



SAPIENZA
UNIVERSITÀ DI ROMA

Delving into the Uncharted Territories of Word Sense Disambiguation

Sapienza Università di Roma

PhD in Linguistics – Cycle XXXIII
Curriculum of Computational Linguistics

Candidate

Marco Maru

ID number 1135442

Thesis Advisors

Prof. Claudia Angela Ciancaglini

Prof. Roberto Navigli

October 2020

Thesis defended on February 26, 2021

in front of a Board of Examiners composed by:

Prof. Patrizia Cordin, Prof. Elisabetta Ježek, Prof. Paolo Poccetti (chairman)

Delving into the Uncharted Territories of Word Sense Disambiguation

Ph.D. thesis. Sapienza – University of Rome

This thesis and its contents are licensed under the CC BY-NC-SA 4.0 License.

This thesis has been typeset by \LaTeX and the Saphesis class.

Author's email: marco.maru@uniroma1.it

*Per aspera
ad astra.*

Abstract

The automatic disambiguation of word senses, i.e. Word Sense Disambiguation, is a long-standing task in the field of Natural Language Processing; an AI-complete problem which took its first steps more than half a century ago, and which, to date, has apparently attained human-like performances on standard evaluation benchmarks.

Unfortunately, the steady evolution that the task experienced over time in terms of sheer performance has not been followed hand in hand by an adequate theoretical support, nor by a careful error analysis. Furthermore, we believe that the lack of an exhaustive bird's eye view which accounts for the sort of high-end and unrealistic computational architectures that systems will soon need in order to further refine their performances could lead the field to a dead angle in a few years.

In essence, taking advantage from the current moment of great accomplishments and renewed interest in the task, we argue that Word Sense Disambiguation is mature enough for researchers to really observe the extent of the results hitherto obtained, evaluate what is actually missing, and answer the much sought for question: “are current state-of-the-art systems really able to effectively solve lexical ambiguity?”

Driven by the desire to become both architects and participants in this period of pondering, we have identified a few macro areas representative of the challenges of automatic disambiguation. From this point of view, in this thesis we propose experimental solutions and empirical tools so as to bring to the attention of the Word Sense Disambiguation community unusual and unexplored points of view. We hope these will represent a new perspective through which to best observe the current state of disambiguation, as well as to foresee future paths for the task to evolve on.

Specifically, 1q) prompted by the growing concern about the rise in performance being closely linked to the demand for more and more unrealistic computational architectures in all areas of application of Deep Learning related techniques, we 1a) provide evidence for the undisclosed potential of approaches based on knowledge-bases, via the exploitation of syntagmatic information. Moreover, 2q) driven by the dissatisfaction with the use of cognitively-inaccurate, finite inventories of word senses in Word Sense Disambiguation, we 2a) introduce an approach based on Definition Modeling paradigms to generate contextual definitions for target words and phrases, hence going beyond the limits set by specific lexical-semantic inventories. Finally, 3q) moved by the desire to analyze the real implications beyond the idea of “machines performing disambiguation on par with their human counterparts” we 3a) put forward a detailed analysis of the shared errors affecting current state-of-the-art systems based on diverse approaches for Word Sense Disambiguation, and

highlight, by means of a novel evaluation dataset tailored to represent common and critical issues shared by all systems, performances way lower than those usually reported in the current literature.

Acknowledgments

The development of this thesis and the projects it contains have been made possible thanks to the support of the ERC Consolidator Grant MOUSSE No. 726487.

Contents

Publications and Thesis Contributions	xiii
Research Questions and Objectives	xv
List of Acronyms and Abbreviations	xvii
1 Background	1
2 Word Sense Disambiguation	3
2.1 Preliminaries	4
2.1.1 Sense Inventories	4
2.1.2 Sense-annotated Corpora	5
2.1.3 Evaluation	7
2.2 A Brief History of WSD	9
2.3 Open Problems	10
2.3.1 Prohibitive Computational Requirements	11
2.3.2 Word Senses as Discrete Entities	13
2.3.3 Inadequate Error Analysis	15
2.3.4 Other Open Problems	17
3 Structured Syntagmatic Information Enhancing Knowledge-based WSD	19
3.1 SyntagNet	20
3.1.1 Related Work	21
3.1.2 A Wide-coverage Lexical-semantic Combination Resource	22
3.1.3 Experimental Setup	25
3.1.4 Experimental Results	27
3.1.5 Impact of LKB Size	29
3.2 SyntagRank	29
3.2.1 Lexical Knowledge Bases	30

3.2.2	Personalized PageRank	30
3.2.3	System Architecture	32
3.2.4	Web Interface	35
3.2.5	Evaluation	38
3.3	Conclusion	39
4	Recasting Word Sense Disambiguation via Contextual Definition Generation	41
4.1	Related Work	42
4.2	Generatory	43
4.2.1	Gloss Generation	44
4.2.2	Discriminative Sense Scoring	45
4.3	Datasets	46
4.3.1	Dictionary Gloss Datasets	46
4.3.2	The Hei++ Evaluation Dataset	47
4.4	Quantitative Experiments	48
4.4.1	Definition Modeling	48
4.4.2	WSD Evaluation	52
4.4.3	Word-in-Context	53
4.4.4	Reproducibility Details	54
4.5	Qualitative Experiment	55
4.5.1	Annotators and Annotation Scheme	55
4.5.2	Results	56
4.6	Generation Examples	58
4.7	Error Analysis	59
4.8	Conclusion	59
5	Dissecting the State of the Art	61
5.1	Systems at Issue	63
5.1.1	Random System Baseline	65
5.2	Analysis of Traditional Benchmarks	65
5.2.1	Quantitative Analysis	65
5.2.2	A Model-agnostic Hard Core	67
5.3	Analysis of 42D: a Challenge Set	74
5.3.1	Corpus and Domain Set	74
5.3.2	Building 42D	75
5.3.3	Dataset Annotation	76
5.3.4	Results and Discussion	76

5.4	Conclusion	81
6	Summary	83
6.1	Perspectives	84
6.1.1	On Knowledge Bases	85
6.1.2	On Definition Generation	85
6.1.3	On Disambiguation Errors and System Performance	86
A	Chapter 4: Supplementary Materials	89
A.1	Additional Results on DM	89
A.2	Perplexity	90
A.3	NLG Measures Details	91
A.4	Generation Examples	91

Publications and Thesis Contributions

This dissertation is the result of a three-years research effort conducted in the field of Word Sense Disambiguation. In what follows, we list the publications which are featured in this thesis as main contributions. Each entry shows thorough referencing details, along with providing a brief description of our individual contributions to each distinct work.

1. **Marco Maru**, Federico Scozzafava, Federico Martelli and Roberto Navigli. *Syn-tagNet: Challenging Supervised Word Sense Disambiguation with Lexical-Semantic Combinations*. Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), Hong Kong, China, November 3-7, 2019 (Maru et al., 2019).

Individual contributions: Main writer; linguistics expertise; methodology development; disambiguation and validation; related work research; support on experiments.

2. Federico Scozzafava, **Marco Maru**, Fabrizio Brignone, Giovanni Torrisi and Roberto Navigli. *Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, July 5-10, 2020 (Scozzafava et al., 2020).

Individual contributions: Main writer; linguistics expertise; interface design and functionality consultant.

3. Michele Bevilacqua*, **Marco Maru***¹, and Roberto Navigli. *Generatory or: “How We Went beyond Word Sense Inventories and Learned to Gloss”*. Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), Online, November 16-20, 2020 (Bevilacqua, Maru, and Navigli, 2020).

¹* indicates equal contribution.

Individual contributions: Writer; concept; linguistics expertise; related work research; dictionary gloss datasets retrieval and processing; qualitative experiment and Hei++ dataset design and realization.

4. **Marco Maru**, Michele Bevilacqua, and Roberto Navigli. *Under the Mask of Word Sense Disambiguation: a Tale of Puzzling Performances*. Excerpt of an article to be submitted at top-tier journal during Q1, 2021 (upcoming).

Individual contributions: Main writer; concept; linguistic expertise; quantitative and qualitative analyses; support on experiments.

Research Questions and Objectives

In this Section we accurately list and describe the research questions and objectives that provide the backbone for this thesis. For each, we briefly report the strategies employed, along with anticipating findings and results. Additionally, we detail the research questions by indicating the corresponding publications, as well as providing the reader with references to the Chapters where these topics are extensively discussed.

“Are there viable alternatives to prevent Word Sense Disambiguation performances from being tied to increasingly prohibitive high-end infrastructures?”

- Objective: identifying a viable alternative to rival the performances of supervised disambiguation systems.
- Results: the injection of syntagmatic information – in the form of disambiguated pairs of lemmas – into an existing Lexical Knowledge Base biased towards paradigmatic knowledge, proved to significantly boost performances of knowledge-based systems, both in the English and multilingual settings. The huge impact brought about by the inclusion of syntagmatic knowledge, along with the performance figures growing steadily accordingly with the number of lexical-semantic relations added, gives evidence of a sound and promising alternative to the use of Language Models in supervised Word Sense Disambiguation.
- References: Chapter 3 (Maru et al., 2019; Scozzafava et al., 2020).

“Are inventories of discrete word senses the best available option to tackle automatic disambiguation?”

- Objective: devise a methodology to perform Word Sense Disambiguation without resorting to a finite sense inventory.
- Results: generating an *ad hoc* contextual definition (gloss) of human-like quality, rather than choosing from a finite list of many, is a suitable way to tackle the task. In

fact, a generative formulation of Word Sense Disambiguation drops the requirement for discreteness in word sense boundaries, while enabling disambiguation on out-of-dictionary items at the same time. From a technical point of view, the use of a pre-trained Encoder-Decoder through fine-tuning allowed us to apply the paradigms of Definition Modeling to Word Sense Disambiguation, with results matching or outperforming the state of the art in both generative and fully discriminative tasks.

- References: Chapter 4 (Bevilacqua, Maru, and Navigli, 2020).

“Are current state-of-the-art systems actually able to solve lexical ambiguity?”

- Objective: perform a detailed error analysis of the state of the art in Word Sense Disambiguation in order to inquire about the actual disambiguation capabilities automatic systems are able to exhibit.
- Results: notwithstanding the algorithm employed, state-of-the-art systems still fail to provide adequate answers when asked to disambiguate instances tagged with infrequent word senses or instances that do not appear in the training sets. Still, several of the systems’ errors can be attributed to the structure of the sense inventory employed, which forces the human annotators to resort to suboptimal choices when asked to produce a gold standard. Our findings are supported by a sound experimental setting, and, particularly, by testing the system’s performances on a newly created dataset, which is tailored to reproduce the environment in which we found systems to struggle the most.
- References: Chapter 5.

List of Acronyms and Abbreviations

As an aid to the reader, with the following, we provide an alphabetically-sorted list of common acronyms and abbreviations as will be used throughout this thesis.

- AI (Artificial Intelligence)
- DL (Deep Learning)
- DM (Definition Modeling)
- ITA (Inter-annotator Agreement)
- LKB (Lexical Knowledge Base)
- ML (Machine Learning)
- MT (Machine Translation)
- NLG (Natural Language Generation)
- NLP (Natural Language Processing)
- NLU (Natural Language Understanding)
- PoS (Part of Speech)
- PPR (Personalized PageRank)
- PWNG (Princeton WordNet Gloss Corpus)
- SOTA (State of the art)
- WiC (Word-in-Context)
- WSD (Word Sense Disambiguation)

Chapter 1

Background

Can computers be used to manipulate and figure out the meaning of written and spoken natural language texts? The answer to this question is a matter of interest for the area of research known as Natural Language Processing (NLP), where researchers aim to exploit established paradigms from theoretical linguistics, neurobiology, computer science, psychology and robotics in order to train machines to perform a set of multifaceted and useful tasks (Chowdhary, 2020).

The number of downstream applications entailed by NLP is huge, with several fields of studies involved, ranging from information retrieval and summarization, up to machine translation, commonsense inference, and speech recognition. In recent years, the dominant architecture for NLP applications has rapidly become the Transformer (Vaswani et al., 2017), a deep learning model designed to handle sequential data processed in random order – hence facilitating efficient parallel training – and able to easily scale according with training data and model size.

The effectiveness of Transformers provoked a significant shift in the field, which led researchers to dismiss alternative neural models such as convolutional and recurrent neural networks (Wolf et al., 2019), so far lying at the foundation of state-of-the-art approaches in many subfields of NLP. Moreover, the brand-new development of pre-trained language models such as BERT (Devlin et al., 2019) or GPT-3 (Brown et al., 2020) by means of hitherto unreasonable amounts of training data has only been made possible thanks to the advent of Transformers.

Considering how (i) language models can be easily fine-tuned in order to perform diverse natural language tasks, and (ii) one of the much sought for goals of NLP is to emulate the human innate ability to identify the correct meaning of a word in context, it is no surprise to witness the widespread use of Transformer architectures in one of the oldest and core application of NLP, namely, Word Sense Disambiguation.

Word Sense Disambiguation, the task of disambiguating a target in context by means of picking a suitable word sense from a finite inventory of many (Navigli, 2009), lies at the core of NLP, and has been deemed as an AI-complete open problem, i.e. a problem that – as many other NLP tasks – requires several specific algorithms and the interaction of different fields of studies to be tackled successfully.

In this thesis we move from the current Transformed-based state of the art in Word Sense Disambiguation (Hadiwinoto, Ng, and Gan, 2019; Vial, Lecouteux, and Schwab, 2019; Huang et al., 2019; Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020) and from the premises that language models seem to have brought WSD to surpass the line traced by human performances, to put this assertion under analysis, and show that, by no means, WSD can be considered a solved problem.

Chapter 2

Word Sense Disambiguation

To infer the correct meaning of homonymous or polysemous words is a fundamental operation for a communicative act to be successful, and a trivial one for most human speakers. On the contrary, the same deed is so hard for a machine to carry out, to be considered an AI-complete problem (Navigli, 2009). Consider the following examples:

1. The whole *table* was entertained.
2. I reserved a *table* at your favorite restaurant.
3. Please make sure to enter the data in the correct *table*.

Word Sense Disambiguation (henceforth WSD) is defined as the computational task of automatically identifying meaning for words in contexts, typically, by picking the most suitable word sense (expressed by means of a gloss, or dictionary definition) from a finite inventory of many. Hence, given the input examples 1, 2 and 3 shown above, with the target word *table* to be disambiguated, we would expect the following outputs:¹

1. A company of people assembled at a table for a meal or game.
2. A piece of furniture with tableware for a meal laid out on it.
3. A set of data arranged in rows and columns.

In the following Sections, we will provide an overview of the WSD task, first, detailing the most commonly employed sense inventories, sense-annotated corpora and evaluation datasets available in the literature (Section 2.1). Secondly, we will retrace the task back to its origins, also, summarizing the evolution of the state of the art until today (Section 2.2).

¹Taken from the WordNet 3.0 sense inventory (Fellbaum, 1998).

Finally, we will point out some of the major criticalities currently affecting WSD, along with introducing the proposed solutions we are going to report in the subsequent Chapters of this thesis as our main contributions (Section 2.3).

2.1 Preliminaries

In this Section we introduce the basic concepts and resources that will be used in this Chapter, and throughout the whole thesis, with reference to WSD.

2.1.1 Sense Inventories

Treating WSD as a classification problem at a computational level has been the most widely adopted strategy to tackle the task so far, with the vast majority of approaches to automatic disambiguation leveraging strict word sense distinctions, as they appear in traditional sense inventories. As of now, the most commonly employed sense inventories for English and multilingual WSD are WordNet (Fellbaum, 1998) and BabelNet (Navigli and Ponzetto, 2012), respectively.

WordNet WordNet² is a large lexical database of English in which distinct concepts are organized by means of groups (sets) of cognitive synonyms called *synsets* (Fellbaum, 1998). Each of the circa 117,000 synsets contained in WordNet (as of its 3.0 release) comprises a set of lemmas sharing the same part of speech, along with a brief definition and, occasionally, one or more usage example (see Figure 2.1). All of the lemmas in a given synset are considered synonyms to denote the specific concept that is expressed by means of the accompanying definition.

Moreover, being interlinked via semantic relations such as hyperonymy (e.g., between the concept for *periodic table* and the concept for *table*) or meronymy (e.g. between *row* and *table*), synsets can be seen as nodes in a semantic network (see Figure 2.2).

BabelNet Similarly to WordNet, BabelNet (Navigli and Ponzetto, 2012) resembles an enhanced thesaurus, grouping and interlinking specific senses of words by means of sets of synonyms (in this case, called *Babel synsets*), but it can be considered a superset of WordNet, in that its network is obtained by automatically integrating information from several resources such as Wikipedia, Wiktionary, the Open Multilingual WordNet (Bond and Foster, 2013), and, in fact, WordNet itself.

²<http://wordnetweb.princeton.edu/perl/webwn>

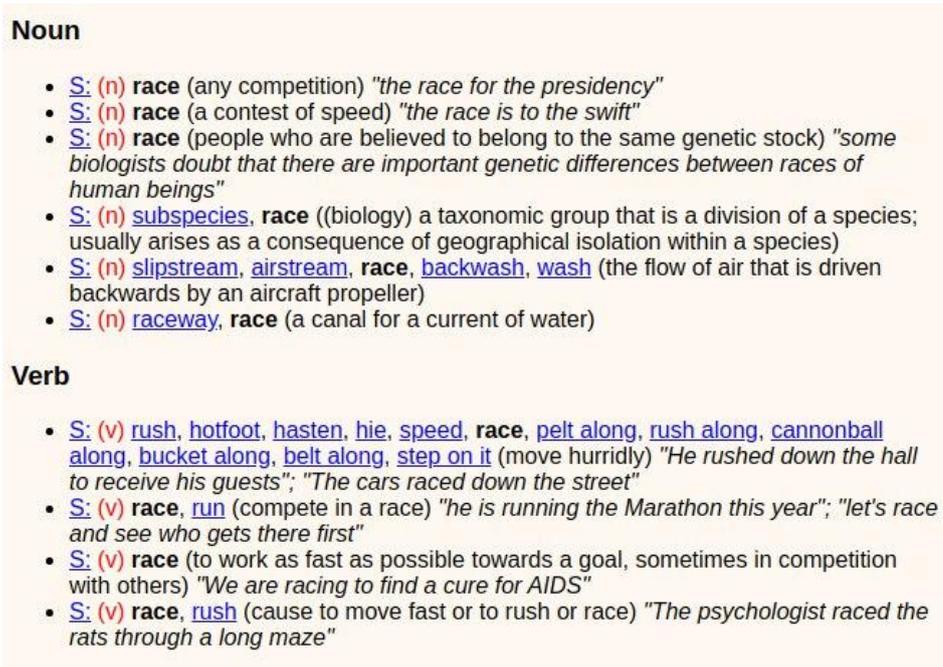


Figure 2.1. Content of the WordNet sense inventory for the word *race*.

At the time of writing, the BabelNet network (in its 4.0 version³) contains about 16 million unique concepts, including both common dictionary concepts and named entities (i.e. real-world objects, such as persons, locations or organizations). Compared to WordNet, BabelNet provides lexicographic and encyclopedic coverage of terms in 284 languages, due to the inherently multilingual nature of the resources it includes (e.g. Wikipedia), along with the aid of automatic translations of concepts (see Figure 2.3).

2.1.2 Sense-annotated Corpora

As will be thoroughly detailed in Section 2.2, the most widespread approaches to WSD exploit machine learning techniques to train a sense classifier on sense-annotated data. In what follows, we describe the largest and most-widely employed corpora of such labeled data, namely, SemCor (Miller et al., 1993) and the Princeton WordNet Gloss Corpus⁴ (henceforth, PWNG).

³<https://babelnet.org/>

⁴<https://wordnetcode.princeton.edu/glosstag.shtml>

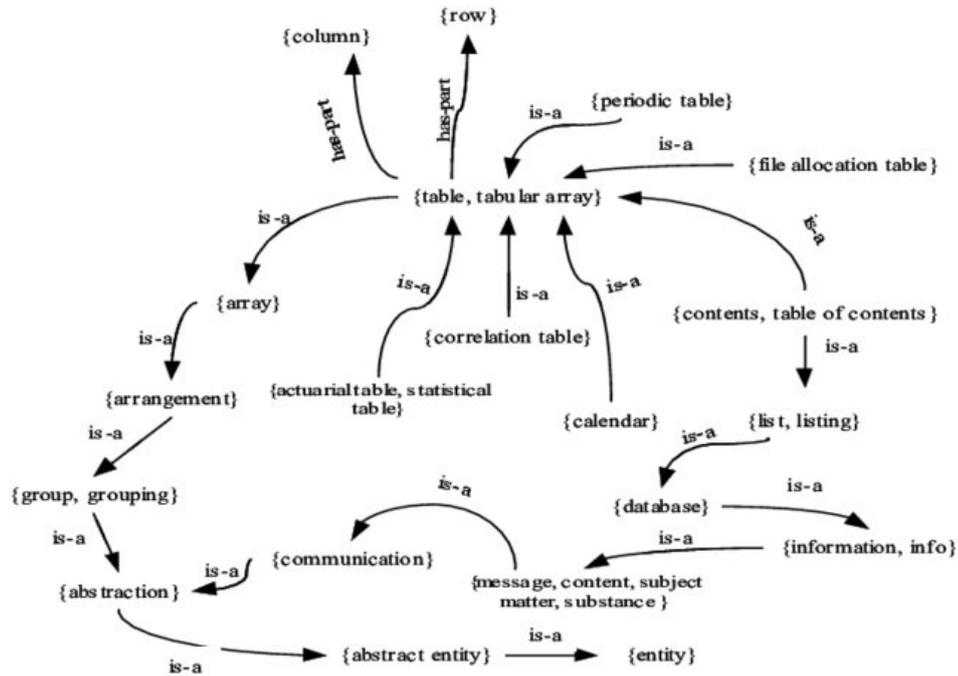


Figure 2.2. Excerpt of the WordNet semantic network (image taken from Al-Saeedan and Menai (2015)).

SemCor SemCor (Miller et al., 1993) is a subset of the Brown Corpus (Francis and Kucera, 1979). It consists of 352 texts whose open-class words have been disambiguated according to the WordNet sense inventory,⁵ for a figure of 226,036 annotated instances. As of now, SemCor is still the most widely employed corpus for WSD⁶ in supervised disambiguation settings (Loureiro and Jorge, 2019; Huang et al., 2019).

The Princeton WordNet Gloss Corpus In the Princeton WordNet Gloss Corpus (PWNG), first released in 2008,⁷ the content words in the definitions (here referred to as “glosses”) of WordNet’s synsets have been semi-automatically disambiguated against the WordNet sense inventory version 3.0. The resulting annotated corpus today comprises 449,355 disambiguated instances (118,856 of which, automatically tagged). Despite having seen a much more limited usage in comparison to SemCor, still, PWNG has had a significant impact on WSD performances, both in knowledge-based (Agirre, de Lacalle, and Soroa,

⁵Out of the whole 352 documents, 186 texts present annotations for verbs only.

⁶Despite having been originally annotated according to the 1.5 version of the WordNet inventory, mappings to the most recent releases of WordNet exist, hence, automatically bringing SemCor annotations up to date.

⁷<http://wordnetcode.princeton.edu/glosstag.shtml>

Italian Arabic Chinese **English** French German Greek Hebrew Hindi + tutte le lingue preferite

Indoeuropeisti, Linguisti svizzeri, Morti il 22 febbraio, Morti nel 1913...
 bn:00025490n • NOME • Entità • Categorie: Indoeuropeisti, Linguisti svizzeri, Morti il 22 febbraio, Morti nel 1913...

IT Ferdinand de Saussure • **De Saussure**

Ferdinand de Saussure è stato un **linguista e semiologo svizzero**. *Wikipedia*

+ Più definizioni

IS A	uomo • linguista
BROTHER	René de Saussure • Léopold de Saussure
CHILD	Raymond de Saussure
COUNTRY OF CITIZENS...	svizzera
EDUCATED AT	Università di Lipsia • Università di Ginevra
EMPLOYER	Ecole pratique des hautes études • Università di Ginevra

+ Più relazioni

Categorie: 1857 births, 1913 deaths, Ferdinand de Saussure, Indo-Europeanists...

EN Saussure • **Ferdinand de Saussure** • **de Saussure**

Swiss linguist and expert in historical linguistics whose lectures laid the foundations for synchronic linguistics (1857-1913). *WordNet*

+ Più definizioni

IS A	human • linguist
BROTHER	René de Saussure • Léopold de Saussure
CHILD	Raymond de Saussure
COUNTRY OF CITIZENS...	Switzerland
EDUCATED AT	Leipzig University • University of Geneva
EMPLOYER	Ecole pratique des hautes études • University of Geneva

+ Più relazioni

Figure 2.3. Babel synset for Ferdinand de Saussure in BabelNet (Italian and English lexicalizations).

2014), as well as in fully supervised settings (Bevilacqua and Navigli, 2020).

2.1.3 Evaluation

As is the case with other NLP tasks, WSD has a long-standing tradition of evaluation exercises aimed at determining which systems perform best. Particularly, the Senseval (now SemEval) evaluation campaign, which took place for the first time in 1998 (Kilgarriff, 1998), nowadays still represents the best reference to observe the evolution and trends of the field. In the following, we briefly report details for every Senseval/SemEval edition that dealt with the task of automatic Word Sense Disambiguation:

Senseval-1 The first edition of Senseval consisted of a simple lexical-sample task – i.e. one in which a system is asked to disambiguate a specific set of target words, typically, one per sentence – for the English, French and Italian languages, according to the HECTOR sense inventory (Atkins, 1992).

Senseval-2 In the occasion of the second Senseval evaluation campaign (Edmonds and Cotton, 2001), the WordNet lexical database (Fellbaum, 1998) became the *de facto* standard

sense inventory for WSD. Both a lexical-sample and an all-words disambiguation task – i.e. one in which a system is asked to provide answers for each content word in a given text – were organized for 12 languages.

Senseval-3 Since the adoption of WordNet, other evaluation exercises for WSD have been devised, with Senseval-3 being the first in time. Particularly, this edition of the Senseval competition took place in 2004 and consisted of 14 tasks for 7 distinct languages, including previously unseen exercises such as semantic role labeling and gloss disambiguation, along with the traditional lexical sample (Mihalcea, Chklovski, and Kilgarriff, 2004) and all-words fine-grained WSD (Snyder and Palmer, 2004) tasks.

SemEval-2007 The SemEval-2007 competition witnessed the organization of 18 unique tasks, including semantic analysis exercises unrelated to WSD (which prompted the name switch from Senseval to SemEval). Among the available tasks, WSD was featured both explicitly (via traditional evaluation exercises such as lexical sample and all-words disambiguation) and, for the first time, implicitly, via new tasks such as lexical substitution (McCarthy and Navigli, 2009) and word sense induction (Agirre and Soroa, 2007). Worth noticing, SemEval-2007 also represented the first attempt at assessing the impact of the sense granularity in existing inventories on WSD performance, particularly, by means of distinguishing coarser sense distinctions (Navigli, Litkowski, and Hargraves, 2007) versus traditional fine-grained sense boundaries (Pradhan et al., 2007).

SemEval-2010 In line with the edition held in 2007, SemEval-2010 comprised 18 tasks ranging from cross-lingual lexical substitution to coreference resolution. In the field of WSD, the SemEval-2010 task 17 (Agirre et al., 2010) addressed the issues entailed by testing over texts characterized by utterly generic domains by proposing a test corpus focused on a single, specific domain.

SemEval-2013 Out of the 14 tasks organized for the SemEval-2013 competition, WSD has been dealt with through the SemEval-2013 task 12 (Navigli, Jurgens, and Vannella, 2013), whose test set consisted of 13 articles obtained from three editions of the workshop on Statistical Machine Translation (WSMT)⁸ and covering domains ranging from sports to financial news. Particularly, this task represented the first attempt at producing a traditional WSD evaluation exercise in a multilingual environment, by exploiting the BabelNet sense inventory.

