Nibbling at the Hard Core of Word Sense Disambiguation

Marco Maru[1], **Simone Conia**[1], **Michele Bevilacqua**[1], and **Roberto Navigli**[2]

Sapienza NLP Group
[1]Department of Computer Science
[2]Department of Computer, Control and Management Engineering
Sapienza University of Rome
firstname.lastname@uniroma1.it

Abstract

With state-of-the-art systems having finally attained estimated human performance, Word Sense Disambiguation (WSD) has now joined the array of Natural Language Processing tasks that have seemingly been solved, thanks to the vast amounts of knowledge encoded into Transformer-based pre-trained language models. And yet, if we look below the surface of raw figures, it is easy to realize that current approaches still make trivial mistakes that a human would never make. In this work, we provide evidence showing why the F1 score metric should not simply be taken at face value and present an exhaustive analysis of the errors that seven of the most representative state-of-the-art systems for English all-words WSD make on traditional evaluation benchmarks. In addition, we produce and release a collection of test sets featuring (a) an amended version of the standard evaluation benchmark that fixes its lexical and semantic inaccuracies, (b) 42D, a challenge set devised to assess the resilience of systems with respect to least frequent word senses and senses not seen at training time, and (c) hardEN, a challenge set made up solely of instances which none of the investigated state-of-the-art systems can solve. We make all of the test sets and model predictions available to the research community at https://github.com/SapienzaNLP/wsd-hard-benchmark.

1 Introduction

In recent years, Natural Language Processing (NLP) has witnessed a quantum leap in benchmark task performance, mainly thanks to the adoption of two major technical innovations: the Transformer architecture (Vaswani et al., 2017) and transfer learning from language models pre-trained on massive amounts of textual data (Devlin et al., 2019; Lewis et al., 2020). The impact of these breakthroughs was so strong that, on many benchmarks, the performance of human non-experts was surpassed (Wang et al., 2019b), prompting researchers to release new, more challenging benchmarks (Wang et al., 2019a).

Word Sense Disambiguation (WSD), the task of automatically assigning a meaning to an ambiguous word in context (Bevilacqua et al., 2021), is undergoing a similar process: current state-of-the-art systems are now capable of attaining and surpassing the F1 score of 80%[1] on standard test datasets (Bevilacqua and Navigli, 2020; Barba et al., 2021a; Conia and Navigli, 2021; Kohli, 2021), a figure often reported as the estimated human performance, because it corresponds to the highest recorded inter-annotator agreement (Edmonds and Kilgarriff, 2002; Navigli et al., 2007; Palmer et al., 2007).

Matching and/or surpassing human performance reasonably triggers the assumption that systems are capable of carrying out tasks in real-world scenarios as effectively as their human counterparts (Kiela et al., 2021), to the point where non-practitioners would regard such tasks as "solved". And yet, once systems are investigated beyond sheer accuracy figures, their flaws become readily apparent (Ribeiro et al., 2016; Belinkov and Bisk, 2018; Ribeiro et al., 2020; Card et al., 2020; Zhou et al., 2020). Following this trend of research, our work provides evidence showing why traditional evaluation measures for WSD, such as the F1 score, should not be taken at face value, hence corroborating the thesis that the problem of disambiguation is far from solved (Emelin et al., 2020; Loureiro et al., 2021).

To provide context, consider the following example, where the sense prediction[2] of the currently state-of-the-art ESCHER model (Barba et al., 2021a) for the word wind is compared with the gold answer from the test set of SemEval-2013 Task 12 (Navigli et al., 2013):

[1]Unless specified, for the remainder of this work, we will use "F1 score" to refer to the micro-averaged F1 score.

[2]According to the most commonly employed sense inventory for WSD, i.e., WordNet 3.0 (Fellbaum, 1998).

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4724 - 4737

context: The banks battling against a strong wind in the USA several years later. Investors and regulators (...)

gold: A tendency or force that influences events.

ESCHER: Air moving (...) from an area of high pressure to an area of low pressure.

Here, the contextual meaning of the word wind is clear to any English speaker, with no cues in the sentence that would lead a human reader to pick the "air" meaning. This is an illustrative case of why, despite having achieved (on paper) *superhuman* performance, systems continue to make mistakes that the inter-annotator agreement would not justify. Similarly, in the context below, another system which breaks the 80% performance ceiling (Conia and Navigli, 2021) makes a trivial mistake on a standard test instance (Snyder and Palmer, 2004), and fails to label the word couple properly:

context: I was just sitting down to meet with some new therapy clients, a couple, and the building started shaking (...)

gold: A pair of people who live together.

Conia and Navigli (2021): A small indefinite number.

With a view to gaining a better understanding of the nature of what systems still fail to disambiguate, in this work we provide the following main contributions: (i) we put forward a detailed quantitative and qualitative analysis of errors shared among seven state-of-the-art systems for English WSD, including systems that have surpassed the 80% human estimate in terms of F1 score (Bevilacqua and Navigli, 2020; Barba et al., 2021a), (ii) we produce an amended version of the English all-words WSD evaluation benchmarks featured in Senseval and SemEval tasks (Agirre et al., 2009; Raganato et al., 2017a), (iii) we devise "42D" (pron. [for·ti·tude]), the first manually-curated test bed made available to the research community after a hiatus of seven years since SemEval-2015 (Moro and Navigli, 2015), and a powerful evaluation tool for estimating system resilience in contexts featuring least frequent word senses, (iv) we establish a new human performance threshold for assessing actual *superhuman* scores on WSD test sets, and propose macro-averaged F1 score as an alternative to micro-averaged F1 score to better account for least frequent word senses in WSD evaluation, (v) we release "hardEN", a challenge set for English all-words WSD on which state-of-the-art systems under investigation achieve exactly 0.0% F1 score, and (vi) we set up an experimental setting to show the impact sense distribution has over the aforementioned datasets.

