The Knowledge Acquisition Bottleneck Problem in Multilingual Word Sense Disambiguation

Tommaso Pasini
Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome
pasini@di.uniroma1.it

Abstract

Word Sense Disambiguation (WSD) is the task of identifying the meaning of a word in a given context. It lies at the base of Natural Language Processing as it provides semantic information for words. In the last decade, great strides have been made in this field and much effort has been devoted to mitigate the *knowledge acquisition bottleneck* problem, i.e., the problem of semantically annotating texts at a large scale and in different languages. This issue is ubiquitous in WSD as it hinders the creation of both multilingual knowledge bases and manually-curated training sets. In this work, we first introduce the reader to the task of WSD through a short historical digression and then take the stock of the advancements to alleviate the *knowledge acquisition bottleneck* problem. In that, we survey the literature on manual, semi-automatic and automatic approaches to create English and multilingual corpora tagged with sense annotations and present a clear overview over supervised models for WSD. Finally, we provide our view over the future directions that we foresee for the field.

1 Introduction

Word Sense Disambiguation (WSD) is at the base of Natural Language Processing (NLP) and aims at associating a word in a given context with one of its possible meanings from a predefined inventory of senses [Weaver, 1949].

WSD approaches may be divided in three different categories depending on the data they require [Navigli, 2009], namely i) supervised [Hadiwinoto *et al.*, 2019; Kumar *et al.*, 2019; Huang *et al.*, 2019], which rely on sense-annotated corpora, i.e., datasets where words in context have been manually tagged with a meaning from a predefined sense inventory; ii) knowledge-based [Moro *et al.*, 2014; Agirre *et al.*, 2014], which drop the requirement on sense-annotated data and address word ambiguity by leveraging the information contained in a semantic network, and iii) unsupervised [Panchenko *et al.*, 2017], also known as Word Sense Induction approaches, which dispose of the knowledge base requirement and employ clustering approaches to create Bag-of-Words representations of meanings. Besides the aforementioned approaches, representation-based methods relying on latent embeddings of senses [Loureiro and Jorge, 2019; Scarlini *et al.*, 2020] proved to attain competitive results with those of classical supervised methods. While unsupervised and knowledge-based systems disposed of manual annotations, they are either difficult to evaluate (unsupervised) or fall behind supervised approaches in terms of performance (knowledge-based), hence making supervised models a better choice. However, one of the problems that mostly affects these latter approaches is the large amount of data needed for achieving satisfactory results. Such datasets are, indeed, exceedingly expensive to produce in terms of both money and time. This issue is better known as the *knowledge acquisition bottleneck* problem. It particularly affects the WSD field since each word in a language vocabulary has its set of possible labels, i.e., senses and, for each of them, one needs to provide a large number of training examples. For instance, consider a language with 200K distinct senses, one will easily end up annotating a 2M-instances dataset to provide 10 examples for each sense. The situation is further worsen by two other issues: i) the fine granularity of word meanings, i.e., senses of the same words are often very similar to each other, and ii) the way senses distribute within a corpus, i.e., by following a Zipfian distribution [McCarthy *et al.*, 2007]. While several efforts have been put in creating sense-annotated datasets for English so as to enrich manually-annotated corpora [Taghipour and Ng, 2015], other languages remained out of scope until recently [Delli Bovi *et al.*, 2017; Scarlini *et al.*, 2019; Barba *et al.*, 2020].

Considering the large amount of works devoted to mitigating the paucity of sense-annotated data for Word Sense Disambiguation and the large interest of NLP community in scaling over multiple languages, in this paper we first introduce the reader to the field of WSD with a brief historical digression. Then, we focus on the *knowledge acquisition bottleneck* problem by giving a broad overview over both manually-curated data that are available, and methods aiming at creating sense-tagged corpora in multiple languages. Throughout the paper, we outline the lessons learned from the presented approaches and provide a view over the possible future directions.