⁸<http://www.statmt.org>

SemEval-2015 The SemEval-2015 task 13 (Moro and Navigli, 2015), with four documents collected from the OPUS project⁹ in three specific domains (biomedical, maths and computers, and social issues), was the only task, among the 17 available, which dealt with WSD, this time, by proposing a joint integration with the Entity Linking task.

On a final note, it is necessary to note how the aforementioned evaluation exercises, held in a time span of 17 years, produced datasets which lacked coherency in terms of format, construction guidelines and sense inventory choice, hence, making WSD evaluation hardly fair. To overcome these issues, Raganato, Camacho-Collados, and Navigli (2017) introduced a common evaluation framework which standardized five all-words WSD datasets (Edmonds and Cotton, 2001; Snyder and Palmer, 2004; Pradhan et al., 2007; Navigli, Jurgens, and Vannella, 2013; Moro and Navigli, 2015) and two training corpora (Miller et al., 1993; Taghipour and Ng, 2015) in order to allow for fair quantitative and qualitative comparisons between systems. The framework is currently employed by all of the state-of-the-art WSD systems (Huang et al., 2019; Vial, Lecouteux, and Schwab, 2019; Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020) and therefore will represent our reference evaluation standard throughout this whole dissertation.

2.2 A Brief History of WSD

First introduced in the context of Machine Translation by Weaver (1955), Word Sense Disambiguation (WSD), i.e. the task of determining the correct meaning of a word in context, was initially addressed merely as a component of a more extensive text understanding problem. Moreover, WSD was right away hindered by the constraining lack of available resources that would have bestowed the task a distinct nature and domain of application.

In light of the above, the following decades saw the flourishing of machine-readable dictionaries and sense-annotated corpora which, in turn, led to the development of advanced techniques and algorithms, such as the dictionary-based approach of (Lesk, 1986), which exploited the overlap between bag of words in the sentence containing the target word to be disambiguated and the bag of words in the definitions for the different senses of the same target word. Still, the birth of modern WSD was not going to occur before the 1990s, when the two key resources for the task were released, namely, the WordNet sense inventory (Miller et al., 1990; Fellbaum, 1998) and the SemCor sense-annotated corpus (Miller et al., 1993).

If, on the one hand, WordNet encouraged further work revolving around the structure of

⁹<http://opus.nlpl.eu/>

a lexical knowledge base (LKB), on the other, SemCor promptly triggered the development of supervised approaches requiring an extensive amount of training data (see also Sections 2.1.1 and 2.1.2).

Another significant turning point in WSD took place thanks to the introduction of the BabelNet sense inventory (Navigli and Ponzetto, 2012), which managed to release the task from the constraint of English as a single language and allowed for studies on semantic similarities shared across different languages, moreover, combining traditional dictionary knowledge and encyclopedic knowledge into a single resource.

The increased availability of high-quality, multilingual lexical-semantic resources provoked a renewed interest towards knowledge-based approaches to WSD, i.e. approaches that exploit the structure of LKBs to determine the correct sense of a word in context (Agirre, de Lacalle, and Soroa, 2014; Moro, Raganato, and Navigli, 2014). However, despite dropping the requirement for huge amounts of training data, and despite being scalable to multilingual environments (thanks to resources such as, e.g. BabelNet), knowledge-based approaches always fell behind their supervised counterparts in terms of sheer performances.

An exemplary of a successful supervised model from the early 2010s is represented by IMS (Zhong and Ng, 2010), which made use of a linear kernel Support Vector Machine (SVM) as a classifier. In fact, according to the common evaluation of framework that was later devised by Raganato, Camacho-Collados, and Navigli (2017), performances on the concatenation of all of the available evaluation datasets for the best configuration of IMS showed an overall score of 69.6 in terms of F1 score (F-Measure), while the best knowledge-based system available as of 2014, namely, UKB (Agirre, de Lacalle, and Soroa, 2014), attained a score of 67.3 (cfr. also Agirre, López de Lacalle, and Soroa (2018)).

The gap between knowledge-based and supervised system performances has been further widened in recent years thanks to the advent of neural networks and language models in the field of Natural Language Processing (NLP) (Devlin et al., 2019). In fact, the automatic learning of features and the use of latent, contextualized representations to encode words in contexts enabled supervised systems to come close (Huang et al., 2019) or even surpass, for the first time, the 80% performance ceiling (Bevilacqua and Navigli, 2020), so far considered to be representative of human-like performance (Edmonds and Kilgarriff, 2002; Palmer, Dang, and Fellbaum, 2007).

2.3 Open Problems

Despite having finally attained performances comparable to those of human annotators, whose inter-annotator agreement has long been considered as an upper bound for the task

(Chklovski and Mihalcea, 2003; Snyder and Palmer, 2004; Palmer, Dang, and Fellbaum, 2007), WSD is still a long way from being a solved problem. In fact, several of the most significant issues that are long known to the research community (Navigli, 2009) have not been dealt with yet, and the current hype over unprecedented performances could conceal the existing difficulties.

With this thesis, we do not aim to provide a fully exhaustive survey and a set of ready-made solutions to tackle all of the existing problems in WSD. Rather, we identify and bring to public attention three focused, often neglected, and pivotal issues which, along with the proposed strategies to address them, can on their own pave the way for more aware, cognitively accurate, and hence effective lines of research on WSD. Such directions will:

1. avoid the need for increasingly prohibitive computational architectures;
2. dispose of finite inventories of discrete word senses;
3. compensate the lack of an adequate error analysis for state-of-the-art WSD systems.

In what follows, we will thus introduce and detail each of these three core issues (Sections 2.3.1, 2.3.2 and 2.3.3), along with providing, for completeness, further information concerning the other open problems in WSD as reported in the available literature (Section 2.3.4).

2.3.1 Prohibitive Computational Requirements

Current Issue The widespread use of Deep Learning (DL) techniques in NLP and, more broadly, in AI, has been for a long time paired with an exponentially growing demand for high-end computational infrastructures (Thompson et al., 2020). The correlation between such pricey architectures and system performances has been exhaustively demonstrated (Soltanolkotabi, Javanmard, and Lee, 2018), and impacts fields ranging from image classification and object detection (Barbu et al., 2019) to machine translation and named entity recognition (Bhatia et al., 2019).

The main reason behind DL being inherently more dependent on computational requirements than other approaches lies in the role of the so-called *overparametrization* phenomenon (Xie et al., 2020) and how this grows as more training data gets used as a means to boost performances; in a way, it can be said that what makes DL-based architectures so ductile and apt to perform different and novel tasks, is also what makes the same architectures so dramatically costly. Bearing in mind how the most preached and successful approaches in WSD are also relying on DL algorithms and techniques (Nithyanandan and Raseek, 2019), along with the fact that researchers are foreseeing a future in which more

computationally-efficient machine learning techniques are needed¹⁰ is definitely a case in point of an impactful open problem in the field.

This is by no means the first time that DL faces such a constraint, with neural networks being affected since 1960 (Minsky and Papert, 2017), and as of 2009, before the introduction of GPUs and ASICs sped-up processes by a factor of 35,¹¹ researchers still had to resort to smaller-scale models and smaller training sets. In fact, the progress is already continuing along dangerous lines, with the burdensome requirement of running models for more time, and on more machines. The steep technical and economical “cliff” is getting climbed by a few powerful companies which have the resources to keep out “smaller players” from the race, to the point of effectively being the only players.

An extreme example is the machine translation system named “Evolved Transformer”, which costed millions dollars to run, and took roughly 2 million GPU hours to be trained (So, Liang, and Le, 2019). More recently, OpenAI published the largest language model ever trained: GPT-3. With an impressive figure of 175 billion parameters, this model would require almost 5 millions of dollars and roughly 355 years to be trained on the cheapest GPU cloud (a Tesla v100) currently available on the market.

On another note, with the computational power required to train the most powerful AI models increased by 300,000 times since 2012,¹² the carbon emissions produced to this end have become alarmingly high. As a matter of example for this “Red-AI” – as opposed to a more eco-friendly Green-AI –, it is estimated that training and developing a single machine translation model with neural architecture search can lead to the production of circa 626,000 lbs of CO₂ (Strubell, Ganesh, and McCallum, 2019).

Proposed Solution In light of the above, and even though we are aware that known theories of DL have yet to fully explore the efficiency of supervised models (Gambetta and Sheldon, 2019), we aim to demonstrate that a paradigm shift is not only possible, but that it may come at a fairly low price, both in terms of effort, environmental requirements and computational infrastructures. Particularly, in Chapter 3, we will turn our attention back to the knowledge-based approaches for WSD, with a focus on the structure of the most commonly employed resources, and note that the simple injection of syntagmatic knowledge into semantic networks with a strong bias towards paradigmatic knowledge, enables knowledge-based WSD systems to attain raw performance boosts with respect to

¹⁰It is worthwhile noting that, with the size of language models growing by a factor of 10 per year, the growth in GPU memory is rapidly falling behind (Amodei and Hernandez, 2018)

¹¹<https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-p100/pdf/tesla-application-performance-guide-us-nv-r18-web.pdf>

¹²<https://openai.com/blog/ai-and-compute/>

state-of-the-art systems up to 20.2% on standard multilingual evaluation benchmarks (see 3.1).

Despite far from enabling performances similar to those of supervised models belonging to the language model era (Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020), we believe that the studies we are going to report in Chapter 3 (Maru et al., 2019; Scozzafava et al., 2020) trace a significant path. On the one hand, they show that knowledge-bases have been largely neglected, to the point of lacking essential semantic features, and hence leaving room for further progress as more analyses are conducted. On the other, they demonstrate that the wide gap in terms of performance with respect to supervised models can (and, in view of the foregoing, should) indeed be narrowed.

2.3.2 Word Senses as Discrete Entities

Current Issue Despite the longstanding warnings on the inaccurate nature of discrete word sense boundaries (Kilgarriff, 1997), researchers dealing with Word Sense Disambiguation still make extensive use of finite inventories of word meanings to tackle the task, both in knowledge-based (Agirre, de Lacalle, and Soroa, 2014; Scarlini, Pasini, and Navigli, 2020), and in supervised settings (Kumar et al., 2019; Huang et al., 2019).

Word senses have been posited to share a prototypical, fuzzy nature (Rosch and Mervis, 1975), to lie within a radial space (Brugman and Lakoff, 1988; Tyler and Evans, 2001), to involve cognitive priming processes (Meyer and Schvaneveldt, 1971; Brown, 1979), to be dependent on the intention of the speakers, and to be subject to conceptual blends (Fauconnier and Turner, 2008). According to Hanks (2000), senses are mere “meaning potentials”, i.e. collections of features that are triggered in relation to the surrounding context. Pustejovsky (1991), on the other hand, suggested a compositional, generative approach, in which senses are related by logical operations which capture semantic regularities. Earlier on, Ruhl (1989) put forward its monosemy position, claiming that words have unitary meanings which assume ad hoc functions (not stored in memory) based on their contexts of use. Finally, the exemplar model of categorisation (Medin and Schaffer, 1978; Nosofsky, 2011) argues that word meanings can be seen as points in dynamic, multidimensional spaces. In this perspective, each dimension represents a relevant perceptual feature of a given word, and word “exemplars” which are plotted closer are assumed to be similar.

To make matters even worse for WSD, as Kilgarriff (2007) argued, understanding of words is peculiar to each individual speaker. Also, one can have a strong awareness of a specific meaning of a given word, despite lacking knowledge of the extent of that same word’s range of polysemy (Talmy, 2000). All of this provides a sound justification as to why inter-annotator agreement figures for Word Sense Disambiguation never managed to

surpass the 80% threshold identified in the literature (Edmonds and Kilgarriff, 2002; Navigli, Litkowski, and Hargraves, 2007; Palmer, Dang, and Fellbaum, 2007) and, furthermore, undermines the reliability entailed by manually-crafted inventories and “gold standard” evaluation datasets (Ramsey, 2017).

In sum, the absence of clear boundaries between word senses, as well as the lack of consensus scholars share when it comes to provide a unified theory of word meaning representation, are fundamental reasons behind the extreme difficulty of dealing with WSD (Jackson, 2019).

For the purpose of illustration, consider the example (1 a), as reported in Lakoff (1987):

- (1) (a) How many windows are there in your room?

The WordNet sense inventory lists eight different senses for the word window, with the two best fitting the example in (1 a) being (2 a) and (2 b), respectively (see below).

- (2) (a) A framework of wood or metal that contains a glass windowpane and is built into a wall or roof to admit light or air.
 (b) A pane of glass in a window.

However, the word window here does not seem to refer explicitly to any of these two senses, but rather, to a combination of both.

Here, it is also worth mentioning the part of the literature whose efforts dealt with reducing the often redundant granularity of traditional sense inventories by means of clustering similar senses (Hovy et al., 2006; Snow et al., 2007) in an attempt to concurrently (i) keep the disambiguation apt at distinguishing word senses which are more clearly separated, and (ii) enabling higher inter-annotator figures, hence preventing systems from being penalized by idiosyncratic gold annotation choices in test sets.

And yet, although the reduction of the granularity of the word senses coincides with higher performances by the systems involved (Zhong, Ng, and Chan, 2008; Lacerra et al., 2020), “looser” and better distinct senses are no less affected than their more granular counterparts from having rigid boundaries between them, and from being constrained by the immutability of their inventories.

Proposed Solution With the work we introduce in Chapter 4, we overcome these limitations by proposing a unified approach to computational lexical semantics that exploits the paradigm of Definition Modeling (DM), i.e. the task of generating a gloss¹³ from

¹³To ensure better readability, here we will use the term “gloss” as a synonym of the traditional dictionary “definition”.

static or contextual embeddings (Noraset et al., 2017). Particularly, the generation of a description (*definiens*) which perfectly fits the contextual meaning of a given word or phrase (*definiendum*), being not constrained to word senses as taken from traditional inventories, represents an easy and effective way to prevent a system from dealing with the issues we detailed above.

We named our generation model *Generatory* – as a portmanteau of the words *generation* and *dictionary* –, and the reasons behind this choice lie in the fact that, by means of our approach, we are not only able to tackle Definition Modeling effectively, but also to deal with discriminative tasks such as WSD or the most recently introduced Word-in-Context (Pilehvar and Camacho-Collados, 2019, WiC).

What makes *Generatory* really stand out in comparison with previous approaches in DM (Gadetsky, Yakubovskiy, and Vetrov, 2018) is its ability to provide contextual definitions for targets of any size, from words, to phrases, to whole sentences. Hence, with the *definiendum* being represented by an arbitrary span, we can easily gloss free word combinations (e.g. *clumsy apology* or *nutty complexion*), which are a case in point of items rarely found in traditional dictionaries. Moreover, treating the target as a span also allows for the exploitation of the huge amount of knowledge contained in a pre-trained Encoder-Decoder model, i.e. BART (Lewis et al., 2019), which, in our scenario, is crucial to boost performances on the DM task, as well as on WSD and WiC.

On a final note, relying on generation – and disposing with the usage of a given sense inventory – enables the use of multiple lexicographic resources at once as training data. As a result, we can build a very resilient model, apt at providing better generalization over different tasks.

2.3.3 Inadequate Error Analysis

As we saw in Section 2.2, the origins of WSD can be traced back to the 1950s, and, notwithstanding a fluctuating trend of interest (Bar-Hillel, 1960), it continued to grow and evolve until now (Huang et al., 2019; Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020). Consequently, and with a few, focused exceptions (Loureiro et al., 2020), it comes as a surprise that no comprehensive error analysis has been carried out to investigate the real nature of system performance in WSD.

The task has received little attention in recent years, with the last traditional evaluation exercise available dating back to 2015 (Moro and Navigli, 2015), but the introduction of the paradigm of Definition Modeling to WSD (Bevilacqua, Maru, and Navigli, 2020), as well as the peak performances attained due to better disambiguation of rare senses (Blevins and Zettlemoyer, 2020) and neural approaches exploiting relational knowledge (Bevilacqua and

Navigli, 2020) are injecting new life to the field.

We believe that the renewed interest that the task is experiencing should also steer researchers to produce more insightful analyses which will in turn prevent blind runs toward increasing performance figures. To this end, it makes sense to raise questions such as “what does it mean that a system is able to disambiguate with an accuracy of 80%?” or “what does it mean for a system to have attained the human inter-annotator agreement accuracy?”

As a matter of fact, the system performance might just be a reflection of its ability to reproduce the behaviour of a few human annotators and hence, to reproduce their individual choices, rather than an actual ability to generalize and correctly infer the right answer.

To the best of our knowledge, an exhaustive error analysis is missing in the literature so far; particularly, one which accounts for:

1. the existence of a shared set of errors among state-of-the-art WSD systems, i.e. the existence of a set of apparently *non-disambiguable* instances;
2. the lexical-semantic nature of such instances, and the existence of common error patterns among them.

We aim to provide an answer to the aforementioned questions with the work that we introduce in Chapter 5,¹⁴ an analysis which shows that state-of-the-art systems in WSD share an impressive 7.4% quota of errors with respect to the standard evaluation benchmarks for English. Moreover, we provide evidence for a dramatic performance drop of more than 25 points in F1 score for the best system when instances annotated with the most frequent sense in WordNet are filtered out from the evaluation datasets, and of more than 35 points when instances in the training set are also filtered out.

To sustain our findings, we propose a sound experimental setting and produce an entirely new and fresh evaluation dataset which reproduces the conditions we identified as being the most difficult for current state-of-the-art systems. This new dataset, 42D (pron. [for-ti-tude]) is thus designed to provide an useful testbed for WSD applications, one that is devised to stress systems and to aid the evaluation of their resilience and flexibility. Additionally, this fair-sized dataset for English WSD is a welcome addition to the traditional evaluation exercise package, the first in 5 years, and also, the first to systematically cover 42 different semantic domains.

¹⁴The Chapter describes the foundational work for our article, here codenamed as *Under the Mask of Word Sense Disambiguation: a Tale of Puzzling Performances*, to be submitted at top-tier journal during Q1 2021.

2.3.4 Other Open Problems

To have an even more exhaustive picture of the WSD scene, in this Section we report a list of other open problems currently affecting the task according to the literature.

Knowledge Acquisition Bottleneck One of the longtime issues concerning WSD is the so-called *knowledge acquisition bottleneck* (Gale, Church, and Yarowsky, 1992). Particularly, with *knowledge acquisition bottleneck* (Pasini, 2020) we refer to the set of problems entailed by gathering annotated information to feed systems (especially supervised systems, given their need for sense-annotated training data). These problems stems from the fact that, in order to effectively exploit the content of a given sense inventory, a training set should contain an adequate number of tagged instances for all of the word senses therein contained, and one that reflects the standard Zipfian skewness (McCarthy et al., 2007). Given the huge efforts required to produce such resources, it is no surprise that, as of now, the few manually annotated datasets available deal with English only (Miller et al., 1993; Hovy et al., 2006).

In order to mitigate the aforementioned issues, semi-automatic (Ng, Wang, and Chan, 2003), as well as fully-automatic methods (Camacho Collados, Pilehvar, and Navigli, 2016; Pasini and Navigli, 2017; Delli Bovi et al., 2017; Scarlini, Pasini, and Navigli, 2019) to create sense-annotated “silver” data have been devised, both exploiting parallel (Taghipour and Ng, 2015) and monolingual corpora (Raganato, Delli Bovi, and Navigli, 2016; Barba et al., 2020).

Low-resource Languages’ Coverage Strictly related to the issue of the knowledge acquisition bottleneck, the paucity of training data in many languages is reflected in the absence of adequate and standardized test tools to assess the quality of systems dealing with low-resource languages, i.e. languages for which few or no machine-readable dictionaries and sense-annotated data exist.

As a matter of fact, the current evaluation benchmark for WSD (Raganato, Camacho-Collados, and Navigli, 2017) supports datasets in English only, whereas the amended version of the multilingual evaluation exercises SemEval-2013 and SemEval-2015¹⁵ allows for testing on German, Spanish, French, and Italian only. Additionally – and related to the issue entailed by assuming word senses have discrete boundaries – is the issue of employing the same inventory to multiple languages, thus expecting sense granularity to be perfectly transversal, whereas literature has provided extended evidence for cases of polysemous morphemes being collapsed into a single morpheme of another language (Talmy, 2000) and

¹⁵Made available at <https://github.com/SapienzaNLP/mwsd-datasets>.

for culturally denoted word senses that are peculiar to single dialects or languages (Hovy and Purschke, 2018).

Language Grounding On a final note, it is worth reporting how most of traditional WSD – similarly to the vast majority of NLP-related tasks – has been carried out without taking into account multi-modality, thus preventing language grounding, and hence having words being “circularly” defined by means of other words (Bender and Koller, 2020).

In substance, it can be argued that WSD has fully embraced an objectivist, algorithmic view of cognition and thought, one that Lakoff (1988) would describe as fully disembodied and involving the manipulation of abstract and meaningless symbols. Even if these symbols – in being mathematically precise – can be advantageous at a computational level,¹⁶ they completely fall short of representing the kind of experience of reality that stems from having bodies and sensory-motor capacities. If meaning should – as objectivists maintain – solely concern the relation between abstract symbols and external reality, we would not be able to represent thoughts processes requiring the projection of reality to abstraction, by means of mechanisms such as metaphor, or categorization, which are crucial in determining the polysemy of lexical items (Dirven and Verspoor, 1998; Malmkjaer, 2005).

Once we recognize how semantic relations between meanings of words only exist by virtue of human perceptual abilities, the way this affects WSD is hence striking, for different senses of a same word may not share any common property, but can be related in a perfectly identifiable way.

¹⁶Or to be listed in an enumerative fashion in traditional repositories of human knowledge, such as dictionaries.

Chapter 3

Structured Syntagmatic Information Enhancing Knowledge-based WSD

As stated in Section 2.3, our first study will bring us to inquire into the unexplored potential of knowledge bases, as well as on their related approaches and techniques. This, in turn, will lead us to reconsider such approaches as potentially valid alternatives to their supervised counterparts, both in terms of sheer performance and computational requirements.

In fact, as a major alternative to supervised WSD, knowledge-based approaches drop the requirement for large amounts of training data and high-end computational infrastructures by drawing on rich Lexical Knowledge Bases (LKB) such as WordNet (Fellbaum, 1998), and allow scaling to multiple languages effortlessly, thanks to multilingual resources such as BabelNet (Navigli and Ponzetto, 2012).

It is widely acknowledged that the performance of a knowledge-based WSD system is strongly correlated with the structure of the LKB employed (Boyd-Graber et al., 2006; Lemnitzer, Wunsch, and Gupta, 2008; Navigli and Lapata, 2010; Ponzetto and Navigli, 2010). Indeed, the knowledge available within LKBs reflects the fact that words can be linked via two types of semantic relations: paradigmatic relations – i.e. the most frequently encountered relations in LKBs – concern the substitution of lexical units, and determine to which level in a hierarchy a language unit belongs by semantic analogy with units similar to it; conversely, syntagmatic relations concern the positioning of such units, by linking elements belonging to the same hierarchical level (e.g. words), which appear in the same context (e.g. a sentence). As a case in point, a paradigmatic relation exists, independently of a given context, between the words $farm_n$ and $workplace_n$ (where a farm is a type of workplace), whereas a syntagmatic relation is entertained between the words $work_v$ and $farm_n$, e.g. in the sentence *'her husband works in a farm as a labourer.'*

While LKBs tend to focus on the paradigmatic dimension of language, such resources fall short in respect of syntagmatic relations, which are also crucial for sense disambiguation due to interconnecting co-occurring words (Navigli and Lapata, 2010; Kolesnikova and Gelbukh, 2012).

In light of the above, we believe that exploiting syntagmatic information is an encouraging research focus to be pursued in an effort to close the gap between knowledge-based and supervised Word Sense Disambiguation performance. In this Chapter, we provide further evidence that the nature of LKBs impacts on system performance: particularly, we show how the injection of syntagmatic relations – in the form of disambiguated pairs of co-occurring words – into an existing LKB biased towards paradigmatic knowledge enables knowledge-based systems to boost their efficiency drastically.

In what follows, we introduce SyntagNet (Maru et al., 2019), a novel resource consisting of manually disambiguated lexical-semantic combinations¹ (Section 3.1). By capturing sense distinctions evoked by syntagmatic relations, SyntagNet enables knowledge-based WSD systems to establish a new state of the art which, moreover, rivals the performances attained by pre-BERT (Devlin et al., 2019) supervised approaches. Moreover, we follow this direction and put forward a next-generation knowledge-based WSD system, SyntagRank (Scozzafava et al., 2020), which we make available via a Web interface and a RESTful API (Section 3.2). SyntagRank leverages the disambiguated pairs of co-occurring words included in SyntagNet to perform state-of-the-art knowledge-based WSD in a multilingual setting. Our service provides both a user-friendly interface, available at <http://syntag.net.org/>, and a RESTful endpoint to query the system programmatically (accessible at <http://api.syntag.net.org/>).

3.1 SyntagNet

Current research in knowledge-based Word Sense Disambiguation indicates that performances depend heavily on the Lexical Knowledge Base employed. Moreover, LKBs tend to have a strong bias towards paradigmatic relations, to the point of being almost completely devoid of syntagmatic knowledge. In this Section we address this deficiency and present, for the first time, a manually-curated large-scale lexical-semantic combination database which associates pairs of concepts with pairs of co-occurring words. Importantly, we prove the effectiveness of our resource by achieving the state of the art in multilingual knowledge-based

¹Here, we will use the term “lexical combination” to refer to both lexical collocations and free word associations (without considering idiomatic expressions) and the term “lexical-semantic combination” to refer to sense-annotated lexical combinations.

WSD and by matching pre-BERT supervised WSD performances when integrated into an LKB made up of WordNet and the Princeton WordNet Gloss Corpus.

3.1.1 Related Work

Several studies on knowledge-based algorithms have indicated that the LKB structure is of vital importance in determining the accuracy of sense disambiguation. In particular, it has been demonstrated that WSD performance improves dramatically when employing an LKB with a larger number of high-quality lexical-semantic relations, i.e. more connections between concepts (Boyd-Graber et al., 2006; Lemnitzer, Wunsch, and Gupta, 2008; Ponzetto and Navigli, 2010). During the last two decades, a certain amount of work has been carried out aimed at enriching LKBs with new lexical-semantic relations. To this end, knowledge has been (semi-)automatically extracted from large collections of data and integrated into lexical resources such as WordNet.

As far as semi-automatic approaches are concerned, Mihalcea and Moldovan (2001) conceived eXtended WordNet, a resource providing disambiguated glosses by means of a classification ensemble combined with human supervision. A set of manually disambiguated glosses, called the Princeton WordNet Gloss Corpus (PWNG), which inherently included syntagmatic content, was subsequently also made available in 2008 (see also 2.1.2).

The rationale behind the creation of such resources was substantiated in a knowledge-based WSD study conducted by Navigli and Lapata (2010), who hypothesized an improvement in performance by several points when enriching a semantic network with tens of lexical-semantic relations for each target word sense. To achieve this demanding goal, endeavors in the literature focused on the fully-automatic production of semantic combinations, such as those obtained by disambiguating topic signatures (Cuadros and Rigau, 2008; Cuadros, Padró, and Rigau, 2012, KnowNet and deepKnowNet) or by disentangling the concepts in ConceptNet (Chen and Liu, 2011). In fact, ConceptNet, among many other resources aimed at representing common-sense or associative knowledge such as the Small World of Words lexicon (De Deyne, Navarro, and Storms, 2013) or the RezoJDM Knowledge Base for the French language (Cousot and Lafourcade, 2017), cannot be directly exploited for WSD purposes by virtue of the lack of sense-annotated concepts accounting for polysemy and homonymy in it.

