Recent Trends in Word Sense Disambiguation: A Survey

Michele Bevilacqua\(^1\), Tommaso Pasini\(^2\), Alessandro Raganato\(^3\) and Roberto Navigli\(^1\)
\(^1\)Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome
\(^2\)Department of Computer Science, University of Copenhagen
\(^3\)Department of Digital Humanities, University of Helsinki
michele.bevilacqua@uniroma1.it, tommaso.pasini@di.ku.dk
alessandro.raganato@helsinki.fi, roberto.navigli@uniroma1.it

Abstract

Word Sense Disambiguation (WSD) aims at making explicit the semantics of a word in context by identifying the most suitable meaning from a predefined sense inventory. Recent breakthroughs in representation learning have fueled intensive WSD research, resulting in considerable performance improvements, breaching the 80% glass ceiling set by the inter-annotator agreement. In this survey, we provide an extensive overview of current advances in WSD, describing the state of the art in terms of i) resources for the task, i.e., sense inventories and reference datasets for training and testing, as well as ii) automatic disambiguation approaches, detailing their peculiarities, strengths and weaknesses. Finally, we highlight the current limitations of the task itself, but also point out recent trends that could help expand the scope and applicability of WSD, setting up new promising directions for the future.

1 Introduction

Word Sense Disambiguation (WSD) is a historical task in Natural Language Processing (NLP) and Artificial Intelligence (AI) which, in its essence, dates back to Weaver [1949], who recognized the problem of polysemous words in the context of Machine Translation. Even today, word polysemy remains one of the most challenging and pervasive linguistic phenomena in NLP. For example, the ambiguous word \(\text{bass}\) refers to two completely disjoint classes of objects in the following sentences: i) “I can hear \(\text{bass}\) sounds”, ii) “They like grilled \(\text{bass}\)”. NLP research has long sought ways to tackle this phenomenon, with the task of WSD being at the forefront of the automatic resolution of polysemy. In WSD, ambiguity is addressed by mapping a target expression to one (or potentially more) of its possible senses, depending on the surrounding context. Indeed, a model should map the word \(\text{bass}\) to the meanings of low-frequency \(\text{tones}\) and \(\text{type}\ \text{of fish}\), in the respective sentences above. WSD systems use the senses that are enumerated by a static, predefined, machine-readable dictionary, i.e., a sense inventory. Sense inventories are mostly concerned with open-class words (nouns, verbs, adjectives and adverbs), as these are the words carrying most of a sentence’s meaning. In WSD, the sense inventory for a language can be very large, i.e., in the order of hundreds of thousands of concepts, but also very sparse, in that each \(\text{lexeme}\)\(^1\) is associated with only a small subset of the sense inventory.

Predefined inventories define the output space for most varieties of past and modern approaches. These exist in many flavors, ranging from purely supervised [Hadiwinoto et al., 2019; Bevilacqua and Navigli, 2019] to knowledge-based [Moro et al., 2014; Agirre et al., 2014; Scozzafava et al., 2020], to hybrid supervised and knowledge-based approaches [Kumar et al., 2019; Bevilacqua and Navigli, 2020; Blevins and Zettlemoyer, 2020; Conia and Navigli, 2021; Barba et al., 2021]. Supervised models, today based on neural architectures, frame the task as a classification problem and take advantage of annotated data to learn the association between words\(^2\) in context and senses. Knowledge-based approaches, instead, often employ graph algorithms on a semantic network, in which senses are connected through semantic relations and are described with definitions and usage examples. Their independence from labeled training data, however, comes at the expense of performing worse than supervised models [Pilehvar and Navigli, 2014; Raganato et al., 2017a; Pasini et al, 2021] which, benefiting from pretrained language models, can now also nimbly scale across different languages. Nonetheless, information in semantic networks, be it unstructured (e.g., definitions) or structured (e.g., relational information), still remains highly relevant. This is demonstrated by hybrid approaches, which, reporting the highest results in literature, are currently attested as the best solution [Barba et al., 2021].

Considering the fast pace at which the field is moving, together with the fact that reference WSD surveys [Nancy and Jean, 1998; Agirre and Edmonds, 2007; Navigli, 2009] are now more than 10 years old, it is hard to have a clear picture as to which the most successful innovations introduced in the last few years may be. In this survey paper we thus provide a comprehensive overview of the literature, summarizing the most effective contributions proposed so far. Specifically, we focus

\(^1\)(\(\text{lemma, part of speech}\) pair.

