec2008/.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61, 2013.

Eduard Hovy, Zornitsa Kozareva, and Ellen Riloff. Toward completeness in concept extraction and classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09) Volume 2*, pages 948–957. Association for Computational Linguistics, 2009.

Eduard H. Hovy, Roberto Navigli, and Simone Paolo Ponzetto. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27, 2013.

Ruihong Huang and Ellen Riloff. Inducing Domain-Specific Semantic Class Taggers from (Almost) Nothing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pages 275–285, Uppsala, Sweden, 2010.

Nancy Ide and Jean Véronis. Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time? In *Proceedings of the Workshop on Knowledge Bases and Knowledge Structures (KB&KS '93)*, pages 257–266, Tokyo, Japan, 1993.

Sean P. Igo and Ellen Riloff. Corpus-based semantic lexicon induction with Web-based corroboration. In *Proceedings of UMSLLS*, pages 18–26, Stroudsburg, PA, USA, 2009. ISBN 978-1-932432-34-3. URL `http://dl.acm.org/citation.cfm?id=1641968.1641971`.

Bärbel Inhelder and Jean Piaget. *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*, volume 22. Psychology Press, 1958.

Rubén Izquierdo, Armando Suárez, and German Rigau. An Empirical Study on Class-Based Word Sense Disambiguation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09)*, pages 389–397, Athens, Greece, 2009.

Hyeju Jang and Jack Mostow. Inferring selectional preferences from part-of-speech N-grams. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*, pages 377–386, Stroudsburg, PA, USA, 2012.

Boris Katz, Jimmy J. Lin, Daniel Loreto, Wesley Hildebrandt, Matthew W. Bilotti, Sue Felshin, Aaron Fernandes, Gregory Marton, and Federico Mora. Integrating Web-based and Corpus-based Techniques for Question Answering. In *Proceedings of The Twelfth Text Retrieval Conference (TREC '03)*, pages 426–435, Gaithersburg, Maryland, 2003.

Dan Klein and Christopher D. Manning. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems 15 (NIPS)*, pages 3–10, Vancouver, British Columbia, Canada, 2003.

Tomáš Kliegr, Václav Zeman, and Milan Dojchinovski. Linked hypernyms dataset-generation framework and use cases. 2014.

Bettina Klimek and Felix Sasaki. D4.7 - third roadmapping workshop report - v1.1, October 2014. URL `http://www.lider-project.eu/sites/default/files/D4.7-v1.1.pdf`.

Kevin Koidl, Dave Lewis, Philipp Cimiano, Matthias Hartung, John McCrae, Christina Unger, Víctor Rodríguez-Doncel, Asunción Gómez-Pérez, Jorge Gracia, Martin Brümmer, Sebastian Hellmann, Bettina Klimek, Sören Auer, Tiziano Flati, Roberto Navigli, Andrea Moro, and Paul Buitelaar. D3.1.1 - linguistic linked data reference architecture – phase i, November 2014. URL `http://www.lider-project.eu/sites/default/files/D3.1.1-v2.0.pdf`.

Zornitsa Kozareva and Eduard Hovy. Learning Arguments and Supertypes of Semantic Relations Using Recursive Patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pages 1482–1491, Uppsala, Sweden, 2010a. URL http://www.aclweb.org/anthology/P10-1150.

Zornitsa Kozareva and Eduard H. Hovy. A Semi-Supervised Method to Learn and Construct Taxonomies Using the Web. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*, pages 1110–1118, Seattle, WA, USA, 2010b.

Zornitsa Kozareva, Ellen Riloff, and Eduard H. Hovy. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. In *Proceedings of ACL/HLT '08*, pages 1048–1056, Columbus, Ohio, 2008.

Sebastian Krause, Hong Li, Hans Uszkoreit, and Feiyu Xu. Large-scale learning of relation-extraction rules with distant supervision from the web. In *Proceedings of the 11th International Semantic Web Conference (ISWC '12), Part I*, pages 263–278, Boston, MA, 2012. ISBN 978-3-642-35175-4. doi: 10.1007/978-3-642-35176-1_17. URL http://dx.doi.org/10.1007/978-3-642-35176-1_17.

Els Lefever, Véronique Hoste, and Martine De Cock. Discovering Missing Wikipedia Inter-language Links by means of Cross-lingual Word Sense Disambiguation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC '12)*, pages 841–846. European Language Resources Association (ELRA), 2012. ISBN 978-2-9517408-7-7. URL http://dblp.uni-trier.de/db/conf/lrec/lrec2012.html#LefeverHC12.

