SAPIENZA
UNIVERSITÀ DI ROMA

Knowledge-based approaches to producing large-scale training data from scratch for Word Sense Disambiguation and Sense Distribution Learning

Scuola di dottorato in Informatica

Dottorato di Ricerca in Informatica – XXXI Ciclo

Candidate

Tommaso Pasini

ID number 1381813

Thesis Advisor

Prof. Roberto Navigli

July 2019

Thesis defended on 16th Semptember 2019
in front of a Board of Examiners composed by:

Prof. Lamberto Ballan (chairman)

Prof. Michele Boreale

Prof. Valeria Cardellini

Reviewers:

Prof. Diana McCarthy

Prof. Alexander Panchenko

Knowledge-based approaches to producing large-scale training data from scratch for Word Sense Disambiguation and Sense Distribution Learning

Ph.D. thesis. Sapienza – University of Rome

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

Version: October 30, 2020

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

Abstract

Communicating and understanding each other is one of the most important human abilities. As humans, in fact, we can easily assign the correct meaning to the ambiguous words in a text, while, at the same time, being able to abstract, summarise and enrich its content with new information that we learned somewhere else. On the contrary, machines rely on formal languages which do not leave space to ambiguity hence being easy to parse and understand. Therefore, to fill the gap between humans and machines and enabling the latter to better communicate with and comprehend its sentient counterpart, in the modern era of computer-science's much effort has been put into developing Natural Language Processing (NLP) approaches which aim at understanding and handling the ambiguity of the human language.

At the core of NLP lies the task of correctly interpreting the meaning of each word in a given text, hence disambiguating its content exactly as a human would do. Researchers in the Word Sense Disambiguation (WSD) field address exactly this issue by leveraging either knowledge bases, i.e. graphs where nodes are concept and edges are semantic relations among them, or manually-annotated datasets for training machine learning algorithms. One common obstacle is the *knowledge acquisition bottleneck* problem, id est, retrieving or generating semantically-annotated data which are necessary to build both semantic graphs or training sets is a complex task. This phenomenon is even more serious when considering languages other than English where resources to generate human-annotated data are scarce and ready-made datasets are completely absent. With the advent of deep learning this issue became even more serious as more complex models need larger datasets in order to learn meaningful patterns to solve the task.

Another critical issue in WSD, as well as in other machine-learning-related fields, is the *domain adaptation* problem, id est, performing the same task in different application domains. This is particularly hard when dealing with word senses, as, in fact, they are governed by a Zipfian distribution; hence, by slightly changing the application domain, a sense might become very frequent even though it is very rare in the general domain. For example the geometric sense of *plane* is very frequent in a corpus made of math books, while it is very rare in a general domain dataset.

In this thesis we address both these problems. Inter alia, we focus on relieving the burden of human annotations in Word Sense Disambiguation thus enabling the

automatic construction of high-quality sense-annotated dataset not only for English, but especially for other languages where sense-annotated data are not available at all. Furthermore, recognising in word-sense distribution one of the main pitfalls for WSD approaches, we also alleviate the dependency on most frequent sense information by automatically inducing the word-sense distribution in a given text of raw sentences.

In the following we propose a language-independent and automatic approach to generating semantic annotations given a collection of sentences, and then introduce two methods for the automatic inference of word-sense distributions. Finally, we combine the two kind of approaches to build a semantically-annotated dataset that reflect the sense distribution which we automatically infer from the target text.

Publications

2019

- **Tommaso Pasini** and Roberto Navigli. *Train-O-Matic: Supervised Word Sense Disambiguation with No (Manual) Effort*. Acceptance at Artificial Intelligence Journal subject to Major Revision. 2019.

- Bianca Scarlini, **Tommaso Pasini** and Roberto Navigli. *Just "OneSeC" for Producing Multilingual Sense-Annotated Data*. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). August 2019.

2018

- **Tommaso Pasini** and Roberto Navigli. *Two Knowledge-based Methods for High-Performance Sense Distribution Learning*. In Proceedings of the 32nd AAAI conference on Artificial Intelligence (AAAI), pages 5374–5381, February 2018.

- **Tommaso Pasini**, Francesco Elia and Roberto Navigli. *Huge Automatically Extracted Training-Sets for Multilingual Word Sense Disambiguation.* In Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC), pages 1694–1698, May 2018.

- José Camacho-Collados, Claudio Delli Bovi, Luis Espinosa Anke, Sergio Oramas, **Tommaso Pasini**, Enrico Santus, Vered Shwartz, Roberto Navigli, Horacio Saggion. *SemEval-2018 Task 9: Hypernym Discovery.* In Proceedings of SemEval at the 16th Annual Conference of the North American

Chapter of the Association for Computational Linguistics (NAACL-HLT), pages 712–724, June 2018.

2017

- **Tommaso Pasini**, Roberto Navigli. *Train-O-Matic: Large-Scale Supervised Word Sense Disambiguation in Multiple Languages without Manual Training Data.* In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 78-88, September 2017.

2016

- Tiziano Flati, Daniele Vannella, **Tommaso Pasini**, Roberto Navigli. *Multi-WiBi: The multilingual Wikipedia Bitaxonomy project.* In Proceedings of the Artificial Intelligence Journal, volume, 241, pages 66–102, 2016.

2014

- Tiziano Flati, Daniele Vannella, **Tommaso Pasini**, 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), pages 945–955, June 2014.

Contents

Chapter 1

Introduction

Word Sense Disambiguation (WSD) is one of the oldest problem which resides at the basis of Natural Language Processing. Indeed, it provides explicit semantic annotations to the words in a text, which may be beneficial to other down-stream applications, such as Machine Translations (Pu et al., 2018), Information Extraction (Delli Bovi, Espinosa Anke, and Navigli, 2015a), etc. Weaver (1949) was the first formulating it as a computational task where, given N tokens surrounding a target word w, a system has to choose the w's meaning that best fits the context. While seemingly simple at first sight, Mallery (1988) and Ide and Véronis (1993) described WSD as an AI-complete problem, i.e. its resolution is central to building intelligent computers. Its complexity mostly derives from the fact that disambiguating a word often requires prior knowledge that is only implicitly expressed in the text (Navigli, 2009). Therefore, to provide machines with such information – hence enabling them to either learn or reason over it – different knowledge representations have been proposed, inter alia: knowledge bases and annotated corpora.

A knowledge base is a graph having concepts as nodes that are in their turn connected via different kinds of connections which can be either typed, i.e., expressing a specific relation such as hypernymy, meronymy, etc., or untyped, i.e., simply expressing relatedness. For example, the concept of MUG (*a large cup, typically cylindrical with a handle.*) may be connected via a *meronymy* relation to HANDLE (*he appendage to an object that is designed to be held in order to use or move it*) and via a *related to* edge to COFFEE (*A beverage consisting of an infusion of ground coffee beans*). Hence, we can distinguish between two types of relations: syntagmatic, which connect mainly co-occuring concepts, and paradigmatic, which,

instead, connect concepts that can be substituted with each other. The latter are encoded into ontologies, i.e., structures where concepts are connected to each other via formal semantic relations. On the contrary, semantic networks (a.k.a. knowledge graphs, semantic graphs, knowledge bases) follow a more relaxed definition and encompass both types of relations. Therefore, building a knowledge base from scratch implies different steps each with its difficulties:

1. Select the set of words and build the set of senses (comprising definition and usage examples) that those words may assume in whatever text.
2. Identify the set of relations that should be included in the resource.
3. Connect the senses previously defined via one or more relations.

Each of these steps, depending on the desired size of the final resource, may take a long time and that is why fully-manual knowledge bases are rare and most of the time of modest size.

WordNet (Fellbaum, 1998a) is one of the most famous and used resources for English comprising $117{,}798$ synsets, i.e. set of synonyms, connected via typed relations[1] for $155{,}287$ distinct lemmas covering the four most important Part-Of-Speech (POS), namely: nouns, verbs, adverbs and adjectives[2]. Since WordNet is fully manual it is of high quality, however it has three main drawbacks: it does not cover most of the named entities, it mainly comprises paradigmatic relations and it covers only English lemmas. On the contrary, Wikipedia, a collaborative electronic encyclopedia, covers mainly nouns and named entities connected only via edges expressing relatedness. Thanks to $270{,}000$ active contributors since 2001, it counts now more than 40M articles in 280 different languages. Hence, it took 18 years and hundred-thousand contributors to reach the current state, nevertheless, it lacks non nominal concepts, such as those expressing actions (there is no Wikipedia article for the concept RUN (*move fast by using one's feet*)), and typed relations. Moreover, Wikipedia does not offer a unified representation for concepts that are expressed in different languages, instead it has separate pages expressing the same concept for each language. To overcome both WordNet and Wikipedia issues, Navigli and Ponzetto (2012) introduced BabelNet, a huge knowledge base created by

[1]The complete list of WordNet relations (pointers) can be found at `https://wordnet.princeton.edu/documentation/wninput5wn`

[2]Statistics for version 3.1 of WordNet

automatically merging many different lexical resources, such as WordNet, Open Multilingual WordNet, Wikipedia, Wikidata, etc., in a unified multilingual encyclopedic dictionary where lemmas representing the same concept in different languages are clustered together in the same synset. Thanks to its intrinsic multilingual nature it makes it easy to scale over new languages as long as they are integrated into the knowledge base. Moreover, the relations extracted from Wikipedia expressing the relatedness between two concepts integrate well WordNet's paradigmatic connections, thus making it a more complete resource.

BabelNet inspired and lays as underlying knowledge base and sense inventory in many works in the field of lexical semantics, inter alia, Moro, Raganato, and Navigli (2014a, Babelfy), Camacho-Collados, Pilehvar, and Navigli (2015, NASARI), Iacobacci, Pilehvar, and Navigli (2015, SensEmbed), Pasini and Navigli (2017, Train-O-Matic), Bovi et al. (2017, EuroSense). Furthermore it paved the way for multilingual and language-agnostic Word Sense Disambiguation, as proved by Moro, Raganato, and Navigli (2014b).

On the other part of the spectrum there are sense-annotated corpora which enclose the knowledge into unstructured texts. The problem, this time, resides mainly in the Zipfian distribution of the senses in a corpus, which makes it unlikely to find sentences expressing the less frequent senses. This leads to build incomplete corpora which fail to cover a meaningful set of senses and words (Kilgarriff and Rosenzweig, 2000; Atkins, 1993; Ng and Lee, 1996). The two biggest and most used corpora are OntoNotes (Hovy et al., 2006) and SemCor (Miller et al., 1993a). The former comprises $233{,}616$ words annotated with roughly $12{,}000$ WordNet senses and $2{,}475{,}987$ tokens in total. The latter, instead, is divided in 352 documents for a total of $802{,}443$ tokens, $226{,}036$ of which are associated with roughly $33{,}000$ WordNet senses. Even though SemCor comprises a slightly lower number of annotated tokens, it covers 3 times more senses than OntoNotes, and, in fact, it is the most used corpus for WSD.

Nevertheless, both corpora have a limited coverage of word senses, thus limiting the supervised models in the number of senses they can classify. Moreover, SemCor has never been updated in terms of covered senses or words since its first release (1993), hence, the sense distribution may not reflect anymore the current one. For example, consider the noun *pipe*, its most frequent sense in SemCor is PIPE (*a tube with a small bowl at one end*), while nowadays it would be more natural to think

about PIPE (*a long tube made of metal or plastic that is used to carry water or oil or gas etc.*). This affects the performance and generalisation power of supervised models, which learn an outdated distribution, and thus, perform poorly on new and unseen data.

The Knowledge Acquisition Bottleneck Building both these kind of knowledge resources, i.e., knowledge bases and sense-annotated corpora, is very expensive in terms of time and money (Ng and Lee, 1996). In fact, this problem is known as the *knowledge acquisition bottleneck* problem. Resources for WSD are particularly affected by this issue for several reasons:

1. The need for linguistic expertise.
2. Each language of interest needs separate efforts and experts.
3. The knowledge acquired for a domain cannot be transferred easily to other domains.
4. The labels, i.e., senses, are typically in the order of the hundreds of thousands, hence, even if one would like to collect just ten examples for each label she would ends up annotating hundreds of thousands of instances.

Therefore, while on the one hand, BabelNet attempts to mitigate the *knowledge acquisition bottleneck* for knowledge bases by taking the best of many worlds (i.e. WordNet, Wikipedia, Omegawiki, etc.) and automatically merging their senses across different languages, on the other hand, the *knowledge acquisition bottleneck* problem for sense-annotated corpora have been tackled by different automatic and semi-automatic approaches that have been proposed to relieve the burden of manual annotations. Taghipour and Ng (2015, OMSTI) and Bovi et al. (2017, EuroSense) exploited parallel data, e.g., United Nations Parallel Corpus (Ziemski, Junczys-Dowmunt, and Pouliquen, 2016) and EuroParl (Koehn, 2005a), for building semantic-annotated data for either English and each language comprised in the parallel corpus respectively. In an alternative approach, Raganato, Delli Bovi, and Navigli (2016, SEW) additionally avoided the need of parallel corpora as well and developed a set of heuristics that, by exploiting the links available in a Wikipedia page, annotate unlinked words with a BabelNet synset. However, these attempts

failed to enable supervised models to perform competitively with knowledge-based methods on multilingual all-words WSD tasks, inasmuch their data were either too noisy (Eurosense) or not available at all for languages other than English (SEW).

Besides the *knowledge acquisition bottleneck* problem, which prevents easy acquisition of statistically significant amount of data for each sense, hence making training data sparse, another important issue affecting the WSD field is the Zipfian shape of sense distributions.

Word Sense Distributions Sense Frequency distributions follow a Zipfian distribution (McCarthy et al., 2007), thus, given a corpus of texts, a word would typically have only one or two of its meanings occurring frequently, while many other meanings would be rare or not appear at all. Moreover, the distribution highly depends on the main topic expressed across the documents within the corpus. Indeed, when changing domain, each sense frequency may drastically change. For example, consider a corpus of texts built from travel magazines, the word *plane* will appear mainly (if not all the time) in its *airplane* sense and very few times (if ever) in its *carpenter tool* or *geometric* sense. On the opposite, when considering a corpus of math books the frequency of *plane* senses changes completely. The *geometric* sense will be the most frequent one while the others will be very rare. Hence, in addition to the paucity of data, WSD models have to face the not easy challenge of handling very skewed distributions of data which may drastically change according to the corpus. This has a huge impact on supervised models, which, while excelling in predicting the most frequent senses of words, achieve poor performance on least frequent ones (Postma, Bevia, and Vossen, 2016a; Postma et al., 2016b). This is a direct effect of sense distribution skewness and data sparsity, which contribute to bias a supervised model towards the most frequent sense of a word, as it is one of the few which has a meaningful amount of annotated data. For example, the sense WORD (*a unit of language that native speakers can identify.*) appears more than 100 times in SemCor, while WORD (*a brief statement.*) less than 20. Hence, a model will be able to effectively build a representation for the first sense of *word*, while, at the same time, fail to correctly model its second meaning and the contexts it appears in.

Another issue deriving from the sense distribution skewness is the way a model is tested. In fact, when the distributions in the training and test sets match, the model will achieve excellent results (Postma, Bevia, and Vossen, 2016a). On the

contrary, when those two diverge, also by a small factor, the performance will be poor as we will show in the experimental part of this thesis. The sense distribution, therefore, plays a key role in supervised as well as in knowledge-based WSD, as, in fact, it influences the knowledge base connectivity, making the most frequent senses more connected than the least frequent ones, hence raising their probability of being chosen by a knowledge-based algorithm.

Despite the large impact that sense distributions have on WSD models, most of the effort has been put only in recognising the predominant meaning of a word (McCarthy et al., 2004; Bhingardive et al., 2015) and when it is advisable to output the most frequent sense of a word (Jin et al., 2009). Bennett et al. (2016), instead, was one of the few in the most recent years proposing a topic modelling approach based on Lau et al. (2014) to estimate the distribution of word senses in a given corpus, providing also a corpus for testing the quality of the induced distributions.

In this thesis we study both the aforementioned problems and propose different approaches for tackling them. We first focus on the *knowledge acquisition bottleneck* problem by studying the manual, semi-automatic and automatic resources available for WSD and then proposing a language-independent approach for generating sense annotations for a hundred-thousand sentences which lead a supervised model to remarkable results, and enabling multilingual WSD for supervised models. Then, we tackle the sense-distribution problem by proposing two knowledge-based approaches for learning the distribution of senses given a corpus of raw texts. We prove they are effective in capturing the high-level semantics of the input corpus and in correctly predicting the probabilities of most word senses. We finally integrate the learned distributions into our automatic approach for generating sense-annotated data, thus making it aware of each sense probability of appearing in the corpus and making it possible to shape the training data on the desired distribution. We show that the empirical results obtained prove that the sense distribution is key for automatically building high-quality sense-annotated datasets, especially when dealing with domain-specific applications of WSD.

The remainder of the thesis is organised as follows: In Chapter 2 we discuss the relevant works for this thesis ranging over several WSD subfields such as knowledge acquisition, supervised and knowledge-based models for WSD. In Chapter 3 we introduce basic concepts we use across the whole manuscript, while, in Chapter 4 we

introduce Train-O-Matic (Pasini and Navigli, 2017), a knowledge-based approach for disposing of human annotations and automatically generating hundreds of thousands of sense-annotated examples for potentially any languages available in BabelNet. In Chapter 5, we present EnDi and DaD, two knowledge-based and language-independent methods for automatically inducing the distribution of words' meanings from a collection of raw texts. In Chapter 6 we couple Train-O-Matic with the two sense-distribution learning approaches that we mentioned above and study the impact that automatically-generated datasets tailored on a specific distribution have on a supervised model. Finally, in Chapter 7 we give in-depth insights on all the data produced in our works and put at the research community disposal and in Chapter 8 we draw the conclusions and the possible future directions this work may lead.

1.1 Objectives

In this Section we briefly summarise and make it clear the objectives of the thesis. In this work, we focus on mitigating the *knowledge acquisition bottleneck* problem in WSD and to provide a valuable solution for overcoming the paucity of sense-annotated data. Then we study the word senses' distribution and introduce two knowledge-based approaches to automatically learn them given only one or more documents of raw texts as input. Finally, we study the impact of automatically induced distribution of senses on the creation of sense-annotated data and the possibility of cutting the dataset distribution on a specific domain or collection of input raw text documents. Therefore, the main objectives of this thesis are:

- Study the *knowledge acquisition bottleneck* problem in English and multilingual WSD, with the aim of mitigating it leveraging knowledge bases (Chapter 4).

- The design of unsupervised approaches for automatically inducing the word sense distribution within a given corpus of raw text (Chapter 5).

- The study of the impact that the word senses' distribution of the training set has on a supervised model (Chapter 6).

1.2 Contributions

This thesis provides the following significant contributions to each objective:

- **Train-O-Matic** (Pasini and Navigli, 2017): we present a knowledge-based approach for automatically generating sense-annotated datasets for any language supported by BabelNet (Chapter 4). We show that a supervised model trained on Train-O-Matic attains better performance across several all-words WSD tasks than when trained on other automatic and semi-automatic alternatives and on the same ballpark with those achieved when using manually-curated data as training set.

- **Sense distribution learning** (Pasini and Navigli, 2018): we present EnDi & DaD, two knowledge-based approaches for automatically inducing the word senses' distribution given as input a corpus of raw text (Chapter 5). Our approaches are language independent and outperforms all their competitors on the intrinsic and extrinsic evaluations.

- **Target-text driven WSD** (Pasini and Navigli, 2019): we incorporated the two methods for inducing the distribution of word senses in Train-O-Matic and fully automatise the process of generating datasets tailored on a given document or set of documents (Chapter 6). We show that, shaping the training set senses' distribution on the one induced by EnDi or DaD results in higher performance of the supervised model. Moreover, DaD proved to be a valid alternative to the WordNet MFS especially on the domain-specific and multilingual benchmarks.

1.2.1 Individual Contributions

I personally contributed to the design and the implementation of all the algorithms and the evaluation setups presented in this thesis, as, in fact, I am the main author and contributor of all the presented approaches and corresponding papers.

Published material not included in this thesis Other works, which did not contribute directly to this thesis or done before starting the Ph.D. program, and are

thus not included but represent valuable effort and contribution, are, in order of publication:

- Two is bigger (and better) than one: the Wikipedia Bitaxonomy project (Flati et al., 2014)
- MultiWiBi: The multilingual Wikipedia Bitaxonomy project (Flati et al., 2016)
- SemEval-2018 Task 9: hypernym discovery (Camacho-Collados et al., 2018).
- Just "OneSeC" for Producing Multilingual Sense-Annotated Data (Scarlini, Pasini, and Navigli, 2019)

Chapter 2

Related Work

2.1 History in Brief

Word Sense Disambiguation (WSD) has been first formally introduced in 1949 by Weaver, in the context of Machine Translation as the task of identifying the correct meaning of a target word given its surrounding tokens. Following this definition, the research community focused on answering the most basic questions raised by this formulation such as the number of surrounding tokens needed to disambiguating the central word correctly (Kaplan, 1950; Koutsoudas and Korfhage, 1956), or the kind of features that would have a positive impact on the task (Reifler, 1955). These preliminary studies laid the foundation for the first approaches to WSD, that, in the beginning, were usually combined with methods for solving more general problems such as text understanding. These leveraged the semantic networks available during the 60s (Quillian, 1961; Masterman, 1961) and were based on psycholinguistics studies. The latter, paved the way to the so-called *connectionist methods* (Ide and Véronis, 1998), i.e., those considering the previously disambiguated words as additional context to infer the meanings of the subsequent tokens (Collins and Loftus, 1975; Anderson, 1976). Despite the excitement towards this new research field, most of the works were based on datasets that were very limited in terms of the number of distinct words, meanings covered and domains of application. Hence, while on the one hand it started becoming evident that WSD models needed increasingly more data in order to generalise over new and previously unseen examples, on the other hand, the *knowledge acquisition bottleneck problem* began to take shape, making

it clear that gathering large sets of manually-annotated data was very expensive in terms of both time and resources. For these reasons, the subsequent two decades (the 80s and the 90s) were devoted to the creation of comprehensive machine-readable resources, such as thesauri, dictionaries and sense-annotated corpora (Miller et al., 1990; Fellbaum, 1998a; Lenat, 1995a; Roget, 2000, first made available in digital format during the 50s).