\(^2\)For ease of reading, we use word to refer to both words and multowrd expressions.
on the most recent and significant models for the task, highlighting their strengths and weaknesses, while, at the same time, outlining possible fruitful directions that lie ahead.

2 Resources for WSD

WSD is a knowledge-intensive task, which needs data of two different kinds: i) sense inventories, i.e., reference computational lexicons which enumerate possible meanings; and ii) annotated corpora, in which a subset of words are tagged with one or more possible meanings drawn from the given inventory. In the following subsections, we review the most popular sense inventories (§2.1) and annotated corpora (§2.2) used for training and testing WSD systems.

2.1 Sense Inventories

Sense inventories enumerate the set of possible senses for a given lexeme. The most popular ones are:

- **Princeton WordNet** [Miller et al., 1990], a large, manually-curated lexicographic database of English and the *de facto* standard inventory for WSD. It is organized into a graph, where nodes are synsets, i.e., groups of contextual synonyms. Each synonym in a synset represents a sense of a word. Synsets and senses are linked to each other through edges representing lexical-semantic relations, such as hypernymy (is-a), and meronymy (part-of), among others. For each synset, WordNet also provides other forms of lexical knowledge, such as definitions (glosses) and usage examples. Most recent works in English WSD use the 3.0 version (released in 2006), containing 117,659 synsets. Recently, English WordNet 2020 [McCrae et al., 2020] extended the original Princeton WordNet by introducing approximately 3,000 new synsets, including slang and neologisms.

- **BabelNet** [Navigli and Ponzetto, 2012], a multilingual dictionary with coverage of both lexicographic and encyclopedic terms obtained by semi-automatically mapping various resources, such as WordNet, multilingual versions of WordNet and Wikipedia, among others. BabelNet is structured as a semantic network where nodes are multilingual synsets, i.e., groups of synonyms lexicalized in several languages, and edges are semantic relations between them. The latest 2021 release, i.e., version 5.0, covers 500 languages and contains more than 20M synsets [Navigli et al., 2021].

Another inventory that has recently been gaining interest [Blevins et al., 2021] is Wiktionary: a collaborative project designed to create a dictionary for each language separately. Each of these inventories suffers from the so-called *fine-granularity problem*, that is, different meanings of the same lexeme are, sometimes, hard to discriminate between even for humans. For example, WordNet enumerates 29 senses for the noun *line*, two of which distinguish between a set of things laid out horizontally and one laid out vertically. To cope with the excessive granularity of word senses and simplify the WSD task, different coarser-grained inventories have been proposed [Hovy et al., 2006; Lacerra et al., 2020], but their use has not yet become mainstream, also due to limited coverage.

Another significant issue is the fact that sense inventories assume that, at least for practical purposes, word meaning can be enumerated in a finite list. However, this also implicitly assumes that language is static and does not change much over time. Unfortunately, this is not the real-case scenario, especially considering how fast new words and senses are introduced online. Alternative approaches like the generative lexicon [Pustejovsky, 1998], which provides a general framework in which word meaning can be produced online, have been proposed in the past, but no large-scale experiments have yet been carried out on them.

2.2 Sense-Annotated Data

As new annotated data are continuously created, in this Section we only describe the standard benchmarks used in WSD, and refer the reader to a recent survey on corpora tagged with sense annotations [Pasini, 2020].

Data for Training

SemCor [Miller et al., 1993] is the largest manually annotated dataset, comprising 200,000 sense annotations using the WordNet sense inventory. Despite the remarkable effort, it only covers 22% of the almost 118,000 WordNet synsets, and, being a subset of the English Brown Corpus from the 1960s, it features a different distribution of senses compared to that of contemporary texts, with numerous meanings that are now commonplace, such as *computer mouse*, being completely absent. To increase the annotation coverage, several works [Vial et al., 2019; Bevilacqua and Navigli, 2020] have recently started using the English Princeton WordNet Gloss Corpus (WNG)\(^4\) as additional data. WNG comprises sense definitions and examples in WordNet, annotated both manually and semi-automatically, covering more than 59,000 WordNet senses.