2 History in Brief

Word Sense Disambiguation (WSD) has been first introduced by Weaver (1949) in the context of Machine Translation as the task of associating a word with one of its possible meanings by considering its surrounding words. Preliminary studies were usually combined with methods for solving more general problems of text understanding and leveraged semantic networks available at that time. However, the amount of data available at that time was very limited in terms of number of distinct words, meanings and domains of application. This immediately highlighted that WSD models needed larger datasets in order to generalise over new and unseen examples. The *knowledge acquisition bottleneck* problem, thus, began to take shape, making it evident that gathering large sets of manually-annotated data were expensive in terms of both time and resources. Therefore, the subsequent decades were mainly devoted to the creation of comprehensive machine-readable resources, such as dictionaries and sense-annotated corpora. This kind of resources enabled the development of more sophisticated approaches to WSD such as Lesk's algorithm [Lesk, 1986], a dictionary-based approach which disambiguates words by considering the overlap between the sentence where the target word appears and the dictionary's definitions of the word's senses. This approach relied only on the local context of a word and on the definitions within a dictionary hence being easy to apply on large collections of texts. Exploiting knowledge bases looked very promising and, to stimulate further research on this topic, Miller *et al.* (1990, WordNet and 1993, SemCor) developed the two resources that would soon become the most used within the Word Sense Disambiguation field and other related areas.

WordNet is a machine readable dictionary where synonyms are grouped into synsets, which, in their turn, are linked via paradigmatic relations (i.e., is-a, part-of, etc.). SemCor, instead, is a corpus of texts where each content word (noun, adjective, adverb or verb) is annotated with its WordNet sense. SemCor, by providing around 200K manual annotations of senses, encouraged the development of supervised approaches for WSD. WordNet, instead, paved the way to the knowledge-based paradigm and steered WSD researchers in the direction of enumerative lexicons, becoming the *de-facto* standard knowledge base across different NLP fields. Nevertheless, WordNet suffers from the *so-called* sense granularity problem, i.e., it makes very fine-grained distinctions between senses. Consider for example the noun *line*, WordNet enumerates 30 different senses distinguishing, among others, between a line organised horizontally or vertically. This kind of fine-grained distinctions are not always needed and which is the best level of granularity to express word meanings is still an open problem. We speculate that, since WSD is an intermediate task, the level of specificity of a sense inventory should not be considered in an absolute way but rather depending on the downstream applications where senses will be used. In this direction, Hovy *et al.* (2006) proposed OntoNotes, a hierarchical sense inventory which provides senses at different granularities.

Despite the several efforts in producing lexical-semantic resources, these were mostly focused on a single language hence neglecting the common semantics that distinct languages may share. Navigli and Ponzetto (2010) , therefore, introduced BabelNet with the aim of unifying encyclopedic and lexicographic knowledge across different languages. BabelNet, indeed, is the result of an accurate and automatic merging of different heterogeneous resources, e.g., WordNet, Wikipedia, Wikidata, etc[1]. Thanks to the creation of this kind of multilingual lexical-semantic databases, in the most recent years the research community could focus on developing novel knowledge-based Word Sense Disambiguation approaches [Moro *et al.*, 2014; Agirre *et al.*, 2014]. These methods largely benefit from multilingual resources and own them their flexibility when it comes to disambiguating texts in different languages. However, knowledge-based approaches are often outperformed by their supervised counterpart, which proved to perform generally better on the all-words WSD English task. One of the first proposed models of this kind was IMS [Zhong and Ng, 2010], an ensemble of distinct SVM classifiers, one for each content word in a language vocabulary. The classifiers relied on hand-crafted features extracted from the target word's context, i.e., surrounding words, surrounding POS tags, etc. In the following years, no significant improvements were made until neural networks started being effectively applied across NLP tasks. Neural models allowed to dispose of hand-crafted features in favor of latent representations learned automatically to represent words in contexts. In the last few years, in fact, contextualized representations [Devlin *et al.*, 2018] breath new life into the NLP field bringing large improvements across tasks and hence proving to encode words more efficiently than their sparse representation counterparts based on manually-selected features. In WSD, models employing contextualized word embeddings attain nowadays state-of-the-art performance exceeding the 80% accuracy ceiling [Bevilacqua and Navigli, 2020] in English.

3 Preliminaries

In this Section we introduce the basic concepts used throughout the paper.