These new resources enabled new and more sophisticated approaches to WSD. Lesk, in fact, introduced a dictionary-based algorithm in 1986 (Lesk, 1986) which, given a sentence and a target word therein, it iterates over the target senses (i.e., the possible senses in a dictionary for the target word) and, for each of them, counts the number of overlapping tokens between the sense's gloss and the input sentence. Finally, the target word is disambiguated with the meaning corresponding to the gloss with the highest number of common words. This approach was more efficient than other expert systems available at that time (Granger, 1982; Sowa, 1983; Walker and Amsler, 1986), as it only relied on the local context of a word and a list of glosses. Moreover, it could be easily applied to a large number of texts. Therefore, exploiting knowledge bases such as thesauri, dictionaries, etc. looked very promising and, to stimulate further research on this topic, Miller et al. and Fellbaum developed the two resources that would become, from a little later, the most used within the Word Sense Disambiguation community and in other related areas: WordNet (Miller et al., 1990; Fellbaum, 1998a) and SemCor (Miller et al., 1993a). WordNet is a lexical database of senses where synonyms are grouped into synsets, which, in turn, are linked via paradigmatic relations (i.e., hypernymy, meronymy, etc.). SemCor, instead, is a subset of the Brown corpus (Kucera and Francis, 1967) where each content word (noun, adjective, adverb or verb) is annotated with its WordNet sense. These two resources, together, paved the way to the knowledge-based paradigm for Word Sense Disambiguation and steered WSD research in the direction of enumerative lexicons, opposing to Pustejovsky (1995), which was proposing a generative lexicon to induce the possible meanings of a word in a specific context by means of hand-crafted rules.

Enumerative inventories, in fact, proved to offer a more flexible framework for Word Sense Disambiguation inasmuch they allow to express the problem as a classification one, where, given a target word and its context, one has to select the most suitable sense among the possible ones. Despite the longstanding debate about which of the two linguistic theories better describe how the human thought

understands word meanings, enumerative lexicons have received more attention and are, nowadays, the most used in WSD as well as in other research areas, e.g., in vision with ImageNet (Deng et al., 2009), question answering (Ferrucci et al., 2010), etc. Nevertheless, WordNet, similarly to other enumerative lexicons, suffers from the sense granularity problem, i.e., it makes very fine-grained distinctions between senses. Just consider the noun *line*, WordNet enumerates 30 different senses distinguishing, among others, between a line organised horizontally (*a formation of people or things one beside another*) or vertically (*a formation of people or things one behind another*). Recently, Pilehvar et al. (2017) investigated the impact of word senses at different granularities on a down-stream application, i.e., text classification. They showed that explicit information of meaning is beneficial to categorise a text. However, Pilehvar et al. proved that super senses, i.e., forty-five coarse-grained labels that cluster WordNet synsets, are more effective than their fine-grained counterpart – at least in the text classification task –. Despite this interesting result, we cannot conclude that coarse-grained senses are always better and, in fact, this is still an open problem which leaves unanswered three critical questions: i) are enumerative lexicons the best way of representing word meanings? ii) do they scale well on different tasks and scenarios? iii) while favouring the enumerative lexicons, what is the granularity level that brings the highest benefit to a task? In the next few years, the WSD research community should focus on answering these three questions in order to let semantics have a more significant impact on downstream applications.

Moving forward to the 2000s, the researchers focused, again, on harvesting new and of better quality data. One of the most significant efforts that contributed largely to NLP research was Wikipedia. Launched in 2001, it aimed at collecting encyclopedic knowledge about the world by putting together the efforts of the whole internet community. Nowadays, it counts more than 50M articles written in 303 distinct languages maintained by more than 200,000 active contributors. It was the base for many projects such as Wikidata, Wiktionary, OmegaWiki[1], DBPedia[2] (Auer et al., 2007), Freebase[3] (Bollacker et al., 2008) and an underlying corpus in many works. One of the most interesting aspects of Wikipedia is its multilingual

[1] `http://www.omegawiki.org/Meta:Main_Page`

[2] `https://wiki.dbpedia.org/`

[3] `https://developers.google.com/freebase`

nature. Indeed, whenever it is possible, an article is linked to its counterpart in another language. Moreover, the text within a page can be linked via hyperlinks to other pages in the same language, hence making Wikipedia interpretable as a semantic network with syntagmatic relations between its pages. These two peculiar aspects led Navigli and Ponzetto to introduce a novel knowledge base named BabelNet (Navigli and Ponzetto, 2010, 2012) with the aim of unifying encyclopedic and lexicographic knowledge by merging together several resources, e.g., WordNet, Wikipedia, Wikidata, etc[4]. BabelNet nodes, i.e., synsets, in fact, can represent both: real entities, such as persons and companies, and abstract concepts, such as THING (*an entity that is not named specifically*) or ATTRIBUTE (*an abstraction belonging to or characteristic of an entity*) and may comprise lexicalisation in multiple languages. Therefore, the same concept is represented with the same node across different languages. Similarly, but focusing mainly on abstract concepts, Pease, Niles, and Li (2002, SUMO) aimed at merging together several upper-level ontologies. YAGO (Suchanek, Kasneci, and Weikum, 2007; Mahdisoltani, Biega, and Suchanek, 2013), instead, exploited data from Wikipedia to build a multilingual semantic network comprising entities and facts about them. Thanks to the proliferation of semantic resources, in the most recent years the research community could focus on developing novel knowledge-based Word Sense Disambiguation approaches (Moro, Raganato, and Navigli, 2014b; Agirre, de Lacalle, and Soroa, 2014; Tripodi and Pelillo, 2017; Chaplot and Salakhutdinov, 2018). However, while this kind of methods largely benefited from the resources above and are very flexible when it comes to disambiguate texts in different languages (as long as they are supported by the underlying knowledge base), supervised approaches proved to perform generally better on the all-words WSD English task at the cost of less flexibility on language dimension. Indeed, by relying on sense-annotated corpora for training, they directly depend on the availability of such datasets on each language of interest. One of the most successful models for supervised WSD was IMS (Zhong and Ng, 2010), a system comprising a distinct SVM classifier for each target word one is interested in disambiguating. It was the state-of-the-art model before the deep learning advent and only relied on the features extracted from the target word context, such as the surrounding words, the surrounding POS tags, etc. More recently, instead, feature-

[4]Refer to `https://babelnet.org/about` for a comprehensive list of resources included in BabelNet.

based approaches left the way to deep learning models which permeated most of the NLP fields inexorably. Thanks to this new paradigm, more and more complex neural models started to spawn here and there showing promising results on WSD. Unfortunately, only a small boost in performance was recorded since 2010, and, thus, deep learning models mainly contributed to confirm what has been already noted by Agirre and Edmonds (2006), i.e., the lack of decent-size sense-annotated datasets for WSD. Indeed, in a 8-year time span (2010-2018) since IMS was first introduced, the overall F-score on the most important WSD benchmarks has increased by only 3 points[5], even though much more complex models (Luo et al., 2018; Vial, Lecouteux, and Schwab, 2018) are now available. Compared to the results attained on the object recognition task in the context of ImageNet (Deng et al., 2009) competitions, this is a disappointing result. Indeed, when AlexNet (Krizhevsky, Sutskever, and Hinton, 2012) was first introduced, it brought an improvement over the previous non-neural state-of-the-art systems of roughly 10 points on the error rate. Moreover, also considering the time span, since neural models were first introduced in computer vision (2012 - 2018), deep networks brought a performance boost of more than 20 points in top-1 prediction accuracy, i.e., from a 63.3 points (Krizhevsky, Sutskever, and Hinton, 2012), to 84.3 points (Huang et al., 2018). This breakthrough was possible not only thanks to the novel paradigm of deep learning but mainly because of ImageNet (Deng et al., 2009). The dataset provided, in fact, 14M images tagged with classes from WordNet and thus created the perfect condition for a deep neural model to shine. Therefore, even though the WSD task is inherently more complex than the image labelling one – trivially by just considering the number of possible classes the models have to deal with –, the impact of neural models for WSD has been unsatisfactory so far. We argue that this has to be attributed to the paucity of sense-annotated data for English, and even more for other languages, which stems the impact of deep neural models on the field, which, now more than ever, needs new data at its disposal.

[5]IMS performs 68.4 (Raganato, Camacho-Collados, and Navigli, 2017), while UFSAC (Vial, Lecouteux, and Schwab, 2018) – the best performing systems nowadays – scores 71.8

2.2 Resources for Word Sense Disambiguation

The need of encoding knowledge has been clear since the 60s; however, large scale semantic networks and sense-annotated corpora have been only introduced in the last 15 years (refer to the previous Section for a historical excursus of WSD and related resources). In what follows, we present the most important works in the context of encoding semantic knowledge either with knowledge bases or sense-annotated corpora.

2.2.1 Knowledge Bases

WordNet (Miller et al., 1990; Fellbaum, 1998a) is a lexical database for English comprising 155,287 unique lexemes (i.e. lemma and Part of Speech pairs) of the English vocabulary organised in more than 117,000 synsets[6] connected through 19 types of paradigmatic relations[7]. For each concept it also provides a definition (i.e., gloss) and one or more examples of usage. WordNet is the *de facto* standard sense inventory for English Word Sense Disambiguation, indeed, it has been employed as tag set in several SemEval and Senseval competitions (Edmonds and Cotton, 2001; Snyder and Palmer, 2004; Pradhan et al., 2007; Navigli, Jurgens, and Vannella, 2013a; Moro and Navigli, 2015). While it mainly encompasses lexicographic knowledge, other resources were built to encode different aspects of semantic knowledge. Lenat (1995b) encoded practical information about concepts in Cyc, a hand-made resource comprising roughly 100K concepts and axioms about them, on the other hand, (Niles and Pease, 2001) aimed at capturing the high-level semantic knowledge into SUMO, an upper-level ontology that can be employed across different domains.

The proliferation of English resources encouraged more and more researchers to transfer English knowledge to other languages. Indeed, many semi-automatic WordNet-like resources were built for each language of interest, inter alia, Emanuele, Luisa, and Christian (2002); Sagot and Fišer (2008) and Huang et al. (2010) built Italian, French and Chinese versions of WordNet. These versions were later merged together by de Melo and Weikum (2009, Universal WordNet – UWN) and Bond and Foster (2013, Open Multi WordNet – OMN). However, even though all these

[6] we will use synsets and concepts interchangeably through this thesis.

[7] A complete list of relations can be found at `https://wordnet.princeton.edu/documentation/wninput5wn`

resources provide a valuable semantic knowledge for each language, they leave out many words of a language vocabulary since a tremendous manual-effort would be required to cover them all. Moreover, even if such a complete resource were available, it would need constant updates to add new words, remove unused words and update the possible meanings of each entry. The language, indeed, is always changing and adapting to the way it is conveyed. This is happening even faster since social networks appeared, hence making it easier to create new words or change the way a word is intended (think about the verb *tweet* before and after the lunch of Twitter). Therefore, recently, automatic approaches for acquiring and mapping knowledge across different sources have been proposed in order to mitigate these issues. Navigli and Ponzetto proposed BabelNet, a method for merging WordNet and Wikipedia in different languages, hence building a large knowledge base of lexicographic and encyclopedic information in 284 languages (Navigli and Ponzetto, 2010, 2012). Differently, from UWN and OMW, BabelNet goal was to enrich WordNet with real world entities and syntagmatic relations. As an alternative to Navigli and Ponzetto, Suchanek, Kasneci, and Weikum (2007); Mahdisoltani, Biega, and Suchanek (2013) proposed YAGO which takes advantage of WordNet taxonomy and encodes factual relation among entities thanks to Wikipedia infoboxes and categories. Similarly, WikiNet (Nastase et al., 2010) is a multilingual knowledge base comprising concepts from Wikipedia and relation extracted by means of textual content, infoboxes, page interlinks and category graph. These resources contributed to advance several fields in NLP, such as Word Sense Disambiguation (Moro, Raganato, and Navigli, 2014b), document classification (Pilehvar et al., 2017), taxonomy extraction (Flati et al., 2014, 2016), textual similarity (Pilehvar and Navigli, 2015), etc. However, supervised WSD methods did not benefit particularly from the information comprised in those resources, and kept relying on sense-annotated corpora which can be directly used as training sets. One recent effort in the direction of including the information comprised by a knowledge-base into a supervised model has been introduced by Luo et al. (2018, GAS). They provided a neural network with a memory of word sense's glosses, i.e., additionally to the context of a word, the model has also access to the word senses' definitions. GAS showed encouraging results beating most of the time the previous state of the art models that used to rely only on the context provided by a sentence only. This prove that lexical-semantic resources still have uncovered potential that can also impact

					English			
	Resource	**Type**	**#Langs**	**#Annotations**	**#Tokens**	**#Annotations**	**Amb**	**Entropy**
SemCor	WordNet	M	1	226,036	802,443	226,036	6.8	0.27
OntoNotes	WordNet	M	1	264,622	1,445,000	264,622	-	-
Princeton Gloss	WordNet	SA	1	449,355	1,621,129	449,355	3.8	0.45
OMSTI	WordNet	SA	1	911,134	30,441,386	911,134	8.9	0.94
SEW	Wikipedia	SA	1	162,614,753	1,357,105,761	162,614,753	7.9	0.40
SenseDefs	BabelNet	A	263	163,029,131	71,109,002	37,941,345	4.6	0.04
EuroSense	BabelNet	A	21	122,963,111	48,274,313	15,502,847	6.5	0.21
Train-O-Matic	BabelNet	A	6	17,987,488	291,550,966	12,722,530	3.6	0.48

Table 2.1. Statistics of the sense-annotated corpora across languages and resources. Type "M" stands for Manual, "SA" stands Semi-automatic, "C" for Collaborative and "A" for Automatic.

the supervised side of Word Sense Disambiguation.

2.2.2 Sense-Annotated Corpora

Sense-annotated corpora tie the concepts represented in a knowledge base with the lexical information contained in a sentence, i.e., one can build a relation between the lexical content and a specific meaning. For example, given the sentence *the plane, after a long flight, landed safely in Rome.* and the annotation PLANE (*an aircraft that has a fixed wing [...]*) for the word *plane*, one can derive that contexts similar to the one in the sentence trigger the *aircraft* meaning of *plane*. Moreover, they also provide syntagmatic relations among synsets appearing within the same contexts. In Table 2.1 we report the statistics for each automatic semi-automatic and manual corpus that we introduce in what follows.

Manually-annotated Corpora

SemCor (Miller et al., 1993a) is a subset of the Brown corpus (Kucera and Francis, 1967) and comprises more than 200,000 tokens annotated with a WordNet sense. It is the corpus with the wider coverage of words and senses[8], hence is the most

[8]Through the thesis we use *sense* as the tuple lemma, POS and synset, thus it is a synset bound to a specific lemma and POS.

obvious choice when it comes to training a supervised model. However, even though the coverage of senses is one of its strengths when compared to other corpora, it is a weakness when considering the absolute numbers of covered senses. Indeed, only the 16% of WordNet senses appear at least in one sentence of the corpus. This issue is compounded with the WordNet's fine granularity and the Zipfian nature of word senses, in fact, it is hard to cover the least common concepts since they occur rarely. Hence, while the fine-granularity problem (that we discussed in the previous Section) is more a problem of WordNet than SemCor itself, the latter suffers its consequences showing a limited coverage. Moreover, the Zipfian distribution of senses also plays an essential role in making SemCor outdated. Indeed, since the corpus dates back to the 60s, the frequency of a word senses has changed (see the example in Section 1), meaning that, a model trained on SemCor is biased towards an outdated distribution that does not reflect the current one anymore.

To overcome some of these issues, Hovy et al. (2006) introduced OntoNotes, a corpus tagged with senses from the Omega ontology (Philpot, Hovy, and Pantel, 2005). The latter is organised hierarchically with an upper-level ontology where the macro senses contain fine-grained specialised senses. Thanks to this structure, Hovy et al. (2006) managed to reach 90% inter-annotator agreement during the annotation of the resource. Therefore, even though OntoNotes resolves in part the fine-granularity problem of SemCor, it is still limited by the number of distinct words covered, in fact, it only comprises 3380 different lemma-pos pairs tagged with at least one sense, a number that is not sufficient for large-scale WSD and does not even cover half the tagged words in the 5 standard WSD benchmarks[9]. Another valuable resource is "Princeton WordNet Gloss Corpus" (2008)[10], i.e., a corpus comprising all WordNet glosses where annotators semantically tagged each content word manually. This has been used to extend SemCor with new annotated data by (Vial, Lecouteux, and Schwab, 2018) and to extract relations between WordNet synsets (Mihalcea and Moldovan, 2001; Cuadros and Rigau, 2008; Espinosa-Anke et al., 2016). More recently, Passonneau et al. (2012) introduced MASC, a manually annotated corpus with senses from WordNet 3.1. Despite being the most recent one,

[9] Senseval-2(Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli, Jurgens, and Vannella, 2013a) and SemEval-2015 (Moro and Navigli, 2015)

[10] `http://wordnetcode.princeton.edu/glosstag.shtml`

it only covers roughly 100 distinct words. Moreover, it is not freely available; thus its contribution to WSD is limited both in terms of coverage and availability.

Even though manually-annotated corpora largely contributed to advance the research in Word Sense Disambiguation and allowed the development of high-performance supervised models – hence establishing supervised WSD as the most effective approach on English all-words WSD tasks – we are now approaching a plateau of model performances which mainly depend on the training corpora. As also noted before, in fact, in the last 18 years supervised WSD performance increased by roughly 3 points despite the growth in models' complexity and the availability of more considerable computational power. This highlight the urge of new and more complete sense-annotated datasets to be fed into supervised models and unleash their power that is still unexpressed.

Automatic and Semi-automatic Corpora

Manually annotating large datasets, as we already discussed extensively, is expensive; therefore different approaches have been developed during the years in order to automatise, in part or completely, this process.

Ng, Wang, and Chan (2003) proposed a semi-automatic approach to align English and Chinese sentences at the word level. Moreover, they exploited the parallel data (i.e., the English words that are aligned to the same word in Chinese are considered unambiguous) and manual annotations to build a sense-annotated training set. While it was already evident even before 2003 that parallel data were an excellent pivotal point for disambiguating words in two languages (Resnik and Yarowsky, 1997), this was one of the first attempts to confirm this intuition empirically. Following this approach, Chan and Ng (2005a) proposed a similar method that considers English words having the same meaning whenever they were translated with the same Chinese word. Differently from Ng, Wang, and Chan (2003), when more than one word-senses were associated with the same Chinese word, they considered valid only the most frequent one according to WordNet. Later, Zhong and Ng (2009) refined the method proposed by Chan and Ng (2005a) and introduced a novel approach to retrieve high-quality Chinese sentences for WordNet senses. Hence, by aligning new English sentences to those retrieved for the target word sense makes it possible to build a semantically annotated dataset by exploiting

bilingual corpora, thus relieving the burden of manual aligning Chinese tokens with WordNet senses.

Taghipour and Ng (2015) summarised the outcome of these aforementioned methodologies, and, by following Chan and Ng (2005a), produced OMSTI, a corpus of 1 Million sense-tagged instances by exploiting Chinese-English part of the MultiUN corpus (Ziemski, Junczys-Dowmunt, and Pouliquen, 2016). Following the same line of research, Bovi et al. (2017) proposed a method that, by exploiting a language agnostic WSD system, i.e., Babelfy (Moro, Raganato, and Navigli, 2014b), disambiguated each set of aligned sentences in Europarl (Koehn, 2005b) at the same time. This allows Babelfy to leverage a broader context coming from all the aligned sentences hence making it more precise in the disambiguation. The very same approach has been then applied to BabelNet glosses in multiple languages (Camacho-Collados et al., 2016) producing a corpus of 35 million definitions for 256 distinct languages.

To finally dispose of the need of aligned corpora, Raganato, Delli Bovi, and Navigli (2016) introduced SEW, a heuristic-based approach to automatically produce sense-annotated corpora from Wikipedia. This approach benefits from the large number of tokens annotated in Wikipedia – those words linked via a hyperlink to a Wikipedia page – and exploit them to propagate their links. As a result, their corpus is large in terms of number of tokens and annotations, however, due to the nature of Wikipedia, it mostly covers named entities and concrete concept while lacks many abstract concepts. A key factor that had been ignored by all the approaches above is the senses' distribution within the corpus, i.e., the probability of a given sense to appear in a sentence. Indeed, they did not even consider this phenomenon, relying on the natural distribution that senses had in the various input corpora (i.e., Wikipedia, Europarl, United Nations Parallel corpus, etc.). However, depending on the corpus, the output dataset may also lack frequent senses or may be biased toward a specific domain. This happens, for example, when using topic-specific corpora such as Europarl or the United Nations Parallel corpus, as well as for encyclopedic resources (e.g., Wikipedia) which are usually full of named entity and concrete concepts but lack many abstract meanings. The importance of controlling the distribution of senses within a corpus has been largely explored by Postma, Bevia, and Vossen (2016a), where they show that, by mimicking the test-set distribution in the training set, supervised models had a boost of dozens of points (more on sense distributions is

presented in the following Section). The most recent effort in this field (which is also part of this thesis) has been carried by Pasini and Navigli (2017, Train-O-Matic) who introduced a knowledge-based approach to automatically generate sense-annotated corpora in potentially any language available in BabelNet. In contrast to all the other approaches, Train-O-Matic – in its extended version (Pasini and Navigli, 2019) – takes directly into account a target sense distribution during the process of generating a dataset, being also able to automatically infer it from a collection of texts given as additional input. Furthermore, it proved to lead a supervised model to results that are sometimes higher or on par with those attained when using a manually-curated resource as training set, i.e., SemCor, and, in general, higher than those achieved when using the best performing corpora among those previously introduced. We will extensively describe Train-O-Matic in Chapter 4.