While English training data is widely available, unfortunately the same does not hold for other languages. Although hand-labeled data are notoriously difficult to obtain on a large scale for many languages, some efforts in the past were directed towards creating manually-translated versions of SemCor [Petrolito and Bond, 2014], but many of these are no longer available. Therefore, several subsequent works proposed automatic methods for producing high-quality sense-annotated data both in English [Taghipour and Ng, 2015; Loureiro and Camacho-Collados, 2020] and other languages by leveraging: information from Wikipedia [Scarlini et al., 2019], the Personalized PageRank algorithm [Pasini and Navigli, 2020], label propagation over comparable texts [Barba et al., 2020] or automatic translations [Pasini et al., 2021].

Data for Testing

Evaluation in WSD is usually carried out using the manually annotated datasets from the Senseval and SemEval evaluation campaigns. English WSD benefits from the evaluation suite of Raganato et al. [2017a] which combines together five all-words gold-standard datasets: Senseval-2 [Edmonds and Cotton, 2001, S2], Senseval-3 [Snyder and Palmer, 2004, S3],

\[^{4}\text{https://wordnetcode.princeton.edu/glosstag.shtml}\]
SemEval-2007 Task 17 [Pradhan et al., 2007, S7], SemEval-2013 Task 12 [Navigli et al., 2013, S13] and SemEval-2015 Task 13 [Moro and Navigli, 2015, S15]. This framework standardized the evaluation in English WSD with the WordNet sense inventory, making it easier to compare systems in a general domain, helping the field to develop increasingly better-performing models. In an attempt to investigate the most common weaknesses among WSD approaches, i.e., poor performance on infrequent senses, Blevins et al. [2021] introduced FEWS, an English benchmark where Wiktionary examples are annotated with Wiktionary definitions. For non-English languages, instead, WSD evaluation datasets have received less attention, as they are often annotated with diverse and outdated inventories. Only very recently, a comprehensive benchmark has been put forward to standardize the evaluation in this setting too [Pasini et al., 2021, XL-WSD].

3.1 Knowledge-Based WSD

Knowledge-based approaches leverage computational lexicons, such as WordNet or BabelNet, especially their graph structure, in which synsets act as nodes and the relations between them as edges. Successful approaches of this kind employ graph algorithms such as random walks [Agirre et al., 2014, UKB], clique approximation [Moro et al., 2014, Babelfy], or game theory [Tripodi and Navigli, 2019]. The richness and quality of the information encoded within their underlying knowledge bases crucially determine the performance of such approaches [Pilehvar and Navigli, 2014; Maru et al., 2019].

The highest-scoring models are two very different models: SyntagRank [Scozzafava et al., 2020] and SREFKB [Wang and Wang, 2020]. SyntagRank is purely graph-based and applies the Personalized PageRank algorithm [Page et al., 1999] on both the WordNet portion of BabelNet augmented with relations from the WNG corpus, and SyntagNet [Maru et al., 2019], a resource providing manually curated relations between synsets whose senses form a collocation. SREFKB, instead, is a vector-based approach leveraging contextualized word representations and sense embeddings to perform disambiguation. Sense vectors are computed by applying BERT [Devlin et al., 2019] on examples and definitions from WordNet, as well as on automatically retrieved contexts from the web. Thanks to BabelNet, SyntagRank showed itself to be able to scale across many different languages, while SREFKB has so far been tested on English only. In addition, SREFKB also does make use of manually-created usage examples from WordNet, which arguably amounts to a form of stronger supervision.

3.2 Supervised WSD

The most successful approaches to WSD are the so-called supervised methods. In abstract terms, these aim to learn a parameterized function \( f_\Theta \) mapping a word \( w \) in a context \( c \) to a sense \( s \in V \) (the vocabulary of senses) using the supervision of a dataset \( D \) of word-context-sense triplets \( \langle w, c, s \rangle \).

In what follows, we focus mainly on neural supervised systems, which over recent years have consistently obtained the best overall results. Most of the methods we discuss exploit transfer learning, with the use of pretrained Transformers being required for state-of-the-art performance.

As the most meaningful classification of the approaches concerns not so much the architecture, but what kind of additional information the model is able to exploit, we group them into (i) purely data-driven models, (ii) supervised models exploiting glosses, (iii) supervised models exploiting relations in a knowledge graph, and (iv) supervised approaches using other sources of knowledge. In what follows we highlight different families of supervised approaches in boldface.