David Lewis, Asunción Gómez-Pérez, Kevin Koidl, and Felix Sasaki. D.4.3.2 -report on lider community and community portal – v2.0, November 2014. URL http://www.lider-project.eu/sites/default/files/D4.3.2-v2.0.pdf.

Mike Lewis and Mark Steedman. Unsupervised Induction of Cross-Lingual Semantic Relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '13)*, pages 681–692, 2013.

Hang Li and Naoki Abe. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244, 1998.

Thomas Lin, Oren Etzioni, et al. Entity linking at web scale. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX '12)*, pages 84–88. Association for Computational Linguistics, 2012.

Maddalen López de Lacalle, Egoitz Laparra, and German Rigau. *Proceedings of the Seventh Global Wordnet Conference (GWA '14)*, chapter First steps towards a Predicate Matrix, pages 363–371. 2014. URL http://aclweb.org/anthology/W14-0150.

Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754, 2009. ISSN 1071-5819. doi: http://dx.doi.org/10.1016/j.ijhcs.2009.05.004.

Yashar Mehdad, Matteo Negri, and Marcello Federico. Towards cross-lingual textual entailment. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT '10)*, pages 321–324. Association for Computational Linguistics, 2010.

Rada Mihalcea and Dan Moldovan. eXtended WordNet: Progress report. In *Proceedings of the NAACL-01 Workshop on WordNet and Other Lexical Resources, Pittsburgh, Penn., June 2001*, pages 95–100, 2001.

Shachar Mirkin, Ido Dagan, and Eyal Shnarch. Evaluating the inferential utility of lexical-semantic resources. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09)*, pages 558–566, Athens, Greece, 2009.

Tom Mitchell. Reading the Web: A Breakthrough Goal for AI. *AI Magazine*, 2005. URL http://www.cs.cmu.edu/~{}tom/pubs/AImagazine-7-2005.html.

Thahir Mohamed, Estevam Hruschka, and Tom Mitchell. Discovering Relations between Noun Categories. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pages 1447–1455, Edinburgh, Scotland, UK., 2011.

Dan Moldovan and Adrian Novischi. Lexical chains for question answering. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING '02), Taipei, Taiwan, 24 August – 1 September 2002*, pages 1–7, 2002.

Mortaza Montazery and Heshaam Faili. Automatic Persian WordNet Construction. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, (COLING '10), pages 846–850, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1944566.1944663.

Andrea Moro and Roberto Navigli. WiSeNet: Building a Wikipedia-based semantic network with ontologized relations. In *Proceedings of the 21th ACM international Conference on Information and Knowledge Management (CIKM '12)*, pages 1672–1676, Maui, HI, USA, 2012.

Andrea Moro and Roberto Navigli. Integrating Syntactic and Semantic Analysis into the Open Information Extraction Paradigm. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI '13)*, Beijing, China, 2013.

Vivi Nastase and Michael Strube. Transforming Wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62–85, 2013.

Vivi Nastase, Michael Strube, Benjamin Boerschinger, Caecilia Zirn, and Anas Elghafari. WikiNet: A Very Large Scale Multi-Lingual Concept Network. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1015–1022, Valletta, Malta, 2010. ISBN 2-9517408-6-7.

Roberto Navigli. Word Sense Disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69, 2009.

Roberto Navigli. A quick tour of word sense disambiguation, induction and related approaches. In *Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, pages 115–129, 2012.

Roberto Navigli and Simone Paolo Ponzetto. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010*, pages 216–225, 2010.

Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012a.

Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012b.

Roberto Navigli and Paola Velardi. Learning domain ontologies from document warehouses and dedicated websites. *Computational Linguistics*, 30(2), 2004.

Roberto Navigli and Paola Velardi. Learning Word-Class Lattices for Definition and Hypernym Extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1318–1327, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Roberto Navigli, David Jurgens, and Daniele Vannella. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM '13)*, volume 2, pages 222–231, 2013.

Patrick Pantel and Dekang Lin. Discovering word senses from text. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD '02)*, pages 613–619, 2002.

Patrick Pantel and Deepak Ravichandran. Automatically labeling semantic classes. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT '13), Boston, Massachusetts, 2–7 May 2004*, pages 321–328, 2004.

Patrick Pantel, Rahul Bhagat, Timothy Chklovski, and Eduard Hovy. ISP: learning inferential selectional preferences. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '07), Rochester, N.Y., 22–27 April*, pages 564–571, Rochester, NY, 2007.