To wrap-up, a plethora of different approaches have been proposed across the years to mitigate the knowledge acquisition bottleneck. Most of them relied on aligned-corpora (Chan and Ng, 2005a; Zhong and Ng, 2009; Taghipour and Ng, 2015; Camacho-Collados et al., 2016; Bovi et al., 2017), on the structure of Wikipedia (Raganato, Delli Bovi, and Navigli, 2016), or on BabelNet, i.e., Train-O-Matic (Pasini and Navigli, 2017; Pasini, Elia, and Navigli, 2018; Pasini and Navigli, 2019). However, even though some of them demonstrated enabling a supervised model to achieve higher performance when merged with SemCor itself (Taghipour and Ng, 2015), or managed to beat SemCor on some of the tested datasets (Pasini and Navigli, 2017) on the English all-words WSD task, all of them failed to become a valid alternative to manually-curated corpora. Their contribution, in fact, can be more appreciated on languages different from English, where it was not even possible to train a supervised model due to the lack of annotated data. This was, therefore, a first step into making richer the set of semantically-tagged corpora available for languages other than English, thus enabling supervised model to compete with knowledge-based and unsupervised approaches on the all-words multilingual tasks of WSD.

2.3 Word Sense Distribution

The distribution of senses, as many other language phenomena, follows the Zipf's distribution, with a few word's senses that are very frequent and a long tail of rare ones. Moreover, the sense distribution may vary drastically from corpus to corpus depending on the topic encompassed by the documents therein. Therefore, ignoring the distribution of senses during the automatic building of an annotated corpus may lead to unbalance the dataset which, in turn, would mislead a supervised model towards a wrong sense distributions. However, despite its central role in creating sense-annotated corpora, only a few efforts have been put on investigating how to automatically infer the distribution of senses or leverage it to bias the dataset towards a specific domain. Inducing the word sense distribution, indeed, is strictly correlated to predicting the topic of a document (Escudero, Màrquez, and Rigau, 2000; Chan and Ng, 2005b) since each topic may have a different distribution. Moreover, due to the skew nature of sense distributions (McCarthy et al., 2004), changing topic means drastically changing the sense distribution in its most important components. This was also the focus of the paper by Escudero, Màrquez, and Rigau (2000). They studied the behaviour of several Word Sense Disambiguation supervised models when trained on a domain-specific corpus and tested on documents from a different domain. As expected, they reported a drop in results of more than 20 points, hence confirming the drastic change in sense distribution from one domain to the other. Furthermore, they investigated the performance of the classifiers when a small portion of annotated data from the target domain were added to the training set, showing an increase in models' accuracy of roughly 12 points. Following a similar intuition, Chan and Ng (2005b) designed two algorithms for inferring the word sense distribution given a corpus of sentences. They fed the learned priors to a naive Bayes WSD classifier and showed improvement in the performance of more than 1 point on WSD standard benchmarks.

A similar line of research was the one followed by McCarthy et al. (2004), Mohammad and Hirst (2006) and McCarthy et al. (2007), who focused on determining the most frequent sense of a word in a given corpus. Due to the sense distribution skewness, determining the most frequent sense means disambiguating correctly the vast majority of a word occurrences; moreover, this can be used as a strong backoff strategy which can also be tuned depending on the texts one expects as input.

More recently, Lau et al. (2014) proposed a topic-modelling-based approach for estimating the word sense distributions in a corpus and was followed two years later by Bennett et al. (2016) who optimised and refined their method and introduced LexSemTM, a sense-annotated corpus to evaluate the induced sense distributions. It has better coverage than SemCor over polysemous words, and better reflects the real distribution of word senses in contemporary English. The last effort in this field, to the best of my knowledge, is Hauer, Luan, and Kondrak (2019). In this work, the authors propose two methods that, given a target word, they can estimate the word sense distribution by looking at the co-occurring senses. Postma et al. (2016b), differently from the other approaches, put the focus on determining when a word is intended with its most frequent sense and when it is not. This allows the authors to focus on the disambiguation of the least frequent senses, improving UKB (Agirre, de Lacalle, and Soroa, 2014) performance, a knowledge-based WSD approach, by roughly 8 points. Following the same philosophy, Postma, Bevia, and Vossen (2016a), showed that bigger datasets do not always lead a WSD model to better performance. Indeed, they proved that a WSD model benefits more from additional automatically-generated training instances that follow the test set distribution, than from adding a more significant number of manually annotated examples with a different distribution.

In this thesis, we contributed to mitigating the knowledge acquisition bottleneck problem in WSD by proposing a method for the automatic generation of sense-annotated corpora in any language that is supported by BabelNet (Pasini and Navigli, 2017; Pasini, Elia, and Navigli, 2018). The proposed approach differs from all the ones mentioned above since it mainly relies on a knowledge base, i.e., BabelNet, and does not need parallel data (Ng, Wang, and Chan, 2003; Chan and Ng, 2006; Taghipour and Ng, 2015; Camacho-Collados et al., 2016) or a set of manually annotated words (Raganato, Delli Bovi, and Navigli, 2016). Moreover, it can be easily applied to any language that is supported by BabelNet (more than 200) and proved to lead a supervised WSD model to results better than those attained by the same model trained on other automatically-, semi-automatically- and sometimes manually-annotated corpora. These works naturally led us to investigate on the sense distribution of the generated dataset and if it would have been beneficial to encode specific distributions based on the final application of the WSD model

(e.g., the various test sets). Therefore, we followed our study by developing two knowledge-based methods for inducing the sense distribution from the raw text (Pasini and Navigli, 2018) and studying the effect of shaping the automatically generated datasets according to the learned distribution (Pasini and Navigli, 2019). The two approaches are language-independent and rely on BabelNet. They differ to the other approaches proposed over the years since, i) they infer the whole distribution of senses and do not compute only the most frequent sense of a word (McCarthy et al., 2004; Mohammad and Hirst, 2006; McCarthy et al., 2007) and ii) they only rely on the underlying semantic network and are language agnostic in the sense that they can be applied to any language supported by BabelNet. Thanks to these approaches it was possible to relieve the need of humans in the process of shaping the training data with respect to the test sense distribution. Moreover, the results confirm on a larger scale the hunch of Escudero, Màrquez, and Rigau that adding annotated examples from the target domain to the training set is beneficial to the overall performance of a WSD model.

We think that, while on the one hand taking advantage of the new technologies in the hardware and machine learning fields is necessary in order to push as forward as possible the performance of WSD models with the currently available data, on the other hand, too little attention is paid to the developing of new and richer sense-annotated dataset. To overcome this limitation, general models that can take advantage from very large amount of unlabelled data have been developed and successfully applied to many different NLP tasks (Howard and Ruder, 2018; Peters et al., 2018; Devlin et al., 2018; Radford et al., 2018, 2019). However, these models are very large and sometimes even too large to fit in a single top-tier gpu, hence, feeding inductive bias directly into the network instead of letting it learn its own rules from enormous amount of data, is another research path that is worth a mention. This kind of approaches, in fact, are catching on also thanks to the recent effort of Luo et al. which aim at providing additional semantic information via sense glosses to a neural model. Nevertheless, even though the performance attained by Luo et al. (2018) are slightly better than the previous state of the art, they are still in the same ballpark, proving that it is not clear which is the best way for exploiting such kind of semantic data into a neural architecture.

Therefore, in this thesis, we focused on the study and development of novel approaches to mitigate the need for human intervention in the process of creating

semantically-annotated datasets despite of building incrementally complicated new models. The results attained show that the automatic generation of training data is a viable way of tackling the knowledge acquisition bottleneck problem in WSD. Indeed they lead to performance that is sometimes better, and in general, in the same ballpark than those attained when the training is performed on a manually-annotated dataset. Moreover, predicting the sense distribution of a target text, and shaping the training data accordingly, proved to be very useful and worthy of further investigation.

Chapter 3

Preliminaries

In this Chapter we provide some important background knowledge that is needed to put the rest of this thesis in context. In what follow, we introduce the resources that have been used across the works presented in the thesis and the algorithms which we leveraged.

3.1 WordNet

WordNet (Miller et al., 1990; Fellbaum, 1998a) is undoubtedly the most used a lexical knowledge resource in NLP and is manually built based on psycholinguistic principles. It is organised in group of words with the same part of speech that share a common meaning, i.e., synonyms, that are called *synset*. For example, the synset comprising the noun *plane* in its *airplane* meaning is the following:

$$\{\text{plane}_n^1, \text{airplane}_n^1, \text{aeroplane}_n^1\}$$

where subscripts and superscripts represent the part of speech and the sense number of each word, respectively. Therefore, all those three words may evoke this specific synsets since all of them share this specific meaning. However, since *plane* is ambiguous and may be used in different contexts, it also appear in four synsets, e.g., the one expressing its mathematical meaning:

$$\{\text{plane}_n^2, \text{sheet}_n^4\}$$

or its more abstract meaning:

$$\{\text{plane}_n^3\}$$

Each synset is then provided with a definition (*gloss*) and a few usage examples. E.g., the *airplane* meaning of *plane* has the following definition: *an aircraft that has a fixed wing and is powered by propellers or jets* and only one examples of use: *the flight was delayed due to trouble with the airplane*.

Synsets are connected to each other via semantic relation which can express different types of connections such as:

- **hypernymy** when a concept is generalised by another one. For example, aircraft^1_n and vehicle^1_n are two parent concept in the hypernymy path from airplane^1_n to the root.
- **meronymy** when a concept is *part of* another one. For example, wing^2_n has a meronymy relation with airplane^1_n.

WordNet counts, in its last version (3.1), $117{,}659$ synsets for all the open class part of speech, i.e., noun, adjective, verb and adverb, and $364{,}569$ lexico-semantic relations between them. It inspired several works in the field and have been largely used as reference knowledge base. For example, the *Princeton WordNet Gloss Corpus*[1] comprises all the WordNet definitions where the content words were manually disambiguated with the WordNet synsets. Moreover, it serves as a pivoting point for several editions in different languages of WordNet, e.g., Italian WordNet (Toral et al., 2010), German WordNet (Hamp and Feldweg, 1997) and many others. Recently, different attempts have been made to conflate all the language-specific versions into a unique multilingual edition of WordNet, it is the case of Open Multilingual WordNet (Bond and Foster, 2013) or MultiWordNet (Pianta, Bentivogli, and Girardi, 2002).

3.2 Wikipedia

Wikipedia is a multilingual encyclopedic resource created thanks to the efforts of hundreds of thousands contributors across the world. It is structured in pages, each representing a concept such as PHILOSOPHY[2], or a real world entity, e.g., BARACK OBAMA[3] and categories, which groups pages that share common properties together,

[1] http://wordnetcode.princeton.edu/glosstag.shtml

[2] `https://en.wikipedia.org/wiki/Philosophy`

[3] `https://en.wikipedia.org/wiki/Barack_Obama`

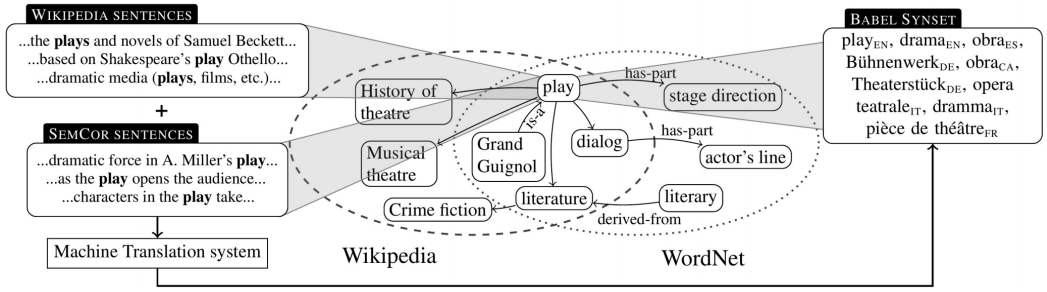

Figure 3.1. An illustration of BabelNet mapping between WordNet and Wikipedia drawn from the original article (Navigli and Ponzetto, 2012). The labeled edges come from WordNet (e.g., Grand Guignol^1_n is-a play^1_n) while the others from Wikipedia hyperlinks (MUSICAL THEATRE is related to play^1_n).

for example, the category ANIMATED COMEDY FILMS comprises all the pages representing an animated comedy. They, in turn, are organised in a Directed Acyclic Graph and are hence connected by some sort of is-a connections. Each Wikipedia article comprise a textual description and several information regarding the concept or the entity it represents. For example, the BARACK OBAMA page comprises not only a brief description of his life and activities, but also many sections describing in details his political and professional career. The text within an article can contains hyperlink to other Wikipedia pages, hence, the resource can be organised in a graph where a page is connected to another one if one of the two has a hyperlink towards the other. Furthermore, Wikipedia can also be seen as a general-purpose corpus of sentences that express concepts or describe named entities across different domains.

3.3 BabelNet

BabelNet (Navigli and Ponzetto, 2010, 2012) is a knowledge graph which encompasses several lexical-semantic resources such as WordNet (Miller et al., 1990; Fellbaum, 1998a), Wikipedia, Wikidata, etc. It was originally created by automatically mapping Wikipedia and WordNet, hence integrating the orthogonal dimensions of the two resources, i.e., the set of nominal and non nominal concepts of WordNet and the large amount of articles dedicated to real world entities. Moreover, while on the one hand, WordNet encompasses lexico-semantic relations that typically describe paradigmatic relations, Wikipedia provide more general connection rep-

resenting relatedness between concepts. Hence, as one can see in Figure 3.1 the information comprised by the two resources are complementary and both contribute to build the rich set of information comprised in a BabelNet synset. The synsets are then organised into a graph structure where edges represent lexico-semantic relations. The latter may be either typed, hence expressing paradigmatic connections, e.g., hypernymy, meronymy, etc., or simply indicating semantic relatedness, thus connecting synsets that may occur together in the same text or sentence. BabelNet nodes group together set of synonyms across languages, therefore, it is possible to represent the same meanings across many different languages (more than 200 at the time of writing) with the same set of concepts. In this thesis, BabelNet has been leveraged to dispose of the need for distinct approaches for each specific language of interest. Specifically we exploited the semantic network and its structure to measure the relatedness among concepts and compute the probability of a meaning to appear in a sentence as later illustrated in Section 4.1.1

3.4 Personalised PageRank

The PageRank (Brin and Page, 1998) is an algorithm that operates on graphs which is designed to compute the importance of each node therein. In fact, a node may be more or less important depending on the number of connections and on the importance of its neighbours. The approach can be approximated by means of random walks on the graph with a restart probability that is uniform between all the nodes. Hence, the PageRank score of node represents the probability of landing on that node when starting the navigation to any other random node in the graph. A variant of this algorithm is the so called Personalised PageRank (PPR) (Haveliwala, 2002), which, instead of uniformly distributing the restart probability across all the nodes, it focuses on one node only which becomes the only possible restarting point. With this change, each PageRank score represents the conditional probability of arriving on a certain node when starting the navigation from the restarting vertex. In NLP and especially in Word Sense Disambiguation, the Personalised PageRank has been largely exploited for different purposes, e.g., to compute the probabilities of each synset in a semantic network given another concept (Moro, Raganato, and Navigli, 2014b; Agirre, de Lacalle, and Soroa, 2014), or, more recently, to compute

dense representation of nodes (Perozzi, Al-Rfou, and Skiena, 2014). In this thesis, we used the PPR implementation by Lofgren et al. (2014)[4] to compute the probability of a concept given a specific context of words (Chapter 4).

[4] https://github.com/plofgren/fast-ppr-scala/

Chapter 4

Gathering Large-Scale and Multilingual Sense Annotations

The semantically-annotated datasets are a rare commodity since they are expensive to create in terms of both time and resources. Indeed, the few corpora that are available for English, i.e., SemCor (Miller et al., 1993a) and OntoNotes (Hovy et al., 2006) are still limited in coverage and outdated. The situation is even more critical when also considering other languages. Note that sense-tagged corpora are missing for Spanish, German, French and Italian, languages that count millions of Wikipedia articles and which have available NLP tools such as POS-taggers, lemmatisers, dependency parsers, etc. This castrated the development of multilingual WSD supervised models and, in general, circumscribed the research mostly to the English language. Similarly, knowledge-based approaches, which leverage graph algorithms on semantic networks, also suffer from the paucity of sense-tagged datasets. Indeed, since the underlying knowledge bases usually lack syntagmatic relations between concepts (i.e., those connecting meanings occurring often together in the same sentence), their performance is limited and usually lower than those attained by supervised models.

Several approaches have been therefore introduced in the past to overcome this scarcity of annotated data. However, most of them failed to scale easily on different languages and to offer a valid alternative to SemCor. Thus, to cope with all the problems mentioned above and dispose of the burden of annotating senses on a large number of texts, we developed a language-independent approach which, by exploiting a knowledge base, can provide hundreds of thousands of semantically-

annotated examples for potentially all the senses of the words in a language. In what follows, we first introduce Train-O-Matic (Pasini and Navigli, 2017) and then evaluate the quality of its automatically-generated corpora by means of several all-words English and multilingual WSD tasks.

4.1 Train-O-Matic

In this Section we present Train-O-Matic, a language-independent approach to the automatic construction of a sense-tagged training set. Train-O-Matic takes as input a corpus C (e.g., Wikipedia) and a semantic network $G = (V, E)$. We assume a WordNet-like structure of G, i.e., V is the set of concepts (i.e., synsets) such that, for each word w in the vocabulary, $Senses(w)$ is the set of vertices in V that are expressed by w, e.g., the WordNet synsets that include w as one of their senses.

Train-O-Matic is divided in the following three steps:

- **Lexical profiling:** for each vertex in the semantic network, we compute its Personalized PageRank (Haveliwala, 2002) vector, which provides its lexical-semantic profile (Section 4.1.1).
- **Sentence scoring:** For each sentence containing a word w, we compute a probability distribution over all the senses of w based on its context by means of two alternative methods: a knowledge-based approach (Section 4.1.2) and a semi-supervised approach (Section 4.1.3).
- **Sentence ranking and selection:** for each sense s of a word w in the vocabulary, we select those sentences that are most likely to use w in the sense of s (Section 4.1.4).

4.1.1 Lexical profiling

In a semantic network, the distance between 2 vertices can be intended as a measure of relatedness between the two concepts. Similarly, the probability of reaching a node v' starting from v can be intended as the probability of finding the concept v' given that we have already found the concept v in a sentence or in a text. We thus define the lexical profile of a vertex v in a graph $G = (V, E)$ as the probability

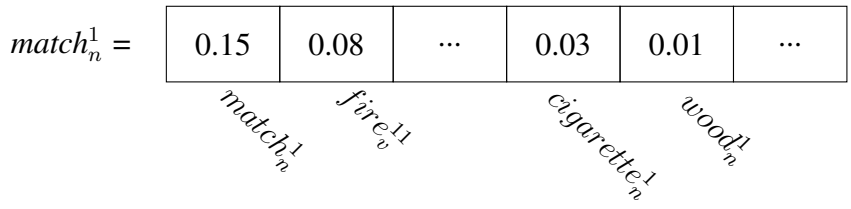

Figure 4.1. Personalized PageRank vector for the lighter sense of *match*.

distribution over all the vertices v' in the graph. Such distribution is computed by applying the Personalized PagaRank (PPR) (Haveliwala, 2002) algorithm, a variant of the traditional PageRank Brin and Page (1998). While the latter is equivalent to performing random walks with uniform restart probability on every vertex at each step, PPR, on the other hand, makes the restart probability non-uniform, thereby concentrating more probability mass in the surroundings of those vertices having higher restart probability. Formally, (P)PR is computed as follows:

$$v^{(t+1)} = (1 - \alpha)v^{(0)} + \alpha M v^{(t)} \tag{4.1}$$

where M is the row-normalized adjacency matrix of the semantic network, the restart probability distribution is encoded by vector $v^{(0)}$, and α is the well-known damping factor usually set to 0.85 (Brin and Page, 1998). If we let $v^{(0)}$ diverge to a unit probability vector $(0, \ldots, 0, 1, 0, \ldots, 0)$, i.e., restart is always on a given vertex, PPR outputs the probability of reaching every vertex starting from the restart vertex after a certain number of steps. This approach has been used in the literature to create semantic signatures (i.e., profiles) of individual concepts, i.e., vertices of the semantic network by Pilehvar, Jurgens, and Navigli (2013), and then to determine the semantic similarity of concepts. As also done by Pilehvar and Collier (2016), we instead use the PPR vector as an estimate of the conditional probability of a word w' given the target sense[1] $s \in V$ of word w:

$$P(w'|s, w) = \frac{\max_{s' \in Senses(w')} v_s(s')}{Z} \tag{4.2}$$

where $Z = \sum_{w"} P(w"|s, w)$ is a normalization constant, v_s is the vector resulting from an adequate number of random walks used to calculate PPR, and $v_s(s')$ is the vector component corresponding to sense s'. For example, given the PPR vector v

[1]Note that we use senses and concepts (synsets) interchangeably, because – given a word – a word sense unambiguously determines a concept (i.e., the synset it is contained in) and vice versa.

of the lighter sense of *match* s (Figure 4.1), we compute the probability of the word *cigarette* given the sense s and the target word *match* by applying Formula 4.2, i.e, by picking the highest scored synset in v which comprise *cigarette* among its senses (i.e., cigarette^1_n). Thus $P(\textit{cigarette} \mid \text{match}^1_n, \textit{match}) = 0.03$. To fix the number of iterations needed to have a sufficiently accurate vector, we follow Lofgren et al. (2014) and set the error $\delta = 0.00001$ and the number of iterations to $\frac{1}{\delta} = 100,000$.

As a result of this lexical profiling step we have a probability distribution over vocabulary words for each given word sense of interest.