Purely Data-Driven WSD

Most supervised WSD models are trained with gradient descent to minimize a cost function \( L(c, s) \) over all \( \langle w, c, s \rangle \in D \) with respect to the parameters \( \Theta \). A popular baseline model, in this case, would be a token tagger, which for each word \( w \) in a context \( c \) produces a probability distribution \( P_w \) over all \( s' \in V \), i.e., over all senses in the vocabulary. Token tagger models for WSD make use of a pretrained embedding, which is usually kept frozen, feed the contextualized representations to either a feedforward network [Hadiwinto et al., 2019] (Eq. 1 below) or a stack of Transformer layers [Bevilacqua and Navigli, 2019; Vial et al., 2019] (Eq. 2), and then multiply the output by a classification layer \( O \):

\[
E_c = \text{Embed}(c) \quad E_c = \text{Embed}(c) \\
H_{c,w} = \text{FFN}(E_{c,w}) \quad H_{c,w} = \text{Transformer}(E_{c,w}) \\
P_{c,w} = \text{Softmax}(H_{c,w}O) \quad P_{c,w} = \text{Softmax}(H_{c,w}O) \\
(1)
\]

where \( \Box_{c,w} \) selects the component that corresponds to the target word \( w \) in \( c \). At inference time, rather than predicting the most likely sense across the whole vocabulary, one predicts the highest among those possible for the given word:

\[
\hat{s} = \arg\max_{s' \in V(w)} P_{c,w}(s') \\
(3)
\]

where \( V(w) \subseteq V \) is the set of possible meanings that \( w \) can take according to the reference sense inventory.
These simple approaches already produce a large improvement over previous mostly randomly-initialized models [Raganato et al., 2017b]. Nevertheless, performances are – at least partially – limited by the categorical cross-entropy that is often used for training. In fact, the binary cross-entropy loss has been shown to be more effective [Conia andNavigli, 2021], as it allows multiple annotations for a single instance that are available in the training set to be taken into account, rather than having to use a single ground-truth sense only.

A simpler approach compared to token taggers is that of the 1-nn vector-based methods [Peters et al., 2018]. This approach creates sense embeddings by averaging the contextual vectors of instances within the training set that were tagged with the same sense:

\[ t^{(c,w)} = \text{Embed}(c)_w \]

\[ t^{(s)} = \frac{1}{|D(w,s)|} \sum_{c' \in D(w,s)} \text{Embed}(c')_w \]  

(4)

where \( t^{(c,w)} \) and \( t^{(s)} \) are the representations for, respectively, a word in context and a sense, and \( D(w,s) \) is the set of contexts where \( w \) appears associated with a sense \( s \) in the dataset \( D \). The predicted sense \( \hat{s} \) is selected as the one with the highest cosine similarity:

\[ \hat{s} = \text{argmax} \text{sim}_{\text{cos}}(t^{(c,w)}, t^{(s)}) \]  

(5)

The approaches presented so far assume that each sense is an opaque class, and the classification architecture cannot exploit any knowledge beyond what can be inferred through the supervision from the training corpus. This issue is not only theoretical but also practical, as many senses do not actually occur in training corpora (§2.2) owing to the extreme class imbalance.

Supervised WSD Exploiting Glosses

One conspicuous source of information in sense inventories consists of textual definitions (also known as glosses). Definitions, mirroring the format of traditional dictionaries, provide a simple human-readable way of clarifying sense distinctions. For example, the concept of nostalgia is defined in WordNet as longing for something past. Glosses have proven themselves quite useful for increasing WSD performances, with multiple ways to exploit them being explored in the literature. Glosses can be encoded as vectors by averaging their tokens’ contextualized representations and easily incorporated into both 1-nn approaches and token tagging architectures. Specifically, 1-nn approaches have been shown to benefit greatly from concatenating gloss vectors to the “supervised” representations (see Eq. 4) [Loureiro and Jorge, 2019, LMMS]. Indeed, glosses are also used in the same manner by more sophisticated 1-nn approaches, such as SensEmBERT [Scarlini et al., 2020a], ARES [Scarlini et al., 2020b] and SREF [Wang and Wang, 2020]. They differ substantially in their approach to automatically retrieving additional contexts in order to build the supervised part of the sense embedding, with ARES attaining the highest performance by leveraging collocational relations between senses to retrieve new example sentences from Wikipedia. Berend [2020] has shown that existing sense embeddings can also be made sparse by applying sparse coding.