Knowledge Base. A knowledge base is a graph where nodes are concepts and edges are semantic relations between them. In lexical resources such as WordNet [Miller *et al.*, 1990] or BabelNet [Navigli and Ponzetto, 2010], each concept is called synset. A synset is a set of words with the same Part-of-Speech tag, and each word can be used to express the same meaning. Each synset features a gloss, i.e., a definition explaining the represented meaning. For example, the concept representing the *necktie*, is defined in WordNet by its set of synonyms, i.e., *tie* and *necktie*, and by the gloss "neckwear consisting of a long narrow piece of material worn (mostly by men) under a collar and tied in knot at the front". Additionally, synsets may also contain usage examples, i.e., sentences where one of the synset's lexicalisations appears in. For instance, "he wore a vest and *tie*" is an example in the *necktie* synset of WordNet showing the usage of *tie* in that meaning. Each pair (lemma, synset) is called sense. It is tied to the word it refers to and uniquely identifies the synset it belongs to. As for senses, we will use the notation l_p^k introduced by [Navigli, 2009] which indicates the k-th meaning in WordNet of the lemma l with POS-tag p.

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

				General Statistics		English Statistics		
	Name	Inventory	Languages	Synsets	Word Types	Instances	Synsets	Word Types
Manual	SemCor	WordNet	1	25,915	38,022	226,036	25,915	38,022
	MASC-AMT	WordNet	1	389	45	1,084,552	389	45
	OntoNotes	WordNet	1	6,534	3,380	233,616	6,534	3,380
Sem-Aut	WNGT	WordNet	1	37,445	31,396	441,656	37,445	31,396
	OMSTI	WordNet	1	3,388	1,149	911,134	3,388	1,149
	SEW	Wikipedia	1	4,098,049	9,293,246	206,475,360	4,098,049	9,293,246
	MuLaN	BabelNet	4	48,000	101,576	-	-	-
Automatic	SenseDefs	BabelNet	263	8,115,401	13,736,019	37,941,345	3,419,661	8,576,183
	EuroSense	BabelNet	21	155,904	453,063	15,441,667	86,881	42,947
	Train-O-Matic	BabelNet	6	53,578	70,250	2,788,763	15,574	11,402
	OneSeC	BabelNet	5	40,041	71,464	888,417	33,721	28,384

Table 1: Statistics of sense-annotated corpora.

Sense Inventory. A sense inventory is the set of meanings that each word in a language vocabulary may take and can be derived from the nodes within a knowledge base. Practically speaking, each lemma-POS pair is mapped to the set of senses that it could express according to the reference knowledge base. For example, the noun *tie* is associated with 9 different senses, among others, *neckwear*, *business relationship*, *a draw in sport*, etc.

Sense-annotated Corpora. Sense-annotated corpora tie the concepts represented in a knowledge base with the lexical information contained in a sentence. In practice, a sense-annotated corpus is a collection of texts where words are tagged with a semantic label, i.e., a sense or a synset, drawn from a given sense inventory. For example, the sentence "I didn't know$^{know_v^2}$ how to tie$^{tie_v^5}$ a tie$^{tie_n^1}$ until I was$^{be_v^1}$ 25." has 4 words annotated with their correct WordNet sense, i.e., $know_v^2$, tie_v^5, tie_n^1 and be_v^1.

4 Sense-Annotated Corpora

Sense-annotated corpora are essential to train supervised models. However, due to the inherent complexity of the task of providing sense annotations, the knowledge acquisition bottleneck problem and to the fine granularity of WordNet senses, annotating large amount of data is both time consuming and expensive. For these reasons, English is the only language where, thanks to SemCor [Miller *et al.*, 1993], part of "Princeton WordNet Gloss Corpus" (WNGT)[2] and OntoNotes [Hovy *et al.*, 2006], manual annotations are available. In all the other languages, instead, one can only rely on a small amount of manually-annotated examples that is usually employed for testing and on automatically-generated datasets for training. In this Section, we detail either existing manually-curated corpora for WSD or semi-automatic and automatic approaches for mitigating the paucity of annotated data. In Table 1 we report the statistics, i.e., the number of languages, unique synsets, unique lemma-POS pairs (word types) and instances for each corpus created automatically, semi-automatically or manually that we introduce in the upcoming Sections.