4.1.2 Knowledge-based Sentence Scoring

The objective of the second step is to score the importance of word senses for each of the corpus sentences which contain the word of interest. Given a sentence $\sigma = w_1, w_2, \ldots, w_n$, that has been tokenised, lemmatised and POS-tagged, we want to compute the probability of a given sense s to appear therein. Therefore, for each given target word w_i in the sentence ($w_i \in \sigma$), and for each of its senses $s \in Senses(w_i)$, we compute the probability $P(s|\sigma, w_i)$. Thanks to Bayes' theorem we determine the probability of sense s of w_i given the sentence as follows:

$$P(s|\sigma, w_i) = \frac{P(\sigma|s, w_i)P(s|w_i)}{P(\sigma|w_i)} \tag{4.3}$$

$$= \frac{P(w_1, \ldots, w_n|s, w_i)P(s|w_i)}{P(w_1, \ldots, w_n|w)}$$

$$\propto P(w_1, \ldots, w_n|s, w_i)P(s|w_i) \tag{4.4}$$

$$\approx P(w_1|s, w_i) \ldots P(w_n|s, w_i)P(s|w_i) \tag{4.5}$$

where Formula 4.4 is proportional to the original probability (due to removing the constant in the denominator) and is approximated with Formula 4.5 due to the assumption of independence of the words in the sentence. Hence, $P(w_j|s, w_i)$ with j ranging in $[1, n]$ is calculated thanks to the PPR vectors (Formula 4.2) and $P(s|w_i)$ is set to $1/|Senses(w_i)|$ (recall that s is a sense of w_i), thus, no bias is given towards any of the word senses. For example, given the sentence $\sigma =$"A match is a tool for starting a fire", the target word $w =$ *match* and its set of senses $S_{match} = \{s^1_{match}, s^2_{match}\}$, where s^1_{match} is the sense of *lighter* and s^2_{match} is the sense of *game match*, we want to calculate the probability of each $s^i_{match} \in S_{match}$

of being the correct sense of *match* in the sentence σ. Following Formula 4.5 we have:

$$
\begin{aligned}
P(s^1_{match}|\sigma, match) &\approx P(\text{tool}|s^1_{match}, match) \\
&\cdot P(\text{start}|s^1_{match}, match) \\
&\cdot P(\text{fire}|s^1_{match}, match) \\
&\cdot P(s^1_{match}|match) \\
&= 2.1 \cdot 10^{-4} \cdot 2 \cdot 10^{-3} \cdot 10^{-2} \cdot 5 \cdot 10^{-1} \\
&= 2.1 \cdot 10^{-9}
\end{aligned}
$$

$$
\begin{aligned}
P(s^2_{match}|\sigma, match) &\approx \\
&P(\text{tool}|s^2_{match}, match) \\
&\cdot P(\text{start}|s^2_{match}, match) \\
&\cdot P(\text{fire}|s^2_{match}, match) \\
&\cdot P(s^2_{match}|match) \\
&= 10^{-5} \cdot 2.9 \cdot 10^{-4} \cdot 10^{-6} \cdot 5 \cdot 10^{-1} \\
&= 1.45 \cdot 10^{-15}
\end{aligned}
$$

As one can see, the first sense of *match* has a much higher probability due to its stronger relatedness to the other words in the context (i.e. *start*, *fire* and *tool*). Note also that all the probabilities for the second sense are at least one magnitude less than the probability of the first sense.

4.1.3 Sentence Scoring with Word Embeddings

The second approach is semi-supervised and exploits the lexical context of synsets available in an annotated corpus (e.g., SemCor) to compute a latent representation for each concept. Such vector is computed as the average of the word vectors across all sentences in the corpus where the synset appears. Thus, for each sentence σ containing a given synset s, we first average the vectors of words in a window surrounding s in σ and then average all the vectors computed for each sentence.

More formally, given a synset s and a set of sentences ξ were s appears as sense of the word w, we compute the vector for s as follows:

$$v(s,w) = \frac{1}{|\xi|}\sum_{\sigma\in\xi} vec(\sigma, w, W) \tag{4.6}$$

$$vec(\sigma, w, W) = \frac{1}{2W+1}\sum_{i=w.index-W}^{w.index+W} word2vec(\sigma[i]) \tag{4.7}$$

where the function $vec(\sigma, w, W)$ is the average of the word vectors surrounding w in the given window W and $word2vec(w)$ returns the vector representation for w. Note that, we can now compare word senses and sentences as they all lie in the same vector space defined by the word vectors. Thanks to this, given a target word w and a new sentence where w appears, we compute its vector as in Formula 4.6 (i.e., by averaging the word vectors in the surroundings of the target word), and we compute the probability distribution over the senses of w by assigning a score defined by the cosine similarity between the sentence and the senses' vectors. More formally we compute the probability of a sense s of a word w in a sentence σ as follows:

$$P(s|\sigma, w) = softmax(sim(vec(\sigma, w, W), v(s, w))) \tag{4.8}$$

where $vec(\sigma, w, W)$ and $v(s, w)$ are the two functions we introduced in Formula 4.6 and 4.7, and $sim(x, y)$ is the cosine similarity between the two input vectors.

4.1.4 Sense-based Sentence Ranking and Selection

We are now ready to rank the sentences where a given sense s appears so that to assign a lower score to those having least reliable disambiguations.

Given word w and a sense $s_1 \in Senses(w)$, we score each sentence σ in which w appears and s_1 is its most likely sense according to a formula that takes into account the difference between the first (i.e., s_1) and the second most likely sense of w in σ:

$$\Delta_{s_1}(\sigma) = P(s_1|\sigma, w) - P(s_2|\sigma, w) \tag{4.9}$$

where $s_1 = \arg\max_{s\in Senses(w)} P(s|\sigma, w)$, and $s_2 = \arg\max_{s\in Senses(w)\setminus\{s_1\}} P(s|\sigma, w)$. We then sort all sentences based on $\Delta_{s_1}(\cdot)$ and return a ranked list of sentences where the word w is most likely to be sense-annotated with s_1. Although we recognise that other scoring strategies could have been used, this was experimentally the most

effective one when compared to alternative strategies, i.e., the sense probability, the number of words related to the target word w, the sentence length or a combination thereof.

4.2 Creating a Denser and Multilingual Semantic Network

In the previous Section we assumed that WordNet was our semantic network, with synsets as vertices and edges represented by its semantic relations. However, while its lexical coverage is high, with a rich set of fine-grained synsets, at the relation level WordNet provides mainly paradigmatic information, i.e., relations like hypernymy (is-a) and meronymy (part-of). It lacks, on the other hand, syntagmatic relations, such as those that connect verb synsets to their arguments (e.g., the appropriate senses of eat_v and $food_n$), or pairs of noun synsets (e.g., the appropriate senses of bus_n and $driver_n$).

Intuitively, Train-O-Matic would suffer from such a lack of syntagmatic relations, as the relevance of a sense for a given word in a sentence depends directly on the possibility of visiting senses of the other words in the same sentence (see Formula 4.5) via random walks as calculated with Formula 4.1. Such reachability depends on the connections available between synsets within the semantic network. Because syntagmatic relations are sparse in WordNet, if it was used on its own, we would end up with a poor estimation of senses' probabilities and ranking of sentences for any given word sense. Moreover, even though the methodology presented in Section 4.1 is language-independent, Train-O-Matic would lack information (e.g. senses for a word in an arbitrary vocabulary) for languages other than English.

To cope with these issues, we exploit BabelNet[2] (Navigli and Ponzetto, 2012) a huge multilingual semantic network obtained from the automatic integration of WordNet, Wikipedia, Wiktionary and other resources, and create the BabelNet subgraph induced by the WordNet vertices. The result is a graph whose vertices are BabelNet synsets that contain at least one WordNet sense and whose edge set includes all those relations in BabelNet coming either from WordNet itself or from links in other resources mapped to WordNet (such as hyperlinks in a Wikipedia

[2] http://babelnet.org

mouse (animal)		mouse (device)	
WordNet	WordNet$_{BN}$	WordNet	WordNet$_{BN}$
mouse^1_n	mouse^1_n	mouse^4_n	mouse^4_n
tail^1_n	little^1_a	wheel^1_n	computer^1_n
hairless^1_a	rodent^1_n	electronic_device^1_n	pad^4_n
rodent^1_n	cheese^1_n	ball^3_n	cursor^1_n
trunk^3_n	cat^1_n	hand_operated^1_n	operating_system^1_n
elongate^2_a	rat^1_n	mouse_button^1_n	trackball^1_n
house_mouse^1_n	elephant^1_n	cursor^1_n	wheel^1_n
minuteness^1_n	pet^1_n	operate^3_v	joystick^1_n
nude_mouse^1_n	experiment^1_n	object^1_n	Windows^1_n

Table 4.1. Top-ranking synsets of the PPR vectors computed on WordNet (first and third columns) and WordNet$_{BN}$ (second and fourth columns) for *mouse* as animal (left) and as device (right).

article connecting it to other articles). The greatest contribution of syntagmatic relations comes, indeed, from Wikipedia, as related articles are interlinked (e.g., the English Wikipedia *Bus* article[3] is linked to *Passenger*, *Tourism*, *Bus lane*, *Timetable*, *School*, and many more).

Because not all Wikipedia (and other resources') pages are connected with the same degree of relatedness (e.g., countries are often linked, but they are not necessarily closely related to the source article in which the link occurs), we apply the following weighting strategy to each edge $(s, s') \in E$ of our WordNet-induced subgraph of BabelNet $G = (V, E)$:

$$w(s, s') = \begin{cases} 1 & (s, s') \in E(\text{WordNet}) \\ WO(s, s') & \text{otherwise} \end{cases} \tag{4.10}$$

where E(WordNet) is the edge set of the original WordNet graph and $WO(s, s')$ is the weighted overlap measure which calculates the similarity between two synsets:

$$WO(s, s') = \frac{\sum_{i=1}^{|S|}(r_i^1 + r_i^2)^{-1}}{\sum_{i=1}^{|S|}(2i)^{-1}} \tag{4.11}$$

[3]Retrieved on April 20th, 2019.

where r_i^1 and r_i^2 are the rankings of the i-th synsets in the set S of the components in common between the vectors associated with s and s', respectively. Because at this stage we still have to calculate our synset vector representation, we use the pre-computed NASARI vectors (Camacho-Collados, Pilehvar, and Navigli, 2015) to calculate WO. This choice is due to WO's higher performance over cosine similarity for vectors with explicit dimensions (Pilehvar, Jurgens, and Navigli, 2013).

As a result, each row of the original adjacency matrix M of G will be replaced with the weights calculated in Formula 4.10 and then normalized in order to be ready for PPR calculation (see Formula 4.1). An idea of why a denser semantic network has more useful connections and thus leads to better results is provided by the example in Table 4.1[4], where we show the highest-probability synsets in the PPR vectors calculated with Formula 4.1 for two different senses of *mouse* (its animal and device senses) when WordNet (left) and our WordNet-induced subgraph of BabelNet (WordNet$_{BN}$, right) are used as the underlying semantic network for PPR computation. Note that, WordNet's top synsets are related to the target synset via paradigmatic (i.e., hypernymy and meronymy) relations, while WordNet$_{BN}$ includes many syntagmatically-related synsets (e.g., *experiment* for the animal, and *operating system* and *Windows* for the device sense, among others).

4.3 Experimental Setup

We now present our setup for evaluating the datasets produced by Train-O-Matic and define, in details, each competitor, external WSD system, test set, etc. that we used to that end.

Corpora for sense annotation We used two different corpora to extract sentences: Wikipedia and the United Nations Parallel Corpus (Ziemski, Junczys-Dowmunt, and Pouliquen, 2016). The first is the largest and most up-to-date encyclopedic resource, containing definitional information, the second, on the other hand, is a public collection of parliamentary documents of the United Nations Parallel Corpus. The application of Train-O-Matic to the two corpora produced two sense-annotated datasets, which we named T-O-M$_{Wiki}$ and T-O-M$_{UN}$, respectively.

[4]We use the notation w_p^k introduced in Navigli (2009) to denote the k-th sense of word w with part-of-speech tag p.

Semantic Network We created sense-annotated corpora with Train-O-Matic both when using PPR vectors computed from vanilla WordNet and when using WordNet$_{BN}$, our denser network obtained from the WordNet-induced subgraph of BabelNet (see Section 4.2).

Gold standard datasets We performed our evaluations using the framework made available by Raganato, Camacho-Collados, and Navigli (2017) on five different all-words datasets, namely: the Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli, Jurgens, and Vannella, 2013b) and SemEval-2015 (Moro and Navigli, 2015) WSD datasets. We focused on nouns only, given the fact that Wikipedia provides connections between nominal synsets only, and therefore contributes mainly to syntagmatic relations between nouns.

Comparison sense-annotated corpora To show the impact of our T-O-M corpora in WSD, we compared its performance on the above gold standard datasets, against training with:

- **SemCor**[5] (Miller et al., 1993b), a corpus containing about $226{,}036$ words annotated manually with WordNet senses.
- **One Million Sense-Tagged Instances**[6] (Taghipour and Ng, 2015, OMSTI), a sense-annotated dataset obtained via a semi-automatic approach based on the disambiguation of a parallel corpus, i.e., the United Nations Parallel Corpus, performed by exploiting manually translated word senses. Because OMSTI integrates SemCor to increase coverage, to keep a level playing field we excluded the latter from the corpus.
- **SenseDefs**[7] (Camacho-Collados et al., 2016), a multilingual corpus built by jointly disambiguating the glosses of the same BabelNet synset in multiple languages. It contains almost 250 million sense annotations in 256 different languages.

[5] http://lcl.uniroma1.it/wsdeval/
[6] http://lcl.uniroma1.it/wsdeval/
[7] http://lcl.uniroma1.it/disambiguated-glosses/

- **EuroSense**[8] (Bovi et al., 2017), similarly to SenseDefs, it exploits the parallel data of EuroParl (Koehn, 2005a) in order to augment the context and increase disambiguation precision. It contains more than 123 million sense annotations in 21 languages.

We note that T-O-M, instead, does not require any parallel corpus.

Reference system In all our experiments, we used It Makes Sense (Zhong and Ng, 2010, IMS), a WSD system based on linear Support Vector Machines, as our reference system for comparing its performance when trained on Train-O-Matic, against the same WSD system trained on other sense-annotated corpora (i.e., SemCor, OMSTI, SenseDefs and EuroSense). Following the WSD literature, unless stated otherwise, we report performance in terms of F1, i.e., the harmonic mean of precision and recall.

We note that it is not the purpose of this paper to show that T-O-M, when integrated into IMS, beats all other configurations or alternative systems, but rather to fully automatize the WSD pipeline with performances which are competitive with the state of the art.

Baseline As a traditional baseline in WSD, we used the Most Frequent Sense (MFS) baseline given by the first sense in WordNet. The MFS is a very competitive baseline, due to the sense skewness phenomenon in language (McCarthy et al., 2004; Navigli, 2009).

Number of training sentences per sense Given a target word w, we sorted its senses $Senses(w)$ by following the WordNet ordering and selected the top k_i training sentences according to Formula 4.9 for the i-th sense, where:

$$k_i = \frac{1}{i^z} * K \tag{4.12}$$

we set $K = 500$ and $z = 2$ by tuning the system on a separate small in-house development dataset[9].

[8] `http://lcl.uniroma1.it/eurosense/`

[9] 50 word-sense pairs annotated manually.

Dataset	T-O-M$_{Wiki}$ BN	T-O-M$_{Wiki}$ WN
Senseval-2	**70.5**	70.0
Senseval-3	**67.4**	63.1
SemEval-2007	**59.8**	57.9
SemEval-2013	**65.5**	63.7
SemEval-2015	68.6	**69.5**
ALL	**67.3**	65.7

Table 4.2. F1 of IMS trained on T-O-M when PPR is obtained from the WordNet graph (WN) and from the WordNet-induced subgraph of BabelNet (BN).

4.4 Results

4.4.1 Impact of syntagmatic relations

The first result we report regards the impact of vanilla WordNet vs. our WordNet-induced subgraph of BabelNet (WordNet$_{BN}$) when calculating PPR vectors. As can be seen from Table 4.2 – which shows the performance of the T-O-M$_{Wiki}$ corpora generated with the two semantic networks – using WordNet for PPR computation decreases the overall performance of IMS from 0.5 to around 4 points across the five datasets, with an overall loss of 1.6 F1 points. Similar performance losses were observed when using T-O-M$_{UN}$ (see Table 4.3). This corroborates our hunch discussed in Section 4.2 that a resource like BabelNet can contribute important syntagmatic relations that are beneficial for identifying (and ranking high) sentences which are semantically relevant for the target word sense. Therefore, in the following experiments, we report only results using WordNet$_{BN}$.

4.4.2 Comparison against sense-annotated corpora

We now move to comparing the performance of Train-O-Matic against corpora which are annotated manually (SemCor) and semi-automatically (OMSTI). In Table 4.3 we show the F1-score of IMS on each gold standard dataset in the evaluation framework and on all datasets merged together (last row), when it is trained with the various corpora described above.

As one can see, OMSTI manages to beat Train-O-Matic when using the United

Dataset	Train-O-Matic$_{Wiki}$	Train-O-Matic$_{UN}$	OMSTI	SemCor	MFS
Senseval-2	70.5	69.0	74.1	**76.8**	72.1
Senseval-3	67.4	68.3	67.2	**73.8**	72.0
SemEval-2007	59.8	57.9	62.3	**67.3**	65.4
SemEval-2013	**65.5**	62.5	62.8	**65.5**	63.0
SemEval-2015	**68.6**	63.5	63.1	66.1	66.3
ALL	67.3	65.3	66.4	**70.4**	67.6

Table 4.3. F1 of IMS trained on Train-O-Matic, OMSTI and SemCor, and MFS for the Senseval-2, Senseval-3, SemEval-2007, SemEval-2013 and SemEval-2015 datasets.

Nations parallel corpus as sentences' source, i.e., the same used by OMSTI itself. However, we argue that Train-O-Matic is more flexible than its competitor since it is not tied to that specific corpus and, instead, can be used to annotate sentences from any source. OMSTI, on the contrary, can only be applied to parallel corpora comprising texts in Chinese and, moreover, needs the manual intervention of human annotators to disambiguate the Chinese words. Hence, when Train-O-Matic is applied to a larger and more complete corpus such as Wikipedia (Train-O-Matic$_{Wiki}$) we show in Table 4.3 that it obtains higher performance than OMSTI (up to 5.5 points above) on three out of 5 datasets, performing overall 1 point above its competitor. The MFS, instead, is in the same ballpark as Train-O-Matic$_{Wiki}$, performing better on some datasets and worse on others.

Interestingly enough, we note that IMS trained on Train-O-Matic$_{Wiki}$ succeeds in surpassing or obtaining the same results as IMS trained on SemCor on SemEval-2015 and SemEval-2013. We view this as a significant achievement given the total absence of manual effort involved in Train-O-Matic.

Because overall Train-O-Matic$_{Wiki}$ outperforms Train-O-Matic$_{UN}$, in what follows we report all the results with Train-O-Matic$_{Wiki}$, except for the domain-oriented evaluation (see Section 4.4.4).

4.4.3 Performance without backoff strategy

IMS uses the MFS as a backoff strategy when no sense can be output for a target word in context (Zhong and Ng, 2010). Consequently, the performance of the MFS

Dataset	OMSTI	EuroSense	SenseDefs	Train-O-Matic	Total
Senseval-2	197	388	399	400	436
Senseval-3	197	420	422	435	469
SemEval-2007	68	126	123	127	127
SemEval-2013	249	628	625	629	751
SemEval-2015	102	218	224	226	253
ALL	456	1311	1334	1350	1557

Table 4.4. Number of nominal tokens for which at least one training example is provided by OMSTI or Train-O-Matic for each English dataset.

Dataset	OMSTI			Train-O-Matic			EuroSense			SenseDefs		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Senseval-2	64.8	28.5	39.6	69.5	65.5	**67.4**	63.1	55.2	58.9	67.4	67.4	**67.4**
Senseval-3	55.7	31.0	39.8	66.1	63.1	**64.6**	49.4	45.1	47.2	61.1	56.3	58.6
SemEval-2007	64.1	35.9	46.0	59.8	59.8	**59.8**	40.4	39.6	40.0	51.7	48.3	49.9
SemEval-2013	50.7	23.4	32.0	61.3	53.3	**57.0**	56.7	48.2	52.1	41.1	39.0	40.0
SemEval-2015	57.0	26.7	36.4	67.0	62.3	**64.6**	56.7	51.2	53.8	50.4	42.2	46.0
ALL	56.5	27.0	36.5	65.1	59.7	**62.3**	56.0	49.3	52.5	55.7	49.5	52.4

Table 4.5. Precision, Recall and F1 of IMS trained on OMSTI, EuroSense, SenseDefs and Train-O-Matic corpus without MFS backoff strategy for Senseval-2, Senseval-3, SemEval-2007, SemEval-2013 and SemEval-2015.

is mixed up with that of the SVM classifier. As shown in Table 4.4, OMSTI is able to provide annotated sentences for roughly half of the tokens in the datasets. Train-O-Matic together with EuroSense and SenseDefs, instead, is able to cover almost all words in each dataset with at least one training sentence. This means that in around 50% of the cases, OMSTI, gives an answer based on the IMS backoff strategy, i.e., the MFS.

To determine the real impact of the different training data, we therefore decided to perform an additional analysis of the IMS performance when the MFS backoff strategy is disabled. Because we suspected the system would not always return a sense for each target word, in this experiment we measured precision, recall and their harmonic mean, i.e., F1. The results in Table 4.5 confirm our hunch, showing that OMSTI's recall drops heavily, thereby affecting F1 considerably. Train-O-

Domain	Backoff	T-O-M$_{Wiki}$			T-O-M$_{UN}$			OMSTI			SemCor	MFS	Size
		P	R	F1	P	R	F1	P	R	F1	F1	F1	
Biology	MFS	63.0	63.0	63.0	65.9	65.9	**65.9**	65.9	65.9	**65.9**	66.3	64.4	135
	-	59.0	53.3	56.0	62.3	56.3	**59.2**	48.1	18.5	26.7	-		
Climate	MFS	68.1	68.1	**68.1**	63.4	63.4	63.4	68.0	68.0	68.0	70.1	67.5	194
	-	63.4	50.0	**55.9**	57.5	45.4	50.7	58.0	24.2	34.2	-		
Finance	MFS	68.0	68.0	**68.0**	56.6	56.6	56.6	64.4	64.4	64.4	63.7	56.2	219
	-	62.1	51.6	**56.4**	48.4	40.2	43.9	57.4	28.3	37.9	-		
Health Care	MFS	65.2	65.2	**65.2**	60.1	60.1	60.1	52.9	52.9	52.9	62.7	56.5	138
	-	61.3	55.1	**58.0**	55.6	50.0	52.6	34.6	18.4	24.0	-		
Politics	MFS	65.2	65.2	65.2	66.3	66.3	**66.3**	63.4	63.4	63.4	69.5	67.7	279
	-	62.5	54.8	58.4	63.9	55.9	**59.6**	54.1	21.5	30.8	-		
Social Issues	MFS	68.5	68.5	**68.5**	63.6	63.6	63.6	65.6	65.6	65.6	66.8	67.6	349
	-	63.1	53.0	**57.6**	57.2	47.9	52.1	54.7	25.2	34.5	-		
Sport	MFS	60.3	60.3	60.3	60.9	60.9	**60.9**	58.8	58.8	58.8	60.4	57.6	330
	-	58.3	54.6	**56.4**	58.1	53.3	55.5	45.0	23.0	30.4	-		

Table 4.6. Performance comparison over SemEval-2013 domain-specific datasets.