Another use of sense embeddings (including gloss information) is in providing the weights for the classification layer (the matrix \( O \) in Eq. 1) of token-tagging architectures. EWISE [Kumar et al., 2019] creates sense representations training a gloss encoder by means of a triplet loss on WordNet (§3.2); EWISE [Bevilacqua andNavigli, 2020], instead, finetunes off-the-shelf sense embeddings based on pretrained language models, i.e., SensEmBERT andLMMS, attaining results close to the state of the art. Finally, BEM [Blevins andZettlemoyer, 2020] fully embraces the idea of jointly training text and sense representations, and puts it into practice by leveraging two separate Transformer models to encode the target word context and its candidate definitions.

Glosses have also been exploited in sequence-tagging approaches [Huang et al., 2019; Yap et al., 2020]. These reframe the WSD task as a sequence classification problem where, given a word \( w \) in a context \( c \), they score the triplet \( (w,c,G(s')) \) for each \( s' \in V^{(w)} \), and select the sense \( \hat{s} \) with the highest score:

\[ \hat{s} = \text{argmax} \Gamma(c, w, G(s')) \]  

(6)

where \( \Gamma \) is a scoring function typically implemented as a fine-tuned Transformer. While attaining competitive performance (§4.2), models of this kind are less efficient than token classifiers since they need to process the same sentence for each content word and for each of its possible definitions.

Finally, a generative variant of the sequence classification approach has been introduced by Bevilacqua et al. [2020] to tackle WSD as a Natural Language Generation (NLG) problem where, given a target word in a sentence concatenated with all its possible definitions, a model has to find the span that best fits the target word use within the sentence. This approach allows a BART-based model [Lewis et al., 2020] to attain state-of-the-art results on the standard English benchmarks while also being able to scale over vocabularies with different granularities.

However, the model is still less efficient than regular token-tagging alternatives, since it needs to run as many forward passes as there are targets to classify in the input sequence.

Finally, a generative variant of the sequence classification approach has been introduced by Bevilacqua et al. [2020] to tackle WSD as a Natural Language Generation (NLG) problem where, given a target word in a sentence concatenated with all its possible definitions, a model has to find the span that best fits the target word use within the sentence. This approach allows a BART-based model [Lewis et al., 2020] to attain state-of-the-art results on the standard English benchmarks while also being able to scale over vocabularies with different granularities. However, the model is still less efficient than regular token-tagging alternatives, since it needs to run as many forward passes as there are targets to classify in the input sequence.

Supervised WSD Exploiting Relations

WordNet offers another rich source of knowledge in the edges that interweave its senses and synsets. Traditionally, this information is exploited by graph knowledge-based systems, for example, those based on Personalized PageRank [Scozzafava et al., 2020]. Nevertheless, many recent supervised systems – either 1-nn or token taggers – also draw benefit from using WordNet as a graph. For example [Loureiro and Jorge, 2019,
LMMS] create representations for those senses not appearing in SemCor by averaging the embeddings of their neighbours in WordNet; Wang and Wang [2020, SREF] employ WordNet hypernymy and hyponymy relations to devise a try-again mechanism that refines the prediction of the WSD model, and Vial et al. [2019] reduce the number of output classes by mapping each sense to an ancestor in the WordNet taxonomy. Among the token-tagger models, EWISE [Kumar et al., 2019] uses the WordNet graph structure to train the gloss embedder offline, while EWISER [Bevilacqua andNavigli, 2020] shows that with a simple modification to Eq. 1 the full graph of WordNet can be directly incorporated into the architecture:

\[ P_{c,w} = \text{Softmax}(H_{c,w}OA) \]  

where \( A \) is a sparse adjacency matrix. A different way to use the same information is proposed by Conia and Navigli [2021], who replace the whole adjacency matrix multiplication with a binary cross-entropy loss where all senses related to the gold one are also considered as relevant.

In general, using relational knowledge is becoming commonplace in supervised WSD, with a gradual hybridization with knowledge-based methods. However, relational knowledge is easily exploited only by token classification and 1-nn approaches, while its integration into sequence classification methods has not yet been investigated.

**Supervised WSD Exploiting Other Knowledge**

WSD models also prove to benefit from using additional sources of knowledge, both internal and external to the knowledge base itself. Luan et al. [2020] leverage translations in BabelNet to refine the output of any arbitrary WSD system by comparing the translation of the output senses with the target’s translations provided by an NMT system.