4.1 Annotations from Humans

SemCor [Miller *et al.*, 1993] is a subset of the Brown corpus (released in 1967) and comprises more than 200,000 tokens manually annotated with a WordNet sense. It is the manually-curated corpus with the wider coverage of words and synsets, hence is the most obvious choice when it comes to training a supervised model. However, even though the coverage of senses is one of its strengths when compared to other corpora of its kind, it is a weakness when considering the absolute numbers of covered senses with respect to WordNet. Indeed, less than 25% of WordNet synsets appears in at least one sentence of the corpus. This issue is the consequence of WordNet's fine granularity and the Zipfian nature of word senses. In fact, it is hard to cover the least common concepts since they describe subtle shades of more common senses and hence occur rarely. Thus, while the fine-granularity problem is more a problem of WordNet than SemCor itself, this latter suffers its consequences showing a limited coverage. Moreover, the Zipfian distribution of senses also plays an important role in making SemCor outdated. Indeed, since the corpus dates back to the 60s, the frequency of a word sense may have changed, e.g., the noun *pipe* appears in SemCor most of the time with its "smoking device" sense, while, nowadays, it is more common to find it with its "metal tube" meaning. This discrepancy could negatively affect the performance and generalisation power of supervised models.

To overcome some of these issues, Hovy *et al.* (2006) introduced OntoNotes, a corpus tagged with senses organised hierarchically, with an upper-level ontology containing macro senses having as child fine-grained sense specialisations. Therefore, even though the OntoNotes's inventory mitigates the fine-granularity problem of WordNet, the corpus is still limited by the number of distinct words covered. Indeed, it only comprises 3,380 different lemma-POS pairs tagged with at least one sense, i.e., three times less than SemCor. Moreover, it does not provide instances for adjectives and adverbs

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

which makes the corpus unsuitable for performing large-scale all-words WSD.

Another valuable resource is "Princeton WordNet Gloss Corpus" (WNGT, 2008), i.e., a corpus comprising all WordNet glosses and many synset examples where content words were tagged either by annotators or with hand-crafted heuristics. More recently, Passonneau *et al.* (2012) introduced MASC, a manually annotated corpus with senses from WordNet 3.1. Despite being the most recent one, it only covers 45 distinct words hence being limited in terms of word and sense coverage (see Table 1).

Even though manually-annotated corpora largely contributed to advance the research in Word Sense Disambiguation and allowed the development of high-performance supervised models – hence to establish supervised WSD as the most effective approach on the English all-words WSD task – we are now approaching a plateau of performance which may depend on the training corpora. Indeed, on the one hand, when models are trained on larger datasets, e.g., the union of SemCor and WNGT, they attain significantly higher results than when trained on SemCor alone. This may suggest that models do already have the expressive power needed to perform the task while lacking annotated data. On the other hand, Huang *et al.* (2019) reported increased performance when combining sense-annotated texts with raw sentences without annotations hence raising the question whether if WSD models are efficiently exploiting the annotated data or if their learning procedure is sub-optimal.

4.2 Annotations from Parallel Corpora

During the years, different approaches have been developed to automatise, in part or completely, the process of creating sense-annotated corpora by exploiting parallel corpora.

Semi-automatic Annotations. In 1997 the intuition that parallel data may be useful to mitigate the paucity of sense-annotated data started arising thanks to Resnik and Yarowsky (1997) . However, no work was done before 2003 to confirm this intuition empirically. In that year, Ng *et al.* proposed a semi-automatic approach which leveraged parallel data and a manual mapping of WordNet senses to Chinese translations. By assigning each English sense to only one Chinese lemma, one can transfer the sense tag from the Chinese word to its aligned English words within the parallel corpus.

Following this idea, Taghipour and Ng (2015) introduced OMSTI, a corpus of 1 Million sense-tagged instances in English, created by exploiting the Chinese-English part of the United Nations corpus. Despite the fact that OMSTI is able to produce a large amount of annotations, we can see from Table 1 that the number of synsets and words it adds to SemCor is modest: around 3K and 1K, respectively. Furthermore, the approach still relies on manually annotating WordNet senses with Chinese words, which, despite being simpler than directly tagging words with meanings, it is still time consuming and needs to be performed for each sense in the knowledge base.

Automatic Annotations. This shortcoming is mitigated by the works of Delli Bovi *et al.* (2017, EuroSense)[3] and Camacho-Collados *et al.* (2016, SenseDefs) [4]. In both works, the authors leveraged the alignment of parallel or comparable datasets (Europarl and BabelNet's glosses, respectively) to provide large and multilingual contexts to Babelfy [Moro *et al.*, 2014], a language-agnostic and knowledge-based approach for WSD. Grouping parallel or comparable sentences in multiple languages together allows Babelfy to exploit a larger context and hence to perform a more precise disambiguation. Finally. they refine the resulting corpora by applying different heuristics to remove the annotations with lower confidence. SenseDefs is the second largest corpus available (Table 1) and showed promising results in English WSD tasks when used as training set for an SVM-based classifier. Indeed, the same model attained higher results when trained on SenseDefs than when trained using SemCor or SemCor and OMSTI together. As for other languages, instead, the quality of SenseDefs and EuroSense corpora remained untested on the standard multilingual WSD benchmarks, i.e., SemEval-2013 task 12 [Navigli *et al.*, 2013] and SemEval-2015 task 13 [Moro and Navigli, 2015].