Matic performances, instead, remain high in terms of precision, recall and F1. This confirms that OMSTI relies heavily on data (those obtained for the MFS and from SemCor) that are produced manually, rather than semi-automatically. EuroSense and SenseDefs, instead, managed to lead IMS to better results than OMSTI but fail to beat Train-O-Matic on all datasets. This is because both approaches only rely on the disambiguation provided by Babelfy without applying any sentence filtering mechanism and hence including more noise in the training data which unavoidably confuses IMS.

4.4.4 Domain-specific Evaluation

To further inspect the ability of T-O-M to enable disambiguation in different domains, we decided to evaluate on specific documents from the various gold standard datasets which could be clearly assigned a domain label. Specifically, we tested on 13 SemEval-2013 documents from various domains[10] and 2 SemEval-2015 documents (namely, maths & computers, and biomedicine) and carried out two separate tests and evaluations of Train-O-Matic on each domain: once using the MFS backoff strategy, and once not using it. In Tables 4.6 and 4.7 we report the results of both T-O-M$_{Wiki}$ and T-O-M$_{UN}$ to determine the impact of the corpus type. As can

[10]Namely biology, climate, finance, health care, politics, social issues and sport.

Domain	Backoff	T-O-M$_{Wiki}$			T-O-M$_{UN}$			OMSTI			SemCor	MFS	Tokens
		P	R	F1	P	R	F1	P	R	F1	F1	F1	
Biomedicine	MFS	76.3	76.3	**76.3**	66.0	66.0	66.0	64.9	64.9	64.9	70.3	71.1	100
	-	76.1	72.2	**74.1**	64.4	59.8	62.0	60.5	26.8	37.2	-		
Maths & Computer	MFS	50.0	50.0	**50.0**	48.0	48.0	48.0	36.0	36.0	36.0	40.6	40.9	97
	-	50.0	47.0	**48.5**	47.8	44.0	45.8	21.2	11.0	14.5	-		

Table 4.7. Performance comparison over the Biomedical and Maths & Computer domains in SemEval-2015.

be seen in the tables, T-O-M$_{Wiki}$ systematically attains higher scores than OMSTI (except for the biology domain), and, in most cases, attains higher scores than MFS when the backoff is used, with a drastic, systematic increase over OMSTI with both Train-O-Matic configurations in recall and F1 when the backoff strategy is disabled. This demonstrates the usefulness of the corpora annotated by Train-O-Matic not only on open text, but also on specific domains. We note that T-O-M$_{UN}$ obtains the best results in the politics domain, which is the closest domain to the UN corpus from which its training sentences are obtained.

4.5 Scaling up to Multiple Languages

4.5.1 Multilingual Experimental Setup

In this section we investigate the ability of Train-O-Matic to scale to low-resourced languages, such as Italian, Spanish, German and French, for which training data for WSD is not available.

Thanks to BabelNet, in fact, Train-O-Matic can be used to generate sense-annotated data for any language supported by the knowledge base. Thus, in order to build new training datasets for the two languages, we ran Train-O-Matic on their corresponding versions of Wikipedia, then we tuned the two parameters K and z on an in-house development dataset[11]. In contrast to the English setting, in order to calculate Formula 4.12 we sorted the senses of each word by vertex degree in BabelNet. Finally we used the output data to train IMS.

[11]We set $K = 100$ and $z = 2.3$ for Spanish and French, $K = 100$ and $z = 2.5$ for Italian and $K = 100$ and $z = 2.0$ for German.

Language	Dataset	Best System	Train-O-Matic		
		F1	P	R	F1
Italian	ALL	56.6	65.1	55.6	**59.9**
	Computers & Math	46.6	52.7	43.3	**47.6**
	Biomedicine	65.9	76.6	67.6	**71.8**
Spanish	ALL	56.3	61.3	54.8	**57.9**
	Computers & Math	42.4	53.3	44.4	**48.5**
	Biomedicine	65.5	71.8	65.5	**68.5**

Table 4.8. Performance comparison between T-O-M and SemEval-2015's best SUDOKU Run.

Language	Best System	Train-O-Matic		
	F1	P	R	F1
German	62.0	65.8	60.8	**63.2**
French	**60.5**	61.1	59.9	**60.5**
Spanish	**71.0**	68.2	65.7	66.9
Italian	65.8	70.8	65.7	**68.2**

Table 4.9. Precision, Recall and F1 of IMS trained on Train-O-Matic, against SemEval-2013's best UMCC-DLSI run.

4.5.2 Multilingual Results

We performed the multilingual evaluation on all the available all-words WSD multilingual tasks, i.e., SemEval-2013 task 12 (Navigli, Jurgens, and Vannella, 2013a) and SemEval-2015 task 13 (Moro and Navigli, 2015). As can be seen from Table 4.8 and 4.9 Train-O-Matic enabled IMS to perform better than the SemEval-2015 best participating system (Manion and Sainudiin, 2014, SUDOKU) in all three settings (All domains, Maths & Computer and Biomedicine). Its performance was in fact, 1 to 3 points higher, with a 6-point peak on Maths & Computer in Spanish and on Biomedicine in Italian. On SemEval-2013, instead, Train-O-Matic manages to lead IMS to state of the art results on most of the datasets, achieving up to 2.2 F1 points more than the best scoring system on the Italian test set. This demonstrates Train-O-Matic to enable supervised WSD systems to surpass state-of-

the-art knowledge-based WSD approaches in most of the test sets without relying on manually curated data for training.

4.6 Conclusion

In this Chapter we presented Train-O-Matic, a knowledge-based approach to the automatic construction of large training sets for supervised WSD in an arbitrary language. Train-O-Matic removes the burden of manual intervention by leveraging the structural semantic information available in the WordNet. Furthermore, we proved that our denser version of WordNet – created by adding connections from BabelNet – is beneficial to Train-O-Matic and led IMS to even higher results across the datasets. What is most interesting, however, is that our automatically-generated training set can lead a supervised model to performance that are higher than those attained when trained on a semi-automatically built dataset, and, sometimes, of those achieved when trained on manually-curated training data. A result that none of the other automatic approaches have managed to achieve. Train-O-Matic, in fact, was shown to provide training data for virtually all the target ambiguous nouns, in marked contrast to alternatives like OMSTI, which covers in many cases around half of the tokens, resorting to the MFS otherwise. When compared to other automatic alternatives, which, differently from OMSTI, were able to provide annotated sentences for most of the tokens, Train-O-Matic proved to provide higher quality data by leading IMS to attain 10 F1 points more than when trained on EuroSense or SenseDefs. Furthermore, the experiments on multilingual all-words WSD, showed that Train-O-Matic can scale well to other languages, for which no manually annotated dataset exists, surpassing the current state of the art of knowledge-based systems and paving the way to supervised all-words WSD on multilingual data. We made available for research purposes all the data generated by Train-O-Matic at `http://trainomatic.org`.

Chapter 5

Inducing Senses' Distribution from Raw Text

Word senses' distributions have been proved to play a fundamental role in Word Sense Disambiguation (Agirre and Martinez, 2004; Postma, Bevia, and Vossen, 2016a; Postma et al., 2016b) either to extract the most frequent sense of a word or to shape training data for supervised WSD models. In this Chapter, we present EnDi and DaD: two language-independent and fully automatic methods for inducing the distribution of words' meanings directly from a raw text (Section 5.2 and 5.3). We show that our learnt distributions are of better quality than those extracted by other automatic methods (Bennett et al., 2016), when evaluated intrinsically (Section 5.5), by means of Kullback-Leibler divergence with a gold distribution, and extrinsically, by means of all-words WSD tasks comprised in the framework of Raganato, Camacho-Collados, and Navigli (2017) (Section 5.6). Finally, we prove that our approaches scale well on different languages and provides a valuable alternative to the MFS baselines across different languages.

5.1 Preprocessing

Before we dive deep in the description of our two approaches for learning the distribution of senses in a given corpus, we introduce a preprocessing step that is shared by both our methods for inducing a sense probability distribution at the sentence level for a given word. It takes as input a lexicon $\mathcal{L}$, a raw corpus

of sentences $\mathcal{C}$ and a semantic network $\mathcal{G} = (V, E)$. We assume a WordNet-like structure for $\mathcal{G}$ (Fellbaum, 1998b), i.e., the vertices in V are synsets that contain different lexicalizations, possibly in multiple languages, of the same concept. Sentence-level sense distribution learning is performed in two steps:

- **Semantic vector computation:** in this step we compute a vector for each synset in the semantic network. Its components are all the synsets in the graph and their values can be interpreted as a measure of relatedness between the starting synset and the corresponding component.
- **Sentence-level word sense distribution:** in this step, for each sentence in $\mathcal{C}$ and for each word $w \in \mathcal{L}$ we compute a probability distribution over its senses by exploiting the lexical vectors computed in the previous step.

The two steps are the same of those introduced in Section 4.1.1 and 4.1.2 and we will briefly review them in what follows for completeness and the sake of clarity.

5.1.1 Semantic Vector Computation

The first step aims at computing a semantic vector for each synset, i.e., node, in the semantic graph that has as components all the others nodes in the graph. Similarly to what explained in Section 4.1.1, we compute the probability value by applying Personalized PageRank (PPR) Haveliwala (2002), a variant of PageRank Brin and Page (1998) in which the uniform restart probability is changed to a custom probability. Also in this case, we concentrate all the restart probability mass onto the synset for which the vector is calculated, so as to increase the probability of reaching nodes in the surroundings of the synset of interest.

Running PPR with restart on a given synset s produces a semantic vector which represents the probability distribution over the synsets in the network (including s itself) of being reachable from, and thus related to, s.

5.1.2 Shallow Sense Distribution Learning

In the second step, each sentence in $\mathcal{C}$ in processed separately by considering all its content words[1] and building, for each of them, a probability distribution over

[1] We filter out non-content words and stopwords.

their senses. Thus, given a word $w \in \mathcal{L}$ contained in a sentence $\sigma \in \mathcal{C}$, we want to score each of the meanings of w (available in the WordNet-like semantic network) with the probability of seeing that sense in the given sentence. Such probability is computed with the following formula:

$$P(s|\sigma, w) = \frac{P(\sigma|s, w)P(s|w)}{P(\sigma|w)} \tag{5.1}$$

$$= \frac{P(w_1, \ldots, w_n|s, w)P(s|w)}{P(w_1, \ldots, w_n|w)} \tag{5.2}$$

$$\propto P(w_1, \ldots, w_n|s, w)P(s|w) \tag{5.3}$$

$$\approx \prod_{w' \in \sigma} P(w'|s, w) \tag{5.4}$$

$$= \prod_{w' \in \sigma} find(\overrightarrow{PPR}_s, w') \tag{5.5}$$

where s is a sense of w, which, in its turn, is contained in σ. We approximate the probability in Equation 5.3 by making the independence assumption between words in the sentence and calculate the probability in Equation 5.4 with the function *find* in Equation 5.5 which returns the highest-probability synset in the first-argument vector v which has the word w as one of its lexicalizations:

$$find(\overrightarrow{v}, w) = \max_{s \in C_w^v} \overrightarrow{v}(s) \tag{5.6}$$

where C_w^v is the set of all the components of $\overrightarrow{v}$ that have w among their lexicalizations.

Once we have applied this procedure to all the sentences in the corpus $\mathcal{C}$ for each given word $w \in \mathcal{L}$, we obtain a sense probability distribution for all the sentences w occurs in. For example, given a sentence σ in the corpus (e.g. *The coordinate plane is a two-dimension surface*), a lexicon $\mathcal{L} = \{plane, surface\}$ and $\mathcal{G}$ = WordNet, the above procedure outputs two distributions (for *plane* and *surface*) as shown in Figure 5.1. Such distributions are used in our two methods, described below, to compute a unified distribution of senses for each word in the lexicon. Hence, in what follows, we assume that the input corpus comes with the senses' distributions annotated by this preprocessing step.

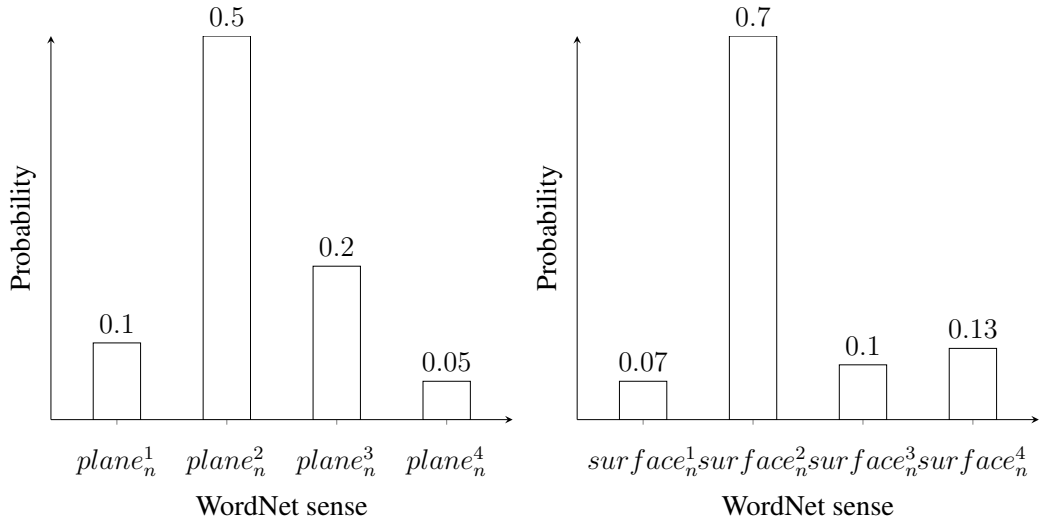

Figure 5.1. Probability of the senses of *plane* (left) and *surface* (right).

5.2 Entropy-Based Distribution Learning (EnDi)

We now introduce the first method for calculating a sense distribution for a given word. This method takes as input the lexicon $\mathcal{L}$, the set of sense distributions $\Gamma_w = \{\gamma_w^\sigma : w \in \sigma\}$ for each word $w \in \mathcal{L}$, which have been computed in the previous step, and a threshold θ.

In order to build a single sense distribution $\mathcal{D}_w$ for each word w in $\mathcal{L}$, we first select the set of sense distributions for all its sentences which have low entropy as follows:

$$\hat{\Gamma}_w = \{\gamma_w^\sigma \in \Gamma_w : \mathcal{H}(\gamma_w^\sigma) \leq \theta\} \tag{5.7}$$

where γ_w^σ is the distribution over w's senses in the sentence σ and $\mathcal{H}(\gamma)$ is the entropy of the input distribution γ:

$$\mathcal{H}(\gamma) = -\sum_{s \in \gamma} \gamma(s) log_2(\gamma(s))$$

As a result $\hat{\Gamma}_w$ contains only skewed sense distributions computed from sentences for which the sense bias is stronger and, therefore, the final decision is clearer. Finally the unified probability mass function $\mathcal{D}_w$ for a word w is computed so as to have, for each sense s of w, the following value:

$$\mathcal{D}_w(s) = \frac{1}{|\hat{\Gamma}_w|} \sum_{\gamma_w^\sigma \in \hat{\Gamma}_w} \gamma_w^\sigma(s) \tag{5.8}$$

Sentence	plane_n^1 (aircraft)	plane_n^2 (geometry)	plane_n^5 (carpentry)
Two people on the **plane** died.	0.92	0.01	0.07
The flight was delayed due to trouble with the **plane**.	0.82	0.07	0.11
Only one **plane** landed successfully.	0.73	0.10	0.17
The cabinetmaker used a **plane** for the finish work.	0.20	0.18	0.62
A catalog of special **plane** curves.	0.10	0.85	0.05
$\mathcal{D}_{plane}$	0.55	0.24	0.21

Table 5.1. A sense distribution computation example for the word *plane*.

For example, let's consider the word *plane* and 5 sentences that contain it (see Table 5.1, left column): we compute its sense distribution by summing the probability of each sense across the sentences and then renormalizing the results by their sum (last row of the Table).

5.3 Domain-Aware Distribution Learning (DaD)

The second method for sense distribution learning, again, takes as input the lexicon $\mathcal{L}$, the set of sense distributions $\Gamma_w = \{\gamma_w^\sigma : w \in \sigma\}$ for each word $w \in \mathcal{L}$, and a semantic network $\mathcal{G} = (V, E)$. The idea, for this second approach, is to exploit the associations between synsets in V and domains from a fixed set $\mathcal{D}$.[2]

Note that each synset might be associated with zero, one or more domains, and that these associations come from an off-the-shelf resource, such as BabelNet domains (Camacho-Collados and Navigli, 2017).

We learn the sense distribution in two steps:

1. **Domain distribution:** we compute the distribution of the domains in the shallow disambiguated corpus.

2. **Sense distribution computation:** we augment the semantic network with domain nodes and connect them to the synsets in the semantic network they are associated with. We then run Personalized PageRank on the augmented semantic graph and obtain a sense distribution over all the synsets in the graph.

[2] The list of domains can be found at:
`http://babelnet.org/javadoc/it/uniroma1/lcl/babelnet/data/BabelDomain.html`

5.3.1 Domain distribution.

Given the set of sense distributions for all the words in $\mathcal{L}$ and all the sentences in the input corpus, we calculate the following probability for each domain $d \in \mathcal{D}$:

$$C(d) = \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} \sum_{\substack{s \in \gamma: \\ d \in domains(s)}} \gamma(s) \tag{5.9}$$

where $domains(s)$ is the set of domains s is associated with. Thus, each sense in a distribution γ contributes to the probability of each domain it belongs to proportionally to its probability in γ. The average of these contributions provides the final probability of domain d in the input corpus.

The hunch behind this step is that, if a corpus is domain-biased, then synsets belonging to that domain should appear more often and with higher probabilities, therefore contributing to increasing the corresponding domain probability. A second hunch is that, even though the shallow distributions inherently come with some unavoidable noise, sometimes due to fine-grained sense distinctions, abstracting synsets with domains enables a coarser, hopefully more accurate and wider coverage, level.

5.3.2 Sense Distribution Computation.

Now that we have a distribution over domains, we add a node for each domain to the original semantic network. We then connect with a direct edge each domain node to all the synsets it is associated with. We finally compute a probability for each synset by applying Personalized PageRank and setting the restart probability vector with the domain distribution that we computed with Formula 5.9. Therefore, given the PageRank formula as defined by Formula 4.1, we set $v^{(0)}$ to 0 except for those components (i.e. nodes) corresponding to each domain d, which, instead are set to the corresponding domain probability $C(d)$. Thus, using the analogy of the random walker, it means that every time the walker decides to restart its walk, it will move to a new domain node according to its probability.

As a result of the PageRank computation we have a distribution over all the semantic network's nodes, i.e., synsets. Note that at this stage the distribution is not specific to a given word, but is general for all synsets in $\mathcal{G}$. Thus the sense distribution of a word can be retrieved by considering the probabilities of all the synsets of

that word (i.e., its senses) in the resulting PageRank vector and normalizing them so as to obtain a sense distribution of the word's senses.

5.4 Experimental Setup

We carried out both intrinsic and extrinsic evaluations in order to have a measure of how well the two methods perform both theoretically and in practice. Both methods for sense distribution learning have some parameters, namely: the semantic network, the corpus and, for the first method, the entropy threshold θ.

Semantic Network We started from BabelNet, which is currently the largest multilingual semantic network, with around 14 million synsets covering hundreds of languages and dozens of domains (Navigli and Ponzetto, 2012). BabelNet is a superset of WordNet, Wikipedia, Wiktionary and other resources and therefore is richer in terms of lexicalizations and semantic relations than any of the resources it integrates. However, because – similarly to alternatives in the literature – we focused only on common nouns, we followed Pasini and Navigli (2017) and chose the WordNet-induced subgraph of BabelNet as the underlying network for the semantic vector computation. In other words, the graph contained only BabelNet synset nodes that comprise at least one WordNet sense but with the considerably larger set of relation edges and multilingual lexicalizations coming from BabelNet.

Corpus We chose Wikipedia as our input corpus because it is available in hundreds of languages and it covers all domains of human knowledge. We used the October 2016 dump of Wikipedia.

Entropy threshold For EnDi we tried different values of the threshold θ, ranging from 0.1 to 4.0 with step 0.1, and tested the results on an in-house development set of 25 lemmas for which we computed the sense distribution in Wikipedia and then selected the value of θ based on the best results in terms of similarity to the SemCor distribution (as explained below). We thus set θ to 1.0.

Comparison sense distributions and gold standard We compared the two sense distribution learning methods against several alternative methods for deriving sense

distributions for words, namely:

- **EuroSense** Bovi et al. (2017): a sense-annotated resource based on the multilingual joint disambiguation of the Europarl corpus Koehn (2005c). For a given word, its sense distribution is obtained by computing the normalized frequency of its senses in the corpus.

- **LexSemTM** Bennett et al. (2016): an approach which builds on top of Lau et al. (2014) and exploits sense glosses and usage examples of the target lemma to build a topic model and then a distribution of the target word's senses.

- **WordNet and BabelNet degree:** we created sense distributions based on the normalized out-degrees of the various senses of the target word in two different semantic networks: WordNet and BabelNet. Given one of the two graphs, we calculated the distribution as follows:

$$\mathcal{D}_w(s) = \frac{\text{out-deg}(s)}{\sum_{s' \in senses(w)} \text{out-deg}(s')}$$

- **SemCor gold-standard sense distribution:** we also compared against the sense distribution computed based on sense frequencies in SemCor Miller et al. (1993b). Notice that this is a gold-standard distribution, as it is the only distribution obtained from manually-annotated data.