In a different direction, Calabrese et al. [2020a] leverage images from the BabelPic dataset [Calabrese et al., 2020b] to build multimodal gloss vectors, which are shown to be stronger than text-only vectors when used to initialize the weights of the classification matrix \( O \) in Eq. 1. Wikipedia and Web search contexts are also used as additional data to create sense embeddings [Scarlini et al., 2020a; Scarlini et al., 2020b; Wang and Wang, 2020] and as an alternative source in order to propagate vectors through the WordNet network, showing higher performance and better representations for rare senses.

4 **Taking Stock of WSD**

In this Section, we review the performance figures of recent WSD models, with details reported in §4.1. In §4.2, we put forward a few high-level guidelines that are meant to help the community to navigate current trends in the field.

4.1 **Evaluation Setting**

The performance of WSD systems is usually assessed in terms of F1 score over held-out test sets. As a performance comparison in WSD, a typical upper bound is given by the inter-annotator agreement (IAA), i.e., the percentage of words tagged with the same sense by two or more human annotators. The IAA over a fine-grained sense inventory is estimated to be around 67-80% accuracy [Navigli, 2009]; these figures, however, call for further studies so as to obtain more centered estimates of human performance, e.g., on up-to-date benchmarks.

We report results (collected from the literature) on the English WSD benchmark of Raganato et al. [2017a] in Table 1. All supervised models therein are trained on SemCor (§2.2). Additionally, we report in Table 2 results on the recent XL-WSD multilingual benchmark [Pasini et al., 2021] including i) a crosslingual 0-shot token-classification baseline (exploiting XLM-R) trained on (English) SemCor, ii) the same baseline trained on the automatically translated silver corpora provided as part of XL-WSD, iii) the best knowledge-based multilingual system, i.e., SyntagRank [Scozzafava et al., 2020].

4.2 **Discussion**

**Pretrained language models.** The use of pretrained language models plays a crucial role in achieving high performance, for both knowledge-based and supervised approaches [Wang and Wang, 2020; Blevins andZettelmoyer, 2020]. The simple model of Hadiwimoto et al. [2019] results in a 2-point improvement over the best model without pretrained contextualized embeddings, i.e., EWISE [Kumar et al., 2019].

**Are knowledge-based methods still relevant?** Pure knowledge-based methods are completely outperformed on English WSD, with a gap of 7.2 points between the best knowledge-based method, i.e., SREFKB, and the best supervised system, i.e., ESCHER. The same trend appears in a recent multilingual benchmark as well [Pasini et al., 2021]. Nevertheless, information within knowledge bases remains valuable and many successful supervised methods are effectively hybridized with knowledge-based methods (§3.2).

**Is it worth it to include other kinds of knowledge?** Additional information is beneficial to boosting the results, with most token classification and 1-nn approaches exploiting knowledge graph information in order to reach competitive performances. We note that different kinds of knowledge are orthogonal to each other and can be exploited in conjunction. For example, token classification models benefit from the logits-adjacency matrix multiplication [Bevilacqua et al., 2020], binary cross-entropy training [Conia andNavigli, 2021], translation-based refinement [Luan et al., 2020] and visual information [Calabrese et al., 2020a].

**Training data.** The addition of more training data, e.g., the WNG corpus (§2.2), increases performance significantly, even though this corpus contains a significant amount of noisy silver annotations. Indeed, multiple works [Bevilacqua andNavigli, 2020; Conia andNavigli, 2021] report that concatenating WNG to SemCor increases the performance of their systems from 1.8 to 2.6 F1 points. This makes it worthwhile investigating whether more advanced techniques for the automatic creation of training corpora can be exploited for further gains.

**What is the best model?** In the standard configuration, i.e., trained on SemCor only and tested in terms of F1 over the Raganato et al. [2017a] English benchmark, the best result is achieved by ESCHER [Barba et al., 2021]. As we recall, ESCHER performs WSD by concatenating all glosses and the
models, especially generative ones, offer zero-shot capabilities over a changing sense inventory [Blevins et al., 2020; Blevins et al., 2021], while 1-nn and token classification approaches are more flexible in terms of integrating task-specific biases, and also more efficient, being able to classify many contexts at once with a single forward pass.