Marius Pasca. Acquisition of categorized named entities for web search. In *Proceedings of the 13th ACM Conference on Information and Knowledge Management (CIKM '04)*, pages 137–145, New York, NY, USA, 2004. ISBN 1-58113-874-1. doi: 10.1145/1031171.1031194. URL `http://doi.acm.org/10.1145/1031171.1031194`.

Marco Pennacchiotti and Patrick Pantel. Ontologizing semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING '06), Sydney, Australia, 17–21 July 2006*, pages 793–800, Sydney, Australia, 2006.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the 1st International Global WordNet Conference (GWC '02), Mysore, India, 21–25 January 2002*, pages 21–25, 2002.

Jan Pomikálek, Milos Jakubícek, and Pavel Rychlỳ. Building a 70 billion word corpus of English from ClueWeb. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC '12)*, pages 502–506, 2012.

Simone Paolo Ponzetto and Roberto Navigli. Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, Cal., 14–17 July 2009*, pages 2083–2088, 2009.

Simone Paolo Ponzetto and Michael Strube. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence (AAAI '07), Vancouver, B.C., Canada, 22–26 July 2007*, pages 1440–1445, 2007.

Simone Paolo Ponzetto and Michael Strube. Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 175(9-10): 1737–1756, 2011.

Hoifung Poon and Pedro Domingos. Unsupervised semantic parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09) Volume 1-Volume 1*, pages 1–10. Association for Computational Linguistics, 2009.

Hoifung Poon, Janara Christensen, Pedro Domingos, Oren Etzioni, Raphael Hoffmann, Chloe Kiddon, Thomas Lin, Xiao Ling, Mausam, Alan Ritter, Stefan Schoenmackers, Stephen Soderland, Dan Weld, Fei Wu, and Congle Zhang. Machine Reading at the University of Washington. In *Proceedings of the 1st International Workshop on Formalisms and Methodology for Learning by Reading in conjunction with NAACL-HLT 2010*, pages 87–95, Los Angeles, California, USA, 2010.

Philip Resnik. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61(1):127–159, 1996.

Alan Ritter, Oren Etzioni, et al. A latent Dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pages 424–434, Uppsala, Sweden, 2010. ACL.

Alan JA Robinson and Andrei Voronkov. *Handbook of automated reasoning*, volume 2. Elsevier, 2001.

Víctor Rodríguez Doncel, Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and María Poveda-Villalón. Licensing patterns for linked data. 2013.

Víctor Rodríguez-Doncel, Asunción Gómez-Pérez, and Nandana Mihindukulasooriya. Rights declaration in linked data. In *COLD*, volume 1034 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013. URL `http://dblp.uni-trier.de/db/conf/semweb/cold2013.html#Rodriguez-DoncelGM13`.

Stephen Roller, Katrin Erk, and Gemma Boleda. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of the Twenty Fifth International Conference on Computational Linguistics (COLING-14), Dublin, Ireland*, pages 1025–1036, 2014.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37^{th} Annual Meeting of the Association for Computational Linguistics (ACL '99)*, pages 104–111, Stroudsburg, PA, USA, 1999. ISBN 1-55860-609-3. doi: 10.3115/1034678.1034703. URL `http://dx.doi.org/10.3115/1034678.1034703`.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In *Advances in Web Intelligence*, volume 3528 of *Lecture Notes in Computer Science*, pages 380–386. Springer Verlag, 2005.

Horacio Saggion. Identifying Definitions in Text Collections for Question Answering. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04), Lisbon, Portugal, 26–28 May 2004*, pages 1927–1930. European Language Resources Association, 2004.

David Sánchez and Antonio Moreno. Web-scale taxonomy learning. In *Proceedings of Workshop on Extending and Learning Lexical Ontologies using Machine Learning (ICML '05)*, pages 53–60, 2005.

Felix Sasaki. D4.5 - first roadmapping workshop athens v1.1, June 2014. URL `http://www.lider-project.eu/sites/default/files/LIDER%20D4.5-First%20Roadmapping%20Workshop%20Athens%20v1.1.pdf`.

Felix Sasaki. D4.8 - fourth roadmapping workshop report - v1.2, 2015. URL `http://www.lider-project.eu/sites/default/files/D4.8-v1.2.pdf`.

Diarmuid O Séaghdha. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pages 435–444, Uppsala, Sweden, 2010.