2 Related Work

WSD has witnessed the creation of many different evaluation benchmarks, most notably as part of the Senseval (now SemEval) evaluation campaigns (Kilgarriff, 1998). Since the release of the popular Unified Evaluation Framework by Raganato et al. (2017a), the experimental setting has become quite standard, with most systems being evaluated on ALL, i.e., the concatenation of Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 Task 1 (Snyder and Palmer, 2004), SemEval-2007 Task 17 (Pradhan et al., 2007), SemEval-2013 Task 12 (Navigli et al., 2013), and SemEval-2015 Task 13 (Moro and Navigli, 2015). Besides reporting results split by part of speech, which has not been particularly insightful, no specific finer-grained analysis is usually performed.[3] This trend runs the risk of promoting a sort of collective hill-climbing behavior, which, in turn, makes it unclear how much the improvement in performance has been due to genuinely stronger generalization power, as opposed to overfitting to increasingly stale test sets.

In opposition to this measure-centered style of evaluation, one possible alternative is that of behavioral testing, as proposed by Ribeiro et al. (2020). In their proposal (which does not address WSD explicitly), the benchmark evaluates separately minimum testable units of behavior, each of which addresses one specific skill required by a usable system. WSD, however, is a tricky problem to address in this way, as it is, in fact, a collection of idiosyncratic, diverse classification problems, which are hard to cluster in a meaningful way.

A different kind of analysis, perhaps more specific to WSD, has tackled the problem of the strong imbalance of sense distributions, which makes learning difficult for automatic algorithms, and monitors how this imbalance affects performance (Calvo and Gelbukh, 2015; Izquierdo et al., 2015; Postma et al., 2016; Wang and Wang, 2021). We

[3] Partial exceptions are Kumar et al. (2019), Bevilacqua et al. (2020), Blevins et al. (2021), Chen et al. (2021), and Barba et al. (2021a), which have paid particular attention to least frequent senses and data efficiency.

follow this line of research in that we also take sense distribution skewness as the core issue in the development of WSD algorithms. Therefore, both in the analysis of current WSD systems and in the creation of our new benchmarks, we check for the excessive influence of the most frequent output classes.

3 Systems at Issue

In an effort to make our analysis as thorough and comprehensive as possible, we consider a set of seven representative cutting-edge approaches for WSD.[4] With the exception of SyntagRank (Scozzafava et al., 2020), all systems are supervised neural architectures exploiting pre-trained language models. Below, we describe each of these systems:[5]

ARES (Scarlini et al., 2020)[6] is a semi-supervised approach to producing contextualized sense embeddings that share the same space as those from BERT (Devlin et al., 2019). It enables a simple 1 Nearest-Neighbour algorithm to attain high performance both in the English and multilingual settings despite relying on English training data only. We use the ARES English vectors freely available at `http://sensembert.org`.

BEM (Blevins and Zettlemoyer, 2020) is a bi-encoder model with high accuracy for the disambiguation of rare word senses. BEM maps the target in context and its word senses (as represented by glosses) independently into a shared embedding space, by means of jointly learned context and gloss encoders. Disambiguation is then performed simply by predicting the sense whose encoding is most similar to that of the target. We employ the model and code available at `https://github.com/facebookresearch/wsd-biencoders`.

ESCHER (Barba et al., 2021a, ESR) frames WSD as a span extraction task similar to SQuAD (Rajpurkar et al., 2016), in which a system is asked to detect the span matching the gloss of the correct sense for a target word from a pseudo-document constructed by concatenating the context of the target word with all the glosses of its possible senses. At the time of writing, ESCHER represents the state of the art in WSD.[7] We employ the model and code available at `https://github.com/SapienzaNLP/esc`.

EWISER (Bevilacqua and Navigli, 2020, EWR) is a WSD classifier that exploits relational information included in WordNet by incorporating a sparse adjacency matrix within the architecture. We employ the model and code available at `https://github.com/SapienzaNLP/ewiser`.

Generationary (Bevilacqua et al., 2020, GEN) reframes WSD as definition modeling, i.e., the task of generating a gloss from static or contextual embeddings (Noraset et al., 2017), therefore recasting disambiguation as a generative problem. We use the GEN-UNI (MBRR) model reported in the original paper. While the only exposure of the model to WordNet-tagged data was through SemCor (Miller et al., 1993), i.e., the most widely employed training set for WSD, the model was also trained on other lexicographic resources, such as the Oxford Dictionary (Chang and Chen, 2019).

GlossBERT (Huang et al., 2019, GLB) formulates WSD as a gloss ranking task, with a cross-encoder scoring context-gloss pairs. The model is trained with a simple learning-to-rank (He et al., 2008) approach, simply predicting whether a gloss is relevant to the context or not. We employ the model and code available at `https://github.com/HSLCY/GlossBERT`.

SyntagRank (Scozzafava et al., 2020, SYN) is a knowledge-based system that jointly exploits the Personalized PageRank algorithm and the wealth of syntagmatic information contained in SyntagNet (Maru et al., 2019) to perform disambiguation in multiple languages. We accessed SyntagRank by means of its APIs which are freely available at `http://api.syntagnet.org/`.