**Multilingual WSD.** In the past, one of the main arguments in favor of knowledge-based WSD was that of scalability. However, as Table 2 shows, this seems no longer to be the case. Overall, thanks to the availability of pretrained multilingual contextualised embeddings, one can train a simple supervised model on just English and get much higher performances compared to a knowledge-based system, even on languages that are very different, such as Basque, Chinese, Hungarian and Korean. In fact, the crosslingual setting works so well that it outperforms language-specific models trained on silver data, which are probably hampered by noise and distribution skewing effects related to the data creation procedure. However, performance figures are still mostly underwhelming compared to those for English WSD, where supervised results on the concatenation of all datasets start from around 74 F1 points (see Table 1).

### Table 1: F1 performance figures of recent WSD systems in the literature

<table>
<thead>
<tr>
<th>Kind</th>
<th>System</th>
<th>ALL</th>
<th>S2</th>
<th>S3</th>
<th>S7</th>
<th>S13</th>
<th>S15</th>
</tr>
</thead>
<tbody>
<tr>
<td>KB</td>
<td>[Scozzafava et al., 2020, SyntagRank]</td>
<td>71.7</td>
<td>71.6</td>
<td>72.0</td>
<td>59.3</td>
<td>72.2</td>
<td>75.8</td>
</tr>
<tr>
<td></td>
<td>[Wang and Wang, 2020, SREF]</td>
<td><strong>73.5</strong></td>
<td>72.7</td>
<td>71.5</td>
<td>61.5</td>
<td>76.4</td>
<td>79.5</td>
</tr>
<tr>
<td></td>
<td><strong>Vector-based 1-nn</strong></td>
<td>[Loureiro and Jorge, 2019, LMMS]</td>
<td>75.4</td>
<td>76.3</td>
<td>75.6</td>
<td>68.1</td>
<td>75.1</td>
</tr>
<tr>
<td></td>
<td>[Berend, 2020]</td>
<td>76.8</td>
<td>77.9</td>
<td>77.8</td>
<td>68.8</td>
<td>76.1</td>
<td>77.5</td>
</tr>
<tr>
<td></td>
<td>[Scarlini et al., 2020b, ARES]</td>
<td><strong>77.9</strong></td>
<td>78.0</td>
<td>77.1</td>
<td>71.0</td>
<td>77.3</td>
<td>83.2</td>
</tr>
<tr>
<td></td>
<td>[Conia and Navigli, 2020, Conception]</td>
<td>76.4</td>
<td>77.1</td>
<td>76.4</td>
<td>70.3</td>
<td>76.2</td>
<td>77.2</td>
</tr>
<tr>
<td></td>
<td>[Luan et al., 2020]</td>
<td>76.4</td>
<td>77.2</td>
<td>77.1</td>
<td>69.2</td>
<td>76.1</td>
<td>77.2</td>
</tr>
<tr>
<td></td>
<td>[Scarlini et al., 2020a, SensEmBERT]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>78.7</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>[Hadiwinoto et al., 2019, GLU]</td>
<td>74.1</td>
<td>75.5</td>
<td>73.6</td>
<td>68.1</td>
<td>71.1</td>
<td>76.2</td>
</tr>
<tr>
<td></td>
<td>[Vial et al., 2019, SVC]</td>
<td>76.7</td>
<td>76.5</td>
<td>77.4</td>
<td>69.5</td>
<td>76.0</td>
<td>78.3</td>
</tr>
<tr>
<td></td>
<td>[Kumar et al., 2019, EWISE]</td>
<td>71.8</td>
<td>73.8</td>
<td>71.1</td>
<td>67.3</td>
<td>69.4</td>
<td>74.5</td>
</tr>
<tr>
<td></td>
<td>[Blevins and Zettlemoyer, 2020, BEM]</td>
<td><strong>79.0</strong></td>
<td>79.4</td>
<td>77.4</td>
<td>74.5</td>
<td>79.7</td>
<td>81.7</td>
</tr>
<tr>
<td></td>
<td>[Calabrese et al., 2020a, EViLBERT]</td>
<td>75.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>[Bivlaquca and Navigli, 2020, EWiSER]</td>
<td>78.3</td>
<td>78.9</td>
<td>78.4</td>
<td>71.0</td>
<td>78.9</td>
<td>79.3</td>
</tr>
<tr>
<td></td>
<td>[Conia and Navigli, 2021]</td>
<td>77.6</td>
<td>78.4</td>
<td>77.8</td>
<td>72.2</td>
<td>76.7</td>
<td>78.2</td>
</tr>
<tr>
<td><strong>Token Classifier</strong></td>
<td>[Huang et al., 2019, GlossBERT]</td>
<td>77.0</td>
<td>77.7</td>
<td>75.2</td>
<td>72.5</td>
<td>76.1</td>
<td>80.4</td>
</tr>
<tr>
<td></td>
<td>[Bivlaquca et al., 2020, Generationary]</td>
<td>76.7</td>
<td>78.0</td>
<td>75.4</td>
<td>71.9</td>
<td>77.0</td>
<td>77.6</td>
</tr>
<tr>
<td></td>
<td>[Yap et al., 2020]</td>
<td>78.7</td>
<td>79.9</td>
<td>77.4</td>
<td>73.0</td>
<td>78.2</td>
<td>81.8</td>
</tr>
<tr>
<td></td>
<td>[Barba et al., 2021, ESCHER]</td>
<td><strong>80.7</strong></td>
<td>81.7</td>
<td>77.8</td>
<td>76.3</td>
<td>82.2</td>
<td>83.2</td>
</tr>
</tbody>
</table>