4.3 Annotations from Monolingual Corpora

Since parallel corpora are a heavy requirement, several works focused on using monolingual datasets only to create sense-annotated corpora.

Semi-automatic Annotations. Raganato *et al.* (2016) introduced SEW[5], a heuristic-based approach to create sense annotations starting from the hyperlinks in Wikipedia. Their heuristics aimed at propagating the hyperlinks information over other untagged words so as to enrich the number of annotations in the Wikipedia corpus. As a result, their corpus counts 206M annotations hence being the largest across the board (Table 1). However, due to the nature of Wikipedia, SEW mostly covers named entities and concrete concepts while lacking abstract concepts. For this reason, the authors focused their evaluation on Entity Linking attaining interesting results which, however, did not improve over those attained by knowledge-based approaches.

This issue is solved in MuLaN[6] [Barba *et al.*, 2020], the most recent approach aiming at creating sense-annotated corpora. MuLaN projects the semantic labels in SemCor to sentences in other languages by exploiting the multilingual representations of BERT and a cross-lingual inventory of meanings, i.e., BabelNet. At the time of writing, the data created by this approach proved to be the best choice to train WSD models on languages other than English, however, it is inherently limited to the senses that appear in SemCor, hence being not able to provide annotated examples for many WordNet's concepts.

The approaches introduced so far do not have any control on the distribution of senses within the produced corpora, hence, they may possibly introduce biases towards certain topics or lack examples for meanings that are frequent in the general domain. Furthermore, they are all limited by the availability of manual annotations.

[3] http://lcl.uniroma1.it/eurosense/

[4] http://lcl.uniroma1.it/disambiguated-glosses

[5] http://lcl.uniroma1.it/sew/

[6] https://github.com/SapienzaNLP/mulan

Automatic Annotations. To cope with both these issues Pasini and Navigli (2020, Train-O-Matic) [7] introduced a knowledge-based approach to generate sense-annotated corpora in potentially any language of BabelNet[8] while taking into account the distribution of word meanings. Indeed, Train-O-Matic assigns a number of annotated examples to each sense depending on its ranking according to either BabelNet, or to automatic methods for inducing the distribution of senses [Pasini *et al.*, 2020]. This also allows Train-O-Matic to customise the built training corpus on a specific distribution.

With the same goal of producing multilingual annotated data, but with a radically different approach, Scarlini *et al.* (2019, OneSeC) proposed a method based on the structure of Wikipedia categories and multilingual lexical-semantic resources such as BabelNet and NASARI [Camacho-Collados *et al.*, 2015]. OneSeC showed to be capable of producing high-quality datasets in 5 different languages[9] (EN, IT, ES, FR and DE) by leading an LSTM-based model trained on its datasets to attain state-of-the-art results on the multilingual tasks of Word Sense Disambiguation. Both Train-O-Matic and OneSeC provide a large number of annotations for many different words and senses (Table 1) while attesting their quality on standard multilingual benchmarks. The main drawback of both these latter approaches is that they can only provide annotated examples for nominal instances.

Wrapping up, integrating different sources of knowledge (semantic networks, manually-annotated corpora, etc.) proved to be the most effective way for automatically producing sense-tagged corpora across languages (MuLaN). Indeed, even though parallel-corpora-based approaches, i.e., OMSTI, EuroSense, SenseGloss, showed to be beneficial to supervised models when merged with SemCor, they either failed to scale over different languages or no quantitative evaluation was performed. On the other hand, knowledge-based methods (Train-O-Matic and OneSeC) have shown to be able to effectively scale to a large number of languages but remained limited to nouns only.

5 Supervised Word Sense Disambiguation

Having largely discussed the sense-annotated corpora available, we now provide an overview of supervised models for WSD that may benefit from them. The supervised approach casts the task as a multi-class classification problem, where, given a sentence and a set of content words therein, the model has to assign to each of them a meaning among their possible ones. The main difference with standard classification tasks is that each word has a different set of labels, i.e., its senses. In what follows we give an overview of existing supervised models that we divided in 3 classes, i.e., feature-based, neural-based and representation-based.