4 The Hard Core

To consider WSD as solved, it would be reasonable to expect disambiguation errors to be little more than mismatches between the reference ground truth and another different, but still reasonable interpretation. For example, if we consider the word

[4] To ensure a fair comparison, we only consider systems/settings that are not exposed to the Princeton WordNet Gloss Corpus (`https://wordnetcode.princeton.edu/glosstag.shtml`).

[5] For an extensive overview of state-of-the-art system backbones and trends in WSD, see Bevilacqua et al. (2021).

[6] For ease of reading, we will henceforth use abbreviations to identify some of the systems under investigation.

[7] Contemporary to this work, ConSeC (Barba et al., 2021c), which extends ESCHER, has now attained the new state of the art.

dataset	#inst	#mono	ARES	BEM	ESR	EWR	GEN	GLB	SYN	gold
ALL	7,253	1,301	71.3%	72.6%	71.2%	72.7%	**69.0%**	74.8%	81.1%	65.2%
ALL_{HC}	541	0	64.7%	71.0%	68.6%	67.8%	**62.7%**	70.6%	80.2%	2.0%
ALL	7,253	1,301	88.2%	87.4%	86.3%	88.8%	**85.9%**	88.6%	88.8%	84.3%
ALL_{HC}	541	0	96.9%	96.7%	96.5%	98.0%	**95.0%**	97.2%	98.3%	67.1%

Table 1: Times (%) systems predict the MFS in WordNet, i.e., WN1st (top), or a sense occurring at least once in SemCor (bottom). Left to right: dataset, number of instances (#inst), number of monosemous instances (#mono), system percentages (ARES, BEM, ESR, EWR, GEN, GLB, SYN), gold standard percentages (gold). **Bold** is closer to gold.

dataset	#inst (#mono)	ARES		BEM		ESR		EWR		GEN		GLB		SYN	
		M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
ALL	7,253 (1,301)	72.9	77.9	73.9	79.0	**76.4**	80.7	73.3	78.3	70.7	76.3	71.3	76.9	64.1	71.7
ALL_{no1st}	2,525 (0)	45.7	50.1	47.8	50.5	**54.2**	55.2	46.8	49.0	45.3	48.4	42.4	45.0	26.9	29.5
ALL_{noSC}	1,138 (448)	60.3	65.3	63.7	67.1	**71.0**	75.0	58.6	64.0	65.5	68.6	57.4	62.2	55.1	61.0
ALL_{HC}	541 (0)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 2: F1 scores for the reported systems on ALL and its subsets analyzed in Section 4.1. Top to bottom: ALL (Raganato et al., 2017a), the subset of ALL with no WN1st instances (ALL_{no1st}), the subset of ALL with no instances whose ground truth is in SemCor (ALL_{noSC}), and the subset of ALL featuring predictions errors shared by all systems (ALL_{HC}). Left to right: dataset, number of instances (#inst) of which monosemous (#mono), system performances (ARES, BEM, ESR, EWR, GEN, GLB, SYN) on macro (M-F1) and micro F1 (m-F1), respectively. **Bold** is M-F1 best.

chestnuts in "my aunt grows chestnuts", the two senses "any of several attractive deciduous trees yellow-brown in autumn" and "edible nut of any of various chestnut trees of the genus Castanea" would both be good, albeit slightly different interpretations, but the sense "the brown color of chestnuts", instead, is clearly not. To show that the current state of the art is nowhere near this level of performance, we select as a case study the set of instances in the Unified Evaluation Framework for English WSD of Raganato et al. (2017a) (ALL) which are wrongly disambiguated by all of the considered systems (see Section 3). We analyze this "hard core" (henceforth, ALL_{HC})—where performances are 0.0% in F1 score across the board by design—from both a quantitative and a qualitative perspective.

4.1 Quantitative Analysis

Sense distribution is a central problem for WSD. In our quantitative study, therefore, we analyze performances on the hard core by dividing test instances into frequency-based partitions. While performances are virtually always computed in terms of micro-averaged F1 scores, here we choose to report macro-averaged F1 (aggregated by sense), as the former gives more weight to frequent senses simply because they occur more often—thus hiding mediocre performances on least frequent senses.

Most Frequent Sense Bias. The most frequent class (in WSD, the most frequent sense, or MFS) can be overpredicted by machine learning algorithms (Postma et al., 2016; Blevins and Zettlemoyer, 2020; Loureiro et al., 2021). To quantify this phenomenon, in Table 1 (top), we report how many times our systems at issue predict the MFS in WordNet (henceforth, WN1st) on ALL_{HC}, as well as on ALL itself.[8]

As can be seen, systems show a clear bias towards WN1st senses on ALL, predicting them much more often (at least 69%) than the WN1st rate on the ground truth (65.2%). The distribution divergence becomes dramatic on ALL_{HC}, where systems predict WN1st at least 62.7% of the times, but the true WN1st rate is now just 2.0%. Overall, systems show a mostly comparable bias towards WN1st, with two notable exceptions: (i) GEN, likely due to the fact that in its UNI setting the system is exposed to multiple resources and hence is less biased; on the other hand, and perhaps counterintuitively (but see Calvo and Gelbukh, 2015) (ii) SYN, which is unsupervised, is the most biased

[8]We consider a test set instance to be a WN1st instance if at least one of the word senses assigned to disambiguate it coincides with the WN1st.

towards WN1st. Finally, we note that ESR, despite being the state of the art, does not behave differently from other systems in this respect, suggesting that there is much room for improvement.