More recently, Espinosa-Anke et al. (2016) aimed at automatically enriching WordNet with collocational information by leveraging the relations between sense-level embedding spaces (ColWordNet), while Simov, Osenova, and Popov (2016) addressed the enhancement of LKBs by exploiting relations over semantically-annotated corpora as contextual information. To the same end, Simov et al. (2018) employed grammatical role embeddings to gather

new syntagmatic relations. The lack of syntagmatic information in semantic networks was also tackled by the extension of a lexical database by means of *phrasets*, i.e. sets of free combinations of words recurrently used to express a concept (Bentivogli and Pianta, 2004).

Unfortunately, due to their (semi-)automatic nature, the aforementioned resources could not inherently offer wide coverage and high precision at the same time. Compared to other resources geared towards knowledge-based WSD, the novel resource we contribute in this work features:

- a) wide coverage with a broad spectrum of possible lexical combinations, and
- b) high precision thanks to being entirely manually curated.

3.1.2 A Wide-coverage Lexical-semantic Combination Resource

In this Section, we present SyntagNet, a knowledge resource created starting from lexical combinations extracted from the English Wikipedia² and the British National Corpus (BNC) (Leech, 1992), and manually disambiguated according to the WordNet 3.0 sense inventory.

3.1.2.1 Methodology

Lexical combination extraction First of all, we employed the Stanford CoreNLP pipeline (Manning et al., 2014) to extract the dependency trees³ for all the sentences in both Wikipedia and the BNC. Then, in order to identify relevant combinations, we determined the strength of correlation between pairs of PoS-tagged, lemmatized content words⁴ w_1, w_2 , co-occurring within a sliding window of 3 words. Each candidate pair (w_1, w_2) was weighted using Dice’s coefficient multiplied by a logarithmic factor of the co-occurrence frequency:

$$score(w_1, w_2) = \log_2(1 + n_{w_1 w_2}) \frac{2n_{w_1 w_2}}{n_{w_1} + n_{w_2}} \quad (3.1)$$

where n_{w_i} ($i \in \{1, 2\}$) is the frequency of w_i and $n_{w_1 w_2}$ is the frequency of the two words co-occurring within a window.

Three filters were then applied in order to slim down the list of pairs:

1. we filtered out English stopwords according to the Natural Language Toolkit (Loper and Bird, 2002, NLTK 3.4);
2. we discarded combinations between verbs and verbs;

²November 2018 English Wikipedia dump.

³According to the Universal Dependencies v2 (<https://universaldependencies.org/u/dep/all.html>).

⁴Restricted to nouns and verbs in the WordNet dictionary.

word 1	word 2	score	sense 1	sense 2
run _v	program _n	18.07	run _v ¹⁹ (carry out a process or program)	program _n ⁷ (a sequence of instructions)
run _v	race _n	11.55	run _v ³⁷ (compete in a race)	race _n ² (a contest of speed)
run _v	farm _n	3.50	run _v ⁴ (direct or control)	farm _n ¹ (workplace with farm buildings)

Table 3.1. Examples of high-ranking lexical (left) and semantic (right) combinations, where each lemma’s subscript and superscript indicate its part of speech and sense number, respectively, in WordNet.

- we discarded combinations not linked by any of the five most frequent dependencies in our list, namely: `compound`, `doobj` (direct object), `iobj` (indirect object), `nsubj` (nominal subject) and `nmod` (nominal modifier).

Finally, we ranked the resulting lexical combination list according to the geometric mean between i) the logarithmic Dice scores and ii) the frequency count of a pair in a given PoS tag/dependency combination. We show some examples with $w_1 = \text{run}_v$, together with their final correlation score in Table 3.1 (left).

We then repeated the whole process described above, with the following changes:

- we set a sliding window of 6 words;
- we removed the constraint on the dependency selection;
- we filtered out all pairs already occurring within the first list;
- we selected only items attested in multiple English monolingual and collocation dictionaries.

Manual Disambiguation We asked eight annotators to manually disambiguate the top-ranking 20,000 lexical combinations from the first list and 58,000 lexical combinations from the second list, i.e. to associate each word in a pair (w_1, w_2) with its most appropriate senses in WordNet (in Table 3.1 (right) we show the senses chosen by the annotators for the corresponding lexical combinations).

The eight annotators shared a background in linguistics (Master’s Degree with a minimum C1 English proficiency level) and were well acquainted with WordNet. In order to facilitate the annotation process, we provided each annotator with a unique batch of lexical combinations in a simple interface; for each pair, the annotators visualized all the synsets for each word of the combination (along with WordNet definitions and examples), and a context

type	#NN	#NV	#total
Lexical combinations	25,568	52,432	78,000
Semantic combinations	26,770	61,249	88,019
Unique lemmas	10,218	3,786	14,004
Unique synsets	14,204	6,422	20,626

Table 3.2. Data for SyntagNet 1.0. Rows: type of data. Columns: number of NOUN-NOUN combinations (#NN), number of NOUN-VERB combinations (#NV), total number of combinations (#total).

of up to 25 random sentences in which the combination was extracted. The annotators were asked to input the sense numbers associated with their chosen synsets for both the words in a given pair. Since the combinations can carry different meanings depending on the context, the annotators were allowed to assign multiple senses to the same word in a given combination (e.g. judge in the “public official” sense vs. the “evaluator” sense in the (judge_n, decide_v) lexical combination).

As a further measure to ensure quality, the annotators were also asked to skip the annotation of lexical combinations

- i) carrying mistakes due to the automatic parsing process;
- ii) for which none of the available senses in WordNet would fit the context, (e.g. (cause_v, flooding_n) with flooding_n being monosemous in WordNet and carrying the “implosion therapy” meaning, but not the “water inundation” one);
- iii) reflecting idiomatic expressions, (e.g. (jump_v, gun_n) as in the “jump the gun” idiom);
- iv) which were multi-word Named Entities. (e.g. (crystal_n, palace_n) referring to the area in South London or to the English professional football club).

Overall, the annotators covered 78,000 lexical combinations, and obtained 88,019 semantic combinations linking 20,626 WordNet 3.0 nodes, i.e. unique synsets, with a relation edge (for a full data overview, see Table 3.2).

We periodically timed the annotators by considering the number of annotations produced on a daily basis, obtaining an average value of 42 disambiguated combinations per hour (1 minute and 26 seconds per word pair). Overall, the annotation process took a period of 9 months. To determine the reliability of the annotations, we calculated the minimum inter-annotator agreement between pairs of annotators on a random sample of 500 combinations.

For each of the 500 lexical combinations used to compute the inter-annotator agreement, the annotators were exceptionally asked to disambiguate the two target words in all of the 25 sentences provided, thus leading to a figure of 25,000 single instances disambiguated per annotator, resulting in a substantial agreement ($\kappa = 0.71$). Moreover, we found that most of the disagreement instances arose out of valid alternative tags, rather than factual errors, due to the fine granularity of the WordNet sense inventory.

3.1.3 Experimental Setup

We now present the setup of our evaluation, carried out to assess the effectiveness of SyntagNet when employed for knowledge-based WSD.

Disambiguation algorithm We performed our experiments employing UKB,⁵ (Agirre, de Lacalle, and Soroa, 2014) a state-of-the-art system for knowledge-based WSD, which applies the Personalized Page Rank (PPR) algorithm (Haveliwala, 2002) to an input LKB. We used its *PPR_{w2w}* single-sentence context disambiguation method, which initializes the PPR vector using the context of the target word in a given sentence, while excluding the contribution of the target word itself.

The outcome of the PPR is a probability distribution over the WordNet synsets, based on the initialization provided by the sentence context. The central idea of this approach is to let the neighboring words determine which sense – among those listed for the target word in the WordNet sense inventory – has more pertinence to the context. Also, in order to account for the fact that some senses of a given polysemous word are more frequent than others, UKB weights the relations within the graph according to the rate with which a specific sense occurs as a tagged instance in various semantically annotated corpora. This information is represented by a frequency counter into the dictionary file which maps the lexical-semantic relations between lemmas and senses in WordNet.⁶

Evaluation benchmarks and measures We used five test sets standardized with WordNet 3.0 (Raganato, Camacho-Collados, and Navigli, 2017) including the English all-words tasks from Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli, Jurgens, and Vannella, 2013) and SemEval-2015 (Moro and Navigli, 2015). To run experiments on multilingual WSD, we used the last two of the foregoing datasets, which also include German, Spanish, French and

⁵Version 3.2 (<http://ixa2.si.ehu.es/ukb/>)

⁶For a fuller discussion of PPR, see Sections 3.2.2 and 3.2.3.2, later in this Chapter.

resource	#relations
WNG (baseline)	671,779
WNG+KnowNet20	+520,682
WNG+deepKnowNet95d	+522,880
WNG+BabelNet 4.0	+9,447,341
WNG+eXtended WordNet	+551,551
WNG+CoIWordNet	+8,424
WNG+SyntagNet	+88,019

Table 3.3. Number of relations (right) added to the WNG (WordNet+PWNG) baseline (top) from each different resource (left).

Italian, employing, as sense inventory, the synset lexicalizations provided in BabelNet 4.0.⁷ Whereas UKB uses WordNet to gather the lemma-sense associations for English, in order to perform multilingual tests, language-specific inventories for each language are needed. Thus, we made use of the BabelNet 4.0 inventory to gather, for each WordNet synset, the corresponding synset lexicalizations in other languages. As a means to ensure high quality at this stage, we avoided lexicalizations coming from automatic translations. Furthermore, for each lexicalization entry in a given language, we assigned a confidence score to each of its synsets, taking into account the number of resources in BabelNet providing that specific lexicalization. Consequently, we were able to produce a sense dictionary matching the structure of the one provided with UKB, preserving the information concerning sense frequencies.⁸

As customary, we computed precision, recall and F1, which in our case coincided, due to UKB always outputting a sense for each target word.

LKBs For the purposes of our evaluation we measured the performance obtained with UKB when combined with different LKBs. As our baseline we used WordNet + PWNG, which is the best configuration of UKB according to its authors. We also evaluated the following LKBs when integrated separately on top of our baseline:

⁷<http://babelnet.org>

⁸Later in this Chapter (Section 3.2.3.2), we will employ again a similar strategy to simulate word sense distribution.

resource	Sens2	Sens3	Sem07	Sem13	Sem15	All
WNG (baseline)	69.2	65.9	54.9	66.8	70.7	67.1
WNG+KnowNet20	67.2	65.8	53.8	67.3	71.5	66.6
WNG+deepKnowNet95d	66.9	64.9	53.6	66.9	71.6	66.2
WNG+BabelNet 4.0	67.5	64.1	53.0	67.6	66.9	65.6
WNG+eXtended WordNet	67.7	65.7	52.3	67.6	71.0	66.7
WNG+ColWordNet	69.2	65.9	54.1	66.7	70.7	67.1
WNG+SyntagNet	71.2	71.6	59.6	72.4	75.6	71.5

Table 3.4. F1 scores (%) for English all-words fine-grained WSD. Each row displays results scored by a specific resource combined with the WNG (WordNet+PWNG) baseline. Statistically-significant differences, according to a χ^2 test ($p < 0.01$), compared to the baseline (first row), are underlined.

- i) the best configurations of KnowNet⁹ and deepKnowNet¹⁰,
- ii) the subgraph of BabelNet 4.0 induced by WordNet 3.0,
- iii) eXtended WordNet¹¹,
- iv) ColWordNet¹²,
- v) SyntagNet (cf. Section 3.1.1).

All of the aforementioned LKBs (whose size in terms of lexical-semantic relations is shown in Table 3.3) are available for download.

3.1.4 Experimental Results

English WSD As shown in Table 3.4, SyntagNet enabled UKB to achieve the best results in the English all-words disambiguation tasks, attaining 4.4 overall points above the WNG baseline, which is the only statistically-significant improvement across LKBs. Furthermore, results for the individual datasets exhibit statistically-significant improvements over the baseline on two out of five datasets. We attribute this result to the fully manual nature of

⁹<http://adimen.si.ehu.es/web/KnowNet>

¹⁰<http://adimen.si.ehu.es/web/deepKnowNet>

¹¹<http://www.hlt.utdallas.edu/~xwn/>

¹²<http://bitbucket.org/luisespinoza/cwn/>

SyntagNet, in contrast to the noisy character of the other LKBs. As a matter of example, most BabelNet relations come from Wikipedia links, which introduces noise and rarely captures relations between verbs and nouns. On the other hand, SyntagNet captures syntagmatic relations systematically: as attested in Table 3.3, BabelNet provides 9,447,331 relations; SyntagNet instead contributes only 88,019, while performing considerably better. A further justification of our results comes from an analysis we performed on relation samples from the various LKBs: we collected 500 random relations for each LKB we experimented with, and manually tagged each of them as syntagmatic or paradigmatic, revealing that their syntagmatic contribution ranges from 39% (deepKnowNet) to 54% (eXtended WordNet). The fully syntagmatic nature of SyntagNet, instead, effectively blends in with the complementary information available in the baseline (63% of the relations in WNG are paradigmatic).

Table 3.5 compares UKB + SyntagNet against the best pre-BERT supervised English WSD systems (Yuan et al., 2016; Melacci, Globo, and Rigutini, 2018; Uslu et al., 2018): none of the differences across datasets between the best performing supervised system and SyntagNet is statistically significant according to a χ^2 test ($p < 0.01$), meaning that SyntagNet enables knowledge-based WSD to rival supervised approaches that do not rely on pre-trained language models.

Finally, to better contextualize these results in a comprehensive comparison framework that shows the extent of the huge discrepancy that still exists between approaches based on LMs and knowledge-based systems, we report in the same Table the results for two of the state-of-the-art approaches for English WSD that currently employ LMs in their architectures (Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020).

Multilingual WSD As regards our multilingual evaluation, SyntagNet enabled UKB to attain the best overall result (see Table 3.6), which is a statistically-significant improvement of 2.1 points over the baseline. With respect to the comparison against the best systems¹³ (Table 3.7), SyntagNet provides a statistically-relevant boost of 4.6 points in relation to the aggregate score of the compared systems (second to last row), attaining state-of-the-art results on five out of the six datasets taken into account. Similarly to the English setting, in order to show the extent of the wide performance gap between knowledge-based systems and systems based on LMs, we also report in Table 3.7 results against a state-of-the-art system for multilingual WSD (Bevilacqua and Navigli, 2020).

¹³As of the time of writing.

system	Sens2	Sens3	Sem07	Sem13	Sem15	All
LSTM \bullet	73.8	71.8	63.5	69.5	72.6	71.5
IMSC2V $_{+PR}$ ∞	73.8	71.9	63.3	<u>68.2</u>	72.8	71.2
fastSense \triangle	73.5	73.5	62.4	<u>66.2</u>	73.2	71.1
UKB+SyntagNet	71.2	71.6	59.6	72.4	75.6	71.5
<i>BEM</i>	<i>79.4</i>	<i>77.4</i>	<i>74.5</i>	<i>79.7</i>	<i>81.7</i>	<i>79.0</i>
<i>EWISER</i>	<i>80.8</i>	<i>79.0</i>	<i>75.2</i>	<i>80.7</i>	<i>81.8</i>	<i>80.1</i>

Table 3.5. F1 scores (%) of UKB+SyntagNet (middle) against the best pre-BERT supervised systems (top) and LM-based systems (*italics*, bottom) for English all-words WSD. Reported systems (pre-BERT, top): \bullet Yuan et al. (2016), ∞ Melacci, Globo, and Rigutini (2018), \triangle Uslu et al. (2018). Reported LM-based systems (*italics*, bottom, first to last): Blevins and Zettlemoyer (2020), Bevilacqua and Navigli (2020). Statistically-significant differences against our results are reported for pre-BERT systems only, and are underlined according to a χ^2 test, $p < 0.01$.

3.1.5 Impact of LKB Size

Finally, we graphed the increase in WSD performance obtained when progressively enriching the baseline UKB graph with random samples of 10, 000 SyntagNet relations at each step. As illustrated in Figure 3.1, the improvements in the English and multilingual settings, respectively, present a growing trend according to a linear regression analysis of the data. This demonstrates that our relations are high-quality and effective for WSD, while leaving room for further improvement as more relations are added in the future.

3.2 SyntagRank

In order to make the results shown in Section 3.1 accessible to the research community, we introduce a Web interface for SyntagRank (Scozzafava et al., 2020), our next-generation knowledge-based multilingual WSD system, which applies the Personalized PageRank (PPR) algorithm (Haveliwala, 2002) to an LKB made up of WordNet, PWNG and the lexical-semantic syntagmatic combinations available in the SyntagNet resource. SyntagRank constitutes a fundamental addition to the family of knowledge-based systems for multilingual WSD, especially, given that (i) it is the first to systematically employ a Lexical Knowledge Base containing explicit and high-quality syntagmatic relations, and (ii) it is freely available

resource	IT _{S13}	ES _{S13}	DE _{S13}	FR _{S13}	IT _{S15}	ES _{S15}	All
WNG (WordNet+PWNG)	71.4	71.2	68.0	69.6	62.2	58.1	67.2
WNG+KnowNet20	71.6	73.1	68.3	70.4	61.4	59.9	67.9
WNG+deepKnowNet95d	71.4	71.9	67.7	70.5	62.4	58.7	67.5
WNG+BabelNet 4.0	73.8	71.6	69.9	67.1	62.4	57.8	67.6
WNG+eXtended WordNet	72.4	71.8	68.5	69.3	62.4	58.9	67.7
WNG+ColWordNet	71.4	71.0	68.0	69.3	61.9	57.8	67.0
WNG+SyntagNet	74.2	73.4	66.9	72.7	65.0	61.2	69.3

Table 3.6. F1 scores (%) for multilingual all-words fine-grained WSD. Each row displays results scored by a specific resource combined with the WNG (WordNet+PWNG) baseline. Statistically-significant differences, according to a χ^2 test ($p < 0.01$), compared to the baseline (first row), are underlined.

online via a user-friendly interface available at <http://syntagnet.org/>.¹⁴

3.2.1 Lexical Knowledge Bases

The disambiguation algorithm employed by SyntagRank relies on an underlying LKB, which can be seen as a graph made up of nodes and links that represent concepts and semantic relations, respectively (see Section 2.1.1). Particularly, SyntagRank’s reference graph is the result of the union of three different LKBs:

1. **WordNet** (see Section 2.1.1);
2. **PWNG** (see Section 2.1.2);
3. **SyntagNet** (see Section 3.1).

3.2.2 Personalized PageRank

The algorithm employed by SyntagRank is the Personalized PageRank (PPR), a variant of the standard PageRank that was first introduced by Brin and Page (1998), and that has already been applied for WSD purposes (Agirre and Soroa, 2009; Agirre, de Lacalle, and Soroa, 2014).

¹⁴Besides the user interface, SyntagRank can also be accessed via a RESTful endpoint at <http://api.syntagnet.org/>.

system	IT _{S13}	ES _{S13}	DE _{S13}	FR _{S13}	IT _{S15}	ES _{S15}	All
BLSTM†	<u>62.0</u>	66.4	69.2	<u>55.5</u>	-	-	-
UMCC-DLSI★	<u>65.8</u>	71.0	62.1	<u>60.5</u>	-	-	-
T-O-M◦	<u>68.2</u>	66.9	63.2	<u>60.5</u>	-	-	-
SUDOKU-RUN1◇	-	-	-	-	59.9	56.0	-
SUDOKU-RUN2◇	-	-	-	-	56.9	57.1	-
Best system‡	<u>68.2</u>	71.0	69.2	<u>60.5</u>	59.9	57.1	<u>64.7</u>
UKB+SyntagNet	74.2	73.4	66.9	72.7	65.0	61.2	69.3
<i>EWISER</i>	<i>77.7</i>	<i>78.8</i>	<i>80.9</i>	<i>83.6</i>	<i>71.8</i>	<i>69.5</i>	<i>77.5</i>

Table 3.7. F1 scores (%) of UKB+SyntagNet (middle) against the best systems for pre-BERT (top) and LM-based (*italics*, bottom) multilingual all-words WSD. Reported systems (pre-BERT, top): † Raganato, Delli Bovi, and Navigli (2017), ★ Gutiérrez Vázquez et al. (2010), ◦ Pasini and Navigli (2017), ◇ Manion (2015), ‡ result obtained by aggregating the outputs of the best systems for each dataset. Reported LM-based system (*italics*, bottom): Bevilacqua and Navigli (2020). Statistically-significant differences of pre-BERT systems against our results are underlined according to a χ^2 test, $p < 0.01$.

In the original PageRank, a graph is traveled from node to node in order to determine the probability that each node has to be reached, starting from another point in the same graph. The relative weight of each node is thus balanced at first and then, by means of several iterations over the graph (also called walks), each node sees its weight adjusted accordingly with the number of ingoing and outgoing connections. In the Personalized PageRank instead, the weight is not distributed equally among all nodes at first, but the initial probability mass is shared by a finite set of nodes, each representing a target word to be disambiguated. The outcome of the PPR algorithm of SyntagRank is therefore represented by a single vector which encodes the probability distributions for each node in the graph, as obtained by starting from a restricted set of nodes.

PPR Implementation Details The damping factor used by SyntagRank is 0.85. The number of iterations (walks) that the algorithm carries out over the graph is instead tied to a threshold. Specifically, as soon as the variation between the scores of any node in the graph (computed at two successive iterations) falls below 10^{-4} , SyntagRank stops performing further iterations.

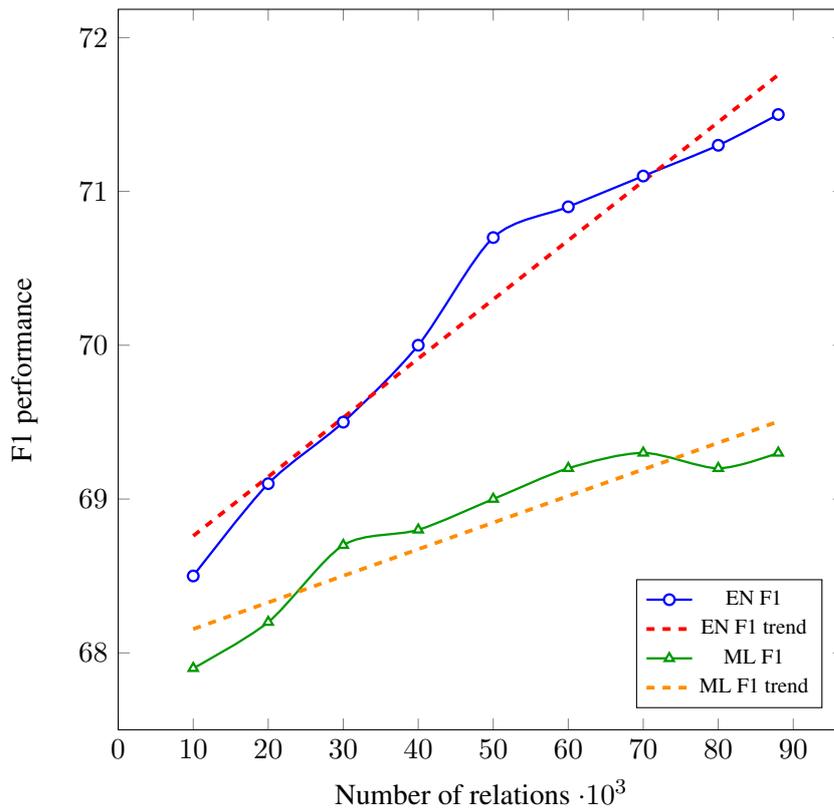


Figure 3.1. Impact of growing samples of SyntagNet over the baseline on overall F1 for English (EN F1 line) and multilingual (ML F1 line) evaluation datasets.

3.2.3 System Architecture

Despite the fact that PPR has already been explored as a possible means to tackle knowledge-based disambiguation (Agirre and Soroa, 2009; Agirre, de Lacalle, and Soroa, 2014), SyntagRank represents an optimized and completely rebuilt system with respect to its predecessors, furthermore, enabling unprecedented performances in a multilingual environment thanks to the exploitation of syntagmatic knowledge.

The architecture of SyntagRank (Figure 3.2) can be explained by means of three different modules, i.e. stages of processing: (i) the multilingual NLP pipeline, (ii) the candidate retrieval stage, and (iii) the disambiguator module.

3.2.3.1 Multilingual NLP Pipeline

The first module of SyntagRank is represented by a multilingual NLP pipeline, by means of which the system can process plain, raw data. In fact, when the user inputs an unprocessed

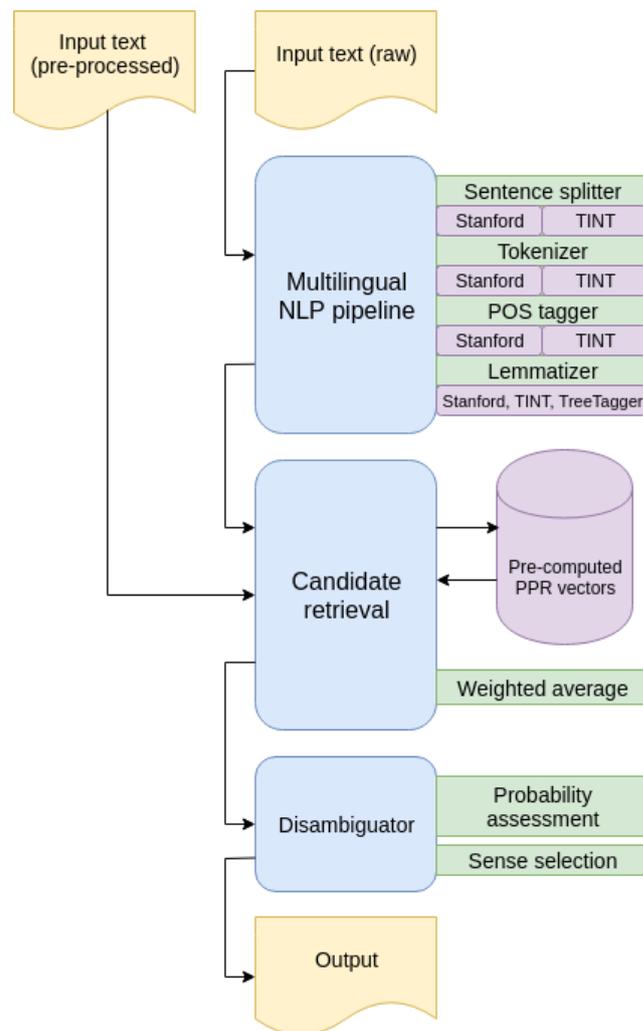


Figure 3.2. Architecture diagram of SyntagRank.

text, SyntagRank uses its multilingual pipeline so as to pre-process it via tokenization, sentence splitting, lemmatization and PoS tagging.

Currently, the system can process texts in five different languages, namely, English, French, German, Spanish, and Italian. According to the input language, SyntagRank will employ one or more of the following resources:

- the Stanford **CoreNLP** suite (Manning et al., 2014);
- **TreeTagger** (Schmid, 1995);
- the models provided by The Italian NLP Tool (Palmero Aprosio and Moretti, 2016, **TINT**).

3.2.3.2 The Candidate Retrieval Stage

English After pre-processing the input by means of the multilingual NLP pipeline (see Section 3.2.3.1), SyntagRank proceeds with identifying the set of WordNet synsets (candidates) for which a lexicalization coincident with one of the content words in the input context exists. At this stage, each content word in the input context, in turn, becomes a target word to be disambiguated by the system. The set of collected synsets gets reduced accordingly, in line with the word-to-word heuristics detailed in Agirre, de Lacalle, and Soroa (2014) (see also Section 3.1.3). Specifically, in order to prevent the most frequent sense of the target word to affect the probability distribution of the graph (see also Calvo and Gelbukh (2015)), the synsets belonging to the target word are temporarily removed from the set. As a result, the set of collected concepts C , which will represent the starting nodes of the PPR algorithm, will include only the synsets of the content words that are not the target word itself.