Feature-based Models. Support-Vector machines (SVM) were among the first supervised models to be used in WSD. Zhong and Ng (2010) introduced It Makes Sense (IMS), an SVM-based model which took into account different features to disambiguate a target word, i.e., its surrounding POS tags, words, lemmas, etc., attaining state-of-the-art results at that time. The model hardly scaled to all the words in a language vocabulary as each word needed a separate SVM classifier (word-expert). Therefore, an ensemble of word-expert classifiers was required to disambiguate all content words within a given text. Moreover, the model relied on language-specific features, hence needing experts to craft new features for each language of interest. This latter issue was mitigated by Iacobacci *et al.* (2016) which replaced the hand-crafted features with learned word embeddings attaining the same or better performance overall. This showed, for the first time, the effectiveness of latent word representations in Word Sense Disambiguation. Subsequent works focused on unified neural architectures that represented words in latent space and could easily scale over all the words in a language vocabulary, therefore removing the need of word experts.

Neural-based Models. A first attempt to build a unified model was made by Kågebäck and Salomonsson (2016), which leveraged bidirectional LSTM to extract latent word features and a classifier for each word that had to be disambiguated. This architecture attained promising results, hence showing that a shared layer can be beneficial to the task. However, the architecture was still relying on word-expert classifiers for the final disambiguation. One year later, Raganato *et al.* (2017) finally disposed of the need of word experts introducing a unified model for Word Sense Disambiguation featuring bidirectional LSTMs, an attention layer and a classifier that was shared across all the vocabulary's words. Despite not showing large increments with respect to word-expert models, the proposed model was more flexible as it could be potentially used to disambiguate also words that were not seen at training time. Furthermore, it was the first model being applied in a 0-shot setting to languages different from that of the training data. Nevertheless, the multilingual setting remained mostly unexplored by the community which instead focused on including external knowledge within neural networks.

To this end, knowledge bases started being considered as additional source of information for training neural models. Luo *et al.* (2018, HCAN) was among the first leveraging such information in the form of synsets' definitions and to attain higher results than previous models, hence empirically proving that knowledge graphs are complementary to sense-annotated corpora. Nevertheless, HCAN was still relying on word-expert classifiers, hence being limited to the words seen at training time. This issue was finally faced and mitigated by Kumar *et al.* (2019, EWISE). EWISE dropped the requirement of one classifier per word by representing synsets' definitions as dense vectors – pre-trained directly from the knowledge graph –, and hence making it possible to classify a word in context by multiplying its embedding with those of its possible meanings. This model showed large improvements on rare words and senses, i.e., those not appearing in SemCor. However, it neglected contextualized word embeddings [Devlin *et al.*, 2018, BERT], i.e., contextual representations of words learned by training the model to fill the gap in an input sentence with the most suitable word. This kind of representations showed

[7] http://trainomatic.org/trainomatic

[8] Annotated datasets available in 6 distinct languages (EN, IT, FR, ES, DE and ZH). Languages in ISO code 639-1.

[9] http://trainomatic.org/onesec

large performance improvements across several NLP tasks and quickly became the standard approach to encode texts. It has been, in fact, a matter of months before BERT-based models for WSD were presented, e.g., GlossBERT [Huang *et al.*, 2019]. GlossBERT reduced the WSD task to a binary classification problem where, given a sentence, a target word therein and the gloss of one of the word's possible meanings, the model had to classify whether the gloss represented the correct meaning for the word in the input sentence or not. This approach proved to be very effective surpassing all the other aforementioned models. Following this trend, Bevilacqua and Navigli [2020] introduced EWISER, which, instead of leveraging sense glosses, it takes advantage of the relations within a knowledge base to enrich the representations of words in contexts. This combinations enables the model to surpass the 80% accuracy on the all-words Word Sense Disambiguation benchmarks for English hence being the state of the art at the moment of writing.