In Table 2, we report both micro- and macro-averaged F1 scores on ALL, a subset of ALL without WN1st instances (ALL_{no1st}), and ALL_{HC}. As a consequence of the reduced importance of frequent senses, macro-averaged F1 scores are always lower than micro-averaged counterparts. Moreover, we can see that the reduced bias on WN1st by GEN results in a partial divergence between the system ranking on ALL and that on ALL_{no1st}, with GEN, which has a much lower WN1st bias, now outperforming GLB on the latter.

Training Dataset Bias. In addition to the WN1st bias, it is also useful to examine how much the lack of extrapolative capabilities is a reason for the existence of such a large set of unanswerable items. Thus, we classify instances and predictions according to whether the sense occurs at least once in SemCor (see also Kumar et al., 2019; Wang and Wang, 2021). Predicting a sense that never occurs at training time not only requires zero-shot capabilities, but also the ability to overcome the bias that a system learns from the training data for other senses of the same word. In Table 1 (bottom), we report the frequency with which our systems at issue predict a word sense that occurs at least once in SemCor. If we look at the raw percentages for ALL, there seems to be a slight bias towards senses that were seen at training time. However, such values do not take into account monosemous words for which the model always outputs the correct answer. In ALL_{HC}, where by construction there cannot be any monosemous sense, occurring senses are predicted at least 95% of the times, while they make up only 67.1% of the ground truth.

We refer back to Table 2 for the F1 scores on ALL_{noSC}, i.e., the subset of ALL with no instances whose gold sense is found in SemCor. The divergence between the ranking on ALL and ALL_{noSC} is even wider than that between ALL and ALL_{no1st}. In this case, GEN, which obtains rather unremarkable results on ALL, becomes the second-to-best on ALL_{noSC}, supporting the notion that gloss modeling is beneficial for WordNet-based WSD, even when using data outside of WordNet. Indeed, the gloss-centric approach of ESR offers the best results across the board, even though its bias on SemCor-attested (and WN1st) senses is still strong—hinting that a possible way forward could be combining ESR (or any equally strong baseline) with strategies meant to mitigate the bias.

4.2 Qualitative Analysis

Determining why a sizeable subset of instances cannot be disambiguated by any of the systems we take into consideration requires a finer-grained, qualitative level of analysis to check whether, i) annotation errors, or ii) gaps in WordNet, are an important factor. At the same time, iii) we also want to see if we replicate previous inter-annotator agreement figures (Edmonds and Kilgarriff, 2002; Navigli et al., 2007; Palmer et al., 2007). In order to achieve these objectives, we ask an expert linguist with extensive experience in tagging with the WordNet inventory[9] to revise the test instances in ALL, the main test set first provided by Raganato et al. (2017a),[10] as well as in the dataset released as part of SemEval-2010 in-domain WSD Task 17 of Agirre et al. (2009), by tagging each instance with one of the following labels:

- **unchanged**, to indicate that the annotator agreed with the existing ground truth;
- **fine-grained**, to indicate that one or more senses need to be added to the ground truth, without removing the existing ones;
- **error:token-lemma**, to indicate that the test instance was originally assigned a wrong lemma, or was improperly tokenized;
- **error:pos**, to indicate that the test instance was originally assigned a wrong part of speech (PoS);
- **error:sense**, to indicate that one or more senses in the ground truth are wrong;
- **error:inventory**, to indicate that the ground truth is wrong, but there is no appropriate sense for the target word in the inventory of WordNet 3.0.

Table 3 showcases an excerpt of instances as tagged by our linguist according to the aforementioned set of labels. Additionally, in Table 4, we

[9] All our annotators have effective operational proficiency in English and received a wage in line with their country of residence. Annotation was carried out by means of a user-friendly, in-house interface.

[10] We exclude SemEval-2007, since this dataset is often used as development set (Pasini et al., 2021).

tag (id)	**fine-grained** (semeval2010.d003.s043.t001)
ctx_tgt	See Map 1 for the boundaries of the realms
old	boundary%1:15:00:: the line or plane indicating the limit or extent of something
new	+ boundary%1:25:00:: a line determining the limits of an area
tag (id)	**error:pos** (senseval3.d001.s022.t007)
ctx_tgt	[...] have become virtually immune to defeat.
old	defeat%2:33:00:: win a victory over (VERB)
new	defeat%1:11:00:: an unsuccessful ending to a struggle or contest (NOUN)
tag (id)	**error:sense** (semeval2013.d003.s013.t002)
ctx_tgt	[...] which have cultivated close ties with the Iraqi Oil Ministry [...]
old	tie%1:11:00:: the finish of a contest in which the score is tied and the winner is undecided
new	tie%1:26:01:: a social or business relationship
tag (id)	**error:inventory** (semeval2010.d003.s059.t001)
ctx_tgt	Mangroves provide nurseries for 85 per cent of commercial fish species [...]
old	nursery%1:06:00:: a building with glass walls and roof; for the cultivation and exhibition of plants [...]
new	(*no suitable word sense featured in WordNet for "nursery"*)
tag (id)	**error:token-lemma** (semeval2015.d002.s021.t005)
ctx_tgt	[...] Italy, the Netherlands and the *United Kingdom*.
old_1	kingdom%1:14:01:: a monarchy with a king or queen as head of state
old_2	kingdom%1:15:01:: a country with a king as head of state
new	united_kingdom%1:15:00:: a monarchy in northwestern Europe occupying most of the British isles [...]