In order to reduce execution times significantly, we pre-computed PPR vectors for each node in SyntagRank’s graph. According to the Linearity Theorem of Jeh and Widom (2003) in fact, the PPR vector computed starting from C is equivalent to the weighted average of the PPR vectors calculated using each of the nodes in C as single starting points. In light of this, SyntagRank can obtain the final PPR vector for an input sentence just by performing the weighted average of the pre-computed vectors for the content word concepts. Thus, the PPR vector for a precise context (i.e. an input sentence) is calculated simply by determining the weighted average of the pre-computed PPR vectors for each of its nodes.

The weight factor $p(w, s)$, for each candidate s associated with a content word w , is computed as follows:

$$p(w, s) = \frac{1}{N * |senses_w|} freq_{ws} \quad (3.2)$$

where N is the number of content words in the input sentence and $senses_w$ is the set of sense candidates associated with w . In addition, since already discussed in the literature (Calvo and Gelbukh, 2015; Postma et al., 2016; Pasini, Scozzafava, and Scarlini, 2020) we accounted for the bias that knowledge-based system share towards the most frequent word senses (according to their rank in WordNet, and, consequently, to their distribution in SemCor), by including the parameter $freq_{ws}$, i.e the normalized value resulting from the number of occurrences for a given word sense in SemCor, divided by the total number of occurrences for all the senses of the same word.

Multilingual So far, we detailed the candidate retrieval process for the English language, for which the WordNet 3.0 inventory is employed. Still, when it comes to other languages,

even though synsets – being representative of concepts – are often assumed to be approximately language agnostic,¹⁵ SyntagRank needs to map words from different languages to WordNet concepts in order to perform disambiguation. To this end, we exploit BabelNet (see Section 2.1.1), which provides alignments for our four non-English working languages (French, German, Spanish and Italian) to WordNet 3.0 synsets.

BabelNet alignments are though induced through automatic methods, hence their quality might often be inadequate for our purpose. Also, the lack of a sense-annotated corpus – such as SemCor – for languages different from English makes the computation of the $freq_{ws}$ parameter we introduced earlier in this Section impossible. Our solution to deal with both of these issues concurrently is to replace the $freq_{ws}$ value with a normalized confidence score that we assigned to each of the resources from which BabelNet retrieves its lexicalizations (e.g. OmegaWiki or Wikidata), based on the quality of such lexicalizations, as determined by an in-house qualitative study we conducted. As a result, SyntagRank can exploit the average confidence score among all the resources providing a given lexicalization, instead of the standard $freq_{ws}$ value (see also Section 3.1.3).

3.2.3.3 Disambiguator

At this stage, SyntagRank has already collected the PPR vectors for each candidate and computed their weighted average as described in Section 3.2.3.2. Thus, the disambiguator module mainly serves to associate a word sense with a given target in context, simply by browsing through the probability values determined by the averaged PPR vector in order to select – as the final system prediction – the sense with the highest value.

3.2.4 Web Interface

The Web interface of SyntagRank, accessible at <http://syntag.net.org/>, is shown in Figure 3.3. Below, we detail each of its components.

A. Query The Web interface allows the user to input raw text in the query field (either single words, phrases, or whole sentences). If the input matches an entry in the SyntagNet database (see Section 3.1), the interface switches to the SyntagNet Explorer (see below, Section 3.2.4.1). Otherwise, SyntagRank proceeds to disambiguate the query.

B. Language Selection The user can select one among the five languages currently supported by SyntagRank to type the query in, namely, English, German, French, Spanish and

¹⁵Even though, as also reported in Section 2.3.4, word sense granularity can significantly differ from language to language.

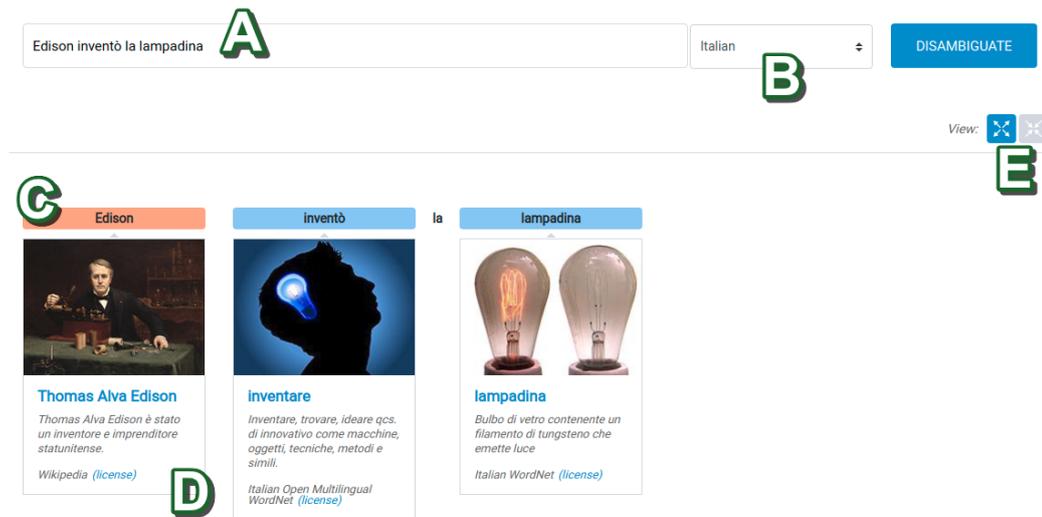


Figure 3.3. User interface of SyntagRank when the Italian language is selected and the sentence ‘Edison inventò la lampadina’ (Edison invented the light bulb) is typed as input query. Disambiguation results are displayed in extended view by default. Overlaying letters over the image are detailed in Section 3.2.4.

Italian.

C. Disambiguated Sentence The disambiguation results are displayed with tokens highlighted in different colors for *Concepts* (blue) and *Named Entities* (orange).

D. Disambiguated Token Each content word that SyntagRank disambiguates is accompanied by a tooltip which shows the image, word sense, and definition, as retrieved from the corresponding entry in BabelNet 4.0.

E. View Selection The disambiguated sentence can be explored either in extended or in compact form. In the extended view, the sentence is displayed as a horizontal slider, and all the tooltip information is visible. Using the compact view instead, tooltip information is hidden, and is only shown when the mouse cursor hovers over a content word.

3.2.4.1 SyntagNet Explorer

The Web interface at <http://syntag.net.org/> provides both access to the SyntagRank knowledge-based disambiguation system, but also to the full SyntagNet resource of lexical-semantic combinations. In fact, by typing into the query bar a word or MWE

mouse COLLOCATE



mouse
noun
Any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails

 <i>n. body</i>	 <i>n. cat</i>	 <i>n. cheese</i>
 <i>n. ear</i>	 <i>n. elephant</i>	 <i>n. embryo</i>
 <i>n. experiment</i>	 <i>n. hole</i>	 <i>n. house</i>
 <i>n. nest</i>	 <i>n. rat</i>	 <i>n. trap</i>



mouse
noun
A hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad

 <i>n. ball</i>	 <i>n. button</i>	 <i>n. click</i>
 <i>n. computer</i>	 <i>n. cursor</i>	 <i>n. pad</i>
 <i>n. trackball</i>	 <i>v. move</i>	 <i>v. operate</i>
 <i>v. roll</i>	 <i>v. snap</i>	 <i>v. use</i>

Figure 3.4. User interface of the SyntagNet Explorer when the English word *mouse* is typed as input query.

which is present in SyntagNet¹⁶ (an autocomplete function will provide the user with search suggestions), the interface will switch to the SyntagNet Explorer (Figure 3.4). The SyntagNet Explorer displays a list of boxes, each containing a sense of the input word/MWE. Senses in the list are ordered according to (i) PoS tag and (ii) sense frequency (in line with BabelNet 4.0). On the left side (blue background), the boxes show information for word senses, along with PoS tags, sense definitions and illustrations. By clicking on a sense name, the corresponding BabelNet entry will open in a separate tab. On the right side (white background), all the lexical-semantic items (collocates) linked with the corresponding word senses via SyntagNet are listed. Further information about collocates is provided by hovering the mouse over each item. Finally, clicking on a collocate will start a new query with the selected word.

¹⁶At the time of writing, the SyntagNet Explorer is available for the English language only.

System	S2	S3	S07	S13	S15	All
Babelfy	<u>67.0</u>	<u>63.5</u>	51.6	<u>66.4</u>	<u>70.3</u>	<u>65.5</u>
UKB	68.8	<u>66.1</u>	53.0	68.8	<u>70.3</u>	<u>67.3</u>
SyntagRank	71.6	72.0	59.3	72.2	75.8	71.7

Table 3.8. F1 scores (%) for English all-words fine-grained WSD. Statistically-significant differences against our results are underlined according to a χ^2 test, $p < 0.01$. Results under “All” refer to the concatenation of the English datasets.

System	IT _{S13}	ES _{S13}	DE _{S13}	FR _{S13}	IT _{S15}	ES _{S15}	All
Babelfy	<u>66.6</u>	69.5	<u>69.4</u>	<u>56.9</u>	-	-	-
SyntagRank	72.1	74.1	76.4	70.3	69.0	63.4	71.2

Table 3.9. F1 scores (%) for multilingual all-words fine-grained WSD. Statistically-significant differences against our results are underlined according to a χ^2 test, $p < 0.01$. Results under “All” refer to the concatenation of the multilingual datasets.

3.2.4.2 RESTful API

SyntagRank can also be queried programmatically via RESTful API. The main difference with the interface we described in Section 3.2.4 lies in the fact that our API allows the user to input already pre-processed text, as well as performing standard, plain text queries. Exhaustive details concerning the usage and parameters description for our RESTful API can be found at <http://syntag.net.org/api-documentation>.

3.2.5 Evaluation

Similarly to the experiments we conducted with UKB and SyntagNet (see Section 3.1.3), we tested the performance of SyntagRank on the five English all-words WSD evaluation datasets standardized according to WordNet 3.0 in the framework of Raganato, Camacho-Collados, and Navigli (2017), i.e. Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli, Jurgens, and Vannella, 2013), and SemEval-2015 (Moro and Navigli, 2015). This time, to appraise SyntagRank in a multilingual setting, we used the German, Spanish, French and Italian annotations available in the amended version of the SemEval-2013 and SemEval-2015

evaluation datasets,¹⁷ which is accordant with the BabelNet API 4.0.1 graph and enables testing on a larger number of instances than hitherto.

We report F1 scores for SyntagRank in the English (Table 3.8), and multilingual (Table 3.9) settings, along with providing comparisons to the best configurations of two distinct graph-based disambiguation systems, namely, Babelfy (Moro, Raganato, and Navigli, 2014) and UKB (Agirre, de Lacalle, and Soroa, 2014).

As can be seen, the performances attained by SyntagRank largely surpass those of its competitors¹⁸ on both the English and multilingual settings. These results once more testify to the significant impact that syntagmatic relations carry with respect to disambiguation performance, particularly, when a PPR algorithm is employed to walk on a graph in which nodes are not exclusively connected by means of paradigmatic relations.

3.3 Conclusion

In this Chapter we presented (a) SyntagNet, a new, wide-coverage, manually-curated resource of lexical-semantic combinations (see Table 3.10 for an excerpt of disambiguated lexical combinations featured in the resource) and (b) SyntagRank, our state-of-the-art knowledge-based system for multilingual Word Sense Disambiguation using syntagmatic information, along with providing information concerning the use of SyntagRank’s Web interface and RESTful API, accessible at <http://syntagnet.org/> and <http://api.syntagnet.org>, respectively.

Wrapping up this Chapter, the injection of syntagmatic knowledge into an LKB biased towards paradigmatic relations has proven a key in enabling knowledge-based systems to significantly boost their performances, to the point of rivaling the pre-BERT supervised system on English, and surpassing the overall performance of multilingual systems by several points. Despite still far from enabling the performances of their supervised counterparts, we have also shown how additional syntagmatic relations could have the power to make knowledge-based systems even more effective, eventually turning them into a viable alternative to supervised systems; an alternative that will, at the same time, curb the need for prohibitive and environmentally perilous infrastructures.

¹⁷Made available at <https://github.com/SapienzaNLP/mwsd-datasets>.

¹⁸For the purpose of these experiments, we set a threshold $T = 0.4$ for the PPR values of any given sense; for values failing to reach the threshold, the most frequent sense (according to WordNet) was chosen instead as the result of the disambiguation.

words	abbey _n	found _v
senses	<i>A monastery ruled by an abbot</i>	<i>Set up or found</i>
context	He later <u>founded</u> the <u>Abbey</u> of Monte Cassino.	
words	ceiling _n	determine _v
senses	<i>An upper limit on what is allowed</i>	<i>Fix conclusively or authoritatively</i>
context	In order to <u>determine</u> the price <u>ceiling</u> on these products [...]	
words	floor _n	sea _n
senses	<i>The bottom surface of a body of water</i>	<i>A division of an ocean</i>
context	Sunlight does not penetrate to the <u>sea floor</u> [...]	
words	gem _n	glitter _v
senses	<i>A crystalline rock polished for jewelry</i>	<i>Be shiny, as if wet</i>
context	The gem <u>glittered</u> nobly in the sunbeams.	
words	load _n	elevator _n
senses	<i>Weight to be borne or conveyed</i>	<i>Lifting device</i>
context	When the <u>load</u> is too much for an <u>elevator</u> to hold [...]	
words	mob _n	shout _v
senses	<i>A disorderly crowd of people</i>	<i>Utter in a loud voice</i>
context	A Somali <u>mob</u> <u>shouts</u> anti-UN slogans [...]	
words	obelisk _n	raise _v
senses	<i>A stone pillar</i>	<i>Construct, build or erect</i>
context	Egyptians might have <u>raised</u> massive <u>obelisks</u> [...]	
words	reaction _n	provoke _v
senses	<i>A bodily process due to a stimulus</i>	<i>Call forth (responses)</i>
context	The doctors are hoping to <u>provoke</u> a <u>reaction</u> .	
words	tea _n	cup _n
senses	<i>A beverage</i>	<i>The quantity a cup will hold</i>
context	How many <u>cups</u> of <u>tea</u> have you been drinking lately?	
words	vibration _n	cause _v
senses	<i>A shaky motion</i>	<i>Give rise to</i>
context	A rotating component will <u>cause</u> <u>vibration</u> [...]	

Table 3.10. An excerpt of 10 disambiguated lexical combinations taken from SyntagNet. Each block shows: lemmas for the lexical combination along with part of speech indication in subscript (words); shortened version of WordNet 3.0 glosses for the senses associated with the lemmas in a given lexical combination (senses); and a sample context in which the lexical-semantic combination occurs (context).

Chapter 4

Recasting Word Sense Disambiguation via Contextual Definition Generation

In the previous Chapter, we provided evidence for the lack of structured and comprehensive information making up existing LKBs, thus providing a means to effectively narrow the gap between knowledge-based architectures and supervised systems which call for increasingly prohibitive computational infrastructures (see also 2.3.1).

Nonetheless, despite being a milestone on its own, the work we described in Chapter 3, as much as any other work carried out by employing traditional knowledge bases such as WordNet, is not exempt from the burden of having to deal with a strictly enumerative approach and hence, to resort to finite inventories of discrete word senses (see 2.3.2).

In this Chapter, we show that it is unnecessary to embrace a discrete view of word senses in order to automatically determine the contextual meaning of an ambiguous target in context.

As a matter of fact, in our model, *Generatory*, we use an innovative span-based encoding scheme, which we employ in order to fine-tune an English pre-trained Encoder-Decoder system to generate definitions. Despite disposing of finite sense inventories, we provide evidence that our model can be employed effectively: in fact, *Generatory* is able to outdo the current state of the art in a generative task such as Definition Modeling while, at the same time, reaching and surpassing the performances of state-of-the-art systems in fully discriminative tasks such as Word Sense Disambiguation and Word-in-Context.

Moreover, we show how *Generatory* attains huge improvements on several zero-shot test beds – including a new dataset of definitions for adjective-noun phrases – due to the

exploitation of training data from multiple inventories at once.

The main contributions of our approach can be summarized as follows:

1. we propose the use of a single conditional generation architecture to perform English DM, WSD and WiC;
2. while dropping the need of choosing from a predefined sense inventory, our model achieves competitive to state-of-the-art results;
3. thanks to our encoding scheme, we can represent the *definiendum* as a span in the context, thus enabling definition generation for arbitrary-sized phrases, and seamless usage of BART (Lewis et al., 2019), a pre-trained Encoder-Decoder system;
4. we release a new evaluation dataset to rate glosses for adjective-noun phrases.

We envision many possible applications for Generationary, such as aiding text comprehension, especially for second-language learners, or extending the coverage of existing dictionaries.

4.1 Related Work

The original goal of Definition Modeling (DM) was to generate a human-readable natural language definition to make the content of static word embeddings explicit (Noraset et al., 2017).¹

In order to take into account polysemy, several approaches for DM later investigated the usage of sense embeddings (Gadetsky, Yakubovskiy, and Vetrov, 2018; Chang et al., 2018; Zhu et al., 2019), but none of these works actively exploited the actual contexts in which the word senses appear. Others have made a fuller use of the sentence surrounding the target, with the goal of explaining the meaning of a word or phrase, as embedded in its local context (Ni and Wang, 2017; Mickus, Paperno, and Constant, 2019; Ishiwatari et al., 2019). However, these approaches have never explicitly dealt with WSD, and showed limits with respect to the marking of the target in the context encoder, preventing an effective exploitation of the context, and making DM overly reliant on static embeddings or surface form information. For example, in the model of Ni and Wang (2017), the encoder is unaware of the contextual target, whereas Mickus, Paperno, and Constant (2019) use a marker embedding to represent targets limited to single tokens.

Finally, Ishiwatari et al. (2019) replace the target with a placeholder, and the burden of representing it is left to a character-level encoder and to static embeddings. This last

¹With one single exception (Yang et al., 2020), DM has been so far concerned only with the English language.

approach is interesting, in that it is the only one that can handle multi-word targets; however, it combines token embeddings via order-invariant sum, thus being suboptimal for differentiating instances such as *pet house* and *house pet*.

Recent approaches have explored the use of large-scale pre-trained models to score definitions with respect to a usage context. For example, Chang and Chen (2019) proposed to recast DM as a definition ranking problem. A similar idea has been applied in WSD by Huang et al. (2019), leading to state-of-the-art results. However, both of these approaches fall back to the assumption of discrete sense boundaries, therefore being unable to define targets outside of a predefined inventory.

With Generatory, by contrast, we are the first to use a single Encoder-Decoder model to perform diverse lexical-semantic tasks such as DM, WSD and WiC. Moreover, we address the issue of encoding the target in context by using a simple, yet effective encoding scheme, which makes use of special tokens to mark the target *span*, producing a complete and joint encoding of the context, without the need for other components. This allows the effective usage of a pre-trained model, which we fine-tune to generate a gloss given the context.

4.2 Generatory

Due to its aptness at generating glosses for arbitrary-sized spans of text in context, Generatory represents a completely new approach to computational lexical semantics, one that has a wider scope than its predecessors, and one that provides a unified method to concurrently overcome the limits of both a generative task such as DM and those of a discriminative task such as WSD.

With respect to DM, our full sequence-to-sequence framing of the task enables us to deal with units having different compositional complexity, from single words, to compounds and phrases. Thus, Generatory can gloss a *definiendum* that is not found in dictionaries, such as *starry sky*, with the appropriate *definiens*, e.g.: ‘The sky as it appears at night, especially when lit by stars’.

As regards WSD, instead, we are no longer bound by the long-standing limits of predefined sense inventories. Thus, it is possible to give (i) a meaningful answer for words that are not in the inventory, and (ii) one that fits the meaning and the *granularity* required by a given context better than any sense in the inventory. Consider the following:

- (3) (a) Why cannot we teach our children to read, write and reckon?
- (b) Mark or trace on a surface.
- (c) To be able to mark coherent letters.

The target word in (3 a) is associated² with the gold gloss (3 b) from WordNet (Fellbaum, 1998), the most used sense inventory in WSD. However, Generationary arguably provides a better gloss (3 c). In what follows, we detail our approach.

4.2.1 Gloss Generation

In this work we address the task of *mapping* an occurrence of a target word or phrase t (in a context c) to its meaning, by reducing it to that of *generating* a textual gloss g which defines $\langle c, t \rangle$. The target t is a span in c , i.e. a pair of indices $\langle i, j \rangle$ corresponding to the first and the last token which make up the target in c .

Formally, we propose to apply the standard sequence-to-sequence conditional generation formulation, in which the probability of a gloss, given a context-target pair, is computed by factorising it auto-regressively:

$$P(g|c, t) = \prod_{k=1}^{|g|} P(g_k | g_{0:k-1}, c, t) \quad (4.1)$$

where g_k is the k th token of g and g_0 is a special start token. By means of this, we can readily perform contextual DM, as well as “static” DM, i.e. when the target encompasses the whole context ($t = \langle 1, |c| \rangle$). Additionally, as we will see, we can do WSD and WiC by using either the token distribution of the model, or the glosses generated by standard decoding (Section 4.2.2).

To learn the function in Eq. (4.1), we employ a recent Encoder-Decoder model, i.e. BART (Lewis et al., 2019), which is pre-trained to reconstruct text spans on massive amounts of data. The use of a pre-trained model is particularly important in our case, as successfully generating a gloss for a wide range of different context-target pairs requires a model which can wield vast amounts of semantic and encyclopedic knowledge. BART can be fine-tuned to perform specific kinds of conditional generation by minimizing the cross-entropy loss on new training input-output pairs.

In our approach we give as input to BART a $\langle c, t \rangle$ pair, and train to produce the corresponding gold gloss g , with $\langle c, t \rangle$ and g being gathered from various sources (see Section 4.3.1). We devise a simple encoding scheme that allows us to make the model aware of the target boundaries, without architectural modifications to BART. Particularly, we encode $\langle c, t \rangle$ pairs as sequences of subword tokens in which the boundaries of the t span in c are marked by two special tokens, i.e. `<define>` and `</define>`. For example, the sentence *I felt like the fifth wheel*, with the phrase fifth wheel as the target, will be encoded as `I felt like the <define> fifth wheel </define>`. We fine-tune BART to generate

²According to the human annotators of the Senseval-2 WSD evaluation dataset (Edmonds and Cotton, 2001).

the corresponding gloss g : (idiomatic, informal) Anything superfluous or unnecessary.

4.2.2 Discriminative Sense Scoring

In this Section we introduce three distinct techniques by means of which Generatory tackles discriminative tasks without additional training.

4.2.2.1 Gloss Probability Scoring

As we recall, the formulation that we use in Eq. (4.1) enables the model to give the probability of some gloss given the context. Thus, if we have readily available glosses, given the one-to-one sense-to-definition mapping available in WordNet 3.0, we can associate each definition with a probability score by using the same teacher forcing input feeding strategy that is used for training, i.e. using as input the next gold token instead of one sampled from the next token distribution.

With Eq. (4.1) we are able to compute the probability of a certain gloss g given a pair $\langle c, t \rangle$. Thus, we can perform classification by picking the sense which is associated with the gloss with the highest probability. Formally, we select:

$$\hat{s} = \operatorname{argmax}_{s \in S_t} P(\mathcal{G}(s) | c, t) \quad (4.2)$$

where $S_t \subset S$ is the set of applicable senses for target t from the full inventory S , and $\mathcal{G} : S \rightarrow G$ is a function mapping senses to glosses (\mathcal{G} , G , S and S_t are determined by the reference dictionary).

To perform standard WSD with WordNet we can just map the sense to the definition associated with its synset. Note that this approach is quite inefficient, as $n = |\{\mathcal{G}(s) | s \in S\}|$ probabilities have to be computed – resulting in a quadratic complexity similar to gloss-based discriminative approaches such as GlossBERT (Huang et al., 2019).

4.2.2.2 Gloss Similarity Scoring

In NLG there is a disconnect between model probability and model performance, the latter often determined by measures such as BLEU (Papineni et al., 2002). In short, often times a better search in the hypothesis space, i.e. the use of more beams, leads to worse scores according to the evaluation metric, e.g. by effect of empty or uninformative outputs (Stahlberg and Byrne, 2019), to which the model assigns relatively high probabilities.

The usage of model gloss probability does not take into account the definitions that are actually generated. Thus, we adopt a simple best match approach where we compute

similarity scores between the system-generated gloss and the glosses associated with the candidates, and we predict the candidate with the highest similarity. We employ a cosine similarity between the gloss vectors produced via the recently introduced Sentence-BERT model (Reimers and Gurevych, 2019, SBERT) which, relying on contextual vectors rather than string matches, counteracts the problem that valid glosses show a relatively high degree of freedom with respect to the conditioning context, and select a predicted sense \hat{s} as follows:

$$\hat{s} = \operatorname{argmax}_{s \in S_t} \operatorname{sim}(\hat{g}, \mathcal{G}(s)) \quad (4.3)$$

where \hat{g} is the most probable output found by beam-search decoding, and sim is the SBERT similarity.

4.2.2.3 Gloss Similarity Scoring with MBRR

Using just the most probable sequence in the decoding process for the best match search is suboptimal, as more probability mass might be cumulatively assigned to a cluster of very similar outputs. To take this into account, we propose the use of a simple approach inspired by Minimum Bayes Risk Reranking (Kumar and Byrne, 2004, MBRR), which considers the mutual (dis)similarity within the set \hat{G} of k generated outputs decoded with beam search. This is done by rescoreing each output as the sum of the dissimilarities over all k outputs, weighted by their conditional probability, as follows:

$$\hat{g} = \operatorname{argmin}_{\hat{g}_i \in \hat{G}} \sum_{\hat{g}_j \in \hat{G}} (1 - \operatorname{sim}(\hat{g}_i, \hat{g}_j)) P(\hat{g}_j | c, t) \quad (4.4)$$

The new prediction \hat{g} is then plugged into Eq. (4.3) as in simple similarity-based scoring.

4.3 Datasets

4.3.1 Dictionary Gloss Datasets

We now move on to describe the datasets which we use to train Generatory models by fine-tuning BART. Each dataset includes $\langle c, t, g \rangle$ triples, which are used as our input and output for training.

CHA (Chang and Chen, 2019) is an online dataset³ of examples and definitions from `oxforddictionaries.com`.

It comes with two settings, each with its own train/dev/test splits: in the *Seen* setting (**CHA_S**), definitions in the training set are also present in the test set, while the *Unseen* setting (**CHA_U**) has a zero-shot test of lemmas not featured in the training set.

³miulab.myDS.me:5001/sharing/lWPBRc8hG

dataset	instances			unique glosses		
	train	dev	test	train	dev	test
CHA _S	555,695	78,550	151,306	78,105	32,953	37,400
CHA _U	530,374	70,401	15,959	73,104	29,540	3958
SEM	333,633	-	-	116,698	-	-
UNI	1,832,302	-	-	947,524	-	-

Table 4.1. Training instances and number of unique glosses in the datasets used.

SEM is a dataset built by exploiting the SemCor corpus (Miller et al., 1993) – which is manually tagged with WordNet senses – to associate sentence-level contexts with definitions. We filtered out NER-like sense annotations (e.g. those mapping proper names such as *Frank Lloyd Wright* to the general sense of *person*). Moreover, to improve coverage, since not all WordNet senses appear in SemCor, we used synonymy information to build additional contexts, e.g. `<define> separate, part, split </define>` → `go one’s own way; move apart.`

UNI is the concatenation of the train splits of SEM and CHA, plus the following: (i) a cleaned-up January 2020 dump of Wiktionary, from which circular definitions (e.g. starting with *synonym of*) were filtered out, and (ii) the training split containing data from the GNU Collaborative International Dictionary of English (GCIDE), included in the dataset of Noraset et al. (2017), which features only “static” pairs, in which the context coincides with the word to be defined.

We have used CHA and SEM as they were employed by state-of-the-art approaches to DM (Chang and Chen, 2019) and WSD (Huang et al., 2019). With UNI, instead, we have brought together diverse sense inventories to create a dataset that is less dependent on the idiosyncrasies of each of its sources. We report data in Table 4.1.