5.5 Intrinsic Evaluation

5.5.1 Evaluation measures

For the intrinsic evaluation we evaluated the similarity between the two distributions learned with our entropy and domain-based methods and all other comparison sense distributions introduced above. We used two different measures for performing the comparison, i.e., Jensen-Shannon divergence and Weighted Overlap. Both measures were computed separately for each pair of distributions and then averaged by the total number of words.

Jensen-Shannon divergence (JSD) This measure is based on the Kullback-Leibler divergence and equals 0 when the two distributions are identical, and is greater than 0 when they are different in some way. It is computed as follows:

$$JSD(\gamma, \gamma') = \frac{D(\gamma, M)}{2} + \frac{D(\gamma', M)}{2} \tag{5.10}$$

where $M = \frac{\gamma+\gamma'}{2}$ and D is the Kullback-Leibler divergence which is given by the following formula:

$$D(\gamma, \gamma') = \sum_s \gamma(s) log\left(\frac{\gamma(s)}{\gamma'(s)}\right) \tag{5.11}$$

where, in our case, s are synsets in our sense distributions.

Weighted Overlap This measure (Pilehvar, Jurgens, and Navigli, 2013, WO) determines how similar are the sense rankings of the two distributions. It is 1 when the two distributions have the same ranking of the components and lower than 1 when they are different. It is defined as follows:

$$WO(\gamma, \gamma') = \sum_{i=1}^{|O|} \frac{(r_i + r'_i)^{-1}}{(2i)^{-1}} \tag{5.12}$$

where O is the intersection of the components of γ and γ' and r_i and r'_i are the ranks of the *i-th* component in the respective distribution γ and γ'. The rank of a component (i.e., sense) of the distribution vector is the position at which the component can be found in the distribution vector when sorted in descending order. The Weighted Overlap is thus a measure that does not consider the value of the components in the distributions, but only their ranking.

These two measures provide different insights about how the sense frequencies of a given word are distributed both numerically and when we only consider the components position when the distributions are sorted by value.

5.5.2 Results

Similarity to SemCor distributions The first experiment we performed aimed at investigating the similarity between the various automatically learned distributions (both with our two methods and the comparison distributions) and the gold-standard SemCor distribution. In Table 5.2 we show the JSD and WO (note that for JSD the lower the better, while for WO the higher the better) averaged among all the words

Method	JSD_{gold}	WO_{gold}	JSD_{sys}	WO_{sys}
EnDi	0.29	0.70	**0.06**	0.89
DaD	0.17	**0.91**	0.12	**0.92**
LexSemTM	0.29	0.67	0.07	0.89
EuroSense	0.60	0.39	0.24	0.75
BabelNet Degree	0.09	0.87	0.09	0.87
WordNet Degree	**0.07**	0.88	0.07	0.88

Table 5.2. Similarity with SemCor in terms of Jensen-Shannon divergence and Weighted Overlap (*gold* evaluates against all words in SemCor; *sys* evaluates only against the words for which each method can provide a sense distribution).

Method	Missing Lemmas
EnDi	2655
DaD	23
LexSemTM	2783
EuroSense	5378
BabelNet Degree	23
WordNet Degree	23

Table 5.3. Lemmas for which a method was not able to build a distribution.

in the test set. Both measures were computed, first, by considering all the lemmas in the test set and assigning 1 and 0 to JSD and WO, respectively, when the method was not able to build a distribution for a given lemma (second and third column), and then considering only the lemmas for which the method was able to build the distribution (fourth and fifth columns of the table). As can be seen both our methods built distributions that are generally most similar to SemCor, in terms of both JSD and WO, than the state-of-the-art LexSemTM and that are either better or on a par with alternative approaches. More in detail, DaD performs best in the gold setting, showing wide coverage of words, but a bit worse in numerical terms according to JSD_{sys}. In contrast, EnDi performs best in terms of JSD_{sys}, due to its ability to prune out noisy sentences, slightly worse in the ranking evaluation and on a par with LexSemTM across the board. Degrees fare well, especially on JSD, but, as we will see, their extrinsic evaluation results turn out to be considerably lower.

Method	JSD	WO
EnDi	**0.099**†	**0.937**
DaD	0.204	0.902
LexSemTM	0.116†	0.932
EuroSense	0.344	0.713
BabelNet Degree	0.224	0.832
WordNet Degree	0.166	0.858
SemCor	0.255	0.837

Table 5.4. Similarity with the gold standard from Bennett et al. (2016) in terms of JSD and Weighted Overlap. Values tagged with † are statistical significant with each other for $p < 0.1$.

We show lemma coverage in Table 5.3: DaD and the degree-based distributions have the highest coverage of words, which is WordNet's, while EnDi and LexSemTM – due to filtering mechanisms – and EuroSense – due to lack of sense annotations – have much lower word coverage.

Similarity to Bennett et al.'s (2016) distributions So far we have shown that our methods produced high-quality sense distributions when compared against SemCor. While this is a good result, we should consider that SemCor dates back to almost 30 years ago and since then sense distributions have surely changed over time for a number of ambiguous words (e.g. *troll*, *tweet*, etc.). To work on more recent data, we performed a second intrinsic evaluation using a gold standard dataset proposed by Bennett et al. (2016), which provides distributions manually annotated for 50 lemmas. In this experiment we also evaluated the SemCor-derived distribution against the 50-lemma gold standard. In Table 5.4 we report the results in terms of JSD and WO on this dataset[3]: our methods have lower JSD values than SemCor distribution. Another interesting result is that both WordNet and BabelNet degree baselines also beat SemCor by 0.09 and 0.03 points, while EuroSense achieved the worst results. LexSemTM, instead, scored pretty well according to both measures, achieving 0.116 on JSD and 0.932 on WO. DaD on the other hand scored better than SemCor but worse than LexSemTM on JSD and slightly worse on WO; in

[3]We note that, here, all the systems were able to generate a distribution for each lemma.

contrast, EnDi turned out to be the best method according to the JSD measure and was equivalent to LexSemTM on WO, achieving the state of the art on this dataset. Note also that JSD values are statistical significant for $p < 0.1$.

5.6 Extrinsic Evaluation

We now move to the extrinsic evaluation, which was performed in the context of all-words Word Sense Disambiguation. It is well known in the literature that always outputting the most frequent sense for each ambiguous word in context – the so-called Most Frequent Sense (MFS) baseline – is a hard-to-beat disambiguation strategy (Navigli, 2009). The MFS for English is usually calculated based on frequencies as reported in WordNet, which exploit those in the SemCor corpus. Therefore, we can evaluate each sense distribution method by i) for each word, identifying the predominant (i.e., highest-probability) sense according to the returned sense distribution, and ii) always outputting that sense every time in a WSD dataset we are required to disambiguate the given word. By applying this procedure, we compared the results of the various approaches against the WordNet MFS and BabelNet and WordNet degree. As test sets, we used the benchmark from Raganato, Camacho-Collados, and Navigli (2017) which is the union of all the past Senseval and SemEval for all-words WSD, namely: Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Navigli, Litkowski, and Hargraves, 2007), SemEval-2013 (Moro and Navigli, 2015) and SemEval-2015 (Moro and Navigli, 2015).

As shown in Table 5.5, both EnDi and DaD beat LexSemTM by several points. Moreover, especially DaD, shows performance close to the WordNet MFS baseline with a gap of 6.7 points. Both BabelNet and WordNet degree strategies, instead, scored lower or equal to LexSemTM. This result corroborates the high performance of DaD in terms of WO that we have shown in the previous Section. Moreover we think that this is a very significant outcome since a fully automatic system was able to learn sense distributions that perform very close to a hard-to-beat baseline obtained from a manually sense-annotated dataset.

This is again a good result, but, as mentioned above, annotated data exists for English from which usable sense distributions can be derived (however, this data is

Method	Precision	Recall	F1
EnDi	54.3	50.1	52.1
DaD	61.0	61.0	61.0
LexSemTM	51.4	47.9	49.6
BabelNet degree	51.3	38.4	43.9
WordNet degree	55.1	44.3	49.1
WordNet MFS	67.7	67.7	**67.7**

Table 5.5. Most Frequent Sense performance averaged on all Senseval/SemEval test sets from Raganato et al. (2017).

outdated and limited in size, while our distributions can be updated over time and cover much of the lexicon). To further show the effectiveness of our methods, we performed experiments in other languages, for which manually annotated data is not available on a reasonable scale. Our methods are indeed language-independent and, thanks to BabelNet lexicalizations, we can apply them to arbitrary languages. We therefore calculated sense distributions from the Italian and Spanish Wikipedia (dumps from October 2016) and compared their performance on the SemEval-2015 all-words multilingual Word Sense Disambiguation task. We set the threshold θ for Italian and Spanish to 1.0 and 0.001 experimentally, equally to how we did for English. The difference is due to the fact that the Spanish part of BabelNet contains more ambiguous data.

Results for Italian are shown in Table 5.6. We compared our two methods against the BabelNet Degree and the BabelNet First Sense (BFS) baseline, a dictionary-based baseline which was used as baseline in the task (due to the lack of a manually annotated dataset from which an MFS could be estimated in Italian). The results show that our method performs better than the current best baseline. In particular, DaD outperforms the BFS by 9 F1 points. The gap is even bigger when looking at precision, where our two methods gain from 6 to 13 points, while recall is increased by 6 points with DaD. BabelNet degree also performed better than the BFS but is anyway less effectively than both our systems. A similar trend is observed for Spanish (Table 5.6), with DaD attaining a 5% F1 improvement over the Spanish BFS.

Language	Method	Precision	Recall	F1
Italian	EnDi	60.3	50.0	54.6
	DaD	66.7	56.1	**60.9**
	BabelNet Degree	57.3	52.4	54.7
	BFS	54.3	50.2	52.2
Spanish	EnDi	57.7	48.1	52.4
	DaD	63.6	54.1	**58.4**
	BabelNet Degree	54.7	52.4	53.5
	BFS	54.5	50.6	52.5

Table 5.6. Comparison of Most Frequent Sense performance for Italian and Spenish on the SemEval-2015 WSD task.

5.6.1 Domain-Specific Evaluation

We next investigated how well our methods were able to produce skewed distributions of senses in specific domains.

Learning domain-specific sense distributions To bias sense distributions towards specific domains, we exploited the 34 domains available in BabelNet (Camacho-Collados and Navigli, 2017). For each domain d, we collected all the synsets in BabelNet tagged with that domain and which contain a Wikipedia page. We then used all the sentences from the retrieved pages to build a corpus for domain d. On each corpus, we then applied EnDi and DaD to obtain sense distributions that were biased towards the domain of interest.

Evaluation We tested our domain-biased sense distributions against domain-specific documents from the same SemEval-2013 and SemEval-2015 test sets used in the above extrinsic experiment, as tagged by the task organizers. In contrast to the above experiments, results are therefore reported individually for each domain where EnDi and DaD used the corresponding distributions learned for that domain[4]. Results are shown in Tables 5.7. EnDi and DaD outperforms LexSemTM on all domains across the two datasets. While the latter performances are always lower

[4]We note that, here, the only additional information provided as input to the methods was the domain label.

Dataset	Domain	EnDi	DaD	LexSemTM	WN MFS
SemEval-2013	Biology	70.9	**79.0**	55.6	61.4
	Climate	52.7	**63.0**	46.6	59.2
	Finance	60.5	**63.9**	49.2	51.9
	Medicine	46.1	**64.3**	41.7	49.6
	Politics	62.1	**66.7**	51.3	63.9
	Social Issues	63.4	**67.5**	52.3	58.2
	Sport	**57.2**	53.6	34.8	56.0
SemEval-2015	Math & Computer	63.1	**66.3**	47.3	46.8
	Biomedicine	63.3	62.6	62.6	**68.0**

Table 5.7. Domain evaluation on SemEval-2013 and SemEval-2015 WSD.

than the WordNet MFS baseline, EnDi is instead able to surpass the baseline on 5 out of 7 domains on SemEval-2013 and on one of the two domains of SemEval-2015. As regards DaD, not only does it consistently beat LexSemTM, achieving up to 23 points higher in F1 on the Biology domain, but it also beats the WordNet MFS on every domain but one of SemEval-2013 and one of the two domains of SemEval-2015, performing on average 8 F1 points higher.

5.7 Conclusions

In this Chapter we introduced the problem of automatically inducing the distribution of senses within a corpus of raw texts and presented EnDi and DaD, two knowledge-based, language-independent methods for tackling such problem. They have been shown to perform well on intrinsic and extrinsic evaluations, outperforming the other baselines. Thanks to effective entropy-based filtering, EnDi outperformed LexSemTM, the previous state of the art in sense distribution learning, in all evaluations for the English language. We also showed that, in contrast to other approaches, both methods scale well to other languages, with DaD surpassing all alternative methods in the two SemEval multilingual disambiguation tasks. Thanks to its domain awareness, not only has DaD proven to generalize well across languages, but also to surpass the F1 performance of the hard-to-beat WordNet MFS on 7 out of 9 domains from two SemEval tasks. LexSemTM is, instead, surpassed by

both methods across all domains. All the data is available for research purpose at `http://trainomatic.org`.

Chapter 6

Train-O-Matic++: Leveraging Sense Distributions

6.1 Coupling Train-O-Matic with EnDi and DaD

In the previous Chapters we introduced Train-O-Matic to automatically building semantically-annotated training data and EnDi and DaD to learning word senses' distributions given a corpus of sentences. Train-O-Matic exploited WordNet sense ordering in order to assign a certain number of sentences (k_i) to a given sense s_i (see Formula 4.12) that we report here for clarity:

$$k_i = \frac{1}{i^z} * K$$

In this Chapter, we aim at coupling Train-O-Matic with EnDi and DaD and present Train-O-Matic++, which, not only can generate training data in any language supported by BabelNet, but it can also adapt the produced datasets to the domain of application or, more in general, to the distribution of senses that we can induce directly from a set of input documents of raw sentences. Therefore, given an input text that we are interested in disambiguating, we first run EnDi or DaD in order to build a distribution for senses of all the target words of the input text. Then, we plug the sense ordering induced by the learnt distributions into Formula 4.12 in order to shape the training data on the sense distribution of the input documents. We therefore dispose of WordNet sense ordering – which depends on the senses' frequencies in SemCor – and remove also the last bit of manual effort that was needed by Train-O-Matic. Indeed, thanks to EnDi and DaD, Train-O-Matic++ can

now induce its own ordering of senses based only on the input text or collection of documents that one is interested in disambiguating. This, for example, may be a sample set of texts that is representative of a target domain in which the final supervised WSD model will be applied. In what follows, we first show the performance attained by IMS trained on Train-O-Matic++, and put them in comparison with those attained by the vanilla Train-O-Matic and when is is coupled with other types of distributions (Section 6.2). Second, we prove that, a good estimation of the target sense distribution is effective only when accompanied by high-quality sense-tagged data. (Section 6.3). Third we show that EnDi and DaD outperforms the manually-curated sense distribution and the MFS strategy provided by WordNet on the domain-oriented evaluation on SemEval-2013 and SemEval-2015 all-words WSD tasks (Section 6.4). Finally, we prove that EnDi and DaD, when used as backoff strategy in the multilingual settings, provide a boost of several points across all the tested datasets (Section 6.5), which crown them as the best backoff strategies when it comes to multilingual all-words WSD.

6.2 Comparing Senses' Distributions

We now study the impact of different training-set-shaping strategies on the final WSD model. Specifically, we generate different datasets by means of Train-O-Matic++ when coupled with different sense distributions that can be either automatically or manually induced.

6.2.1 Experimental Setup

Reference System We use IMS (Zhong and Ng, 2010) as WSD system across all the experiments to measure the impact of each training set on a supervised model.

Comparing Distributions Across the section we put in comparison the following distributions:

1. **Uniform:** we assign the same number of training samples to each meaning. In the case Train-O-Matic could not find enough sentences for a given sense, then we associate it with all the samples that Train-O-Matic was able to find.

2. **WordNet:** we rank the senses of each word according to WordNet ordering and then apply the Zipf's law (see Formula 4.12) to compute the number of sampled sentences for each synset according to its rank.

3. **Automatic:** we rank the word senses of each word according to either DaD or EnDi (see Section 5) and then apply Formula 4.12 as described before.

4. **Random:** we assign a random number of examples to each synset of a word. When the random number is higher than the available examples, we assign the maximum number of examples available.

5. **Oracle:** we rank the word senses according to the sense frequency in the test set. Later, we assign a number of examples to each sense following a Zipfian distribution as in Formula 4.12.

6. **Oracle 1-2:** acts as Oracle but exchanges the first and the second sense in the synset ranking. For example, being $[s_3, s_1, s_2]$ the ordering of the senses for the word w according to the Oracle distribution, then, Oracle 1-2 would provide the following distribution for w: $[s_1, s_3, s_2]$, i.e., where s_1 and s_3 have been swapped.

7. **Oracle 2-3:** acts as Oracle 1-2 but exchanges the second and the third sense in the synset ranking. Therefore, given the Oracle distribution $[s_3, s_1, s_2]$, Oracle 2-3 would output the distribution $[s_3, s_2, s_1]$.

The Oracle variants are designed to prove the importance of identifying the predominant sense of a word, and especially, how much mistaking the first sense for the second one would affect the performance (Oracle 1-2) with respect to mistaking the second and the third ones (Oracle 2-3).

We measure the impact of each distribution to the final reference model (i.e., IMS) by building 7 distinct datasets with Train-O-Matic, each following one of the aforementioned distributions. Then we use a all-words WSD task to measure the performance of IMS trained on each dataset.

Test set We use the ALL dataset comprised in the Raganato, Delli Bovi, and Navigli (2016) evaluation framework which is the concatenation of all the past Senseval and SemEval datasets.

Measure	WordNet	EnDi	DaD	Uniform-100	Random	Oracle	Oracle 1-2	Oracle 2-3
Precision	65.1	48.0	60.3	47.1	45.3	81.3	47.6	79.8
Recall	59.7	43.5	54.3	42.7	41.0	73.5	43.1	72.2
F1	62.3	45.6	57.3	44.6	43.0	77.2	45.3	75.8

Table 6.1. Evaluation on the ALL dataset of the same WSD system trained on training sets balanced in different ways.

6.2.2 Results

In Table 6.1 we report the results of IMS trained on different training data, each following a different sense distribution that is induced by one of the aforementioned approaches. We note that, the results are computed without the backoff strategy, therefore, there may be words for which the classifier does not output any answer and thus precision and recall and F1 may be different. The *Random* method, that assigns a random number of examples to each synset, is the one that performed the worst with 43 points in F1. Interestingly we note that the *Uniform* strategy (which assigns the same number of examples to each sense) lead IMS to achieve only 1.6 points more then the *Random* approach, hence proving that, uniform balancing is not a viable options when dealing with senses. This is because the classifier learns an unnatural distribution that will never be encountered in real-world scenarios (which are usually represented in the test sets). Moreover, with a perfectly balanced training set, the classifier tries to model each sense equally, while, given the Zipfian nature of meanings' distribution, it would be more optimal to learn a good representation for the most frequent senses and a less accurate one for the others. Another interesting result is the one achieved with the Oracle strategy. As one can see, in fact, the vanilla Oracle leads IMS to 77.2 F1 points on the ALL dataset which is almost 10 points higher than the current state-of-the-art WSD model (Luo et al., 2018) on the same dataset. This raise an important question about the generalisation capabilities of the WSD models presented so far in the literature. In fact, it seems like, IMS, and all the models presented so far, are not taking into account some important aspect of a sentence semantic which, instead, may allow them to better shape a word senses' distribution based on the given context. This is further corroborated by the results attained when the Oracle 1-2 and Oracle 2-3 strategies are followed to shape the training data. With Oracle 1-2, in fact, IMS performs 45.3 F1 points, only 1 point

higher than the Uniform strategy, while, in strong contrast, Oracle 2-3 performed more than 30 F1 points than the Uniform strategy and just 2 points less than the plain Oracle. This shows how sensitive the model is to the sense distribution in the training data and, more interestingly, how identifying the most common sense of a word is actually more important than correctly estimating the rest of the distribution. In fact, swapping the senses in the distribution from the second position onwards, does not degrade the performance to the same extent as when the first sense is swapped with the second one.

Together with all these baselines, we also report the performance achieved by IMS when we use the sense distribution induced by WordNet, EnDi and DaD.

As expected the WordNet ordering outpeformed the two automatic baselines as it takes advantage of the manual annotations available in SemCor. While we note that EnDi performed poorly, slightly more than the Oracle 1-2, we stress the fact that DaD, even though fully unsupervised, enables IMS to achieve 57.3 F1 points, 5 less than when WordNet is used.

These outcomes are encouraging and show the impact that a good estimation of the target sense distribution may have on a supervised WSD model, hence proving that, better distribution learning techniques could highly affect the results on the WSD task. Moreover, considering the huge gap between the learnt distributions and the Oracle one, there is still a large room for improvement which should motivate researchers to further contribute to this field.

6.3 Word Sense Distribution vs. Data Quality

We now want to study whether the distribution of senses encodes enough information to steer a supervised WSD model to achieve decent results. Therefore, to give a measure of how good a model could perform by only having information regarding the sense distribution, we evaluate the results of a supervised model, i.e., IMS, trained on data shaped according to the best possible distribution (i.e., the one computed on the test set (Oracle)) and collected at random, i.e., we assign a random sense to each target word in each sentence.

Method	Precision	Recall	F1
Oracle-T-O-M 500	81.3	73.5	77.2
noisy-T-O-M 100	51.7	48.8	**50.2**
noisy-T-O-M 200	51.1	48.4	49.7
noisy-T-O-M 500	49.0	46.4	47.6

Table 6.2. Evaluation of IMS with Oracle distribution when random senses are assigned to each word occurrence.

6.3.1 Experimental Setup

Reference System We use IMS (Zhong and Ng, 2010) as supervised model to measure the effectiveness of each produced dataset.

Comparing Datasets We compare the results of IMS trained on four different training set automatically generated with Train-O-Matic:

- **Oracle T-O-M 500:** It is generated by following the Oracle distribution and with each target word disambiguated according to Train-O-Matic pipeline (see Chapter 4). We set K, i.e., the number of sentences assigned to the most frequent sense of each word, to 500.
- **noisy-T-O-M 100/200/500:** It is generated by disambiguating each target word at random and then sampling the set of sentences for each sense according to ordering given by the Oracle distribution. We produced three variants of this corpus with K set to 100, 200 and 500.