Table 3: Error analysis excerpt. In each block (top to bottom): (i) error **label** and instance identifier (tag(id)); (ii) original context and target (ctx_tgt); (iii) old ground truth (old); (iv) new ground truth (new). + indicates that a new sense has been added. *Italics* indicates the correct tokenization for the **error:token-lemma** case reported.

dataset	**#inst**	**unch.**	**fine**	**token**	**pos**	**sense**	**inv.**
ALL-	5,523	72.6	9.4	2.9	**0.3**	8.0	6.8
ALL_{NS}-	5,023	**75.4**	8.3	2.9	0.0	7.0	6.1
ALL_{HC}-	500	44.6	**20.4**	3.0	0.0	**17.8**	14.2
S10-	1,251	62.4	7.6	**4.7**	0.0	8.2	**17.1**

Table 4: Times (%) a label type is assigned to test set instances during the qualitative evaluation. **Bold** is highest.

provide a broader look and report the frequency of appearance (percentage) for each label, as assigned to (a) the concatenation of datasets in Raganato et al. (2017a) with the exception of monosemous words and SemEval-2007 instances (ALL-), (b) its subset of shared errors making up the hard core described in Section 4 (ALL_{HC}-), (c) ALL- not including instances featured in ALL_{HC}- (ALL_{NS}-), and (d) SemEval-2010 with no monosemous instances (S10-).

Two interesting results emerge from this analysis. On the one hand, the hard core seems to be "hard" for the human annotator too, since the majority of instances are labeled as either disambiguation errors (**error:sense**), or as lacking equally valid word senses (**fine-grained**). Indeed, the shared error subset (ALL_{HC}-) features the lowest level of **unchanged** instances and, at the same time, the highest rate of **error:sense** instances, meaning that the linguist had a significantly higher disagreement with respect to the original test set in ALL_{HC}- than in ALL_{NS}-. Furthermore, the percentage of cases in which the linguist deemed necessary the use of (i) additional word senses to disambiguate a certain instance (**fine-grained**) or (ii) the use of a word sense not featured in the inventory (**error:inventory**) is more than double that of the rest of the dataset. On the other hand, if we sum the percentage of **unchanged** instances with that of **fine-grained**, and exclude from the set of all instances the samples where disagreements do not depend on disambiguation choices (**error:pos**, **error:token-lemma**, **error:inventory**), the agreement of the linguist with respect to the gold standard is far superior to what is traditionally reported in the literature, reaching a high ceiling of 91.1%, more than 10% above traditional estimates (Edmonds and Kilgarriff, 2002; Navigli et al., 2007;

Palmer et al., 2007). Indeed, **fine-grained** instances do not involve a disambiguation error, but merely extend the instance with additional possible meanings. This can only *increase* performances, since the standard evaluation scorer provided as part of the framework of Raganato et al. (2017a) gives the system full score if the predicted sense is in the ground truth set.

5 New Benchmarks

Results from the quantitative and qualitative analysis carried out on the hard core reveal two main reasons why F1 scores can be potentially misleading indicators of the actual capabilities of current systems: (i) scores are actually a long way from estimated human performance when observed in challenging, but nevertheless real-world scenarios, and (ii) errors found in traditional test beds compromise insightful model evaluations. Against this background, we put forward a set of evaluation tools to enable a more robust appraisal of system performance in English WSD, namely, (i) 42D, a multi-domain challenge set, (ii) amended versions of ALL (ALL_{NEW}) and SemEval-2010 Task 17 (S10_{NEW}), and (iii) the new hardEN/softEN benchmark.

5.1 42D

Thus far, we have only considered existing evaluation benchmarks for WSD. In view of this—and with the purpose of showing that the issues highlighted in Section 4.1 are not artifacts of the data taken into account, but a general problem with current WSD systems—we introduce "42D", a novel test set for English WSD, built from scratch by manually annotating paragraphs taken from the British National Corpus (Leech, 1992, BNC).[11] 42D, with its 370 test instances, is specifically designed to be a challenge set (Belinkov and Glass, 2019), since for each of the instances the ground truth, i) does not occur in SemCor, and ii) is not the first sense in WordNet. In addition to this, 42D's source texts are sampled so as to be representative of different text domains, specifically, the 42 domains defined in BabelNet[12] 4.0 (Navigli and Ponzetto, 2012; Navigli et al., 2021).[13]

5.2 ALL_{NEW} and S10_{NEW}

With the aim of providing a cleaner test set, one in which non-system-dependent issues have been removed, we ask the same linguist who performed the error analysis of Section 4.2 to complete the task by also updating the instances from ALL and SemEval-2010 based on the labels assigned during the first phase: additional word senses are assigned for instances labeled as **fine-grained** and existing annotations are amended for **error:sense** cases; PoS tagging, lemmatization, and tokenization errors are fixed, and the instance updated with suitable word senses (see Table 3 for an excerpt of changes applied to the original test sets).

As a result, we obtain two test sets: ALL_{NEW}, featuring $4,917$ polysemous instances amending the original ALL dataset of Raganato et al. (2017a)[14], and S10_{NEW}, with 955 polysemous test instances amending the original SemEval-2010 Task 17 of Agirre et al. (2010).