4.3.2 The Hei++ Evaluation Dataset

Free phrases (e.g. *exotic cuisine*) are not commonly encountered in traditional dictionaries. Considering how they can represent a key case of items which are not featured in standard training sets, along with the fact that no dataset to test the quality of definition generation on such items currently exists, we devised Hei++: a dataset which associates human-made definitions with adjective-noun phrases.

With Hei++ we can assess Generatory’s ability to generate glosses, in a zero-shot

setting, for items which are not featured in the training set.

As a first step in building Hei++, we retrieved the test split of the HeiPLAS dataset (Hartung, 2016),⁴ which we chose as our starting point since it contains commonly used adjective-noun phrases. After removing duplicates and discarding ill-formed phrases, we asked an expert lexicographer to write a single definition for each adjective-noun pair, choosing the most salient sense (according to the intuition of the lexicographer) in the event of an ambiguous phrase (e.g. *new car*, a phrase that can refer both to a newly manufactured car, or to a recently bought car). At the end of the annotation process, we obtained a dataset made up of 713 adjective-noun phrases with their definitions to be used as a gold standard.

4.4 Quantitative Experiments

In what follows, we perform a threefold automatic evaluation to test the strengths of Generationary in different settings.

On the one hand, we assess its ability to produce suitable definitions by testing the generation quality on the DM setting (Section 4.4.1). On the other, we aim to further appraise how well the generated outputs describe the contextual meaning, by evaluating the performance they bring about on the discriminative benchmarks of WSD (Section 4.4.2) and WiC (Section 4.4.3). Additionally, we report a full description of hyperparameters in Section 4.4.4.

4.4.1 Definition Modeling

In this experiment we use different NLG measures to automatically assess how well generated definitions match gold glosses.

We evaluate on the *Seen* (CHA_S) and *Unseen* (CHA_U) test splits of CHA, which is the largest contextual DM benchmark released so far. Moreover, we report results on our Hei++ (HEI) dataset of adjective-noun phrases. We did not include results on the datasets of Noraset et al. (2017) and Gadetsky, Yakubovskiy, and Vetrov (2018), as the first only includes targets with no surrounding context, and the second is largely included in CHA.⁵

⁴www.cl.uni-heidelberg.de/~hartung/data

⁵Results on these datasets are reported in Appendix A.1.

4.4.1.1 Systems

For each evaluation dataset D we test two Generatory models: one trained on the corresponding train split (Gen- D), and one trained on UNI (Gen-UNI).⁶

We compare against (i) a random baseline obtained by predicting, for each test item, a random definition taken from the same test set; (ii) the model of Ishiwatari et al. (2019), which we have re-trained on the same data as Generatory (Ishiwatari- D), and (iii) the state-of-the-art approach of Chang and Chen (2019) (Chang).

On HEI, which has no training split, we only evaluate Gen-UNI and the random baseline, since Ishiwatari-UNI generates strings consisting of mostly unknown word placeholders (`<unk>`), and Chang and Chen (2019) cannot handle multi-word targets.

4.4.1.2 Measures

Previous approaches have employed both perplexity (PPL) and string-matching measures (e.g. BLEU) for scoring DM systems. PPL is very appropriate as in DM, there are many possible “good” answers.⁷ PPL, however, produces a score just on the basis of a pre-existing gold definition, by collecting teacher forcing probabilities, without taking into account any actual output generated through beam-search decoding, thus not assessing the quality of the generation.

To evaluate the latter, BLEU and ROUGE-L (Lin, 2004) are also reported. Note, however, that these two measures are based on simple string matches which, in many cases, are not good indicators of output quality. To counteract this problem, we also report results with METEOR (Banerjee and Lavie, 2005) – which uses stemming and WordNet synonyms – and BERTScore (Zhang et al., 2019), which uses vector-based contextual similarities.⁸

Finally, to present a complete comparison against the ranking-based approach of Chang and Chen (2019), we report results (precision@ k) on their retrieval task of recovering the correct definition, for the target in context, from the whole inventory of 79,030 unique glosses in their dataset. We rank definitions by applying the MBRR plus cosine similarity strategy described in Section 4.2.2.3.

4.4.1.3 Results

As shown in Table 4.2, Generatory models outperform competitors in every setting. On CHA_S , our specialized model (Gen- CHA_S) shows much better results than Gen-UNI,

⁶To ensure a fair comparison, when evaluating on the *Unseen* setting of CHA, we have removed lemmas appearing in the CHA_U test set from the UNI training set.

⁷See Appendix A.2 for details on perplexity computation.

⁸See Appendix A.3 for configuration details.

	model	ppl↓	BL↑	R-L↑	MT↑	BS↑
CHA _S	Random	-	0.2	10.8	3.2	68.1
	Chang	-	74.7	78.3	-	-
	Ishiwatari-CHA _S *	-	6.2	28.2	11.1	74.2
	Ishiwatari-UNI*	-	3.0	23.2	8.2	72.6
	Gen-CHA _S	1.2	76.2	78.9	54.8	93.0
	Gen-UNI	1.4	66.9	72.0	47.0	90.7
CHA _U	Random	-	0.3	11.0	3.2	68.2
	Chang	-	7.1	19.3	-	-
	Ishiwatari-CHA _U *	-	2.1	19.9	7.1	71.7
	Ishiwatari-UNI*	-	2.1	19.7	6.7	71.5
	Gen-CHA _U	20.3	8.1	28.7	12.7	76.7
	Gen-UNI	15.4	8.8	29.4	13.5	76.8
HEI	Random	-	1.6	12.7	0.4	73.4
	Gen-UNI	16.0	6.3	26.3	15.1	78.9

Table 4.2. DM evaluation results. Columns: perplexity, BLEU, Rouge-L, METEOR, BERTScore (ppl/BL/R-L/MT/BS). Row groups are mutually comparable (**bold** = best). ↑/↓: higher/lower is better. *: re-trained.

because NLG measures give high scores to glosses which are lexically similar to the gold, while multi-inventory training will instead expose the model to many other, differently formulated, but equally valid definitions. Note, moreover, that our Gen-CHA_S model outperforms both Ishiwatari et al. (2019) and Chang and Chen (2019), even though the latter, being a ranking model, is obviously at an advantage, since it gets a perfect score when it ranks the gold definition first.

In CHA_U we observe that the Gen-UNI model reaches higher performances than Gen-CHA_U, indicating that, when ‘overfitting’ on the inventory is factored out, multi-inventory training enables the model to generalize better on a zero-shot setting. Furthermore, figures for HEI are in the same ballpark as those on CHA_U, demonstrating that Generatory can easily deal, not only with unseen lemmas, but also with entirely different kinds of target.

Additionally, we report the results of the precision@*k* evaluation in Table 4.3 when macro-averaging on lemmas (left) and senses (right). Figures on the two different splits of CHA show very different trends.

On the CHA_S setting, in which the definitions in the test set are also in the training set, the base model from Chang and Chen (2019) achieves, in most cases, the highest recovery

model	P@ <i>k</i> (lemmas)			P@ <i>k</i> (senses)			
	1	5	10	1	5	10	
CHA _S	Chang (base)	74.8	83.3	85.5	63.3	74.0	77.1
	Chang (large)	73.9	82.6	84.9	62.4	73.2	76.3
	Gen-CHA _S	73.0	77.7	79.4	67.9	72.9	74.7
	Gen-UNI	63.0	70.2	72.7	55.5	63.1	65.8
CHA _U	Chang (base)	3.3	9.6	14.4	2.3	7.4	11.4
	Chang (large)	3.5	10.5	15.6	2.5	8.2	12.4
	Gen-CHA _U	7.8	19.9	25.5	6.5	16.8	22.0
	Gen-UNI	9.3	21.3	27.7	7.4	18.0	23.8

Table 4.3. Macro precision@*k* (lemmas and senses) on the retrieval task of Chang and Chen (2019). Row groups are mutually comparable (**bold** = best).

model	S2	S3	S7	S13	S15	ALL	ALL ⁻	0-shot	N	V	A	R
LMMS ₂₃₄₈	76.3	75.6	68.1	75.1	77.0	75.4	75.9*	66.3*	78.0*	64.0*	80.7*	83.5*
GlossBERT	77.7	75.9	72.1	76.8	79.3	77.0	77.2*	68.7*	79.7*	66.5*	79.3*	85.5*
Gen-SEM (Prob.)	76.9	73.7	69.2	74.6	78.2	75.3	75.7	60.6	77.5	65.0	78.4	87.6
Gen-SEM (Sim.)	77.5	76.4	71.6	76.8	77.4	76.7	77.0	63.3	80.1	64.8	79.1	85.0
Gen-SEM (MBRR)	78.0	75.4	71.9	77.0	77.6	76.7	77.0	65.0	79.9	64.8	79.2	86.4
Gen-UNI (MBRR)	77.8	73.7	68.8	78.3	77.6	76.3	76.8	73.0	79.8	63.3	80.1	84.7

Table 4.4. Results on the WSD evaluation. Row groups: (1) previous approaches; (2) Generatory. Columns: datasets in the evaluation framework (S2 to S15), ALL w/ and w/o the dev set (ALL/ALL⁻), zero-shot set (0-shot), and results by PoS on ALL (N/V/A/R). F1 is reported. **Bold**: best. *: re-computed with the original code.

rate. However, with $k = 1$, which is the most realistic case, Gen-CHA_S outperforms the competitor by 4.6 points when macro-averaging on senses, i.e. items with the same gold definition.

On the more challenging zero-shot CHA_U setting, both Generatory models strongly outperform Chang (large), more than doubling the performance on $k = 1$ and showing an improvement of more than 75% on $k = 10$. Gen-UNI, that was underperforming Gen-CHA_S in the *Seen* setting, now achieves better results across the board, since it can exploit the supervision of a wide array of different glosses from multiple inventories.

4.4.2 WSD Evaluation

Even though Generatory drops the need for a fixed sense inventory, we still want to assess its capabilities in a fully-discriminative setting such as that of WSD. The reason is simple: we want to show how dropping strict sense boundaries does not undermine performances in a traditional “pick-one-among-many” type of task either.

We test on the five datasets collected in the evaluation framework of Raganato, Camacho-Collados, and Navigli (2017), namely: Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli, Jurgens, and Vannella, 2013), SemEval-2015 (Moro and Navigli, 2015), which are annotated with WordNet 3.0 senses.

We denote with ALL and ALL⁻ the concatenation of all evaluation datasets, including or excluding, respectively, SemEval-2007, which is our development set for this experiment. Moreover, we test on the subset of ALL⁻ containing instances whose lemmas are not covered in SemCor (0-shot).

4.4.2.1 Systems

To choose a possible sense from WordNet and perform WSD, we evaluate the techniques presented in Section 4.2.2, i.e. probability scoring (Prob.), simple similarity scoring (Sim.), and similarity scoring with MBRR.

We evaluate our Gen-SEM, which is trained on examples specifically tagged according to the WordNet inventory, and Gen-UNI, which includes definitions from many different inventories. We compare against recent WSD approaches which make use of gloss knowledge, i.e. LMMS (Loureiro and Jorge, 2019) and the state-of-the-art approach of GlossBERT (Huang et al., 2019).

From a technical point of view, LMMS (Loureiro and Jorge, 2019) creates sense-level embeddings covering the whole WordNet inventory, so as to enable a simple Nearest Neighbors model to perform state-of-the-art disambiguation. GlossBERT (Huang et al., 2019), on the other hand, proposes the construction of context-gloss pairs to recast WSD as a sentence-pair classification task. To do so, the authors fine-tune the pre-trained BERT (Devlin et al., 2019) model with SemCor, in an attempt to have a supervised neural system exploit gloss knowledge at best.

As regards LMMS, we include the best model reported in the original paper for our experimental setup, i.e. LMMS₂₃₄₈ (BERT), and retrieve all the reproduction materials from <https://github.com/danlou/LMMS>. The model we employ features the concatenation of (i) BERT contextual embeddings created from SemCor, (ii) dictionary

embeddings for all word senses in WordNet, and (iii) fastText embeddings (Bojanowski et al., 2017), for an overall figure of 2,348 embedding dimensions.

With respect to our experimental setup of GlossBERT, we use its best configuration, namely GlossBERT(Sent-CLS-WS), which uses context-gloss pairs with weak supervision as input and a classification label which highlights the target word. We implement the original system, as available at <https://github.com/HSLCY/GlossBERT>.

4.4.2.2 Results

We report the results of the WSD evaluation in Table 4.4.

The MBRR scoring strategy proves to be the most versatile, with Gen-SEM (MBRR) achieving a higher F1 than Gen-SEM (Prob.) on almost every dataset, and outperforming Gen-SEM (Sim.) on the 0-shot set. As both Sim. and MBRR outscore Prob., it is clear that generating a gloss and ranking candidates with similarity is a better strategy than directly ranking with model probability, which leaves room for further improvements as better similarity measures are developed.

On another note, Gen-SEM (MBRR) achieves performances which are overall comparable with those of the state of the art (GlossBERT) without having been explicitly trained to perform WSD. Compared to Gen-SEM (MBRR), Gen-UNI (MBRR) sacrifices 0.4 and 0.2 points on, respectively, ALL and ALL⁻, but obtains 8 points more on the zero-shot set, also improving over GlossBERT by 4.3 points. This demonstrates that, when using Generatory with data from multiple inventories, (i) performances remain in the same ballpark as those of a state-of-the-art system, and (ii) much improved generalizability is achieved, as shown by the state-of-the-art results on the zero-shot setting.

4.4.3 Word-in-Context

In the task of Word-in-Context (WiC) Pilehvar and Camacho-Collados (2019), predefined sense inventories are not required and meaning identification is reduced to a binary problem in which, given two contexts, both featuring an occurrence of the same lemma, a model has to predict whether the two targets have the same meaning. We compare against Chang and Chen (2019), which is the only DM approach reporting results for WiC, following their setting in which no task-specific training is performed and the training set for the task is used for testing. Results are reported for both Gen-CHA_S, which is trained on the same data as Chang and Chen (2019), and Gen-UNI.⁹

⁹In this experiment we have excluded Wiktionary, which was used to build the WiC dataset, from the UNI training set.

To perform the task, for each pair in the WiC dataset we generate two sets, γ and γ' , each of 10 glosses, for the two respective sentences in the pair. Then, for each generated gloss $\hat{g} \in \gamma$, we compute the score $z_{\hat{g}}$ as the mean SBERT similarity between \hat{g} and the 10 generated glosses in γ' . Analogously, we compute $z_{\hat{g}'}$ as the mean similarity between $\hat{g}' \in \gamma'$ and the glosses in γ . For each gloss g we normalize z_g by subtracting an approximate mean similarity of g with random glosses, computed as the mean similarity between g and all other unrelated glosses in the batch. If the mean score $(\sum_{\hat{g} \in \gamma} z_{\hat{g}} + \sum_{\hat{g}' \in \gamma'} z_{\hat{g}'})/20$ exceeds a threshold t (tuned on the dev set), we predict that a WiC pair shares the same sense.

Gen-CHA_S, with an accuracy of 69.2, outperforms Chang and Chen (2019), which achieves 68.6, demonstrating the strength of our approach in this setting. Moreover, Gen-UNI, which attains a result of 71.1, outscores both Gen-CHA_S and the competitor, once again bearing witness to the versatility of training on multiple inventories.

4.4.4 Reproducibility Details

To train our models we employ the `fairseq` library. Generationary has the same number of parameters as BART Lewis et al. (2019), i.e. ca. 458M. For fine-tuning, we use the same hyperparameters used in Lewis et al. (2019) for summarization,¹⁰ except that:

- the learning rate is set to 5×10^{-5} on the basis of preliminary experiments;
- due to memory concerns, we feed the input in batches of 1024 tokens, updating every 16 iterations;
- we use inverse square root learning rate scheduling, which does not require to set a maximum number of iterations a priori;
- we double the number of warmup steps to 1000.

Training is performed for at most 50 epochs. We employ a single NVIDIA GeForce RTX 2080 Ti GPU to perform all the reported experiments, with average runtimes per epoch of BART fine-tuning ranging from ca. 50 minutes (Gen-SEM) to >120 minutes (Gen-UNI).

On the DM task, we evaluate on the best epoch, i.e. the one with the lowest cross-entropy loss on the dev set, with no hyperparameter tuning. On the WSD task, instead, we perform minimal hyperparameter tuning, with search trials just on beam size (testing with values of 1, 10, 25, and 50), choosing as the best configuration the one with the highest F1 on our dev set, SemEval-2007; with simple similarity scoring, the best Gen-SEM has a beam size of 10,

¹⁰<https://github.com/pytorch/fairseq/blob/master/examples/bart/README.summarization.md>

while, with MBRR similarity scoring, the best Gen-SEM had a beam size of 25. We have used only MBRR with Gen-UNI, with a beam size of 10, resulting in the best performance on the development set. On the WiC task we have only performed tuning of the threshold on the dev set, by trying every value in range between the lowest and the highest z score, with a minimum step of 0.025. We compute similarities in batches of 125 pairs.

For training and prediction of the models of Ishiwatari et al. (2019), we use the code provided by the authors.¹¹ We use the same hyperparameters, except that we increase the vocabulary size to 39,000, which results in much improved performances on our benchmarks.

4.5 Qualitative Experiment

Given that the ability of Generationary to produce fluent and meaningful definitions is its key asset, in addition to the automatic evaluation reported in Section 4.4, we devised a qualitative experiment on two distinct datasets we constructed.

While previous experiments shed light upon the quality of Generationary in comparison with other automatic systems, now, we employ human annotators to compare definitions produced with our approach against glosses written by human lexicographers. The datasets that we use in this experiment are (i) our Hei++ dataset of definitions for adjective-nouns phrases (Section 4.3.2) and (ii) SamplEval, i.e. a sample of 1000 random instances made up of 200 items¹² for each of the five WSD datasets included in ALL (see Section 4.4.2), with at most one instance per sense.

With Hei++ we assess the ability of Generationary to accurately gloss complex expressions, such as free phrases (e.g. *wrong medicine* or *hot forehead*), that are rarely covered by traditional dictionaries. With SamplEval, instead, we test whether generated glosses can improve over gold definitions associated with gold senses in WordNet.

4.5.1 Annotators and Annotation Scheme

For each context-target pair in Hei++ and SamplEval we have two definitions: a gold one, written by a lexicographer (in Table 4.5 we provide an excerpt of 20 random entries sampled from the Hei++ dataset), and one generated by Gen-UNI, which is not tied to any specific inventory and has proven the most versatile model across tasks.

We hired three annotators with Master’s Degree in Linguistics and effective operational proficiency in English and, in a similar fashion to Erk and McCarthy (2009), we asked them to assign a graded value to the definitions based on their pertinence to describe the target

¹¹<https://github.com/shonosuke/ishiwatari-naacl2019>

¹²We did not sample instances annotated with many senses.

phrase	definition
small-time actor	An actor of minor importance.
implicit anger	Anger that remains unexpressed.
domestic animal	An animal that is raised or adapted to live at home.
powerful argument	An argument that is very convincing.
narrow bridge	A bridge with a narrow deck.
new car	A recently bought car.
abnormal circumstance	A circumstance that deviates from what is ordinary.
noble deed	An act that is admirable and motivated by compassion.
powerful drug	A drug that has a powerful effect.
short flight	A flight that reaches its destination in a brief time.
big group	A group comprising a large number of members.
good joke	A joke that is considered funny and witty.
short month	Any month containing less than 31 calendar days.
natural phenomenon	Any phenomenon brought about by a natural cause.
free port	A port in which goods are exempt from customs duty.
formal requirement	A requirement set out by prescribed norms.
empty seat	An available seat that can be occupied (e.g. in a theatre).
hot stove	A stove with fuel burning, dangerous to touch.
adequate training	Training that achieves its intended purpose.
new year	The year to come.

Table 4.5. Random sample of 20 entries collected from Hei++. Columns, left to right: adjective-noun phrase, as originally featured in HeiPLAS (Hartung, 2016); gold definition provided by our lexicographer.

t in c , according to a five-level Likert scale. In Table 4.6 we show one of the annotation examples that were provided to the annotators to be used as interpretation guidelines.

Definitions for each sentence were presented in shuffled order. The ITA was substantial, with an average pairwise Cohen’s κ of 0.69 (SampleEval) and 0.67 (Hei++).

4.5.2 Results

As can be seen in Table 4.7, the quality of Generatory glosses in the SampleEval dataset is comparable to those drawn from WordNet. Note that, although it would be expected

	Was he going to be saddled with a creep for a <u>bar-buddy</u> ?
1	<i>Wrong gloss. May refer to a homonym of the target.</i> A heating element in an electric fire.
2	<i>Wrong gloss. Captures the domain of the target.</i> A counter where you can obtain food or drink.
3	<i>Correct gloss. Utterly vague and generic.</i> A person with whom you are acquainted.
4	<i>Correct gloss. Fits the context, but misses some details.</i> A close friend who accompanies his buddies in their activities.
5	<i>Correct gloss. Perfectly describes the target in its context.</i> A friend who you frequent bars with.

Table 4.6. Annotation guidelines excerpt. Rows: Likert score, *explanation* and example definition for target.

dataset	gold	Gen.	\geq
Hei++	4.43	3.58	29.9
SampleEval	3.75	3.62	51.3

Table 4.7. Qualitative evaluation results. Columns: dataset, average Likert for gold and Generationary, % of Generationary scores equal or better than gold (\geq).

for gold annotations to come close to the top of the scale, this is not the case, as they received an average score of 3.75 out of 5, demonstrating the suboptimal nature of “ready-made” meaning distinctions. We report analogous scores on the Hei++ dataset. The gap with respect to gold definitions here is wider, probably because (i) Generationary is not specifically trained on complex expressions and (ii) the gold score is higher since phrases are less ambiguous than single words.

Interestingly, the annotators rated Generationary glosses at least as high as their gold counterparts on 51.3% and on 29.9% of the total cases on SampleEval and Hei++, respectively; a result that provides evidence for the reliability of Generationary definitions as valid alternatives to glosses taken from established inventories of discrete word senses.

4.6 Generation Examples

In Table 4.8 we show a sample of definitions generated via our Gen-UNI model for various spans in context.¹³ As can be seen, the glosses \hat{g}_1 and \hat{g}_2 (extracted from SamplEval and Hei++, respectively) demonstrate that Generatory can indeed provide better, more specific definitions than gold standard ones.

The following reported examples show the strength of our model on contexts which do not resemble those it is trained on: Generatory is proficient at (i) handling fixed or semi-fixed idioms of different lengths (\hat{g}_3 , \hat{g}_4) and (ii) defining non-conventional words and phrases (\hat{g}_5 , \hat{g}_6); interestingly, Generatory is also able to (iii) provide high-level explanations for whole figurative contexts (\hat{g}_7 , \hat{g}_8), which goes well beyond what is commonly referred to as *glossing*. This might result in interesting applications beyond the scope of this work, e.g. for paraphrase generation and metaphor interpretation (Rai and Chakraverty, 2020).

c_1	[...] I <u>scooted</u> them into the dog run.
\hat{g}_1	Cause to move along by pushing.
g_1	Run or move very quickly or hastily.
c_2	<u>Exotic cuisine</u> .
\hat{g}_2	A style of cooking that is out of the ordinary and unusual (as if from another country).
g_2	Cuisine involving unfamiliar foods.
c_3	He was never the same after the <u>accident</u> .
\hat{g}_3	Indicates that a person has lost the good qualities that were present before the accident.
c_4	Sam is in a better <u>place</u> now.
\hat{g}_4	A phrase used to express that one has learned about another’s death.
c_5	Yesterday I had to undergo a <u>beardectomy</u> .
\hat{g}_5	The surgical removal of the beard.
c_6	You’ve got a hard coconut to <u>smash</u> here, Dr. Yang!
\hat{g}_6	Something difficult to deal with.
c_7	The mind is haunted by the <u>ghosts of the past</u> .
\hat{g}_7	People’s memories of the past are still present in their mind, even after they have ceased to exist.
c_8	The fault, dear Brutus, is not in our stars, but in our <u>ourselves</u> .
\hat{g}_8	The responsibility for a problem lies with the people who cannot see it themselves.

Table 4.8. Sample of Generatory definitions (\hat{g}) for several targets in context (c). g : gold definition.

¹³See Appendix A.4 for further samples of generated glosses.

4.7 Error Analysis

To have a broader picture of the quality of the outputs produced by means of Generation-ary, we perform behavioural testing for our Gen-UNI model, in the spirit of Ribeiro et al. (2020). As a result, we can identify two main trends behind failures to generate an appropriate contextual definition, which we refer to as *disambiguation errors* and *hallucinations*, respectively.

Disambiguation errors When the model predicts a perfectly good definition for the target, but one that fits another common context of occurrence, a disambiguation error arises. For instance, given the $\langle c, t \rangle$ pair in (4 a), with the word pupil as the target, the model fails to identify the “aperture in the iris of the eye” sense, and instead produces an output gloss which is compatible with the meaning of the homograph (4 b):

- (4) (a) The teacher stared into the pupils of her pupil.
 (b) A person receiving instruction, especially in a school.

Hallucinations Other errors stem from the fact that the model can only rely on the knowledge about possible *definienda* that it is able to store in the parameters during the pre-training and training stages. Thus, if the contextual knowledge is not sufficient to extrapolate a definition, the model – which is required to always generate an output – will hallucinate an answer on the basis of contextual clues, incurring the risk of introducing non-factualities. This particularly concerns named entities and domain-specific concepts, but the clearest examples can be seen with targets that do not correspond to any existing, fictional or non-fictional entity. For example, given the input sentence (5):

- (5) Squeaky McDuck wasn’t happy about it,

the model outputs the following:

- (6) The title character in the “Squeaky Squeakety-Squeakiness” cartoon series.

In this case, the model picked the cue of the *cartoonish* Squeaky McDuck character, and hallucinated the name of a plausible cartoon series for it. Note that neither Squeaky McDuck nor the cartoon series actually exist.

4.8 Conclusion

With this Chapter, we showed that generating a definition can be a viable alternative to the traditional use of sense inventories in computational lexical semantics, and one that

better reflects the non-discrete nature of word meaning. We introduced Generationary, an approach to automatic definition generation which, thanks to a flexible encoding scheme, can (i) encode targets of arbitrary length, and (ii) exploit the vast amount of knowledge encoded in the BART pre-trained Encoder-Decode, through fine-tuning.

From two points of view, Generationary represents a unified approach: first, it exploits multiple inventories at once, hence going beyond the quirks of each one; second, it is able to tackle both generative (Definition Modeling) and discriminative tasks (Word Sense Disambiguation and Word-in-Context), obtaining competitive or state-of-the-art results, with particularly strong performances on zero-shot settings. Finally, human evaluation showed that Generationary is often able to provide a definition that is on a par with or better than one written by a lexicographer.

We make the code needed to reproduce the experiments, along with a new evaluation dataset of definitions for adjective-noun phrases (Hei++), available at <http://generationary.org>.

Chapter 5

Dissecting the State of the Art

Since the very beginning of this dissertation, we aimed to provide answers to a few, focused, and often overlooked problems affecting Word Sense Disambiguation. First, in Chapter 3, we provided evidence for the benefits of injecting syntagmatic knowledge into existing LKBs, and secondly, in Chapter 4, we introduced a revolutionary way of recasting WSD, so as to get rid of the constraints imposed by finite sense inventories. Moving forward, one of the three questions we posed in Section 2.3 still remains unanswered: “can state-of-the-art systems for WSD actually disambiguate?”