Zero-Shot Cross-Lingual WSD Models. While several works attempted to create language-specific training sets, only a few efforts have been put in cross-lingual 0-shot WSD, i.e., training a model only in English and testing its WSD capabilities in other languages. Indeed, EWISER, together with the BiLSTM model introduced by Raganato *et al.* 2017, are the only two approaches that have been tested in this setting. While Raganato *et al.*'s model attained only modest performance, EWISER, instead, thanks to the massive multilingual pre-trained language model it relies on (multilingual BERT), shows results that are comparable to those attained by language-specific models, hence renewing the interest and opening to further research in this direction.

Representation-based Models. Since pre-trained language models proved to effectively encode words in context, a new line of research spawned leveraging this property and building sense embeddings laying in a vector space that is comparable to the one of their reference language model. This makes it possible to perform WSD through a simple 1-Nearest-Neighbour algorithm computing the distance between a sense embedding and a contextualized word embedding and choosing the sense corresponding to the embedding that minimizes the distance with the target words' representations. The general approach relies on a pre-trained language model to encode words in contexts and then on an aggregation function that combines the representations of words that express a given sense. Peters *et al.* (2018) encoded the tagged words in SemCor by means of ELMo and then, for each sense s, averaged the embeddings of the words tagged with s. While showing interesting results, the approach was limited to only those senses appearing in SemCor. Therefore, Loureiro and Jorge (2019) elaborated more on this idea and, first replaced ELMo with BERT and second, extended the sense coverage to all WordNet's meanings by exploiting the relations within the knowledge base to propagate vectors. The resulting sense embeddings were then used for WSD in a 1-NN algorithm with BERT contextualized embeddings showing surprisingly high performance surpassing all classic supervised approaches at the time of its publication.

Wrapping up, supervised models attained surprisingly high performance in the last two years surpassing the 80% accuracy ceiling on the English WSD datasets thanks to the availability of large pre-trained language models which can effectively encode words' semantic features and to the clever usage of knowledge from semantic networks. Furthermore, large pre-trained language models contributed to mitigate the paucity of sense-annotated data across languages as they enabled WSD models trained on data only in English to attain performance in other languages that are comparable to those of language-specific models.

6 Conclusions & Future Directions

In this paper we introduced the reader to the *knowledge acquisition bottleneck* problem in Word Sense Disambiguation and provided a survey on manual, semi-automatic and automatic approaches to mitigate such issue while detailing their strengths and the weaknesses. Furthermore, we presented an overview of the most recent development on supervised methods for WSD which can benefit from the large amount of sense-annotated data that can nowadays be automatically generated. Considering the landscape that we painted in this paper we foresee the followings directions:

1. **Active Learning.** An interesting direction may be to couple knowledge-based approaches for producing sense-annotated data with human annotations by leveraging active learning techniques so as to create datasets with human-level quality while reducing the annotation cost.
2. **Multilingual Gold Standards.** The field lacks large-scale datasets to test WSD models on low-resourced languages. Therefore, it will be worth in the near future to focus on generating multilingual gold standards at scale so as to encourage and enable the development of WSD systems in many languages.
3. **Sense-Annotating as a Game.** A topic that has remained under-explored is the use of reinforcement learning in WSD, where no efforts, to the best of our knowledge, have been spent to formalise WSD in terms of this paradigm. Indeed, by starting form the work of Tripodi and Pelillo [2017], it seems reasonable to formulate the WSD problem in reinforcement learning terms, i.e., defining an agent, a policy an environment and a feedback function to solve the word ambiguity.

Acknowledgments

The author wishes to thank José Camacho-Collados for his remarks on earlier versions of the work, Bianca Scarlini for proofreading the manuscript and Roberto Navigli for his advice and encouragement.

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

References

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

[Barba *et al.*, 2020] Edoardo Barba, Luigi Procopio, Niccolò Campolungo, Tommaso Pasini, and Roberto Navigli. MuLaN: Multilingual Label propagatioN for Word Sense Disambiguation. In *Proceedings of IJCAI*, 2020.

[Bevilacqua and Navigli, 2020] Michele Bevilacqua and Roberto Navigli. Breaking through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information. In *Proc. of ACL*, 2020.

[Camacho-Collados *et al.*, 2015] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. NASARI: a Novel Approach to a Semantically-Aware Representation of Items. In *Proc. of NAACL*, pages 567–577, 2015.

[Camacho-Collados *et al.*, 2016] José Camacho-Collados, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli. A Large-Scale Multilingual Disambiguation of Glosses. In *Proc. of LREC*, pages 1701–1708, 2016.