5.3 hardEN and softEN

Besides an analysis of the current WSD evaluation datasets, in this paper we also want to make available one comfortable-to-use benchmark that addresses the discussed issues. For this reason, we derive a new intersection of 476 test instances that the systems at issue were not able to solve, this time, from the concatenation of the amended sets ALL_{NEW} and S10_{NEW}, as well as 42D. We name this challenge set "hardEN", in contrast to its counterpart, "softEN", which, instead, features the remaining $5,766$ test instances for which at least one system is able to provide a correct prediction. The hardEN/softEN benchmark is useful in that it sets a new "starting line" for WSD systems, one that concurrently accounts for what they still fail to do, while keeping track of what they can already do.

5.4 Evaluation

Table 5 compares the results obtained on our revised ALL_{NEW} dataset by the current state-of-the-art systems, with respect to the original ALL test set of Raganato et al. (2017a)—filtered to include only instances featured in ALL_{NEW} ($\text{ALL}*$), showing that the ranking of the systems taken into account

[11] This work was endorsed by the BNC staff via the official inquiry mail (ota@bodleian.ox.ac.uk) on October 15, 2019 and it complies with the BNC Licence for the use of paragraphs and other fragments (http://www.natcorp.ox.ac.uk/faq.xml?ID=licensing).

[12] BabelNet is freely available for research purposes at https://babelnet.org.

[13] See Appendix A for a full description of the building and annotation process of 42D.

[14] With the exception of SemEval-2007 instances.

dataset	#inst	ARES		BEM		ESR		EWR		GEN		GLB		SYN	
		M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
ALL$*$	4,917	69.3	75.5	69.9	76.2	**73.1**	78.3	70.0	76.0	66.1	73.1	67.7	74.4	57.9	66.9
ALL$_{NEW}$	4,917	75.2	79.0	75.6	79.5	**78.7**	81.6	75.6	79.2	72.2	76.7	73.2	77.4	61.4	68.5
S10$_{NEW}$	955	77.9	81.4	77.1	82.2	**78.0**	82.1	76.1	81.1	72.3	77.0	75.8	80.4	64.0	66.7
42D	370	41.8	37.8	53.2	47.8	**58.9**	54.1	43.9	40.8	50.2	48.9	45.7	41.9	32.8	28.1
softEN	5,766	78.7	83.3	80.3	84.5	**83.7**	86.8	79.2	85.0	76.4	82.3	77.1	82.0	63.4	71.3
hardEN	476	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 5: F1 scores for the reported systems on the datasets described in Section 5. Left to right: dataset/subdataset (dataset), number of instances (#inst), system performances (ARES, BEM, ESR, EWR, GEN, GLB, SYN) measured using both macro (M-F1) and micro F1 (m-F1). **Bold** is M-F1 best. $*$ indicates the subset of ALL (Raganato et al., 2017b) that includes only those instances that are also featured in ALL$_{NEW}$.

does not change as a result of the amending process. However, we can appreciate the significant difference in terms of performance when this is measured using the macro-averaged F1 score as opposed to the micro-averaged F1 score used in the literature. For example, the performance of ESCHER drops by almost 3 points on ALL$_{NEW}$, from 81.6% to 78.7%. Indeed, the macro-averaged F1 score is better suited to highlighting the weaknesses of a system with imbalanced class distributions, as is the case for word senses, whose distribution follows Zipf's Law. We argue, therefore, that future systems should also report their results using this measure in order to better enable their strengths and weaknesses to be determined.

Table 5 also shows the performance of each system on our revised SemEval-2010 (S10$_{NEW}$), 42D, and the hardEN/softEN benchmark. 42D is of particular interest as it showcases how the state of the art still struggles in challenging settings, including rare word senses and out-of-domain instances: the best system, ESR, only manages to score 54.1% in micro F1, a value that is very distant from the 80% figure originally estimated for human experts. As a last remark, it is worth noting how the performances on softEN for EWR and ESR reach and surpass the threshold of 85%, hence showing figures closer to the new, higher human performance ceiling we described in Section 4.2.

6 Where to go?

In this work, we dived deep into what the current state of the art in WSD can achieve and what the main roadblocks to overcome in the future are. With hardEN as the new frontier to surpass and softEN as a milestone to preserve, in this Section, we take the opportunity to briefly discuss possible directions for achieving both ends.

dataset	ESCHER		Uniform E.		Ranked E.	
	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
ALL$_{NEW}$	78.7	81.6	77.8	81.6	**78.8**	82.3
S10$_{NEW}$	78.0	82.1	79.5	83.7	**80.7**	84.9
42D	**58.9**	54.1	50.9	46.8	53.2	48.9
softEN	**83.7**	86.8	82.7	87.6	83.4	88.3
hardEN	0.0	0.0	0.0	0.0	0.0	0.0

Table 6: Macro- (M-F1) and micro-averaged F1 (m-F1) scores of our Uniform and Ranked ensemble strategies compared against the best performing systems, ESCHER. Best macro-averaged F1 scores are in **bold**.

Joining forces. One might wonder whether putting together multiple systems can be a viable approach for achieving progress in WSD, as preliminarily explored in the past by (Brody et al., 2006). Here we provide a provisional answer by investigating two simple ensemble strategies with the aim of understanding if it is possible to improve the results by making different and diverse systems agree. In the first ensemble strategy, i.e., uniform ensemble, we apply majority voting among the predictions of each of the seven systems; in the second strategy, i.e., ranked ensemble, each voting system is ranked according to its performance rank on ALL$_{NEW}$, e.g., the vote of ESCHER (the best system on ALL$_{NEW}$) is worth seven times that of SyntagRank (the seventh and worst system), in order to favor systems that are more likely to predict correct senses.