**5 Beyond Word Sense Disambiguation**

While the WSD task has benefited from recent breakthroughs in transfer learning, even surpassing its expected upper bound, there are certain limits intrinsic to the task itself. The choice of using a discrete sense inventory, while it is computationally convenient, prevents scaling to newer and more creative uses of words, and constrains systems to a given sense granularity, which may be suboptimal for the chosen application.

For these reasons, Pilehvar and Camacho-Collados [2019] chose to eschew discrete meanings altogether by putting forward the Word-in-Context (WiC) task, a tool for evaluating the semantic competence of models without the need for predefined inventories. WiC requires a model to take as input two contexts featuring the same target words, and to predict whether those words are used with the same meaning. Building WiC datasets is easier than building ones for WSD, and indeed large-scale benchmarks are also available for non-English languages [Raganato et al., 2020; Martelli et al., 2021].

In a different direction, but with the same purpose of dropping the need for predefined inventories, the task of lexical substitution [McCarthy and Navigli, 2009] requires models to disambiguate a word in context by searching for meaning-preserving substitutes. For example, given the con-
models reach top performance, the WSD task is still not solved [Navigli, 2018; Blevins et al., 2021] and this opens up new exciting directions.

With the breaching of this glass ceiling, current benchmarks are really starting to show their inadequacy. This calls for the construction of new challenging test sets (possibly through adversarial techniques) to shed light on what remains problematic for WSD. Indeed, the behavior of current models in out-of-domain sense distributions should be studied further in the near future, in order to build WSD approaches that are more robust to domain shift and reliable with Web text, e.g., from social media. Moreover, multilingual WSD lacks a comprehensive investigation to assess model capabilities in non-English languages. While the recent cross-lingual evaluation suite, i.e. XL-WSD [Pasini et al., 2021], is a first step towards a large-scale multilingual WSD benchmark, more effort is needed to create training or testing data for as many languages as possible in the coming years.

An additional avenue for research is the integration of WSD with the related task of Entity Linking [Sevgili et al., 2021], in which the model is required to associate mentions with entities in a knowledge base such as Wikipedia. While the existence of BabelNet provides a unified repository that allows one to perform both tasks [Moro et al., 2014], the recent literature has not taken up this path. It is worth exploring whether recent approaches which efficiently classify over the huge output space of Entity Linking [Cao et al., 2021] can be combined with the techniques for the exploitation of glosses and relations developed within the WSD community.

Since WSD systems now work fairly well, it is time to employ them in other applications too, e.g., boosting semantic-intensive downstream tasks such as Machine Translation, Semantic Role Labeling, and Question Answering. Finally, WSD could help pretrained language models to ground word representations onto a knowledge base [Pappas et al., 2020], providing the semantics they seem to lack [Bender and Koller, 2020], and a gateway to other information sources and perceptive domains, such as vision: a whole new realm that NLP, with approaches such as Vokenizer [Tan and Bansal, 2020], is just now starting to exploit, and in doing so may finally break out of its sandbox!

**Acknowledgments**

The authors gratefully acknowledge the support of the ERC Consolidator Grants MOUSSE No. 726487, and FoTran No. 771113 under the European Union’s Horizon 2020 research and innovation programme.

This work was supported in part by the MIUR under the grant “Dipartimenti di eccellenza 2018-2022” of the Department of Computer Science of Sapienza University and by the Innovation Fund Denmark under the LEGALESE project.

**References**