Test Set We test IMS trained on each noisy-Train-O-Matic on the ALL dataset included in Raganato, Delli Bovi, and Navigli (2016).

6.3.2 Results

In Table 6.2 we prove that the quality of the contexts provided by a training set matter and that mimicking the test set distribution of senses is not enough to build a well-performing classifier. As one can see, in fact, when we use Train-O-Matic with the Oracle distribution and $K = 500$ to generate the dataset (Train-O-Matic 500),

IMS achieves 77.2 F1 points. However, when we generate noisy datasets, i.e., by assigning random examples for a target sense – instead of those that are most likely to contain the word with the given sense – the performance drops heavily. Indeed, noisy-Train-O-Matic 100 attains 27 points less than Oracle-T-O-M 500 proving that, having a good sense distribution estimator would not be enough if not coupled with high-quality data. This is also confirmed by the fact that, increasing the number of examples each synset can receive (from 100 to 500), makes the performance decrease. Intuitively this happens because more noise is added to the dataset and thus the classifier gets increasingly confused. To conclude, this further corroborate our claim that Train-O-Matic produces high-quality examples since, if they were too noisy, then even when following the Oracle distribution the training set would have not led IMS to attain results 10 points higher than the state-of-the-art model reported by Raganato, Camacho-Collados, and Navigli (2017).

6.4 Domain-oriented WSD

Now that we proved the importance of sense distribution when coupled with high-quality data, we show the EnDi and DaD distributions impact on the final WSD model when tested on a domain-specific setting. Inter alia, we are interested in showing that, by including specific sense distributions into the training data, we bring two main benefits to both Train-O-Matic and supervised models: i) we make Train-O-Matic more flexible, hence enabling it to build training sets tailored to a specific distribution and ii) we boost the performance of a supervised model which can now be trained on data specific for a target domain (see the Test Sets section of this Chapter for the complete list of domains comprised by the two SemEval).

6.4.1 Experimental Setup

Comparing Systems For each testing domain we compared three different configurations of Train-O-Matic++:

1. **+MFS:** Original Train-O-Matic is used to build the training corpus and the MFS learnt by DaD or EnDi is set as backoff strategy, and therefore the data are sorted according to WordNet ordering.

Domain	Backoff	T-O-M			T-O-M++$_{DaD}$			T-O-M++$_{EnDi}$			SemCor	WN MFS	Size
		P	R	F1	P	R	F1	P	R	F1	F1	F1	
Biology	+MFS	63.0	63.0	63.0	63.7	63.7	63.7	63.7	63.7	63.7	66.3	64.4	135
	+SD	59.0	53.3	56.0	69.8	65.2	67.4	54.8	51.1	52.9			
	+MFS+SD	63.0	63.0	63.0	71.1	71.1	**71.1**	57.0	57.0	57.0			
Climate	+MFS	68.1	68.1	68.1	66.0	66.0	66.0	65.5	65.5	65.5	**70.1**	67.5	194
	+SD	63.4	50.0	55.9	64.5	56.2	60.1	52.7	45.9	49.0			
	+MFS+SD	68.1	68.1	68.1	66.5	66.5	66.5	55.7	55.7	55.7			
Finance	+MFS	68.0	68.0	**68.0**	58.9	58.9	58.9	58.9	58.9	58.9	63.7	56.2	219
	+SD	62.1	51.6	56.4	62.0	58.0	59.9	52.2	48.9	50.5			
	+MFS+SD	68.0	68.0	**68.0**	63.9	63.9	63.9	54.8	54.8	54.8			
Health Care	+MFS	65.2	65.2	65.2	58.0	58.0	58.0	58.7	58.7	58.7	62.7	56.5	138
	+SD	61.3	55.1	58.0	66.4	60.1	63.1	50.4	45.7	47.9			
	+MFS+SD	65.2	65.2	65.2	68.1	68.1	**68.1**	54.3	54.3	54.3			
Politics	+MFS	65.2	65.2	65.2	72.4	72.4	72.4	71.3	71.3	71.3	69.5	67.7	279
	+SD	62.5	54.8	58.4	70.2	65.9	68.0	59.9	55.9	57.7			
	+MFS+SD	65.2	65.2	65.2	72.0	72.0	**72.0**	60.9	60.9	60.9			
Social Issues	+MFS	68.5	68.5	68.5	67.6	67.6	67.6	67.9	67.9	67.9	66.8	67.6	349
	+SD	63.1	53.0	57.6	72.2	63.9	67.8	59.5	52.7	55.9			
	+MFS+SD	68.5	68.5	68.5	73.9	73.9	**73.9**	63.0	63.0	63.0			
Sport	+MFS	60.3	60.3	60.3	55.5	55.5	55.5	55.5	55.5	55.5	**60.4**	57.6	330
	+SD	58.3	54.6	56.4	53.4	50.6	51.9	46.3	43.9	45.1			
	+MFS+SD	60.3	60.3	60.3	53.6	53.6	53.6	47.0	47.0	47.0			

Table 6.3. Performance comparison over SemEval-2013 domain-specific datasets when WordNet EnDi and DaD are used to compute MFS (+MFS) and sense distribution (+SD).

2. **+SD:** We used sense distribution learnt by EnDi and DaD to order the word senses and decide the amount of training examples to assign to each one. We note that in this setting no backoff strategy is used.

3. **+MFS+SD:** We used both: the EnDi and DaD learnt MFS and sense distribution as backoff strategy and sense distribution to shape the training set respectively.

Original Train-O-Matic (T-O-M) follows, by design, the WordNet sense ordering to shape the training data, therefore, the results in its +MFS and +MFS+SD rows are identical. The two versions of Train-O-Matic++ (T-O-M++ $_{EnDi}$ and T-O-M++ $_{DaD}$), instead, use WordNet ordering in the +MFS setting to shape the training data, and either EnDi or DaD in the +MFS+SD version.

Domain	Backoff	T-O-M			T-O-M++$_{DaD}$			T-O-M++$_{EnDi}$			SemCor	MFS	Size
		P	R	F1	P	R	F1	P	R	F1	F1	F1	
Biomedicine	+MFS	76.3	76.3	**76.3**	68.0	68.0	68.0	68.0	68.0	68.0	70.3	71.1	100
	+SD	76.1	72.2	**74.1**	75.0	71.1	73.0	48.9	45.4	47.1			
	+MFS+SD	76.3	76.3	**76.3**	75.3	75.3	75.3	51.5	51.5	51.5			
Maths & Computer	+MFS	50.0	50.0	50.0	51.0	51.0	**51.0**	50.3	50.3	50.3	40.6	40.9	97
	+SD	50.0	47.0	48.5	62.1	59.0	**60.5**	40.9	38.0	39.4			
	+MFS+SD	50.0	50.0	50.0	62.0	62.0	**62.0**	41.4	41.4	41.4			

Table 6.4. Performance comparison over the Biomedical and Maths & Computer domains in SemEval-2015 when DaD and EnDi are used to predict the sense distribution of the test set..

Test Sets We tested on the 13 documents of SemEval-2013 belonging to different domains, namely: biology, climate, finance, health care, politics, social issues and sport, and 2 documents of SemEval-2015 in 2 different domains, i.e., Maths & computers and biomedicine.

6.4.2 Results

The results in Table 6.3 show that EnDi and DaD are more effective when used to shape the dataset (+SD) than as backoff strategy (+MFS). By only shaping the training data with the distribution learnt by DaD, in fact, we see the highest gain in results with respect to the plain Train-O-Matic. IMS trained on Train-O-Matic++$_{DaD}$, in fact, attains from 3.5 to 10.2 F1 points more than Train-O-Matic (+SD rows), proving that DaD captured a sense distribution that better represent the testing documents than the WordNet one. Even higher results are achieved when using both the learnt sense distributions and the learnt most frequent sense (+MFS+SD). In this last configuration, in fact, IMS trained on Train-O-Matic++$_{DaD}$ achieved the highest scores across the board with a boost over vanilla Train-O-Matic between 2.9 and 8.1. Especially, when compared to Train-O-Matic, Train-O-Matic++$_{DaD}$ shows consistent improvements over the *Biology*, *Health Care*, *Politics* and *Social Issues* domains, while performing worse on the 3 remaining domains. We also note that Sport is the only domain where the Train-O-Matic++$_{DaD}$ has a significant decrease in performance. Therefore, we further investigated this case and noticed that the most misclassified word is $game$. This happened because two reasons: i) The domain distribution built by DaD in its first phase is skewed towards *Sport*

and recreation, that received the highest probability (0.19), and *Philosophy and Psichology* with a probability of 0.18. Hence, there is not a clear domain for the document and thus, the synsets close to each of the two domains will receive similar probabilities, ii) the Sport sense of game in BabelNet is not associated with the Sport domain, therefore it did not benefit from the direct connection with the domain during the propagation step of DaD (see Section 5.3) so it does not result as one of the most frequent senses of game in the document even if it should be. We further note that a few mistakes may correspond to big differences in performance inasmuch the domain datasets comprise only 300 instances each on average. Moreover, a few words, such as *game*, are very frequent (23 occurrences in the sport test set), hence several F1 points can be gained or lost by simply classifying correctly or not one or two distinct word types.

Finally, when compared against IMS trained on SemCor, Train-O-Matic coupled with both the learnt MFS and distribution is able to beat it on all domains but 2 (*Climate* and *Sport*) by 0.2 to 7.1 points. A similar behaviour is observed in Table 6.4 where Train-O-Matic +MFS +SD always beats IMS trained on SemCor. When compared to WordNet, instead, DaD lead IMS to attain more than 10 F1 points on the Biomedicine domain, at the very small cost of only 1 F1 point less on the Maths & Computer domain. In summary, our domain-specific evaluation confirmed the intuition that WSD is more effective when domain information is provided. While we note that there might be many different ways to provide such information, we prove that, balancing the training set according to the learned distribution of word senses, which in DaD directly depends on the domain distribution of the document, is an effective way for providing domain information to the classifier and to increase performance by several F1 points without any human intervention, nor explicitly providing domain labels to the classifier.

6.5 Multilingual WSD

We now measure the benefits carried by EnDi and DaD to IMS in the multilingual setting of the all-words WSD task.

Dataset	Language	Method	Best System	Train-O-Matic		
			F1	P	R	F1
SemEval-2013	Italian	Vanilla	65.8	69.6	65.7	**67.6**
		+MFS$_{DaD}$	65.8	71.0	71.0	**71.0**
		+MFS$_{EnDi}$	65.8	70.9	70.9	**70.9**
	Spanish	Vanilla	**71.0**	68.0	65.6	66.8
		+MFS$_{DaD}$	**71.0**	69.0	69.0	69.0
		+MFS$_{EnDi}$	**71.0**	68.6	68.6	68.6
	French	Vanilla	**60.5**	61.1	59.9	60.5
		+MFS$_{DaD}$	60.5	61.4	61.4	**61.4**
		+MFS$_{EnDi}$	60.5	61.1	61.1	**61.1**
	German	Vanilla	62.1	65.9	60.8	**63.2**
		+MFS$_{DaD}$	62.1	67.5	67.5	**67.5**
		+MFS$_{EnDi}$	62.1	67.2	67.2	**67.2**
SemEval-2015	Italian	Vanilla	56.6	65.1	55.6	**59.9**
		+MFS$_{DaD}$	56.6	65.9	65.9	**65.9**
		+MFS$_{EnDi}$	56.6	64.7	64.7	**64.7**
	Spanish	Vanilla	**56.6**	53.3	53.3	53.3
		+MFS$_{DaD}$	56.3	62.6	62.6	**62.6**
		+MFS$_{EnDi}$	56.3	61.0	61.0	**61.0**

Table 6.5. Performance comparison between T-O-M and SemEval-2013's best UMCC-DLSI Run.

6.5.1 Experimental Setup

Reference Model We use IMS (Zhong and Ng, 2010) as reference model to evaluate extrinsically the quality of the generated multilingual training data.

Comparing Systems We compare the best performing system on SemEval-2013 and 2015 with Train-O-Matic in 3 different settings:

- **Vanilla:** The plain Train-O-Matic as introduced in Chapter 4 with the BabelNet Most Frequent Sense as IMS backoff strategy.
- **+MFS$_{DaD}$** Train-O-Matic++ coupled with the backoff strategy computed

according to DaD distribution.

- **+MFS**$_{EnDi}$ Train-O-Matic++ coupled with the backoff strategy computed according to EnDi distribution.

We note that, we do not test the +SD setting since we are not evaluating each domain separately, hence, for the multilingual setting, we only keep the distribution induced by the BabelNet ordering of synsets.

Test Sets We tested IMS trained on different settings of Train-O-Matic on all the available multilingual all-words WSD tasks, i.e., SemEval-2013 task 12 and SemEval-2015 task 13.

6.5.2 Results

As can be seen from Table 6.5 Train-O-Matic enabled IMS to perform better than the best participating system to SemEval-2013 (Manion and Sainudiin, 2014, SUDOKU) and SemEval-2015 (Fernández-López, Gómez-Pérez, and Suárez-Figueroa, 2013, UMCC-DLS) across most of the datasets. In fact, IMS trained on Train-O-Matic++ with either EnDi or DaD consistently achieves higher performance than its competitor with a boost that ranges between 0.6 (on SemEval-2013 French dataset) and 9.3 (on SemEval-2015 Italian dataset) points. Furthermore, we note that EnDi and DaD always improve the performance of Train-O-Matic which uses BabelNet to determine the most frequent sense of each lemma, hence proving to be a valuable alternative across different languages.

6.6 Conclusion

In this Chapter we showed that the sense distribution that we automatically induce from an input corpus of raw text by means of EnDi and DaD can be very valuable for Train-O-Matic, hence enabling it to generate datasets that are tailored to a specific distribution and domain. We first proved that there is still a large margin of improvements over automatic approaches for word sense distribution learning. Indeed, the Oracle distribution extracted from the test set, allow the supervised model to achieve very high performance, almost 10 points higher than the current

state of the art in supervised WSD (Luo et al., 2018). Moreover, we proved that learning the most frequent sense of a word in a corpus is way more important than learning the whole distribution. In fact, when confusing the first and the second most frequent sense of a word, we noticed a drop of more than 30 F1 points in the model performance, which, instead, did not happen when confusing the second and the third senses. We than proved that, by only relying on good estimation of sense distributions and totally ignoring the quality of the annotated data, does not even lead a supervised model to decent results, hence showing that the two properties, i.e., high-quality data and good sense distribution estimator, are complementary to each other. We continued our experiments by comparing Train-O-Matic++ coupled with EnDi and DaD with the original version of Train-O-Matic (i.e., which makes use of WordNet sense ordering for shaping the training data). The results proved that Train-O-Matic++ training sets, which are tailored to the sense distribution automatically extracted from the target documents, lead IMS to consistently attain higher performance across the domains. Moreover, we showed that using the most frequent sense identified by DaD as backoff strategy for IMS is beneficial and further improve the supervised model performance. Finally, we leverage EnDi and DaD to extract the Most Frequent Sense information from the test set documents and used it as backoff strategy of IMS on the all-words multilingual WSD setting. The results proved that they are a better alternative than the BabelNet ordering (i.e., based on the node degree) boosting the classifier performance by several points across all the languages.

Train-O-Matic++ proved to be a valuable alternative to manually-curated datasets (e.g., SemCor) when it comes to disambiguate texts in specific domains. In fact, thanks to EnDi and DaD it is able to tailor the training data on a sense distribution that can be automatically induced from a collection of raw texts. Therefore, it makes it possible to customise the training data according to the expected application domain of a supervised WSD model.

Chapter 7

Data Produced

In this Chapter we summarise the data produced by each of the presented works and that which we believe represent a valuable and tangible contribution of this theses. All the data produced were publicly released to the community for research purposes.

1. **Train-O-Matic dataset.** This is a sense-annotated dataset which has been automatically produced by Train-O-Matic[1]. It comprises $1{,}012{,}021$ annotations for English, Italian and Spanish which cover 2844 distinct lexemes (those used for as testing instances in all the past all-words English WSD Senseval and SemEval tasks and SemEval-2015 multilingual all-words WSD task.). This data were produced in the context of Pasini and Navigli (2017).

2. **Train-O-Matic dataset large.** This is an extension of the Train-O-Matic dataset, which also comprises annotated data for Chinese, German and French and an improved coverage[2]. Indeed, it contains $8{,}053{,}721$ semantic annotations for $94{,}591$ distinct lexemes. In Table 7.1 are reported the statistics for each covered language. This data were produced in the context of Pasini, Elia, and Navigli (2018).

3. **EnDi and DaD annotations.** This dataset comprises the distributions computed by both EnDi and DaD on the SemEval-2013 and SemEval-2015 and each of the domains therein in three languages, i.e., English, Italian and Span-

[1] Available at `http://trainomatic.org`
[2] Available at `http://trainomatic.org`

ish[3]. This data were produced in the context of Pasini and Navigli (2018) and Pasini and Navigli (2019).

Language	Annotations	Unique Lexemes
EN	2,788,763	11,402
FR	1,597,230	25,690
DE	1,213,634	22,300
IT	1,037,253	19,192
ES	935,713	14,596
ZH	481,128	12,897
TOT.	8,053,721	94,591

Table 7.1. Statistics by language for Train-O-Matic enlarged dataset.

[3]Available at `http://trainomatic.org`

Chapter 8

Conclusion

In this thesis we addressed the long standing problem of the *knowledge acquisition bottleneck* in Word Sense Disambiguation. This hampers the development of semantic resources and WSD models and hinders the performance of both, knowledge-based and supervised approaches, on the all-words WSD task. Furthermore, despite the many efforts put in solving this problem, English has always been the main focus of most of the previously proposed approaches (Ng, Wang, and Chan, 2003; Chan and Ng, 2005a; Zhong and Ng, 2009; Taghipour and Ng, 2015). This made it completely impossible to have supervised models for languages other than English, where, indeed, no sense-annotated corpora with decent coverage of a language lexicon were available until the most recent years when Camacho-Collados et al. (2016); Bovi et al. (2017); Pasini, Elia, and Navigli (2018); Pasini and Navigli (2019); Scarlini, Pasini, and Navigli (2019) introduced novel methods for the automatic harvesting of sense-annotated datasets for WSD.

Therefore, as primary effort in this direction we introduced Train-O-Matic, a knowledge-based and language-independent approach for generating hundreds of thousands sentences where a target word is annotated with one of its possible meanings. Train-O-Matic relies only on the BabelNet structure for computing the probability of a sense to appear in a given sentences, and on the sense ordering of WordNet to determine the number of sentences to be associated with each sense (see Chapter 4). Its main novelty with respect other approaches that were previously introduced in the literature, inter alia, (Ng, Wang, and Chan, 2003; Chan and Ng, 2005a; Zhong and Ng, 2009; Taghipour and Ng, 2015; Camacho-Collados et al., 2016; Bovi et al., 2017), is that it does not rely nor on parallel data neither on ready-

made WSD systems. Moreover, it comprises a scoring mechanism to filter out the sentences which are the least reliable and takes advantage of WordNet ordering of sense to mimic the sense distribution one can find in manually-annotated data such as SemCor. This, especially, was never directly took into account by the aforementioned approaches, which, instead, were limited to follow the distributions drawn by the input corpora. Differently from all the parallel-corpus-based approaches, Train-O-Matic, can also scale easily to any language supported by BabelNet and for which a POS-tagger and a lemmatiser are available[1]. Moreover, we proposed a semi-supervised variant of Train-O-Matic to represent, both sentences and senses, in a shared latent space so that we could score a sentence with respect to a sense directly by computing their cosine similarity. This approach showed encouraging results, sometimes better than those achieved by the knowledge-base one.

We then focused our efforts on studying the distribution of word senses within a corpus (Chapter 5) . As was noted also before by McCarthy et al. (2007), it has a Zipfian shape and hence for a given word there are usually a few senses that are very frequent in a corpus and most of the others that are very rare. This makes it harder to find a good balancing of the training corpus since, on the one hand, flattening the distribution and making it uniform would not serve our purpose as the classifier would learn the wrong senses' distribution, on the other hand, maintaining a skewed distribution towards one or two meanings of each target word may bias the model too much and it would not be able to scale over different domains. In fact, changing application domain, often coincides with changing distribution of senses, at least for the word that are most peculiar for that specific domain. Therefore, to mitigate these issues, we proposed EnDi and DaD, two knowledge-based approaches for inducing the senses' distribution given a corpus of raw texts, hence allowing to shape the training data according to it, and to learn the MFS of a word and using it as backoff strategy in a general-purpose model applied to a specific domain. EnDi (Chapter 5.2)only exploits the local distribution of senses computed at sentence level to filter out those that have greater entropy and hence more noisy. Then, by relying only on the most informative ones, it counts the number of times a word meaning has been ranked first in all the distribution associated to each sentence in the given corpus. DaD (Chapter 5.3), instead, aggregates the sense distribution computed at

[1]We note that, nowadays, POS-taggers and lemmatisers are available for a plethora of languages thanks to the Universal Dependencies (Nivre et al., 2016)

the sentence level into the domains corresponding to each concept. Therefore, it propagates the new and coarser distribution of the domains, over the entire WordNet graph thanks to the PageRank algorithm, which assigns a probability to each synset. Finally, given a word, it retrieve its meanings with their associated probabilities and compute the final distribution. We proved, both intrinsically and extrinsically, that our approaches are valuable and achieved excellent results across all the experiments and especially on the domain-specific and multilingual evaluation. In fact, they proved to achieve a higher similarity, in terms of Kullback-Leibler divergence, with manually-annotated distributions and to be a good estimator of the most frequent sense of a word in a given corpus of raw texts.

Lastly, we studied the impact of shaping the training data according to a specific senses' distribution on a supervised model (Chapter 6). To this end, we coupled Train-O-Matic with EnDi and DaD (hence building Train-O-Matic++) so that it can decide, according to either EnDi or DaD distributions, the number of sentences to assign to each word meaning. This allows Train-O-Matic++ to build training data that are cut on a specific distribution, which, eventually can be induced either on the target text that one is interested in disambiguating, either from a set of documents which represent the specific domain of application where the final WSD system will be applied. This added a huge flexibility to Train-O-Matic, which is now able to customise the training data on the need of its user. Moreover, a supervised model trained on Train-O-Matic++ showed higher performance than when trained on the vanilla Train-O-Matic on both domain-specific and multilingual settings.