[Delli Bovi *et al.*, 2017] Claudio Delli Bovi, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. Eurosense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proc. of ACL*, pages 594–600, 2017.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of Arxiv*, 2018.

[Hadiwinoto *et al.*, 2019] Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. Improved word sense disambiguation using pre-trained contextualized word representations. In *Proc. of EMNLP*, 2019.

[Hovy *et al.*, 2006] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: The 90% solution. In *Proc. of NAACL*, pages 57–60, 2006.

[Huang *et al.*, 2019] Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proc. of EMNLP*, 2019.

[Iacobacci *et al.*, 2016] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Embeddings for word sense disambiguation: An evaluation study. In *Proc. of ACL*, pages 897–907, 2016.

[Kågebäck and Salomonsson, 2016] Mikael Kågebäck and Hans Salomonsson. Word Sense Disambiguation using a Bidirectional LSTM. In *Proc. of CogALex*, pages 51–56, 2016.

[Kumar *et al.*, 2019] Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. Zero-shot Word Sense Disambiguation using Sense Definition Embeddings. In *Porc of ACL*. ACL, 2019.

[Lesk, 1986] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proc. of Systems Documentation*, 1986.

[Loureiro and Jorge, 2019] Daniel Loureiro and Alípio Jorge. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proc. of ACL*, 2019.

[Luo *et al.*, 2018] Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. Leveraging Gloss Knowledge in Neural Word Sense Disambiguation by Hierarchical Co-Attention. In *Proc. of EMNLP*, pages 1402–1411, 2018.

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

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

[Miller *et al.*, 1993] George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. A semantic concordance. In *Proc. of DARPA*, pages 303–308, 1993.

[Moro and Navigli, 2015] Andrea Moro and Roberto Navigli. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proc. of SemEval*, pages 288–297, 2015.

[Moro *et al.*, 2014] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL*, 2:231–244, 2014.

[Navigli and Ponzetto, 2010] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: Building a very large multilingual semantic network. In *Proc. of ACL*, pages 216–225, 2010.

[Navigli *et al.*, 2013] Roberto Navigli, David Jurgens, and Daniele Vannella. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Proc. of SemEval*, pages 222–231, 2013.

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

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

[Panchenko *et al.*, 2017] Alexander Panchenko, Fide Marten, Eugen Ruppert, Stefano Faralli, Dmitry Ustalov, Simone Paolo Ponzetto, and Chris Biemann. Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguation. In *Proc. of EMNLP*, 2017.

[Pasini and Navigli, 2020] Tommaso Pasini and Roberto Navigli. Train-o-matic: Supervised word sense disambiguation with no (manual) effort. *Artificial Intelligence*, 279:103215, 2020.

[Pasini *et al.*, 2020] Tommaso Pasini, Federico Scozzafava, and Bianca Scarlini. CluBERT: A Cluster-Based Approach for Learning Sense Distributions in Multiple Languages. In *Proc. of ACL*, 2020.

[Passonneau *et al.*, 2012] Rebecca J Passonneau, Collin Baker, Christiane Fellbaum, and Nancy Ide. The masc word sense sentence corpus. In *Proc. of LREC*, 2012.

[Peters *et al.*, 2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. NAACL*, pages 2227–2237, 2018.

[Raganato *et al.*, 2016] Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia. In *Proc. of IJCAI*, pages 2894–2900, 2016.

[Raganato *et al.*, 2017] Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. Neural sequence learning models for word sense disambiguation. In *Proc. of EMNLP*, 2017.

[Resnik and Yarowsky, 1997] Philip Resnik and David Yarowsky. A Perspective on Word Sense Disambiguation Methods and Their Evaluation. In *Proc. of SIGLEX*, pages 79–86, 1997.

[Scarlini *et al.*, 2019] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. Just "OneSeC" for Producing Multilingual Sense-Annotated Data. In *Proc. of ACL*, 2019.

[Scarlini *et al.*, 2020] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proc. of AAAI*, 2020.

[Taghipour and Ng, 2015] Kaveh Taghipour and Hwee Tou Ng. One million sense-tagged instances for word sense disambiguation and induction. In *Proc. of CoNLL*, pages 338–344, 2015.

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

[Weaver, 1949] Warren Weaver. Translation. In *Machine Translation of Languages*, pages 15–23, 1949.

[Zhong and Ng, 2010] Zhi Zhong and Hwee Tou Ng. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proc. of ACL*, pages 78–83, 2010.