In what follows, we will answer this question and, to this end, we will start by briefly reviewing the already available literature. In fact, during recent years, a few Natural Language Processing tasks have apparently achieved and surpassed the estimated human performance to the point of being regarded as solved (Rajpurkar et al., 2016; Wang et al., 2019). Even Word Sense Disambiguation, the task of automatically selecting the proper sense for an ambiguous word in context (Navigli, 2009), has seen brand-new systems achieve F1 scores near or above 80 (Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020), a value long considered to be a hard ceiling in being representative of the agreement between human annotators (Edmonds and Kilgarriff, 2002; Navigli, Litkowski, and Hargraves, 2007; Palmer, Dang, and Fellbaum, 2007).

But what does it really mean to match or even surpass human performance? We could legitimately expect systems capable of carrying out a task to the extent of being indistinguishable from its human counterparts, but as soon as the real capabilities of such systems are investigated beyond a mere accuracy score, their shortcomings become immediately apparent (Zhou et al., 2020). Studies criticizing the use of sheer accuracy figures as the only possible measures to evaluate system performance are increasing in the literature (Ribeiro, Singh, and Guestrin, 2016; Belinkov and Bisk, 2018; Ribeiro et al., 2020; Card et al., 2020), and the picture that emerges is anything but that of “solved” tasks (Rajpurkar, Jia, and Liang,

2018). In actual fact, evidence suggests that systems are not able to perform human-like generalization at all, but are simply getting better at overfitting the training data by learning surface patterns devoid of semantics (Weissenborn, Wiese, and Seiffe, 2017; Bender and Koller, 2020).

The work introduced in this Chapter somehow follows this trail and acts as an analysis tool in demonstrating how traditional evaluation measures such as the F1 score in the context of WSD could be misleading factors that fail to reflect the actual capabilities of a system and, therefore, should not be simply taken at face value. Particularly, we begin by selecting an heterogeneous set of systems for English WSD that employ sense embeddings (Loureiro and Jorge, 2019), neural-based classification models (Huang et al., 2019; Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020), graph-based (Scozzafava et al., 2020) or generative approaches (Bevilacqua, Maru, and Navigli, 2020) to attain current state-of-the-art results in terms of F1 score. We then proceed to test these systems by means of a series of ablation experiments conducted on traditional Senseval (Edmonds and Cotton, 2001; Snyder and Palmer, 2004) and SemEval test sets (Pradhan et al., 2007; Navigli, Jurgens, and Vannella, 2013; Moro and Navigli, 2015) and on their concatenation (Raganato, Camacho-Collados, and Navigli, 2017), as well as identifying and analyzing qualitatively the set of instances that none of the aforementioned systems can disambiguate: an impressive 7.4% quota out of the whole set of gold annotations.

As a result of this analysis, we highlight three main reasons why state-of-the-art performance does not coincide with actual disambiguation capability. On the one hand, we provide sound empirical support to the literature claiming that (i) all systems, regardless of the approach used, suffer from a critical bias towards the most frequent word senses (Calvo and Gelbukh, 2015; Postma et al., 2016; Raganato, Camacho-Collados, and Navigli, 2017) and – in the case of trained models – towards the instances featured in the training set (Loureiro et al., 2020). On the other hand, our analysis evidences how (ii) different annotation guidelines for tagging “gold standards”, along with the idiosyncratic understanding of word meanings by different annotators (Kilgarriff, 2007; Passonneau et al., 2010), can be root causes of numerous disambiguation errors on the part of automatic systems. Finally, harming current performance figures, are (iii) all the errors inherent in the commonly employed test sets, ranging from disambiguation mistakes committed by the annotators, up to faulty PoS-tag labels and lack of a proper number of multiple answers in the gold test sets.

Moreover, as a means to further validate our findings, and in order to prevent our analyses from reflecting artifacts which might only belong to already existing datasets, we created a fresh test set: “42D” (pron. [for-ti-tude]).

42D is a manually-annotated evaluation dataset for English WSD, encompassing the wide range of 42 domains covered in BabelNet 4.0 (Navigli and Ponzetto, 2012) with labeled paragraphs taken from the British National Corpus (Leech, 1992, BNC).¹ The purpose of 42D is to act as a very difficult testing ground, a so-called challenge set (Belinkov and Glass, 2019), representative of the cases that put the automatic disambiguation systems to the most severe test. For this reason, none of the word senses appearing in 42D – which have been assigned according to the WordNet 3.0 inventory (Fellbaum, 1998) – are featured in the most-widely employed training set for WSD, i.e. SemCor (Miller et al., 1993), nor do they appear as first senses according to the distribution of WordNet itself, which ranks senses according to their frequency counts within sense-tagged corpora.

System performance on 42D shows comparable trends with the ablation studies conducted on traditional evaluation sets, thus supplying further evidence concerning the actual and alarming impact of biases towards training data and sense skewness in WSD.

42D represents a fresh test-bed, and the first that will be made available to the research community after a five-year hiatus since SemEval-2015 (Moro and Navigli, 2015).² Despite being an evaluation exercise based on the use of the F1 score, 42D portrays a very different picture of traditional system performance given the same evaluation measure, and hence represents a complementary evaluation tool compared to the existing ones, useful for observing the resilience of systems in contexts that are currently arduous for automatic disambiguation, and yet utterly ordinary in natural language.

5.1 Systems at Issue

In an attempt to make our analysis as across-the-board as possible, we have simultaneously examined six of the most relevant systems for automatic disambiguation, an assortment which currently constitutes a rather exhaustive spectrum of the approaches to perform state-of-the-art in WSD. Below, we introduce and detail each of these six systems along with describing an additional random baseline (in Section 5.1.1) which we will use to establish a further comparison in some of our experimental settings.

BEM (Blevins and Zettlemoyer, 2020) is a bi-encoder model that embeds the target word (with its context) and its sense definitions – concurrently and independently – in the same representation space. The two encoders are thus optimized together, in a way that

¹This work complies with the BNC Licence for paragraphs and other fragments, as endorsed by the BNC staff via the official inquiry mail ota@bodleian.ox.ac.uk on October 15, 2019. For further information, see also <http://www.natcorp.ox.ac.uk/faq.xml?ID=licensing>.

²Upon journal acceptance of the article this Chapter provides the foundation for.

allows BEM to carry out disambiguation simply by assigning the label of the nearest sense embedding to a given target word embedding. A key feature of BEM, which is trained on SemCor (Miller et al., 1993), is its significant error reduction capabilities with respect to low-frequency word senses. For the purposes of this work, we implemented the original, full BEM model, freely-available at <https://github.com/facebookresearch/wsd-biencoders>.

EWISER (Bevilacqua and Navigli, 2020) is a neural supervised architecture for WSD that is able to exploit not only pre-trained synset embeddings – in a purely supervised fashion –, but also the relational information included in widely-used lexical knowledge bases, hence enabling the network to predict synsets outside the training set. Particularly, EWISER brings together structured and unstructured knowledge in a unique architecture for WSD where (i) implicit knowledge, in the form of synset embeddings, is used to initialize the output embeddings, and (ii) explicit knowledge, in the form of relational knowledge integrated via a WordNet adjacency matrix, is concurrently added on top of the same baseline neural classifier. In our analysis, we implemented the best performing model reported in the original paper, i.e. $EWISER_{hyper}$,³ which is trained on several resources at once, namely, SemCor, the set of the untagged WordNet glosses, the tagged glosses of the Princeton WordNet Gloss Corpus, and WordNet examples as well. All data is freely-available at <https://github.com/SapienzaNLP/ewiser>.

Generatory (Bevilacqua, Maru, and Navigli, 2020) is the generative approach to WSD that we authored, and that has been introduced in Chapter 4. For the purpose of the following experiments, we employed its Gen-UNI (MBRR) configuration, as detailed in Section 4.3.1.

GlossBERT (Huang et al., 2019) is the system we already introduced in Section 4.4.2.1, here used in the same, aforementioned configuration.

LMMS (Loureiro and Jorge, 2019) is the system we already introduced in Section 4.4.2.1, here used in the same, aforementioned configuration.

SyntagRank (Scozzafava et al., 2020) is the current state-of-the-art system for English knowledge-based WSD, and one of the main contribution of this thesis. We conducted the experiments in this Chapter using the setting described in Section 3.2.5.

³For the remainder of this Chapter, we will refer to this model simply as EWISER.

5.1.1 Random System Baseline

In addition to the systems just described, during the experiments we also employ a random baseline to provide a further term of comparison indicative of the significance of the results. Specifically, our random baseline is represented by a simple system that randomly selects one of the possible senses for each instance in the test sets. Its results are reported as the average of 1,000 runs, together with their relative sample standard deviation.

5.2 Analysis of Traditional Benchmarks

The first step in our investigation, aimed at determining how commonly-used evaluation measures for WSD can be misleading and not indicative of the systems' real disambiguation capacity, starts from traditional evaluation benchmarks. Hence, to begin with, we first observe the variations in terms of F1 score when the systems under analysis are tested on ablations of well-established evaluation exercises.

Specifically, in Section 5.2.1 we take into consideration the concatenation of the test sets (ALL) reported by Raganato, Camacho-Collados, and Navigli (2017), and empirically verify the impact of two critical factors affecting the performance of the systems: (i) the bias towards the most frequent word sense, and (ii) the bias towards the data featured in the training set.

Next, in Section 5.2.2, we identify the subset of ALL that none of the systems investigated is currently able to disambiguate, and analyze its characteristics by means of a precise qualitative analysis.

5.2.1 Quantitative Analysis

To carry out our quantitative analysis, we simply test the systems listed in Section 5.1 on three different settings, i.e. ablations of the ALL test set, which represents our baseline. Following, we report details for each setting:

1. **SemCor**: in this setting, we filtered out from ALL every test instance which featured at least one gold sense annotation that was also featured in the SemCor training set.
2. **WN1st**: in this setting, we instead filtered out from ALL every test instance which featured at least one gold sense annotation that appears in WordNet 3.0 as first sense. As a consequence, this ablation set also filters out all monosemous instances.
3. **SemCor+WN1st**: in this setting, we filtered out from ALL every test instance which featured at least one gold sense annotation appearing either in SemCor, or in WordNet

ablation	#inst (mono)	random \pm sstd	BEM	EWISER	Gener.	Gloss.	LMMS	Syntag.
-	7253 (1301)	36.3 \pm 0.44	79.0	80.1	76.3	76.9	75.4	71.7
SC	1138 (448)	53.5 \pm 0.94	67.1	70.4	68.6	62.2	61.7	61.0
WN1st	2525 (0)	19.6 \pm 0.73	50.5	54.7	48.4	45.0	52.6	29.5
SC+WN1st	562 (0)	25.5 \pm 1.72	41.3	43.6	41.6	29.7	28.3	26.5

Table 5.1. Ablation experiments on the concatenation of the datasets (ALL), as reported in Raganato, Camacho-Collados, and Navigli (2017). Columns, left to right: subset of annotations removed from ALL (ablation), where SC is SemCor; total number of annotations (#inst) of which monosemous instances are indicated between round brackets (mono); F1 score for a random baseline and its sample standard deviation (random \pm sstd); F1 scores for BEM, EWISER, Generatory (Gener.), GlossBERT (Gloss.), LMMS and SyntagRank (Syntag.). **Bold** is best.

3.0 as first sense. As for the WN1st ablation, this setting too does not include monosemous instances.

5.2.1.1 Results

Results for our ablation experiments on ALL are shown in Table 5.1.

In the SemCor ablation (SC) the impact of the data present in the training set is immediately evident, despite the higher proportion of monosemous instances (which are normally taken into account when computing F1 scores) compared to ALL. This effect is best mitigated by Generatory, which uses many other training data and loses the least points with respect to ALL (7.7), but curiously affects also SyntagRank, which is a knowledge-based system, and therefore is not trained on SemCor. This outcome probably stems from the fact that 4,152 instances from ALL appear simultaneously in SemCor, as well as being also listed as most frequent senses according to WordNet, hence representing notoriously more interlinked nodes in the knowledge-base graph (Calvo and Gelbukh, 2015).

The WN1st setting is even more impactful, and although all systems have no difficulty in outperforming the random baseline, none come close to an F1 score of 60 points, not even EWISER, whose score of 54.7 makes it the best performer in this setting.

Finally, in the SemCor+WN1st hybrid setting (SC+WN1st), consisting of 562 instances, only three systems significantly surpass the random baseline, while GlossBERT, LMMS and SyntagRank lie dangerously close to it.

On the one hand, therefore, these simple ablation studies show us empirically that no system is currently able to generalize outside the training set, a fact highlighted among other

things by the fact that the random baseline shows an inverse trend with respect to every other system, since it has less difficulty in solving the SemCor+WN1st setting, where it attains a mean score of 25.5 with a sample standard deviation of 1.72, but attains only 19.6 points in the one that excludes only the WN1st. On the other hand, these results provide evidence that the disambiguation performance is much lower than the one normally reported, especially when the system cannot rely on its knowledge of the most frequent sense, be it because of the sense distribution learned in the training set, or because of the structure of the graph employed.

5.2.2 A Model-agnostic Hard Core

Beyond demonstrating once more how WSD systems are prone to significant biases, it is also interesting to analyze in detail if critical cases in traditional test sets exist, such that none of the current state-of-the-art systems is able to provide a correct answer for. We have therefore identified the intersection of the errors made by our six systems on ALL, and obtained a subset made up of 536 instances, i.e. a 7.4% quota of the whole dataset, and proceeded to study it in order to understand the reasons behind the apparent impossibility of solving this “model-agnostic hard core”.

5.2.2.1 Observations

At first glance, the subset of shared errors (hereafter, ALL_{SE}) provides further support for the results already shown in Section 5.2.1.1. ALL_{SE} , in fact, seems to exemplify to a large extent the most critical conditions for current disambiguation systems. As shown in Figure 5.1, the average percentage of times systems predict a sense seen in SemCor rises from 87.7% on ALL to 95.3% on ALL_{SE} , while the gold annotations containing at least one word sense featured in SemCor for the same instances drop from 84.3% to 62.9%.

Even more noticeable is the case of WN1st senses, where the system predictions remain rather linear, going from 72.3% to 66.7%, but the gold annotations show an impressive drop from 65.2% on ALL, to 2.2% on ALL_{SE} .

These trends therefore seem to highlight how the decrease in terms of instances annotated with WN1st senses in the gold answers does not correspond to a decrease in the predictions of the systems. Moreover when dealing with senses featured in SemCor, system predictions and gold answers follow opposite trends: the higher the number of gold answers that do not feature SemCor senses, the more the systems will predict word senses seen in the training set.

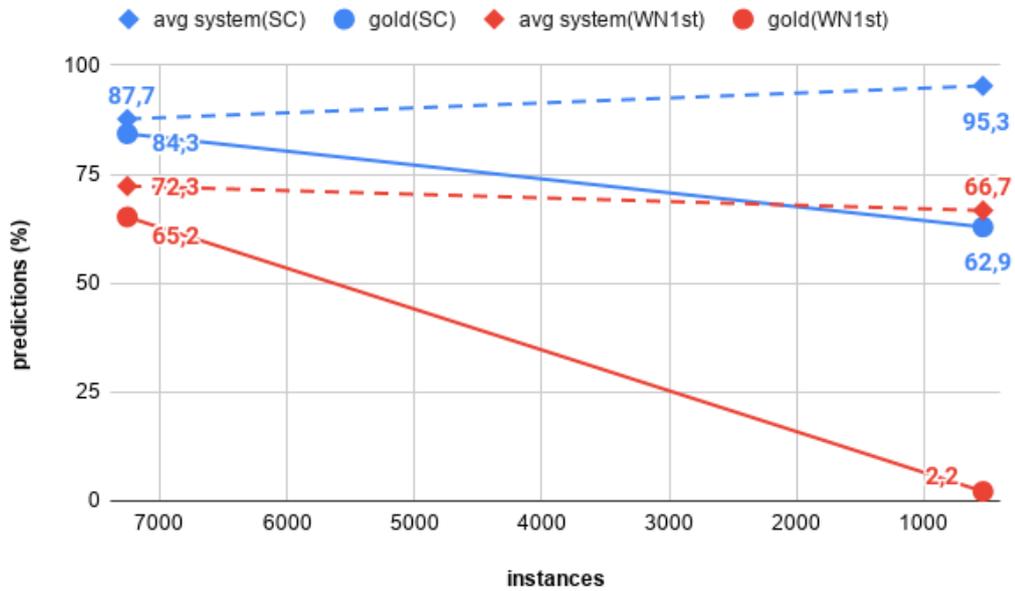


Figure 5.1. Trendlines showing the change in the percentage of senses belonging to SemCor (blue lines) or to WN1st (red lines), both for the average number of system predictions (dashed lines) and gold annotations (solid lines), in the ALL dataset (left) and in the subset of the shared errors on ALL (right), respectively.

Diachronic Evaluation In an attempt to substantiate the results shown in Figure 5.1, we carried out a further investigation considering additional systems besides the ones already described in Section 5.1. In particular, we verified the impact on ALL_{SE} given the inclusion of predictions from three more systems that constituted the state of the art in WSD at the time when the common evaluation framework of Raganato, Camacho-Collados, and Navigli (2017) was released, namely:

- **Babelfy** (Moro, Raganato, and Navigli, 2014);
- **Context2Vec** (Melamud, Goldberger, and Dagan, 2016), in its setting trained both on SemCor and on OMSTI (Taghipour and Ng, 2015);
- **IMS+embeddings** (Iacobacci, Pilehvar, and Navigli, 2016) in its setting trained on SemCor and OMSTI which uses all IMS original default features (Zhong and Ng, 2010) and excludes surrounding words.

As shown in the Table 5.2, a quota of shared errors is resolved thanks to the inclusion of new systems in the comparison, bringing the size of ALL_{SE} from 536 instances to 413.

setting	#inst	avg system	gold
$ALL_{SE}(\text{SC})$	536	95.3%	62.9%
$ALL_{SE+}(\text{SC})$	413	96.3%	58.8%
	-123	+1.0%	-4.1%
$ALL_{SE}(\text{WN1st})$	536	66.7%	2.2%
$ALL_{SE+}(\text{WN1st})$	413	66.0%	0.2%
	-123	-0.7%	-2.0%

Table 5.2. Percentages variation for SemCor and WN1st predictions when further systems are introduced in the comparison. Columns, left to right: setting; number of instances; mean prediction percentage for the original six systems in our comparison; percentage of gold annotations. Middle and last row show variations from the original shared error subset (ALL_{SE}) to the shared error subset when further systems are taken into account (ALL_{SE+}).

Interestingly, this new subset of shared errors (henceforth ALL_{SE+}) highlights the trends already noted in Section 5.2.2.1. Essentially, the higher the number of systems included in the comparison, the bigger the proportion of gold annotations that are not WN1st senses, nor featured in SemCor, in the subset of their shared errors.

5.2.2.2 Qualitative Error Analysis

Determining the reasons why a particular subset of instances taken from traditional WSD evaluation benchmarks is so hard to disambiguate requires an additional level of analysis: an error assessment aimed at identifying the possible presence of regular error patterns. Such an analysis do not, to the best of our knowledge, exist in the available literature, and we strongly believe it could be useful for understanding the path forward to improve Word Sense Disambiguation applications.

In this regard, we asked an expert lexicographer with extensive knowledge of WordNet and high proficiency in English to individually analyze each instance in the ALL_{SE} subset, starting from the contexts and target words appearing in the original datasets, and comparing the predictions of all the systems investigated in Section 5.1 with the annotations present in the gold standard for the same instances. Specifically, we asked the lexicographer to mark each instance among the 536 present in ALL_{SE} with one of the following labels:

- **<factual>**, which indicates that all systems provided predictions which are factual

errors, whereas the gold annotations correctly fit the ambiguous target in context;

- **<multi>**, which indicates that the annotators who devised the gold standard could have assigned multiple senses to disambiguate the target in context, with the constraint that one of this additional sense annotation would have allowed at least one of the systems to solve the instance;⁴
- **<ge_pos>**, which indicates an error in the original test set, in this case, a PoS-tagging error;
- **<ge_dis>**, which indicates an error in the original test set, in this case, a wrong sense selection by the annotators;
- **<ge_mis>**, which indicates an error in the original test set, in this case a sub-optimal sense choice due to the lack of a proper sense in the WordNet inventory;
- **<ge_t/l>**, which indicates an error in the original test set, in this case, a lemmatization or a tokenization error.

In addition to the qualitative error analysis on ALL_{SE} , we asked our lexicographer to perform the same analysis also on random samples of the same size as ALL_{SE} , but taken from the set of remaining instances in ALL without shared errors (ALL_{NS}) and from SemCor.⁵

Results of this qualitative assessment are shown in Table 5.3 and suggest a different picture from the one highlighted so far. In this case, in fact, the idea we get is that of underestimated system performances, and the only factor that remains common with the analyses in Section 5.2.1 and 5.2.2, respectively, is therefore the imprecise nature of the results we currently use to assess WSD systems.

As a matter of fact, the ALL_{SE} subset shows a very high number of errors inherent in the original test set (118) which, together with cases in which multiple annotations would have been beneficial to the systems (176), makes up a figure greater than that of “real” errors, which have been labeled as factual on just 242 occurrences. Fortunately, this proportion does not affect the ALL_{NS} setting in the same way, and a total figure of 70 annotations were labeled as either gold errors or possible multiple annotations. Also, it is worth noting how the sample taken from the training set itself, SemCor, is not devoid of errors. Here, in fact,

⁴For further studies dealing with multiple sense annotations and graded sense assignments for WSD, see Erk and McCarthy (2009) and Erk, McCarthy, and Gaylord (2009).

⁵Before sampling SemCor, in order to maintain full comparability with the other samples, we discarded annotations of auxiliary verbs and Named Entities annotated with generic lemmas (person, location and organization) and relative senses.

setting	#inst	sample	amb+m	sks	factual	multi	ge_pos	ge_dis	ge_mis	ge_t/l	ge_tot
ALL _{SE}	536	536	9.4	401	242	176	16	66	28	8	118
ALL _{NS}	6,717	536	5.5	499	151	45	4	9	6	6	25
SemCor	226,036	536	6.5	504	-	-	5	19	2	0	26

Table 5.3. Qualitative error analysis breakdown. Each row shows a different setting: the shared errors on ALL (ALL_{SE}), the set of remaining instances of ALL not included in the shared errors (ALL_{NS}), and SemCor. Columns, left to right: setting; number of instances (#inst); sample size (sample); ambiguity including monosemous instances (amb+m); unique sense keys (sks); factual errors of the investigated systems (factual); number of instances for which multiple gold annotations could have been provided (multi); PoS tagging errors in the gold standard (ge_pos); disambiguation errors in the gold standard (ge_dis); gold errors due to missing senses in the WordNet inventory (ge_mis); tokenization or lemmatization errors in the gold standard (ge_t/l); total number of errors in the gold standard (ge_tot).

the analysis has been aimed only at identifying errors on the part of the original annotators, and an overall number of 26 total errors have been identified, therefore making up the 4.9% of the 536 investigated instances.

In Table 5.4 we finally provide an excerpt of instances that the lexicographer marked according to the labels described above.

The case of the verb “say” A further interesting datum has also emerged from the close observation of the error instances, one which indicates that the personal interpretation of the meaning of an ambiguous word in context by different annotators could not only have an impact on inter-annotator agreement figures, but be a root cause for several errors committed by disambiguation systems.

In fact, contextual information is known to be of fundamental importance for humans in the act of disambiguating a word (Chatterjee et al., 2012), but working with a finite inventory of word senses forces the annotator to fit a personal interpretation of meaning into at least one of the senses provided for in the inventory, which therefore acts as a filter. This operation can be anything but linear, and can turn into one of the fundamental causes of low agreement figures between different annotators (Passonneau et al., 2010; Martínez Alonso et al., 2015; Oba et al., 2020). As a case in point, eye tracking studies have shown that annotator fixation times, before a disambiguation choice is made, are extremely long on the list of glosses provided by the inventory (Joshi, Kanojia, and Bhattacharyya, 2013). Indeed, together with the examples and words that make up a synset, glosses represent the most important clues

LSI	factual (ALL_{SE}) semeval2007.d000.s016.t005
c-t	[...] recent increase in the number of persons living below the poverty level.
gold	lead a certain kind of life; live in a certain style
sys	inhabit or live in; be an inhabitant of
LSI	multi (ALL_{SE}) semeval2007.d000.s028.t008
c-t	[...] self-serving groups that "know a good thing when they see it" [...]
gold	a vaguely specified concern
LMMS	a special situation
LSI	ge_pos (ALL_{SE}) senseval2.d001.s045.t001
c-t	He assumed the missing [VERB] piece contained a gene [...]
LSI	ge_dis (ALL_{NS}) semeval2013.d003.s013.t002
c-t	[...] which have cultivated close ties with the Iraqi Oil Ministry [...]
gold	the finish of a contest in which the score is tied and the winner is undecided
Syntag.	a social or business relationship
LSI	ge_mis (ALL_{SE}) semeval2015.d001.s051.t004
c-t	You can set several graphs on the same view .
gold	outward appearance
LSI	ge_t/l (ALL_{SE}) senseval3.d001.s013.t011
c-t	[...] to prevent other legislators from " <u>bringing home</u> the bacon"

Table 5.4. Excerpt of labels assigned while carrying out the error analysis. Each block can show: label type, subset, and original instance identifier (LSI); original context and **target** (c-t); WordNet gloss for the sense chosen as gold answer (gold); gloss for the sense chosen from a given system (LMMS, Syntag.); gloss for the sense chosen by all systems (sys). Underlined text indicates the expected tokenization in the ge_t/l case reported.

on the basis of which the annotator makes his final choice (Kanojia et al., 2014).

In the course of our qualitative analysis of errors, numerous cases have been found in which different interpretations assigned to different ambiguous words have made it impossible for the system to perform proper disambiguation. By way of example, we report the analysis conducted on the verb “say”, for which we have identified a sense distribution totally disjoint in the test set with respect to the training set, particularly, with respect to the use of the verb as an introductory verb of direct speech.

In this case, we asked our lexicographer to review all instances of the verb “say”, both

sense	gloss	ALL	ALL _{SE}	SemCor
1	<i>express in words</i>	3/8	-	966/1,687
2	<i>report or maintain</i>	0/14	0/1	21/219
3	<i>express a supposition</i>	0/1	0/1	2/27
4	<i>have a certain wording</i>	-	-	4/8
5	<i>give instructions to</i>	-	-	0/8
6	<i>utter in a certain way</i>	-	-	1/4
7	<i>express nonverbally</i>	-	-	0/2
8	<i>utter aloud</i>	16/18	14/16	0/1
9	<i>state as one's opinion</i>	0/3	0/3	-
10	<i>recite a fixed text</i>	-	-	-
11	<i>indicate</i>	-	-	-
		19/44	15/21	994/1,956

Table 5.5. WordNet 3.0 sense distribution for the instances of the verb “say” in ALL, ALL_{SE} and in SemCor. Columns, left to right: number of sense according to the WordNet inventory (sense); gloss for the WordNet sense (gloss); following columns indicate the number of instances in which the verb appear as introductory to a direct speech (before the slash), versus the overall number of instances (after the slash). Bottom row shows aggregate values.

appearing in SemCor and in the ALL test set, and to mark the verb when it appeared as introductory to direct speech.

As can be seen in Table 5.5, the sense distribution of “say” changes radically within the training set (SemCor) and test set (ALL). Particularly, we observe how, in almost all cases in which the verb appear as introductory to direct speech, the test set makes use of the WordNet sense glossed as *utter aloud*, which appears only once in the training set, and not as an introductory verb to direct speech. In the training set, on the other hand, the vast majority of instances introducing direct speech appear labeled with the WordNet’s first sense *express in words*.

Not surprisingly, almost all the instances labeled with the *utter aloud* sense in the test set are erroneously disambiguated by all systems (in 16 out of the overall 18 cases), which, in the case of supervised architectures, have never been exposed to the *utter aloud* sense assigned to the verb “say” in an introductory position to direct speech. Our best guess to

explain this discrepancy may therefore lie in the existence of different annotation guidelines which establish precise sources of cues to exploit when uncertainty occurs, or simply in different interpretations given by the annotators to different linguistic phenomena. In the case reported, test set annotators, unlike training set annotators, have probably exploited the cue present in the WordNet examples, and noted that the only example where the verb is used as introductory to direct speech is precisely in the sense identified with the gloss *utter aloud*: “she said ‘Hello’ to everyone in the office.”