Interestingly, as Table 6 shows, even though re-

dataset	SemCor		K1		SemCor+K1	
	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
ALL_{NEW}	**78.7**	81.6	61.0	60.8	75.9	80.0
$S10_{NEW}$	**78.0**	82.1	68.5	67.4	76.2	80.1
42D	58.9	54.1	63.0	60.3	**65.2**	60.5
softEN	**83.7**	86.8	65.1	64.3	80.4	84.6
hardEN	0.0	0.0	**35.3**	33.6	16.8	14.5

Table 7: Macro- (M-F1) and micro-averaged F1 (m-F1) scores of ESCHER: trained only on SemCor, only on K1 (automatically-generated dataset containing one example per sense), and on SemCor + K1. Improving on hardEN decreases scores on softEN. Best macro-averaged F1 scores are in **bold**.

sults for ALL_{NEW} are slightly higher when using ranked ensembling, this strategy appears to be impairing performance in challenging settings such as 42D. Furthermore, by construction, if hardEN features all and only those instances that all the systems at issue fail to provide a correct answer for, then ensembles cannot represent a solution for hardEN, no matter the strategy employed.

Data augmentation. A renowned problem in WSD is the knowledge acquisition bottleneck: we have thousands of senses for which we have no available training data, but manual sense tagging is an expensive process (Pasini, 2020). What happens when a system is trained with automatically generated usage examples? To find out, we employ the examples generated via the EXMAKER encoder-decoder architecture (Barba et al., 2021b), to train ESCHER in two configurations: the first, in which the system is trained only with one automatically generated example per sense (K1), and the second, in which ESCHER is trained on the concatenation of SemCor and K1 (SemCor+K1).

As shown in Table 7, although ESCHER, when using K1, successfully "nibbles" at hardEN (achieving 35.3% in terms of macro-averaged F1 score), it does so at the expense of its performance on softEN (dropping more than 18% in macro-averaged F1 score), which is clearly undesirable. This is further proof that flattening the sense distribution on the training set is not sufficient to deal with hard test instances while at the same time preserving performance on the easier ones (see also Postma et al. (2016) and Loureiro et al. (2021)).

7 Conclusion

Although traditional metrics indicate that WSD systems have attained human-level performances, the actual capabilities of state-of-the-art models are poorly reflected by the current evaluation benchmarks. In this paper, we analyzed the intersection of errors made by a heterogeneous set of seven state-of-the-art systems for English WSD from a quantitative and qualitative perspective, detailing two main reasons why they still falter when compared to their human counterparts, namely, their strong bias towards most frequent word senses and towards senses featured in the training data, as well as the presence of an array of lexical and semantic fallacies in traditional evaluation benchmarks. With the aim of providing a test bench that is more effective in reflecting the actual capabilities of WSD systems, we introduced (i) an amended version of the most popular test bed for WSD, and (ii) the 42D challenge set. As a result of the aforementioned work, we also present the hardEN/softEN benchmark, a unified test bed aimed at moving forward with the disambiguation of so far unresolved instances, while keeping track of the current strong points of WSD systems. We make our test sets and model predictions available at https://github.com/SapienzaNLP/wsd-hard-benchmark.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487, the ELEXIS project No. 731015 under the European Union's Horizon 2020 research and innovation programme, and the European Language Grid project No. 825627 (Universal Semantic Annotator, USeA).

This work was partially supported by the COST Action CA18209 - NexusLinguarum "European network for Web-centred linguistic data science".

References

Eneko Agirre, Xabier Arregi, and Arantxa Otegi. 2010. Document expansion based on WordNet for robust IR. In *Coling 2010: Posters*, pages 9–17, Beijing, China. Coling 2010 Organizing Committee.

Eneko Agirre, Oier Lopez de Lacalle, Christiane Fellbaum, Andrea Marchetti, Antonio Toral, and Piek

Vossen. 2009. SemEval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 123–128, Boulder, Colorado. Association for Computational Linguistics.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021a. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

Edoardo Barba, Luigi Procopio, Caterina Lacerra, Tommaso Pasini, and Roberto Navigli. 2021b. Exemplification modeling: Can you give me an example, please? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3779–3785. ijcai.org.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021c. ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Michele Bevilacqua, 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. Association for Computational Linguistics.

Michele Bevilacqua 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. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.

Terra Blevins, Mandar Joshi, and Luke Zettlemoyer. 2021. FEWS: Large-scale, low-shot word sense disambiguation with the dictionary. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 455–465, Online. Association for Computational Linguistics.

Terra Blevins 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. Association for Computational Linguistics.

Samuel Brody, Roberto Navigli, and Mirella Lapata. 2006. Ensemble methods for unsupervised WSD. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 97–104, Sydney, Australia. Association for Computational Linguistics.

Hiram Calvo 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.

Jose Camacho-Collados 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. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Ting-Yun Chang 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. Association for Computational Linguistics.

Howard Chen, Mengzhou Xia, and Danqi Chen. 2021. Non-parametric few-shot learning for word sense disambiguation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1774–1781, Online. Association for Computational Linguistics.

Simone Conia and Roberto Navigli. 2021. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. In *Proceedings of the 16th Conference of the European*

Chapter of the Association for Computational Linguistics: Main Volume, pages 3269–3275, Online. Association for Computational Linguistics.

Jacob Devlin, 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. Association for Computational Linguistics.