8.1 Future Work and Perspectives

Train-O-Matic, together with EnDi and DaD, lay the foundation for building high-quality sense-annotated corpora in potentially any language. The key contributions brought by this thesis represent, in fact, a step forward to the resolution of the problems they address, namely, *the knowledge acquisition bottleneck* and the *word sense distribution learning*, and open up new paths in the Word Sense Disambiguation field. Train-O-Matic, indeed, can provide semantically-annotated resources for languages other than English that were not available before and hence shedding the light on scenarios for multilingual Word Sense Disambiguation that were before

neglected from the supervised systems point of view. A short-term evolution of this work is, therefore, to measure the performance of more complex models trained on our datasets and, eventually, develop new approaches for getting rid of the possible noise that is encompassed by the automatically-generated datasets. Train-O-Matic method, moreover, can be put at the disposal of human annotators and, by coupling it with an active learning mechanism, it could speed up by a large margin the process of annotating a corpus of raw texts. Especially, it would be a valuable approach to retrieving and ranking the sentences for a given sense which could then be kept or discarded by simply specifying whether they were correctly tagged or not.

EnDi and DaD contributed to revalue and bring again to the attention of the community the importance of the word sense distributions and showed the high impact that they have on the newly opened scenario of supervised multilingual Word Sense Disambiguation. Indeed, most of the WSD approaches neglected the information provided by the sense distributions. Therefore, a short-term extension of EnDi and DaD is to develop new supervised and unsupervised approaches to WSD for taking directly into account the expected distribution of meanings instead of relying only on the implicit bias given by the training data.

Mid-term Perspectives Some mid-term directions that are worth to mention and that are inspired by the work presented in this thesis are the followings:

1. **Multilingual Semi-supervised Train-O-Matic.** Semi-supervised Train-O-Matic showed encouraging results, however, due to its need of annotated data for building the sense representation it was limited to the English language only. Therefore, a natural extension would be to study more complex model to build senses and sentences embeddings that lay in the same latent space that is shared across languages. One starting point for this direction could be the MUSE vectors (Conneau et al., 2017) which provide word embeddings for multiple languages in the same space. Moreover, instead of simply averaging the word embeddings of the words in a sentence, one could think of more complex aggregating functions which could be learnt by means of neural networks. Thus, finding a model for building multilingual latent representations of words and sentences that lay within the same latent space, would be highly beneficial for developing a new and more powerful version of Train-O-Matic.

2. **Neural Sense Distribution Induction.** The senses' distributions proved to be a key point for effective WSD, being able to boost, by several points, the performance of a supervised model. Therefore, an interesting direction would be to encode the distribution information in a latent space, so that it could be directly used within another WSD neural model. A starting point for this line could be the variational eutoencoder architecture which is already capable of encoding the gaussian distribution of the data by means of two latent vectors which represent μ and σ. This could lead to neural models that can easier generalise to new domains while at same time reducing the number of annotated data that a general purpose WSD model may need. Moreover, similar to what we showed with EnDi and DaD, it would allow to customise the WSD model by simply changing the distribution encoder without the need of retraining the whole model.

3. **Language-agnostic Document-level Embeddings** Since the domain information is nowadays discarded by supervised WSD, a promising and interesting research topic could involve the creation of document level emebddings, which would naturally bring the high-level semantic information carried by the whole text, and their integration with supervised WSD model. This would carry document-level information at the service of supervised models, which, nowadays, only rely on sentence-level context to disambiguate each word. Similarly to the previous point, this could help the systems to better generalise over new topics without the need of additional semantically-annotated training data.

4. **Sense-annotating as a Game.** A topic that has remained underexplored is the use of reinforcement learning in NLP and especially in WSD, where no efforts, to the best of our knowledge, have been ever spent to formalise WSD in terms of this paradigm. However, while it is hard to express the WSD puzzle in terms of the reinforcement learning method (i.e., with an agent, a policy, environment and feedback), it seems more reasonable to formulate the problem of creating a sense-annotated dataset in its terms. Indeed, selecting a sentence as example for a sense and scoring the decision seems reasonable if a very small amount of sense-annotated data are available. Therefore, this track would be worth at least a try, considering the important results attained

recently thanks to this technique (Silver et al., 2016).

5. **ImageNet for WSD** Least but not last, an important and maybe essential step for WSD, is a newer, updated, and multilingual SemCor, in other words, an ImageNet for WSD. Semantically annotating parallel corpora may be one of the best solutions in order to build a large, multilingual, and at the same time, parallel dataset for WSD, which could highly benefit the field and therefore affect also related down-stream application such as machine translation. Indeed, not only machine translation models would have semantic annotations across languages at their disposal, but also, foreseeing a considerable increase in the WSD models performance, they could also benefit from higher accuracy automatic disambiguations thus finally closing the loop with the idea of Weaver, that, back in 1949, formulated the Word Sense Disambiguation task to solve the machine translation problem.

Long-term Perspectives Looking forward to a more distant future and considering the advancements brought by the work described in this thesis, we expect that the WSD field will catch up with other NLP tasks[2] where, thanks to the advent of deep learning, important milestones have been achieved in terms of results. Therefore, by starting from what has been introduced in Chapter 4 and Chapter 6, in order to pursue this goal, it will be crucial to develop and refine methodologies for augmenting the volume of semantically-annotated data not only for English but also for other languages, while, at same time, maintaining a high quality of the annotations. What mostly differentiate WSD from other tasks, in fact, is the availability of sense-tagged data, which, nowadays, play a key role in the training of deep neural networks which proved to be capable of effectively capturing different aspects and phenomena of a language.

A parallel path which we explored with Train-O-Matic++, EnDi and DaD (Chapter 5 and 6), would be to develop new methods for providing prior knowledge about a language or a corpus to supervised models. We want to further extend this concept, not only to models that leverage the knowledge available in a semantic network, but also to those approaches which aim at encoding explicitly inductive bias in their architecture. An interesting direction would be, in fact, to brush up

[2]Sentiment analysis, paraphrase detection, textual similarity, natural language inference.

the generative lexicon theory (Pustejovsky, 1995) and build a parallelism with the current neural language models which are capable of dynamically creating word representations that are dependent on the context. Pustejovsky's theory, in fact, could directly benefit and leverage these models thanks to its dynamic nature of inducing word senses from the context of a word, and, on the other way around, help to encode the dynamic structure of a sentence semantics directly into a neural network.

To conclude, while the first direction seems more tempting since deep neural networks already proved to be effective in exploiting a large amount of data, encoding linguistic structure within a model, as shown by the self attention mechanism (Bahdanau, Cho, and Bengio, 2015; Vaswani et al., 2017; Devlin et al., 2018), proved to be successful and beneficial to a large set of different tasks in NLP. Being able to encode some linguistic priors, in fact, could help to reduce the amount of data needed for training and increase the generalisation capabilities of the models. Therefore, instead of providing prior knowledge only in the form of training data as we did with Train-O-Matic++, we should decouple the training observations from the linguistic biases and move the latter within the learning model. Hence, based on the outcomes shown along this thesis which proved the effectiveness of linguistic priors, such as sense distribution, to a supervised model, we should encode this knowledge in the architecture itself, hence providing the latter with human-like capabilities of structured reasoning and making it a less superficial learner.

Bibliography

Agirre, Eneko and Philip Edmonds, editors. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer.

Agirre, Eneko, Oier Lopez de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Agirre, Eneko and David Martinez. 2004. Unsupervised wsd based on automatically retrieved examples: The importance of bias. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

Anderson, John Robert. 1976. *Language, Memory, and Thought*. Lawrence Erlbaum and Associates, Hillsdale, NJ.

Atkins, S. 1993. Tools for corpus-aided lexicography: the hector project. *Acta Linguistica Hungarica*, 41:5–72.

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ive. 2007. 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.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bennett, Andrew, Timothy Baldwin, Jey Han Lau, Diana McCarthy, and Francis Bond. 2016. Lexsemtm: A semantic dataset based on all-words unsupervised

sense distribution learning. In *Proceedings of the 54nd Annual Meeting of the Association for Computational Linguistics*, pages 1513 – 1524, Berlin, Germany.

Bhingardive, Sudha, Dhirendra Singh, V Rudramurthy, Hanumant Redkar, and Pushpak Bhattacharyya. 2015. Unsupervised most frequent sense detection using word embeddings. In *Proc. of NAACL*, pages 1238–1243.

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the International Conference on Management of Data (SIGMOD '08)*, pages 1247–1250, New York, NY, USA.

Bond, Francis and Ryan Foster. 2013. 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.

Bovi, Claudio Delli, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. 2017. Eurosense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 594–600.

Brin, Sergey and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

Camacho-Collados, José, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli. 2016. A Large-Scale Multilingual Disambiguation of Glosses. In *Proceedings of LREC*, pages 1701–1708, Portoroz, Slovenia.

Camacho-Collados, José, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli. 2016. A large-scale multilingual disambiguation of glosses. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*, pages 1701–1708.

Camacho-Collados, Jose, Claudio Delli Bovi, Luis Espinosa Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 task 9: Hypernym discovery. In *Proceedings*

of The 12th International Workshop on Semantic Evaluation, pages 712–724, Association for Computational Linguistics, New Orleans, Louisiana.

Camacho-Collados, Jose and Roberto Navigli. 2017. Babeldomains: Large-scale domain labeling of lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 223–228.

Camacho-Collados, José, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Nasari: a novel approach to a semantically-aware representation of items. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 567–577, Association for Computational Linguistics, Denver, Colorado.

Chan, Yee Seng and Hwee Tou Ng. 2005a. Word sense disambiguation with distribution estimation. In *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30 - August 5*, pages 1010–1015.

Chan, Yee Seng and Hwee Tou Ng. 2005b. Word sense disambiguation with distribution estimation. In *IJCAI*, volume 5, pages 1010–5.

Chan, Yee Seng and Hwee Tou Ng. 2006. Estimating class priors in domain adaptation for word sense disambiguation. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*, pages 89–96.

Chaplot, Devendra Singh and Ruslan Salakhutdinov. 2018. Knowledge-based word sense disambiguation using topic models. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5062–5069.

Collins, Allan M and Elizabeth F Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Cuadros, Montse and German Rigau. 2008. KnowNet: Building a Large Net of Knowledge from the Web. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 161–168, Manchester, United Kingdom.

Delli Bovi, Claudio, Luis Espinosa Anke, and Roberto Navigli. 2015a. Knowledge Base Unification via Sense Embeddings and Disambiguation. In *Proc. of EMNLP*, pages 726–736.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, Ieee.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Edmonds, Philip and Scott Cotton. 2001. Senseval-2: overview. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Association for Computational Linguistics.

Emanuele, Pianta, Bentivogli Luisa, and Girardi Christian. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Escudero, G., L. Màrquez, and G. Rigau. 2000. On the portability and tuning of supervised word sense disambiguation. In *Proceedings of the joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, EMNLP/VLC 2000*, pages 172–180, Hong Kong, China.

Escudero, Gerard, Lluís Màrquez, and German Rigau. 2000. An empirical study of the domain dependence of supervised word sense disambiguation systems. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the*

38th Annual Meeting of the Association for Computational Linguistics-Volume 13, pages 172–180, Association for Computational Linguistics.

Espinosa-Anke, Luis, Jose Camacho-Collados, Sara Rodríguez Fernández, Horacio Saggion, and Leo Wanner. 2016. Extending WordNet with Fine-Grained Collocational Information via Supervised Distributional Learning. In *Proceedings of COLING 2016: Technical Papers. The 26th International Conference on Computational Linguistics*, pages 900–910, Osaka, Japan.

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

Fellbaum, Christiane, editor. 1998b. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

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

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

Flati, Tiziano, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2014. 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, Association for Computational Linguistics, Baltimore, Maryland.

Flati, Tiziano, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2016. Multiwibi: The multilingual wikipedia bitaxonomy project. *Artificial Intelligence*, 241:66–102.

Granger, Richard H. 1982. Scruffy text understanding: design and implementation of'tolerant'understanders. In *Proceedings of the 20th annual meeting on Association for Computational Linguistics*, pages 157–160, Association for Computational Linguistics.

Hamp, Birgit and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for german. *Automatic information extraction and building of lexical semantic resources for NLP applications*.

Hauer, Bradley, Yixing Luan, and Grzegorz Kondrak. 2019. You shall know the most frequent sense by the company it keeps. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 208–215, IEEE.

Haveliwala, Taher H. 2002. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526, ACM.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Companion Volume to the Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, New York, N.Y., 4–9 June 2006*, pages 57–60.

Howard, Jeremy and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339.

Huang, Chu-Ren, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese Information Processing*, 24(2):14–23.

Huang, Yanping, Yonglong Cheng, Dehao Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, and Zhifeng Chen. 2018. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *arXiv preprint arXiv:1811.06965*.

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 95–105.

Ide, Nancy and Jean Véronis. 1993. 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.

Ide, Nancy and Jean Véronis. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics*, 24(1):2–40.

Jin, Peng, Diana McCarthy, Rob Koeling, and John Carroll. 2009. Estimating and exploiting the entropy of sense distributions. In *Proc. of NAACL*, pages 233–236.

Kaplan, Abraham. 1950. An experimental study of ambiguity and context6. In *Mechanical Translation*, volume 2, pages 39–46.

Kilgarriff, Adam and Joseph Rosenzweig. 2000. Framework and results for english senseval. *Computers and the Humanities*, 34(1-2):15–48.

Koehn, Philipp. 2005a. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Koehn, Philipp. 2005b. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Koehn, Philipp. 2005c. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*.

Koutsoudas, Andreas K. and R. Korfhage. 1956. M.t. and the problem of multiple meaning. In *Mechanical Translation*, volume 2, pages 46–51.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kucera, H. and W. N. Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press.

Lau, Jey Han, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and

identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 259–270.

Lenat, Douglas B. 1995a. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11).

Lenat, Douglas B. 1995b. Cyc: a large scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Lesk, Michael. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual Conference on Systems Documentation, Toronto, Ontario, Canada*, pages 24–26.

Lofgren, Peter A, Siddhartha Banerjee, Ashish Goel, and C Seshadhri. 2014. Fast-ppr: Scaling personalized pagerank estimation for large graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1436–1445, ACM.

Luo, Fuli, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018. Incorporating glosses into neural word sense disambiguation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2473–2482.

Mahdisoltani, Farzaneh, Joanna Biega, and Fabian M. Suchanek. 2013. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *CIDR*, Asilomar, United States.

Mallery, J. C. 1988. *Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers, Ph.D. Thesis*. M.I.T. Political Science Department, Cambridge, MA.

Manion, Steve L and Raazesh Sainudiin. 2014. An iterative "sudoku style" approach to subgraph-based word sense disambiguation. In *In Proceedings of the Third Joint Conference on Lexical and Computational Se- mantics (*SEM 2014)*, pages 40–50, Dublin, Ireland.

Masterman, Margaret. 1961. Semantic message detection for machine translation, using an interlingua. In *Proc. 1961 International Conf. on Machine Translation*, pages 438–475.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 21–26 July 2004*, pages 280–287.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590.

de Melo, Gerard and Gerhard Weikum. 2009. 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.

Mihalcea, Rada and Dan Moldovan. 2001. eXtended WordNet: Progress Report. In *Proc. of the NAACL Workshop on WordNet and Other Lexical Resources*, pages 95–100.

Miller, George A., R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. 1990. WordNet: an online lexical database. *International Journal of Lexicography*, 3(4):235–244.

Miller, George A., Claudia Leacock, Randee Tengi, and Ross Bunker. 1993a. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308, Plainsboro, N.J.

Miller, George A., Claudia Leacock, Randee Tengi, and Ross Bunker. 1993b. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308, Plainsboro, N.J.

Mohammad, Saif and Graeme Hirst. 2006. Determining word sense dominance using a thesaurus. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Moro, Andrea and Roberto Navigli. 2015. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 288–297.

Moro, Andrea, Alessandro Raganato, and Roberto Navigli. 2014a. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Moro, Andrea, Alessandro Raganato, and Roberto Navigli. 2014b. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Nastase, Vivi, Michael Strube, Benjamin Boerschinger, Caecilia Zirn, and Anas Elghafari. 2010. 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.

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

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

Navigli, Roberto, David Jurgens, and Daniele Vannella. 2013b. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the* 7^{th} *International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, pages 222–231, Atlanta, USA.

Navigli, Roberto, Kenneth C. Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: Coarse-grained English all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23–24 June 2007*, pages 30–35.

Navigli, Roberto and Simone Paolo Ponzetto. 2010. 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.

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

Ng, Hwee Tou and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 40–47, Association for Computational Linguistics.

Ng, Hwee Tou, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for Word Sense Disambiguation: an empirical study. In *Proc. of ACL-03*, pages 455–462.

Niles, Ian and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems, Ogunquit, Maine, 17-19 October 2001*, pages 2–9.

Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

Pasini, Tommaso, Francesco Elia, and Roberto Navigli. 2018. Huge automatically extracted training-sets for multilingual word sensedisambiguation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*, pages 1694 – 1698.

Pasini, Tommaso and Roberto Navigli. 2017. Train-o-matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 78–88.

Pasini, Tommaso and Roberto Navigli. 2018. Two knowledge-based methods for high-performance sense distribution learning. In *Proceedings of the Thirty-Second*

AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 5374–5381.

Pasini, Tommaso and Roberto Navigli. 2019. Train-o-matic: Supervised word sense disambiguation with no (manual) effort. *Accepted with Major Revision at Artificial Intelligence Journal*.

Passonneau, Rebecca J, Collin Baker, Christiane Fellbaum, and Nancy Ide. 2012. The masc word sense sentence corpus. In *Proceedings of LREC*.

Pease, Adam, Ian Niles, and John Li. 2002. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In *In Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, page 2002.

Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, ACM.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237.

Philpot, Andrew, Eduard Hovy, and Patrick Pantel. 2005. The omega ontology. In *Proceedings of IJCNLP workshop on Ontologies and Lexical Resources (OntoLex-05)*, pages 59–66, Jeju Island, South Korea.

Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi. 2002. 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.

Pilehvar, Mohammad Taher, Jose Camacho-Collados, Roberto Navigli, and Nigel Collier. 2017. Towards a seamless integration of word senses into downstream nlp

applications. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1857–1869.

Pilehvar, Mohammad Taher and Nigel Collier. 2016. De-conflated semantic representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1680–1690, Austin, TX.

Pilehvar, Mohammad Taher, David Jurgens, and Roberto Navigli. 2013. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of ACL*, pages 1341–1351.

Pilehvar, Mohammad Taher and Roberto Navigli. 2015. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95–128.

Postma, Marten, Ruben Izquierdo Bevia, and Piek Vossen. 2016a. More is not always better: balancing sense distributions for all-words word sense disambiguation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3496–3506.

Postma, Marten, Ruben Izquierdo, Eneko Agirre, German Rigau, and Piek Vossen. 2016b. Addressing the mfs bias in wsd systems. *Recall*, 20(40):60.

Pradhan, Sameer S, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task 17: English lexical sample, srl and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 87–92.

Pu, Xiao, Nikolaos Pappas, James Henderson, and Andrei Popescu-Belis. 2018. Integrating weakly supervised word sense disambiguation into neural machine translation. *CoRR*, abs/1810.02614.

Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press, Cambridge, Massachusetts.

Quillian, M Ross. 1961. A design for an understanding machine. In *Communication presented at the colloquium Semantic problems in natural language*.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1:8.

Raganato, Alessandro, Jose Camacho-Collados, and Roberto Navigli. 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of EACL*, pages 99–110, Valencia, Spain.

Raganato, Alessandro, Claudio Delli Bovi, and Roberto Navigli. 2016. Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia. In *Proceedings of IJCAI*, pages 2894–2900, New York City, NY, USA.

Reifler, Erwin. 1955. The mechanical determination of meaning. In *Machine Translation of Languages*, pages 136–164.

Resnik, Philip and David Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pages 79–86.

Roget, Peter Mark. 2000. *Roget's Thesaurus of English words and phrases*. Harmondsworth, U.K.: Penguin. New ed. / completely revised, updated and abridged by E.M. Kirkpatrick.

Sagot, Benoît and Darja Fišer. 2008. Building a free french wordnet from multilingual resources. In *OntoLex*.

Scarlini, Bianca, Tommaso Pasini, and Roberto Navigli. 2019. Just "onesec" for producing multilingual sense-annotated data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1.

Silver, David, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484.

Snyder, Benjamin and Martha Palmer. 2004. The english all-words task. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain.

Sowa, John F. 1983. *Conceptual structures: information processing in mind and machine*. Addison-Wesley Pub., Reading, MA.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge. In *Proc. of WWW-07*, pages 697–706.

Taghipour, Kaveh and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015, Beijing, China, July 30-31, 2015*, pages 338–344.

Toral, Antonio, Stefania Bracale, Monica Monachini, Claudia Soria, et al. 2010. Rejuvenating the italian wordnet: upgrading, standardising, extending. *Citeseer*.

Tripodi, Rocco and Marcello Pelillo. 2017. A Game-Theoretic Approach to Word Sense Disambiguation. *Computational Linguistics*, 43(1):31–70.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010.

Vial, Loïc, Benjamin Lecouteux, and Didier Schwab. 2018. Ufsac: Unification of sense annotated corpora and tools. In *Language Resources and Evaluation Conference (LREC)*.

Walker, Donald E and Robert A Amsler. 1986. The use of machine-readable dictionaries in sublanguage analysis. *Analyzing language in restricted domains: Sublanguage description and processing*, pages 69–83.

Weaver, Warren. 1949. Translation. In *Machine Translation of Languages: Fourteen Essays (written in 1949, published in 1955)*, pages 15–23, W. N. Locke and A. D. Booth, Eds. Technology Press of MIT, Cambridge, MA, and John Wiley & Sons, New York, NY.

Zhong, Zhi and Hwee Tou Ng. 2009. Word sense disambiguation for all words without hard labor. In *Twenty-First International Joint Conference on Artificial Intelligence*.

Zhong, Zhi and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Association for Computational Linguistics, Uppsala, Sweden.

Ziemski, Michał, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3530 – 3534, European Language Resources Association (ELRA), Portoroz, Slovenia.