Dan Shen and Mirella Lapata. Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning (EMNLP-CoNLL '07), Prague, Czech Republic, 28–30 June*, pages 12–21. ACL, 2007. URL `http://dblp.uni-trier.de/db/conf/emnlp/emnlp2007.html#ShenL07`.

Amit Singhal. Introducing the Knowledge Graph: Things, Not Strings. Technical report, Official Blog (of Google). Retrieved May 18, 2012, 2012.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning Syntactic Patterns for Automatic Hypernym Discovery. In *Advances in Neural Information Processing Systems 17 (NIPS)*, pages 1297–1304, Cambridge, Mass., 2004.

Rion Snow, Dan Jurafsky, and Andrew Ng. Semantic taxonomy induction from heterogeneous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL '06)*, pages 801–808, 2006.

Philipp Sorg and Philipp Cimiano. Enriching the crosslingual link structure of wikipedia - a classification-based approach -. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (WikiAI'08)*, To appear, 2008.

Asher Stern and Ido Dagan. BIUTEE: A Modular Open-Source System for Recognizing Textual Entailment. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL '12), System Demonstrations*, pages 73–78, Jeju Island, Korea, 2012.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A large ontology from Wikipedia and WordNet. *Journal of Web Semantics*, 6(3):203–217, 2008.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL '03)*, pages 8–15, Stroudsburg, PA, USA, 2003. doi: 10.3115/1075096.1075098. URL `http://dx.doi.org/10.3115/1075096.1075098`.

Geoff Sutcliffe, Martin Suda, Alexandra Teyssandier, Nelson Dellis, and Gerard de Melo. Progress Towards Effective Automated Reasoning with World Knowledge. In *FLAIRS Conference*, 2010.

M. Thelen and E. Riloff. A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '02)*, pages 214–221, Salt Lake City, UT, USA, 2002.

Daniele Vannella, Tiziano Flati, and Roberto Navigli. WoSIT: A Word Sense Induction Toolkit for Search Result Clustering and Diversification. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL 14), System Demonstrations*, pages 67–72, Baltimore, Maryland, 2014. Association for Computational Linguistics.

Paola Velardi, Stefano Faralli, and Roberto Navigli. OntoLearn Reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3): 665–707, 2013.

Daniel Vila-Suero, Asunción Gómez-Pérez, Elena Montiel-Ponsoda, Jorge Gracia, and Guadalupe Aguado-de Cea. Publishing linked data on the web: The multilingual dimension. In *Towards the Multilingual Semantic Web*, pages 101–117. Springer, 2014.

Piek Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer, Dordrecht, The Netherlands, 1998.

Zhichun Wang, Juanzi Li, Zhigang Wang, and Jie Tang. Cross-lingual knowledge linking across wiki knowledge bases. In *Proceedings of the 24th International World Wide Web Conference (WWW 2012)*, pages 459–468. ACM, 2012. ISBN 978-1-4503-1229-5. URL http://dblp.uni-trier.de/db/conf/www/www2012.html#WangLWT12.

Dominic Widdows. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL '03)*, pages 197–204, Edmonton, Canada, 2003.

Yorick Wilks. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, 6(1):53–74, 1975.

Larry Wos, Ross Overbeck, Ewing Lusk, and Jim Boyle. Automated reasoning: introduction and applications. 1984.

Fei Wu and Daniel S. Weld. Open Information Extraction Using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pages 118–127, Uppsala, Sweden, 2010.

Akane Yakushiji, Yusuke Miyao, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. Automatic construction of predicate-argument structure patterns for biomedical information extraction. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*, pages 284–292, Stroudsburg, PA, USA, 2006. ISBN 1-932432-73-6. URL http://dl.acm.org/citation.cfm?id=1610075.1610116.

David Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the* 33^{th} *Annual Meeting of the Association for Computational Linguistics (ACL '95)*, pages 189–196, Cambridge, MA, USA, 1995.

Alexander Yates and Oren Etzioni. Unsupervised methods for determining object and relation synonyms on the web. *Journal of Artificial Intelligence Research*, 34 (1):255, 2009.

Alexander Yates, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen Soderland. TextRunner: open information extraction on the web. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007) - Demonstrations*, pages 25–26, Stroudsburg, PA, USA, 2007. URL http://dl.acm.org/citation.cfm?id=1614164.1614177.

Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. Statsnowball: a statistical approach to extracting entity relationships. In *Proceedings of the 18th*

international conference on World Wide Web (WWW '09), pages 101–110. ACM, 2009.