5.3 Analysis of 42D: a Challenge Set

The analysis we have conducted starting from traditional benchmarks for WSD has highlighted rather specific causes that justify the hypothesis that the current performance of disambiguation systems is not really representative of their actual capabilities. And yet, the phenomena discussed so far could be the reflection of peculiarities inherent in the analyzed datasets.

In order to have further proof of our findings, we therefore decided to build a new evaluation test set from scratch. Specifically, the dataset we present, called 42D, constitutes an innovation in the field of WSD evaluation exercises, acting in fact as a harder benchmark than its predecessors, or, quoting Belinkov and Glass (2019), a challenge set. 42D is thus the first challenge set for WSD, an English test set systematically covering a wide range of 42 domains, designed to provide a different and complementary evaluation tool for those systems that want to test their resilience towards those that currently represent critical factors for disambiguation, i.e. biases towards most frequent word senses and towards senses featured in the training data.

In what follows, we first describe the process by means of which we built and annotated 42D (Sections 5.3.1, 5.3.2 and 5.3.3), then, we proceed with the discussion concerning the results of the experiments we conducted on this new challenge set (Section 5.3.4).

5.3.1 Corpus and Domain Set

As our source corpus, we chose the British National Corpus (Leech, 1992, BNC), which, considering its well-balanced range of genres and topics, constituted a sound alternative to other corpora such as Gigaword (Parker et al., 2011) or Wikipedia. As regards our domain set, we narrowed down candidates to BabelDomains (Camacho-Collados and Navigli, 2017) and WordNet Domains (Bentivogli et al., 2004), but we opted for the former in order to avoid an exceedingly fine granularity of the labels. In fact, BabelDomains provides a mapping between synsets (concepts) in BabelNet (and, consequently, in WordNet) and its wide array

of 42 domain labels⁶ (ranging from `food_and_drink` and `history` to `nautics` and `meteorology`).

5.3.2 Building 42D

In order to build our dataset, we assigned domains to paragraphs in the BNC through a simple unsupervised technique that exploits counts of WordNet lexicalizations, which can be taken as cues that the text belongs to that domain, as they are associated with a synset mapped to a BabelDomain by Camacho-Collados and Navigli (2017).⁷

We automatically labeled each paragraph with the highest occurring domain among its tokens. As a starting point, we collected the set of senses from WordNet that satisfy the following criteria: (i) the sense is associated with some domain, and (ii) the lexicalization is polysemous and associated with at least one other domain⁸. We then employed the Stanford CoreNLP pipeline (Manning et al., 2014) to perform tokenization, lemmatization, PoS tagging, and sentence splitting in order to pre-process the BNC raw text, and split the corpus into paragraphs containing no more than 250 adjacent tokens (including punctuation). At this stage, we assigned each paragraph to a specific domain, simply by determining which, among the 42 domains, showed the highest number of distinct lexicalizations within a paragraph. After that, we assembled and ranked 25 paragraphs for each domain, also prioritizing paragraphs with the least number of repetitions among the domain lexicalizations. Finally, for each domain, we manually checked each of its 25 paragraphs in order to (i) ensure no paragraph duplicates were being included, and (ii) select the most suitable ones to represent and fit the domain,⁹. The dataset was therefore assembled as a result of the concatenation of the 42 chosen paragraphs, with an average paragraph length of 208 tokens (including punctuation), hence complying with the BNC Licence for paragraphs and other fragments (see also footnote 1 in Chapter 5).

⁶<http://babelnet.org/4.0/javadoc/it/uniroma1/lcl/babelnet/data/BabelDomain.html>

⁷Each BabelDomain label is associated with some set of synsets, with each synset having its words (e.g. `car`, `automobile`, and `machine`, for the WordNet synset “a motor vehicle with four wheels”). Hence, we can see each domain as a set of words obtained from its synsets.

⁸The second requirement was included to ensure a significant inter-domain ambiguity in our dataset.

⁹To prevent automatically-labeled paragraphs from being off-topic with respect to their domain, or unrepresentative of usual running text, the top 25 paragraphs per domain (i.e. the ones with more domain words) were manually validated and ranked by the annotators. Paragraphs containing lists – e.g. ingredients in a recipe, or neighbouring countries – were discarded in the process.

5.3.3 Dataset Annotation

The manual annotation of the paragraphs was performed by a single expert linguist with previous experience of tagging with WordNet 3.0, i.e., the sense inventory we chose to disambiguate 42D. The annotator reviewed the Part-of-Speech (PoS) tags¹⁰, along with automatic tokenization and lemmatization (see Section 5.3.2), while compounding multi-word expressions that are present in the WordNet inventory and resorting to multiple sense annotations for cases in which several WordNet senses could properly fit the context. The annotator was asked to annotate each content word in the paragraphs and, as a result of the whole annotation process, we collected an overall figure of 3,688 annotated instances.

At this stage, we proceeded to transform 42D from a standard all-words test set into a challenge set representative of the criticalities identified in Section 5.2.1. To this end, we have automatically excluded from 42D all those sense annotations which were featured in SemCor or in WordNet as most frequent senses. As a result, we obtained a challenge set containing exactly 451 unique instances.

Once collected, we asked a second expert linguist to perform a blind annotation over a sample of 350 polysemous instances taken across all domains from the challenge set. Our evaluation resulted in an agreement of 79.6%, which is in line with other evaluation datasets.¹¹ The reason we chose to employ no more than two highly-trained linguists to carry out the annotation (as previously done in SemEval-2007) stems from the fact that well-known issues in reaching an adequate inter-annotator agreement when employing non-expert annotators exist (Pradhan et al., 2007), as well as from accounting for the hypothesis that performance depends on the quality of the word sense inventory, rather than the number of annotators (Passonneau et al., 2012).

Finally, to assess the accuracy of the annotations, we performed a standard statistical evaluation: according to the Cochran's Sample Size Formula, out of a population of 451 polysemous instances in 42D, and with a confidence interval of 5, a number of 350 instances constitutes a statistically representative sample size to compute inter-annotator agreement and thus ensure the reliability of the tagging.

5.3.4 Results and Discussion

We first carried out a standard evaluation on 42D, simply testing all of the systems we analyzed in this Chapter (see Section 5.1) on this new challenge set.

¹⁰According to the coarse-grained PoS tags in the universal PoS tagset of Petrov, Das, and McDonald (2012).

¹¹We report the raw inter-annotator agreement (in line with other WSD evaluation datasets such as SemEval-2007 and SemEval-2015), given that Cohen's Kappa is not well defined when multiple tags are allowed (Palmer, Dang, and Fellbaum, 2007).

setting	#inst (mono)	random \pm sstd	BEM	EWISER	Gener.	Gloss.	LMMS	Syntag.
ALL	7253 (1301)	36.3 \pm 0.44	79.0	80.1	76.3	76.9	75.4	71.7
42D	451 (0)	24.7 \pm 1.94	40.6	46.3	43.5	35.9	29.3	23.3

Table 5.6. F1 scores on the concatenation of the datasets (ALL) and on 42D. Columns, left to right: test set (setting); total number of annotations (#inst) of which monosemous instances are indicated between round brackets (mono); F1 score for a random baseline and its sample standard deviation (random \pm sstd); F1 scores for BEM, EWISER, Generatory (Gener.), GlossBERT (Gloss.), LMMS and SyntagRank (Syntag.). **Bold** is best.

setting	BEM	EWISER	Gener.	Gloss.	LMMS	Syntag.	gold
ALL(SC)	87.4%	87.0%	85.9%	88.6%	88.5%	88.8%	84.3%
42D(SC)	49.4%	41.7%	45.7%	55.4%	56.8%	65.2%	0.0%
ALL(WN1st)	72.6%	70.3%	69.0%	74.8%	65.9%	81.1%	65.2%
42D(WN1st)	36.4%	33.3%	31.9%	41.0%	42.1%	59.0%	0.0%

Table 5.7. Times (percentages) an investigated system predicts a sense that is (top) featured in SemCor or (bottom) appears as WordNet 1st sense. Results are reported both on ALL and on 42D. Right column shows the times (percentages) the gold standard annotations on the test sets coincide with senses featured in SemCor or appearing as WordNet 1st senses. **Bold** is closest to gold.

The results shown in Table 5.6 empirically validate the existence of significant biases affecting all WSD systems, regardless of the approach used, both towards the most frequent senses and towards the word senses featured in the training set. Similarly to what we observed with the ablation studies reported in Table 5.1, performances for state-of-the-art WSD systems on 42D fail to reach a score of 50, with EWISER and Generatory being the most resilient models so far, owing to their more heterogeneous training sets.

From another point of view, as shown in Table 5.7, it is possible to observe how, in the context of ALL, the percentage of system predictions for senses featured in SemCor or as WordNet first senses, is closer to the real extent of the senses used in the gold standard. As a matter of example, 88.8% of SyntagRank’s sense predictions, which is the one showing the wider gap with respect to the gold standard, are also featured in SemCor, but the percentage of gold annotations using SemCor senses is just at a 4.5 percentile points below. The same cannot be said in the case of 42D, where the closest gap with respect to the gold annotations is represented by Generatory, which predicts only 31.9% of the times a WordNet first

setting	monosem	avg flexibility	avg solving	amb+mono	amb-0-mono
ALL	17.9%	1.060	0.39	5.9	6.9
Senseval-2	19.1%	1.058	0.40	5.5	6.6
Senseval-3	16.8%	1.025	0.36	6.8	7.9
SemEval-2007	5.7%	1.009	0.26	8.5	8.9
SemEval-2013	20.7%	1.009	0.41	4.9	5.9
SemEval-2015	18.4%	1.236	0.42	5.5	6.5
ALL _{SE}	<u>0%</u>	1.026	<u>0.18</u>	<u>9.4</u>	<u>9.4</u>
SemCor	16.6%	<u>1.003</u>	0.35	6.9	8.1
42D	<u>0%</u>	1.044	0.26	6.5	6.5

Table 5.8. Data breakdown for different training, test and ablation sets. Columns, left to right: setting; percentage of monosemous instances in the set (monosem); average number of gold annotations available per polysemous instance (avg flexibility); average random solving chance for instances in the set, computed as the mean value between all instances in the set, each obtained by dividing the number of gold senses associated with a given lemma instance by the whole number of senses associated to that lemma in WordNet (avg solving); ambiguity including monosemous instances (amb+mono); ambiguity not including monosemous instances (amb-0-mono). **Bold** is lowest difficulty. Underlined is highest difficulty.

sense. In a similar way to what we already found out with the subset of shared errors ALL_{SE} (Section 5.2.2), even though gold annotations feature few or no SemCor and/or WordNet first senses, systems will still continue to predict them in a large percentage of cases.

We then devised another small experiment to determine if there are any objective criteria which make 42D more challenging than other traditional datasets, particularly, in order to verify if the difficulty of 42D is actually due to the fact of representing a stressing environment for WSD systems, and not caused by other accidental and undesired factors. To this end, we tested all of the traditional evaluation datasets (Senseval-2, Senseval-3, SemEval-2007, SemEval-2013 and SemEval-2015), their concatenation (ALL), and the set of the shared errors (ALL_{SE}), as well as 42D and the SemCor training set, to determine the following properties:

- The percentage of monosemous instances in the investigated dataset (monosem);
- The average number of word senses that the dataset provides as gold annotations for polysemous instances (avg flexibility);

- The average random solving chance for each instance in the dataset D , computed as the mean value between all instances in D , each obtained by dividing the number of gold senses associated with a given lemma instance by the whole number of senses associated to that lemma in WordNet (avg solving);
- The ambiguity level, including monosemous instances. Computed, as in Raganato, Camacho-Collados, and Navigli (2017), as the total number of candidate senses for the lemmas in D , divided by the number of sense annotations in D (amb+mono).
- The ambiguity level of D , this time, computed without accounting for monosemous instances (amb-0-mono).

The results shown in Table 5.8 show that – notwithstanding its intended lack of coverage for monosemous instances – none of the factors examined justifies the higher difficulty of 42D with respect to other traditional datasets. As can be noted from the results, both ambiguity levels (amb+mono and amb-0-mono) of 42D are in fact perfectly comparable to those of other test sets, if not lower, as is the case of Senseval-3 or SemEval-2007. The same goes for the average random solving chance (avg solving), which is equal to that of SemEval-2007, even if the results in terms of F1 score for the best system in the literature, i.e. EWISER, are 75.2 on SemEval-2007, but only 46.3 on 42D.

This, and the fact that the flexibility in assigning multiple labels (avg flexibility) of 42D is even higher than in Senseval-3, SemEval-2007 and SemEval-2013, clearly support the idea that no other unwanted factor plays a significant role in determining the difficulty of our challenge set.

To conclude our roundup of experiments on 42D, in Table 5.9 we report the PoS breakdown data for 42D, along with comparisons to ALL and ALL_{SE}. Though, the most interesting value that we have decided to include in this Table is that showing the percentage of predictions shared by the systems (“shared preds”), i.e. the percentage of cases in which the systems agree when choosing a specific sense to disambiguate an instance in a given test set (calculated as the pairwise average of the systems’ responses). As it can be seen, the shared predictions reach a 77.5 percentage on ALL (considering all parts of speech) and a 69.6 percentage on the subset ALL_{SE}. On 42D instead, although the challenge set is less ambiguous than ALL_{SE}, they only reach a percentage of 56.6. This fact is likely attributable to the heterogeneous nature of the domains (and hence, contexts) covered in 42D, a feature not found in traditional evaluation datasets.

In fact, (i) the second Senseval evaluation campaign (Edmonds and Cotton, 2001) released an evaluation corpus not adjusted for domain-specific WSD (being a collection of texts from various sources), and among the other evaluation exercises that have been devised

setting	pos	#inst	percentage	amb+mono	unique_sks	unique_syns	shared_preds
SemCor	ALL	226,036	-	6.9	33,362	25,916	-
ALL	ALL	7,253	-	5.9	3,669	3,239	77.5%
ALL	NOUN	4,300	59.3%	4.8	1,943	1,745	79.9%
ALL	VERB	1,652	22.8%	10.4	991	838	67.0%
ALL	ADJ	955	13.2%	4.0	543	498	81.8%
ALL	ADV	346	4.8%	3.1	192	158	86.3%
ALL _{SE}	ALL	536	-	9.4	401	386	69.6%
ALL _{SE}	NOUN	255	47.6%	6.7	182	179	72.1%
ALL _{SE}	VERB	203	37.9%	14.0	153	145	61.8%
ALL _{SE}	ADJ	59	11.0%	6.8	53	49	81.7%
ALL _{SE}	ADV	19	3.5%	5.8	13	13	80.4%
42D	ALL	451	-	6.5	395	376	56.6%
42D	NOUN	273	60.5%	5.6	236	225	57.3%
42D	VERB	91	20.2%	10.7	81	77	50.3%
42D	ADJ	64	14.2%	5.0	62	58	60.1%
42D	ADV	23	5.1%	4.6	16	16	64.1%

Table 5.9. Part-of-speech breakdown for SemCor, ALL, the shared set of errors on ALL (ALL_{SE}) and 42D. Columns, left to right: setting; part-of-speech (pos); number of instances (#inst); ambiguity value (amb+mono), including monosemous instances, as seen in Raganato, Camacho-Collados, and Navigli (2017); number of unique sense keys in the subset, according to WordNet 3.0 (unique_sks); number of unique synsets in the subset, according to WordNet 3.0 (unique_syns); percentage of predictions shared by analyzed systems on the subset (shared_preds).

so far, (ii) the Senseval-3 task 1 (Snyder and Palmer, 2004) featured a test data consisting of three texts representing three distinct genres (editorial, news story, and fiction), (iii) the SemEval-2007 task 17 (Pradhan et al., 2007) covered a 3,500 words section of the Wall Street Journal corpus (Paul and Baker, 1992), (iv) the SemEval-2010 task 17 (Agirre et al., 2010) had a test corpus focused exclusively on the environment domain, (v) the SemEval-2013 task 12 (Navigli, Jurgens, and Vannella, 2013) had a test set consisting of 13 articles obtained from three editions of the workshop on Statistical Machine Translation (WSMT)¹² and covering domains ranging from sports to financial news, and (vi) the SemEval-2015 task 13 (Moro and Navigli, 2015) featured four documents collected from the OPUS project¹³ in three specific domains (biomedical, maths and computers, and social issues).

¹²<http://www.statmt.org>

¹³<http://opus.nlpl.eu/>

So, to date, no dataset for WSD has been specifically designed to encompass a wide and systematic selection of semantic domains such as 42D. This fact, which seems to substantiate why the shared prediction percentages on our challenge set are significantly lower than in previous test sets, makes 42D an even more effective tool to probe into the quirks and error classes pertaining only to specific systems.

5.4 Conclusion

In this Chapter we provided an answer to a much sought for question, i.e.: “are current state-of-the-art WSD systems, today apparently performing on par with humans, actually capable of disambiguating?”

We started by analyzing existing evaluation sets and, by means of different ablation studies, we showed how the disambiguation performance can be severely overestimated and hence, should not be merely taken at face value. Specifically, we provided further empirical evidence to support the theories that see all systems, regardless of the underlying approaches, suffering from biases, both towards the data featured in the training set, and towards the most frequent word senses. Secondly, we analyzed the subset of traditional evaluation exercises composed of the errors shared by all systems, and noted two more factors that make current performance figures unreliable, this time, resulting in underestimated system scores: on the one hand, set lower by the numerous and different errors inherent in the test sets, and on the other, undermined by interpretation problems which arise as a result of the structure of the employed inventory.

So what does it mean to surpass the inter-annotator agreement figure? Does this mean that current disambiguation systems are really capable of performing the task like their human counterparts? More likely, it means that these systems are simply becoming more efficient at mimicking specific groups of annotators and their particular pigeonholing strategies, apt at fitting meaning representations into the rigid scheme provided by a fixed sense inventory. To further validate our findings, we have also created the first challenge set for WSD: 42D. This new dataset – the first to be released after five years – contains no annotations for word senses featured in SemCor, nor appearing as most frequent senses in WordNet, and shows a very different picture of the current disambiguation performances that we know, with F1 scores for best performers still lingering behind 50 points.

We will make the 42D English WSD challenge set freely available, knowing that it can be employed as a fundamental and complementary tool to already existing datasets, particularly, for probing the actual disambiguation capabilities of systems with respect to those that we identified as the most difficult obstacles to overcome.

Chapter 6

Summary

As unequivocally stated since its very title, with this thesis we delved into what we refer to as the “uncharted territories” of WSD. In fact, we have seen how a long list of open problems has yet to be dealt with before the research community will be able to call the AI complete task of WSD finally solved. Furthermore, we also made clear why the albeit astonishing performances attained by supervised systems could easily mislead researchers and the general public into thinking that machines have actually rivaled human performance.

WSD is far from being resolved, and with this thesis we have paved the way for a better understanding of the reasons why. To prove our hypothesis, we started by identifying three critical challenges affecting the field and we provided as many possible strategies and solutions.

First of all, we have shown how the structure of commonly used LKBs is largely overlooked, to the point of being devoid, almost completely, of a fundamental semantic relation such as that brought about by syntagmatic combinations (see Chapter 3). Considering how impactful and viable a reevaluation of knowledge-based approaches could be in a moment in which models employ DL techniques requiring increasingly prohibitive architectures (see also Section 2.3.1), it is clear why our findings can be of immediate interest. Particularly, the simple injection of structured syntagmatic knowledge into a pre-existing LKB biased towards paradigmatic relations have led knowledge-based systems to attain hitherto out of reach results, with a significant boost of 4.4 F1 points over the previous state of the art in English, and equivalent results in a multilingual setting (Maru et al., 2019).

From a different perspective, we analyzed the flaws underlying the usage of traditional sense inventories for WSD and proposed a generative formulation to recast the task. As a matter of fact, several of the issues currently affecting WSD stem from the widespread use of sense inventories (see also Section 2.1.1) which, despite providing a computationally handy solution in being finite lists of possible options, are often inadequate in terms of, among

other, sense granularity, interpretability and lack of coverage. We hence borrowed from DM the technique of generating textual definitions, but with a significant difference. While DM aims to provide a means to make the content of an embedding explicit and human-readable, we employed its paradigms to provide contextual, *ad hoc* definitions for arbitrarily-sized targets in context (see Chapter 4). By means of this, not only we are not anymore limited to the word senses as enumerated in a given sense inventory, but we can generate definitions for previously unseen words and phrases (Bevilacqua, Maru, and Navigli, 2020).

Finally, in line with the above, we conducted a detailed study to further highlight the inadequacy of traditional word sense inventories in WSD (see Chapter 5). To do so, we performed a series of ablation studies on traditional evaluation exercises, and collected the set of shared mistakes that several SOTA systems commit with respect to the standard English evaluation framework of Raganato, Camacho-Collados, and Navigli (2017).

If our objective was to use these errors to better understand the “actual” state of things, and to have a means to better interpret the idea of WSD systems performing on par with human annotators, our results have given evidence of three main causes why the reported system performances might substantially differ in reality from the same systems’ actual performances.

On the one hand, in fact, independently of the methodology employed, we have shown how (i) systems share a bias towards what – according to the training set and the sense inventory used – are the most frequent senses, moreover, demonstrating how supervised systems’ performance is way lower when instances included in the training set are excluded from the test set. On the other hand, we have claimed reported performances are further distorted in that (ii) the structure of a sense inventory forces human annotators to make suboptimal choices when devising a gold standard (see also Section 5.2.2.2), choices that are not tied to common or shared knowledge, but rather dictated by the information as presented in the inventory itself, and also in that (iii) a significant number of errors are featured in traditional evaluation benchmarks, hence undermining the reliability of evaluation figures.

To further demonstrate the findings reported in Chapter 5, we also devised an entirely new evaluation exercise, a challenge set named 42D, specifically tailored to reproduce the harshest conditions for current state-of-the-art WSD systems.

6.1 Perspectives

As thoroughly anticipated, the number of issues related to WSD is large (see also 2.3) and their minute description and review is beyond the scope of this work. Our main aim was rather to prove that WSD has still a lot of work to be carried out, and in this thesis, we

highlighted three, focused problem, along with providing as many solutions so as to support our statement. For all this, we firmly believe our findings should not be considered as finishing lines, but rather, as start points. In light of this, in what follows, we build on top of our work, and foresee potential developments and future works for each of the topic we dealt with in this thesis.

6.1.1 On Knowledge Bases

Despite having attained so far unprecedented performances thanks to the inclusion of syntagmatic information (Maru et al., 2019; Scozzafava et al., 2020), still, knowledge-based systems linger behind supervised systems by several points according to standard evaluation benchmarks (Bevilacqua and Navigli, 2020). Nonetheless, it should be noted that such improvement has been brought about by the inclusion of less than 90,000 lexical-semantic combinations between nouns and verbs only. This, and the fact that the performance keeps growing as a factor of the increasing number of relations (see also 3.1.5), indicate that results can still improve as more manually-disambiguated combinations are added to the LKB. Additionally, syntagmatic relations have not been type-labeled so far (Maru et al., 2019), and their categorization could be a useful tool to perform bootstrapping over resources containing non-disambiguated combinations (Chen and Liu, 2011; Cousot and Lafourcade, 2017), thus triggering a virtuous circle which could lead syntagmatic relations to become the keystone to finally match supervised systems' performance.

6.1.2 On Definition Generation

To generate a definition instead of picking one from a finite inventory of many is a huge innovation in the field of WSD, one that opens up new and unexplored research scenarios. With our work, we introduced our methodology with a focus on the English language only (Bevilacqua, Maru, and Navigli, 2020). A first step forward would therefore be represented by an attempt to transfer the same approach to other languages. The major obstacle in doing this could lie in the paucity of freely available resources and machine-readable dictionaries in other languages.

Another interesting line of research could also focus on (i) the generation of precise definitions for named entities, whereas Generatory focuses primarily on common concepts, as well as on (ii) discontinuous targets, i.e. when the target to be defined is separable, such as in a phrasal verb (e.g. “you can take Jacob on”, where the underlined tokens should make up a single target).

It is also worth noting and commenting here two potential critical points that can

undermine this approach, particularly, in the context of the statements made in this same thesis. On the one hand, it is legitimate to ask how a characterization of senses no longer based on predefined inventories can be used downstream, and on the other hand, it is appropriate to observe how, while an approach such as the one described in Chapter 4 poses a possible solution to the problem of fixed sense inventories, it still employs those very paradigms we sought alternatives for with the re-evaluation of knowledge-based systems in Chapter 3.

To respond to the first potential observation, the approach described in Chapter 4, unlike traditional approaches for WSD, can already be viewed as an application in itself, for example, as an integrative tool in ebook readers, particularly useful for learners of a language, or as a support tool in the creation of new dictionaries. Furthermore, representing word senses by means of glosses generated in a *continuum* does not preclude the possibility of using this data in downstream applications, such as, *inter alia*, the use of definition embeddings as a means to boost BERT performances (Pappas, Mulcaire, and Smith, 2020).

As for the potential criticisms directed towards the fact that an approach of this type implies the use of Language Models such as those whose usage we have invited to exercise caution with since the introduction of this thesis (see also Section 2.3.1), it is primarily good to reiterate how the spirit of this work is to place the first stones along different tracks, albeit distant from each other, each representative of a distinct issue under the wide scope of WSD applications.

Having said that, it should be noted that we do not wish to condemn the use of Language Models at all, but rather invite the research community to quickly figure out alternative solutions and to exercise caution with *overparametrization* as the only way to progress the state of the art. To make mention of our own case, Generationary and the important results derived from its application did not require out-of-the-ordinary infrastructures (see Section 4.4.4). And yet, it is certainly advisable to right away direct future efforts in an attempt to prevent such an approach to see its further improvements exclusively tied to the use of increasingly prohibitive and potentially hazardous infrastructures.

6.1.3 On Disambiguation Errors and System Performance

Similarly to Generationary, the analysis of the shared errors in WSD we detailed in Chapter 5 has been specifically conducted on the English language only, given that it is the only language featured in the standard evaluation benchmarks (Raganato, Camacho-Collados, and Navigli, 2017). Nonetheless, it would be interesting to see both how the common issues we identified in our study affect systems dealing with other languages, as well as to eventually discover language-specific phenomena.

Finally, a similar error analysis should be likewise conducted to investigate systems' performance when coarser sense inventories are employed (Snow et al., 2007; Lacerra et al., 2020), so as to determine whether reducing the fine granularity of word senses would also help mitigating the impact of rare and out-of-training senses in the test sets, without going to the detriment of interpretability.

Appendix A

Chapter 4: Supplementary Materials

A.1 Additional Results on DM

In Table A.1 we report results, for the Definition Modeling evaluation described in Section 4.4.1, on two additional datasets.

NOR (Noraset et al., 2017) includes data from the GCIDE and WordNet. It features only “static” pairs, in which the context coincides with the word to be defined. Nonetheless, each lemma can be associated with multiple definitions.

GAD (Gadetsky, Yakubovskiy, and Vetrov, 2018) collects context-target pairs and definitions from `oxforddictionaries.com`.