Udaya Raj Dhungana and Subarna Shakya. 2015. Hypernymy in wordnet, its role in wsd, and its limitations. In *2015 7th International Conference on Computational Intelligence, Communication Systems and Networks*, pages 15–19. IEEE.

Philip Edmonds 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. Association for Computational Linguistics.

Philip Edmonds and Adam Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering*, 8(4):279–291.

Denis Emelin, Ivan Titov, and Rico Sennrich. 2020. Detecting word sense disambiguation biases in machine translation for model-agnostic adversarial attacks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7635–7653, Online. Association for Computational Linguistics.

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

Chuan He, Cong Wang, Yi-Xin Zhong, and Rui-Fan Li. 2008. A survey on learning to rank. In *2008 International Conference on Machine Learning and Cybernetics*, volume 3, pages 1734–1739. Ieee.

Luyao Huang, 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. Association for Computational Linguistics.

Ruben Izquierdo, Armando Suarez, and German Rigau. 2015. Word vs. Class-Based Word Sense Disambiguation. *Journal of Artificial Intelligence Research*, 54:83–122.

Salil Joshi, 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, Atlanta, Georgia. Association for Computational Linguistics.

Diptesh Kanojia, 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, Tartu, Estonia. University of Tartu Press.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124.

Adam Kilgarriff. 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.

Harsh Kohli. 2021. Training bi-encoders for word sense disambiguation. *ArXiv preprint*, abs/2105.10146.

Sawan Kumar, 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. Association for Computational Linguistics.

Geoffrey Neil Leech. 1992. 100 million words of English: the British National Corpus (BNC). *Language Research*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for Word Sense Disambiguation. *Computational Linguistics*, pages 1–55.

Christopher Manning, 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, Maryland. Association for Computational Linguistics.

Marco Maru, 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 3534–3540, Hong Kong, China. Association for Computational Linguistics.

George A. Miller, Claudia Leacock, Randee Tengi, 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*.

Andrea Moro 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, Colorado. Association for Computational Linguistics.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of BabelNet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Roberto Navigli, 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, Georgia, USA. Association for Computational Linguistics.

Roberto Navigli, 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. Association for Computational Linguistics.

Roberto Navigli 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.

Thanapon Noraset, 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, February 4-9, 2017, San Francisco, California, USA*, pages 3259–3266. AAAI Press.

Martha Palmer, 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.

Tommaso Pasini. 2020. The knowledge acquisition bottleneck problem in multilingual word sense disambiguation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4936–4942. ijcai.org.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: an extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13648–13656. AAAI Press.

Marten Postma, 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. European Language Resources Association (ELRA).

Sameer Pradhan, 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. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017a. 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. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017b. 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. Association for Computational Linguistics.

Pranav Rajpurkar, 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, Austin, Texas. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Marco Tulio Ribeiro, 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. Association for Computational Linguistics.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Online. Association for Computational Linguistics.

Federico Scozzafava, 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, Online. Association for Computational Linguistics.

Benjamin Snyder 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. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ming Wang and Yinglin Wang. 2021. Word sense disambiguation: Towards interactive context exploitation from both word and sense perspectives. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5218–5229, Online. Association for Computational Linguistics.

Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. 2020. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8215–8228, Online. Association for Computational Linguistics.

A Building and Annotating 42D

Building 42D. As a first step, we pre-processed the whole BNC raw text by means of the Stanford CoreNLP pipeline (Manning et al., 2014). Then, we split the corpus into chunks of less than 250 adjacent tokens (including punctuation). We exploited a straightforward unsupervised technique to automatically tag paragraphs from the BNC with domain labels from BabelDomains (Camacho-Collados and Navigli, 2017). Given that each BabelDomain label is associated with a set of synsets, with each synset having its own lexicalizations (e.g., car, automobile, and machine, for the WordNet synset "a motor vehicle with four wheels"), 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.[15] As the the automatic domain classification method is error-prone, we asked a linguist to check whether the top chunk for each domain, ranked by highest number of lexicalizations, was fluent and representative of conventional descriptive or narrative discourse, e.g., filtering out lists of countries for the `geography_geology_and_places` domain. 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).

Annotating 42D. We asked a linguist to annotate the pre-processed data from the BNC. For the annotation process, the linguist was asked to consider *all* the lexical clues available in WordNet, namely, lexicalizations, glosses, examples, and hypernymy/hyponymy relations, which often act as complementary sources of evidence (Joshi et al., 2013; Kanojia et al., 2014; Dhungana and Shakya, 2015). As a case in point, WordNet 3.0 defines two senses of the verb *say* as "utter aloud" and "express in words", respectively. Such glosses can be deemed similar when the verb is used to introduce direct speech. However, it is by looking at the usage examples that it can be noted how the direct speech is only featured for the word sense glossed as "utter aloud". In view of the above, the annotator was asked to (i) tag all content words in 42D, (ii) use multiple sense tags where appropriate, (iii) manually fix errors caused by the automatic nature of the pre-processing stage, and (iv) treat multiwords that appear in WordNet as a single instance. Finally, annotations featuring WN1st or monosemous senses were discarded. As a result, we collected an overall total of 370 manually annotated, challenging instances.

Once collected, we asked a second linguist to perform a blind annotation over the whole dataset and, consequently, computed a raw agreement of 79.6%. While this figure is lower than that computed for the ALL test set by Raganato et al. (2017a), 42D is much harder, as evidenced in Section 5.4.

[15] To ensure a significant inter-domain ambiguity, we only considered lexicalizations featured in more than one domain.