The target lemma is not present in all contexts, so in these cases we have prepended the lemma according to the following template: ‘*lemma: context*’.¹

	model	ppl↓	BL↑	R-L↑	MT↑	BS↑
	Random	-	0.2	6.3	1.9	69.0
NOR	Noraset et al. (2017)	48.2	-	-	-	-
	Ishiwatari-NOR*	-	1.9	15.7	5.0	72.9
	Gen-NOR	28.6	3.8	17.7	8.1	72.9
	Random	-	0.2	8.7	2.8	68.6
GAD	Gadetsky, Yakubovskiy, and Vetrov (2018)	43.5	-	-	-	-
	Mickus, Paperno, and Constant (2019)	34.0	-	-	-	-
	Ishiwatari-GAD*	-	2.5	18.7	7.0	72.8
	Gen-GAD	12.3	9.9	28.9	12.8	77.9

Table A.1. DM evaluation results. Columns: perplexity, BLEU, Rouge-L, METEOR, BERTScore (ppl/BL/R-L/MT/BS). Row groups are mutually comparable (**bold** = best). ↑/↓: higher/lower is better. *: re-trained.

¹The train/dev/test splits of NOR and GAD are disjoint in the lemma of the target words.

A.2 Perplexity

Perplexity captures the confidence of the model in outputting a certain sequence.

In approaches with word-level tokenization, evaluated at word-level, perplexity can be computed by exponentiating the negative log-likelihood that is used for training:

$$PPL_w^w = \exp\left(-\sum_{w \in V} P(w|c, t, \bar{h}) \ln \hat{P}(w|c, t, \bar{h})\right) \quad (\text{A.1})$$

$$= \exp\left(-\ln \hat{P}(\bar{w}|c, t, \bar{h})\right) \quad (\text{A.2})$$

where c is the context, t is the target, V is the vocabulary, \bar{w} is the gold word, and \bar{h} is the gold history of previous tokens.

Generational models employ subword-level tokenization, but we can still obtain the word-level probabilities by applying the chain-rule of conditional probability:

$$PPL_w^s = \exp\left(-\ln \prod_{i=1}^{|\bar{w}^*|} \hat{P}(\bar{w}_i^*|c, t, \bar{h}, \bar{w}_{1:i-1}^*)\right) \quad (\text{A.3})$$

where \bar{w}^* is the n -ple that is the subword split of \bar{w} , e.g. $\langle \text{token}, \text{\#ization} \rangle$ for tokenization.

Do we maintain full comparability? There are two issues here. The first stems from the fact that, thanks to the application of the chain rule, the vocabulary is open, i.e. the support of the subword model is the set of possible words, so that every item receives non-zero probability.

On the contrary, a word-level model without some kind of backoff strategy has a closed vocabulary. If the evaluation set includes a word outside V , the closed vocabulary model has a special $\langle \text{unk} \rangle$ token, on which it is trained to concentrate all the probability mass that the open vocabulary model, instead, would spread over all the possible words which are not in V . This entails an unfavorable advantage of the closed vocabulary model over the open vocabulary.

Moreover, there is an additional complication arising from the fact that, while the subword tokenizers are usually deterministic, i.e. any word is always split in the same way, there might be multiple legal subword splits according to the vocabulary, and to obtain the probability of the word, we would need to marginalize over all splits. In other words, we would need to marginalize by summing the probability of $\langle \text{token}, \text{\#ization} \rangle$, $\langle \text{token}, \text{\#iz}, \text{\#ation} \rangle$, $\langle \text{to}, \text{\#ken}, \text{\#ization} \rangle$ and so on. This is very burdensome, and practically we only consider the deterministic split produced by the tokenizer. In doing this, we underestimate the probability of the word and thus, overestimate the perplexity of the subword level model.

A.3 NLG Measures Details

In order to ensure comparability, here we report the BLEU, ROUGE, METEOR, and BERTScore configurations that we used. A scorer is available as part of the provided software.

BLEU We employed the reference implementation of corpus BLEU provided in the `sacrebleu` package (Post, 2018)². We use default parameters.

Signature:

```
BLEU+case.mixed+numrefs.1+smooth  
.exp+tok.13a+version.1.3.6.
```

ROUGE We have employed the Python `rouge` library (<https://github.com/pltrdy/rouge>).

METEOR We have employed the Java `meteor` library (<https://www.cs.cmu.edu/~alavie/METEOR>), version 1.5. METEOR is calculated using the `-norm` and `-noPunct` flags.

Signature:

```
meteor-1.5-wo-en-norm_nopunct-  
0.85_0.2_0.6_0.75-ex_st_sy_pa-1.0  
_0.6_0.8_0.6
```

BERTScore We evaluate using the Python `BERTScore` (https://github.com/Tiiiger/bert_score) library, with the `roberta-large-mnli` model and default parameters.

Signature:

```
roberta-large-mnli_L19_no-idf  
_version=0.3.0(hug_trans=2.8.0)
```

A.4 Generation Examples

In the evaluation of NLG systems, human qualitative assessment is very important. Therefore, we have chosen to report a fair number of non cherry-picked, zero-shot generation examples, produced by means of our GEN-UNI model.

²<https://github.com/mjpost/sacreBLEU>

c_1 : <u>Good news.</u>
\hat{g}_1 : (New Testament) The gospel as revealed by Jesus to the apostles.
g_1 : Any news that arouses feelings of joy or eases anxiety.

c_2 : <u>Uneven margin.</u>
\hat{g}_2 : A margin that is not uniform.
g_2 : A margin that is not perfectly leveled.

c_3 : <u>Early diagnosis.</u>
\hat{g}_3 : The diagnosis of a condition before symptoms appear.
g_3 : A diagnosis that is made at an initial stage of a disease.

c_4 : <u>Sincere friendship.</u>
\hat{g}_4 : A friendship that is not based on deceit or hypocrisy.
g_4 : Friendship marked by genuine feelings of benevolence.

c_5 : <u>Painful performance.</u>
\hat{g}_5 : A performance of a piece of music that is difficult to play.
g_5 : A performance that is exceptionally bad.

c_6 : <u>Courageous heart.</u>
\hat{g}_6 : A heart that is strong enough to endure adversity.
g_6 : The feelings of a person that is not afraid of getting hurt.

c_7 : <u>Inaccurate thermometer.</u>
\hat{g}_7 : A thermometer that is inaccurate in measuring temperature.
g_7 : A thermometer that indicates the wrong temperature.

c_8 : <u>New friend.</u>
\hat{g}_8 : A friend who has recently come into one’s life.
g_8 : A recently made friend.

c_9 : <u>Familiar guest.</u>
\hat{g}_9 : A person who is a regular customer or client of a hotel, restaurant, etc.
g_9 : A well known guest.

c_{10} : <u>Vivacious hostess.</u>
\hat{g}_{10} : A woman who entertains guests at their home and makes them feel welcome.
g_{10} : A woman host who shows liveliness.

Table A.2. Random sample of Generatory definitions (\hat{g}) for Hei++ contexts (c). g : gold definition.

In Table A.2 we show Generatory outputs and gold definitions for 10 randomly sampled phrases in the Hei++ dataset. In addition, in Table A.3 we report gloss generation examples for random words and noun phrases taken from the `webtext` corpus included in

c_1 : Hipster guys #2 & #3.	\hat{g}_1 : Members of the hipster subculture.
c_2 : Because you have <u>no estrogen</u> ...you could not possibly understand what chocolate means to a woman.	\hat{g}_2 : The absence of estrogen in the body.
c_3 : <u>Mozilla Firebird bookmark toolbar</u> doesn't hide command click to open link in new tab.	\hat{g}_3 : A toolbar in the Mozilla browser.
c_4 : Guy #2: On <u>that table</u> over there.	\hat{g}_4 : The (usually specified) table on which something is located.
c_5 : Man, why do <u>chip manufacturers</u> always put the crumbs at the bottom of the bag?	\hat{g}_5 : A company that manufactures potato chips.
c_6 : Install new theme without <u>an internet connection</u> .	\hat{g}_6 : A connection to the internet.
c_7 : It's a <u>true story</u> .	\hat{g}_7 : A fictional story based on real events.
c_8 : Seeking <u>female friend</u> , 40s, attractive 5 ft 5+.	\hat{g}_8 : A friend who is not a man.
c_9 : Clicking Get New Extensions in <u>Options Dialog</u> opens a new browser...	\hat{g}_9 : A dialog box in a web browser, allowing the user to select from a number of options.
c_{10} : And he goes to <u>my other roommate</u> for advice.	\hat{g}_{10} : The other person with whom one shares a room.

Table A.3. Generatory definitions (\hat{g}) for random targets and contexts (c) excerpted from `webtext`.

the NLTK suite Loper and Bird (2002). We have excluded swear words, slurs, numbers, and noun phrases consisting entirely of named entities. Moreover, every sampled item whose target was featured in our training set was filtered out.

Bibliography

- Agirre, Eneko, Oier Lopez De Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. Semeval-2010 task 17: All-words Word Sense Disambiguation on a Specific Domain. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 75–80, Uppsala, Sweden.
- Agirre, Eneko, Oier López de Lacalle, and Aitor Soroa. 2018. The risk of sub-optimal use of open source NLP software: UKB is inadvertently state-of-the-art in knowledge-based WSD. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 29–33, Association for Computational Linguistics, Melbourne, Australia.
- 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 Aitor Soroa. 2007. SemEval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Association for Computational Linguistics, Prague, Czech Republic.
- Agirre, Eneko and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41, Athens, Greece.
- Al-Saeedan, Wojdan and Mohamed El Bachir Menai. 2015. Swarm intelligence for natural language processing. *International Journal of Artificial Intelligence and Soft Computing*, 5(2):117–150.
- Amodei, Dario and Danny Hernandez. 2018. AI and Compute. Retrieved from <https://blog.openai.com/ai-and-compute>.
- Atkins, Beryl TS. 1992. Tools for computer-aided corpus lexicography: the Hector project. *Acta Linguistica Hungarica*, 41(1/4):5–71.

- Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, MI, USA.
- Bar-Hillel, Yehoshua. 1960. The present status of automatic translation of languages. In *Advances in computers*, volume 1. Elsevier, pages 91–163.
- Barba, Edoardo, Luigi Procopio, Niccolo Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. MuLan: Multilingual Label propagation for Word Sense Disambiguation. In *Proceedings of IJCAI*.
- Barbu, Andrei, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, pages 9453–9463.
- Belinkov, Yonatan and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.
- Belinkov, Yonatan and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.
- Bender, Emily M. and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online.
- Bentivogli, Luisa, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the WordNet Domains Hierarchy: Semantics, Coverage and Balancing. In *Proc. of MLR*, pages 101–108.
- Bentivogli, Luisa and Emanuele Pianta. 2004. Extending WordNet with Syntagmatic Information. In *Proceedings of second global WordNet conference*, pages 47–53, Brno, Czech Republic.
- Bevilacqua, Michele, Marco Maru, and Roberto Navigli. 2020. Generationary or: “how we went beyond word sense inventories and learned to gloss”. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online.

- Bevilacqua, Michele and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online.
- Bhatia, Parminder, Busra Celikkaya, Mohammed Khalilia, and Selvan Senthivel. 2019. Comprehend Medical: a Named Entity Recognition and Relationship Extraction Web Service. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1844–1851.
- Blevins, Terra and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online.
- Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Bond, Francis and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria.
- Boyd-Graber, Jordan, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding Dense, Weighted Connections to WordNet. In *Proceedings of the third international WordNet conference*, pages 29–36, South Jeju Island, Korea.
- 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.
- Brown, Alan S. 1979. Priming effects in semantic memory retrieval processes. *Journal of experimental psychology: Human learning and memory*, 5(2):65.
- Brown, Tom B, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Brugman, Claudia and George Lakoff. 1988. Cognitive topology and lexical networks. In *Lexical ambiguity resolution*. Elsevier, pages 477–508.
- Calvo, Hiram and Alexander Gelbukh. 2015. Is the Most Frequent Sense of a Word Better Connected in a Semantic Network? In *Proc. of ICIC*, pages 491–499.

- 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*, pages 223–228, Valencia, Spain.
- Camacho Collados, José, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.
- Card, Dallas, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. *arXiv preprint arXiv:2010.06595*.
- Chang, Ting-Yun and Yun-Nung Chen. 2019. What does this word mean? Explaining contextualized embeddings with natural language definition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6064–6070, Hong Kong, China.
- Chang, Ting-Yun, Ta-Chung Chi, Shang-Chi Tsai, and Yun-Nung Chen. 2018. xSense: Learning sense-separated sparse representations and textual definitions for explainable word sense networks. *arXiv preprint arXiv:1809.03348*.
- Chatterjee, Arindam, Salil Joshi, Pushpak Bhattacharyya, Diptesh Kanojia, and Akhlesh Kumar Meena. 2012. A study of the sense annotation process: Man v/s machine. In *GWC 2012 6th International Global Wordnet Conference*, page 79.
- Chen, Junpeng and Juan Liu. 2011. Combining ConceptNet and WordNet for Word Sense Disambiguation. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 686–694, Chiang Mai, Thailand.
- Chklovski, Timothy and Rada Mihalcea. 2003. Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation. In *Proc. of RANLP*.
- Chowdhary, KR. 2020. Natural language processing. In *Fundamentals of Artificial Intelligence*. Springer, pages 603–649.
- Cousot, Kévin and Mathieu Lafourcade. 2017. Explicative Path Finding in a Semantic Network. In *Proceedings of the Computing Natural Language Inference Workshop*.

- Cuadros, Montse, Lluís Padró, and German Rigau. 2012. Highlighting relevant concepts from topic signatures. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3841–3848, Istanbul, Turkey.
- 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.
- De Deyne, Simon, Daniel J Navarro, and Gert Storms. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior research methods*, 45(2):480–498.
- Delli Bovi, Claudio, 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)*, pages 594–600, Vancouver, Canada.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.
- Dirven, R and M Verspoor. 1998. Cognitive explorations of language and linguistics. *Amsterdam/New York*.
- Edmonds, Philip and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France.
- Edmonds, Philip and Adam Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering*, 8(4):279–291.
- Erk, Katrin and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 440–449, Singapore.
- Erk, Katrin, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 10–18.

- 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.
- Fauconnier, Gilles and Mark Turner. 2008. *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.
- Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.
- Francis, W. Nelson and Henry Kucera. 1979. Brown Corpus Manual. Technical report, Department of Linguistics, Brown University.
- Gadetsky, Artyom, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 266–271, Melbourne, Australia.
- Gale, William A., Kenneth W. Church, and David Yarowsky. 1992. A method for disambiguating word senses in a corpus. *Computers and the Humanities*, 26:415–439.
- Gambetta, J and S Sheldon. 2019. Cramming more power into a quantum device. *IBM Research Blog*.
- Gutiérrez Vázquez, Yoan, Antonio Fernandez Orquín, Andrés Montoyo Guijarro, and Sonia Vázquez Pérez. 2010. UMCC-DLSI: Integrative Resource for Disambiguation Task. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 427–432, Uppsala, Sweden.
- Hadiwinoto, Christian, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. *arXiv preprint arXiv:1910.00194*.
- Hanks, Patrick. 2000. Do word meanings exist? *Computers and the Humanities*, 34(1–2):205–215.
- Hartung, Matthias. 2016. *Distributional Semantic Models of Attribute Meaning in Adjectives and Nouns*. Ph.D. thesis, Institut für Computerlinguistik Ruprecht-Karls-Universität Heidelberg.

- Haveliwala, Taher H. 2002. Topic-Sensitive PageRank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526, Honolulu, HI, USA.
- Hovy, Dirk and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium.
- Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA.
- Huang, Luyao, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China.
- Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany.
- Ishiwatari, Shonosuke, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3467–3476, Minneapolis, MN, USA.
- Jackson, Philip C. Jr. 2019. I do believe in word senses. *Proceedings ACS*, 321:340.
- Jeh, Glen and Jennifer Widom. 2003. Scaling Personalized Web Search. In *Proceedings of the 12th international conference on World Wide Web*, pages 271–279, Budapest, Hungary.
- Joshi, Salil, Diptesh Kanojia, and Pushpak Bhattacharyya. 2013. More than meets the eye: Study of human cognition in sense annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 733–738, Association for Computational Linguistics, Atlanta, Georgia.

- Kanojia, Diptesh, Pushpak Bhattacharyya, Raj Dabre, Siddhartha Gunti, and Manish Shrivastava. 2014. Do not do processing, when you can look up: Towards a discrimination net for WSD. In *Proceedings of the Seventh Global Wordnet Conference*, pages 194–200, University of Tartu Press, Tartu, Estonia.
- Kilgarriff, Adam. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.
- Kilgarriff, Adam. 1998. Senseval: An exercise in evaluating word sense disambiguation programs. In *Proc. of the first international conference on language resources and evaluation*, pages 581–588.
- Kilgarriff, Adam. 2007. Word senses. In Eneko Agirre and Phillip Edmonds, editors, *Word Sense Disambiguation*. Springer, Dordrecht, pages 29–46.
- Kolesnikova, Olga and Alexander Gelbukh. 2012. Semantic relations between collocations: A spanish case study. *Revista Signos*, 45(78):44–59.
- Kumar, Sawan, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy.
- Kumar, Shankar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 169–176, Boston, MA, USA.
- Lacerra, Caterina, Michele Bevilacqua, Tommaso Pasini, and Roberto Navigli. 2020. Csi: A coarse sense inventory for 85% word sense disambiguation. In *Proceedings of the 34th Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence.
- Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press.
- Lakoff, George. 1988. Cognitive semantics. *Meaning and mental representations*, 119:154.
- Leech, Geoffrey Neil. 1992. 100 million words of English: the British National Corpus (BNC). *Language Research*.

- Lemnitzer, Lothar, Holger Wunsch, and Piklu Gupta. 2008. Enriching GermaNet with verb-noun relations - a case study of lexical acquisition. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008, May 28-30, 2018*, pages 156–160, Marrakech, Morocco.
- Lesk, Michael. 1986. Automatic Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proc. of SIGDOC*, pages 24–26.
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.
- Loper, Edward and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1, ETMTNLP '02*, pages 63–70, Stroudsburg, PA, USA.
- Loureiro, Daniel and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy.
- Loureiro, Daniel, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Camacho-Collados Jose. 2020. Language models and word sense disambiguation: An overview and analysis. *arXiv preprint arXiv:2008.11608*.
- Malmkjaer, Kristen. 2005. The linguistics encyclopedia second edition. *London: Taylor & Francis E-Library*.
- Manion, Steve L. 2015. SUDOKU: Treating Word Sense Disambiguation & Entity Linking as a Deterministic Problem-via an Unsupervised & Iterative Approach. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pages 365–369, Denver, CO, USA.
- Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In

- Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, MD, USA.
- Martínez Alonso, Héctor, Anders Johannsen, Oier Lopez de Lacalle, and Eneko Agirre. 2015. Predicting word sense annotation agreement. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 89–94, Association for Computational Linguistics, Lisbon, Portugal.
- Maru, Marco, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. SyntagNet: Challenging Supervised Word Sense Disambiguation with Lexical-Semantic Combinations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3532–3538, Hong Kong, China.
- McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590.
- McCarthy, Diana and Roberto Navigli. 2009. The english lexical substitution task. *Language resources and evaluation*, 43(2):139–159.
- Medin, Douglas L and Marguerite M Schaffer. 1978. Context theory of classification learning. *Psychological review*, 85(3):207.
- Melacci, Stefano, Achille Globo, and Leonardo Rigutini. 2018. Enhancing Modern Supervised Word Sense Disambiguation Models by Semantic Lexical Resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, May 7-12, 2018, Miyazaki, Japan*.
- Melamud, Oren, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany.
- Meyer, David E and Roger W Schvaneveldt. 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, 90(2):227.
- Mickus, Timothee, Denis Paperno, and Matthieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the First NLP Workshop on Deep Learning for Natural Language Processing*, pages 1–11, Turku, Finland.

- Mihalcea, Rada, Timothy Chklovski, and Adam Kilgarriff. 2004. The Senseval-3 English Lexical Sample Task. In *Proc. of Senseval*, pages 1–4.
- 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, Pittsburgh, PA, USA.
- 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 Teng, and Ross T. Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.
- Minsky, Marvin and Seymour A Papert. 2017. *Perceptrons: An introduction to computational geometry*. MIT press.
- 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 2015)*, pages 288–297, Denver, CO, USA.
- Moro, Andrea, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL*, 2:231–244.
- Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.
- Navigli, Roberto, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, GA, USA.
- Navigli, Roberto and Mirella Lapata. 2010. An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692.
- Navigli, Roberto, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval-2007 task 07: Coarse-grained English all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35, Prague, Czech Republic.

- Navigli, Roberto and Simone P. Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence Journal*, 193:217–250.
- Ng, Hwee Tou, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 455–462, Sapporo, Japan.
- Ni, Ke and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)*, pages 413–417, Taipei, Taiwan.
- Nithyanandan, Sandeep and C Raseek. 2019. Deep learning models for word sense disambiguation: A comparative study. In *Proceedings of the International Conference on Systems, Energy & Environment (ICSEE)*, Kerala, India.
- Noraset, Thanapon, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3259–3266, San Francisco, CA, USA.
- Nosofsky, Robert M. 2011. The generalized context model: An exemplar model of classification. In Emmanuel M. Pothos and Andy J. Wills, editors, *Formal Approaches in Categorization*. Cambridge University Press, Cambridge, pages 18–39.
- Oba, Daisuke, Shoetsu Sato, Satoshi Akasaki, Naoki Yoshinaga, and Masashi Toyoda. 2020. Personal semantic variations in word meanings: Induction, application, and analysis. *Journal of Natural Language Processing*, 27(2):467–490.
- Palmer, Martha, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163.
- Palmero Aprosio, Alessio and Giovanni Moretti. 2016. Italy goes to Stanford: a collection of CoreNLP modules for Italian. *ArXiv e-prints*.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

- Pappas, Nikolaos, Phoebe Mulcaire, and Noah A. Smith. 2020. Grounded compositional outputs for adaptive language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1252–1267, Online.
- Parker, Robert, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition. *Linguistic Data Consortium, LDC2011T07, Vol. 12*.
- Pasini, Tommaso. 2020. The knowledge acquisition bottleneck problem in multilingual word sense disambiguation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-20, Yokohama, Japan*.
- Pasini, Tommaso and Roberto Navigli. 2017. Train-O-Matic: Large-Scale Supervised Word Sense Disambiguation in Multiple Languages without Manual Training Data. In *Proc. of EMNLP: System Demonstrations*, pages 78–88, Copenhagen, Denmark.
- Pasini, Tommaso, Federico Scozzafava, and Bianca Scarlini. 2020. CluBERT: a Cluster-Based Approach for Learning Sense Distributions in Multiple Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, WA, USA.
- Passonneau, Rebecca J, Vikas Bhardwaj, Ansa SALLEB-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252.
- Passonneau, Rebecca J., Ansa SALLEB-Aouissi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.
- Paul, Douglas B. and Janet M. Baker. 1992. The Design for the Wall Street Journal-based CSR Corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.
- Petrov, Slav, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey.
- Pilehvar, Mohammad Taher and Jose Camacho-Collados. 2019. WiC: the Word-in-Context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the*

- 2019 Conference of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologies (NAACL-HLT)*, pages 1267–1273, Minneapolis, MN, USA.
- Ponzetto, Simone P. and Roberto Navigli. 2010. Knowledge-Rich Word Sense Disambiguation Rivaling Supervised Systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1522–1531, Uppsala, Sweden.
- Post, Matt. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.
- Postma, Marten, Ruben Izquierdo, Eneko Agirre, German Rigau, and Piek Vossen. 2016. Addressing the MFS bias in WSD systems. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1695–1700, Portorož, Slovenia.
- Pradhan, Sameer, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic.
- Pustejovsky, James. 1991. The generative lexicon. *Computational Linguistics*, 17(4).
- Raganato, Alessandro, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 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 *Proc. of IJCAI*, pages 2894–2900.
- Raganato, Alessandro, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167, Copenhagen, Denmark.
- Rai, Sunny and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.
- Rajpurkar, Pranav, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the*

- Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Association for Computational Linguistics, Melbourne, Australia.
- Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Association for Computational Linguistics, Austin, Texas.
- Ramsey, Rachel. 2017. *An Exemplar-Theoretic Account of Word Senses*. Ph.D. thesis, Northumbria University.
- Reimers, Nils and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China.
- Ribeiro, Marco, Sameer Singh, and Carlos Guestrin. 2016. “why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California.
- Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online.
- Rosch, Eleanor and Carolyn B Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605.
- Ruhl, Charles. 1989. *On monosemy: A study in linguistic semantics*. SUNY Press.
- 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*, pages 699–709, Florence, Italy.
- Scarlini, Bianca, Tommaso Pasini, and Roberto Navigli. 2020. SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, New York, NY, USA.
- Schmid, Helmut. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *Proc. of the ACL SIGDAT-Workshop*, pages 47–50.

- Scozzafava, Federico, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli. 2020. Personalized PageRank with syntagmatic information for multilingual word sense disambiguation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–46, Association for Computational Linguistics, Online.
- Simov, Kiril, Petya Osenova, and Alexander Popov. 2016. Using Context Information for Knowledge-Based Word Sense Disambiguation. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 130–139, Varna, Bulgaria.
- Simov, Kiril, Alexander Popov, Iliana Simova, and Petya Osenova. 2018. Grammatical Role Embeddings for Enhancements of Relation Density in the Princeton WordNet. In *Proceedings of the 9th Global WordNet Conference*, pages 287–295, Singapore.
- Snow, Rion, Sushant Prakash, Daniel Jurafsky, and Andrew Y. Ng. 2007. Learning to merge word senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1005–1014, Prague, Czech Republic.
- Snyder, Benjamin and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain.
- So, David R, Chen Liang, and Quoc V Le. 2019. The evolved transformer. *arXiv preprint arXiv:1901.11117*.
- Soltanolkotabi, Mahdi, Adel Javanmard, and Jason D Lee. 2018. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769.
- Stahlberg, Felix and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3356–3362, Hong Kong, China.
- Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*.
- Taghipour, Kaveh and Hwee Tou Ng. 2015. Semi-supervised word sense disambiguation using word embeddings in general and specific domains. In *Proceedings of the 2015 Con-*

- ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–323, Denver, Colorado.
- Talmy, Leonard. 2000. *Toward a cognitive semantics - volume 1: Concept structuring systems*.
- Thompson, Neil C, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. 2020. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*.
- Tyler, Andrea and Vyvyan Evans. 2001. Reconsidering prepositional polysemy networks: The case of over. *Language*, 77(4):724–765.
- Uslu, Tolga, Alexander Mehler, Daniel Baumartz, Alexander Henlein, and Wahed Hemati. 2018. FastSense: An Efficient Word Sense Disambiguation Classifier. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, May 7-12, 2018, Miyazaki, Japan*.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Vial, Loïc, Benjamin Lecouteux, and Didier Schwab. 2019. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Proc. of Global Wordnet Conference*.
- Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3266–3280.
- Weaver, Warren. 1955. Translation. *Machine translation of languages*, 14(15-23):10.
- Weissenborn, Dirk, Georg Wiese, and Laura Seiffe. 2017. Making neural QA as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280, Vancouver, Canada.
- Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.
- Xie, Qizhe, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

- Yang, Liner, Cunliang Kong, Yun Chen, Yang Liu, Qinan Fan, and Erhong Yang. 2020. Incorporating sememes into Chinese definition modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Yuan, Dayu, Ryan Doherty, Julian Richardson, Colin Evans, and Eric Altendorf. 2016. Semi-supervised Word Sense Disambiguation with Neural Models. In *Proc. of COLING*, pages 1374–1385, Osaka, Japan.
- Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.
- 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, Uppsala, Sweden.
- Zhong, Zhi, Hwee Tou Ng, and Yee Seng Chan. 2008. Word sense disambiguation using OntoNotes: An empirical study. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1002–1010, Honolulu, Hawaii.
- Zhou, Xiang, Yixin Nie, Hao Tan, and Mohit Bansal. 2020. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. *arXiv preprint arXiv:2004.13606*.
- Zhu, Ruimin, Thanapon Noraset, Alisa Liu, Wenxin Jiang, and Doug Downey. 2019. Multi-sense definition modeling using word sense decompositions. *arXiv preprint arXiv:1909.09483*.