ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension

Edoardo Barba and **Luigi Procopio** and **Roberto Navigli**
Sapienza NLP Group, Sapienza University of Rome
{edoardo.barba, luigi.procopio, roberto.navigli}@uniroma1.it

Abstract

Supervised systems have nowadays become the standard recipe for Word Sense Disambiguation (WSD), with Transformer-based language models as their primary ingredient. However, while these systems have certainly attained unprecedented performances, virtually all of them operate under the constraining assumption that, given a context, each word can be disambiguated individually with no account of the other sense choices. To address this limitation and drop this assumption, we propose CONtinuous SEnse Comprehension (CONSEC), a novel approach to WSD: leveraging a recent re-framing of this task as a text extraction problem, we adapt it to our formulation and introduce a feedback loop strategy that allows the disambiguation of a target word to be conditioned not only on its context but also on the explicit senses assigned to nearby words. We evaluate CONSEC and examine how its components lead it to surpass all its competitors and set a new state of the art on English WSD. We also explore how CONSEC fares in the cross-lingual setting, focusing on 8 languages with various degrees of resource availability, and report significant improvements over prior systems. We release our code at `https://github.com/SapienzaNLP/consec`.

1 Introduction

Being able to understand the meaning of the various words within a particular text is a crucial problem in Natural Language Processing (NLP), with Word Sense Disambiguation (WSD) as arguably its most famous framing: given a word in context, this task aims to pair it with its most suitable meaning, chosen from a fixed sense inventory (Bevilacqua et al., 2021). Similarly to the vast majority of other NLP tasks, the advent of Deep Learning has significantly affected the landscape of WSD and has led to supervised neural models becoming its primary actors. In their simplest flavour, these approaches essentially frame this task as a multi-label classification problem over a large vocabulary of discrete senses (Raganato et al., 2017b; Hadiwinoto et al., 2019). However, although effective and straightforward, this formulation suffers from a number of pitfalls, most notably i) senses are only defined via their training set occurrences, with their actual linguistic meaning not explicitly embedded within the neural model, and ii) these architectures either behave poorly on rare and unseen senses, or cannot handle them at all. In order to address these issues, recent literature has proposed more sophisticated forms of supervision where definitions of senses, i.e. glosses (Kumar et al., 2019; Blevins and Zettlemoyer, 2020), and relational knowledge coming from the sense inventory (Bevilacqua and Navigli, 2020; Conia and Navigli, 2021) are integrated within the neural models.

Although performances have been rising steadily and are now beyond the 80% barrier on the established framework of Raganato et al. (2017a), virtually all modern approaches work under the limiting operational hypothesis that the target word's explicit meaning does not depend upon those of its surrounding words, differently from pre-neural strategies (Navigli and Velardi, 2004; Cuadros and Rigau, 2008). Indeed, while the commonly used pre-trained Transformer architectures and their self-attention mechanism (Vaswani et al., 2017; Devlin et al., 2019) certainly help in contextualizing a word on its context and latently model sense information, the actual disambiguation is performed *independently*, that is, without taking into consideration the explicit senses assigned to nearby words.[1] This assumption, likely a heritage from the original classification formulation, creates an intrinsic difference between the behavior of humans and that of systems.

[1] Henceforth, unless otherwise specified, we will always use the word *independent* with this connotation when referring to the disambiguation process.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1492–1503
November 7–11, 2021. ©2021 Association for Computational Linguistics

In this work, we focus on this shortcoming and propose CONtinuous SEnse Comprehension (ConSeC), a novel approach to WSD that exploits a feedback loop strategy to condition the disambiguation process also on the senses of co-occurring words. In particular, given an input text, we define an ordering of the words contained therein and disambiguate each word conditioning not only on its context and possible meanings, but also on the senses assigned to those words already classified. As the underlying neural architecture, inspired by the recent re-framing of WSD presented by Barba et al. (2021a) and its nimble adaptability to our setting, we leverage a Transformer model trained with a text extraction objective: given as input a text with a target word, its possible sense definitions and the list of already disambiguated words along with their chosen glosses, the model has to learn to extract the text span associated with the sense definition that best expresses the target word's meaning. Backed by several experiments on the English all-words WSD task (Raganato et al. (2017a), we show the benefits of our formulation, which surpasses the prior state of the art by 1.3 F1 points, and perform a complete ablation of the various components that lead ConSeC to achieve unprecedented performances. Furthermore, we also examine the scalability of our approach on the cross-lingual setting, evaluating ConSeC on the recently proposed framework of Pasini et al. (2021), and report significant improvements over prior systems. The contributions of this work are therefore as follows:

- We put forward ConSeC, a novel approach to WSD where the disambiguation process is conditioned not only on context and possible meanings, but also on the explicit senses assigned to nearby words.
- Our formulation surpasses all modern approaches on both English and cross-lingual WSD tasks by significant margins.
- We carry out a detailed analysis of different aspects of our approach, including an ablation over ConSeC components.

We release our code and models at `https://github.com/SapienzaNLP/consec`.

2 Related Work

Word Sense Disambiguation (WSD) is the task of associating words in context with their most suitable meaning in a fixed sense inventory (Bevilacqua et al., 2021), which is usually a dictionary-like lexical resource where a word's meanings (*senses*) are enumerated and defined via definitions (*glosses*) and usage examples. Nowadays dominated by supervised systems, with WordNet (Miller et al., 1990) and SemCor (Miller et al., 1993) acting, respectively, as the *de facto* standard sense inventory and training corpus for the English language, this task is generally approached as a multi-label classification problem with a number of different neural formulations.

Early neural approaches (Kågebäck and Salomonsson, 2016; Raganato et al., 2017b) focused on architectures where WSD was framed as token classification over WordNet senses. While already effective, these architectures displayed a number of shortcomings, especially with regard to modeling rare and unseen senses. To cope with these, many works started to complement the training data by exploiting different forms of lexical knowledge stored in WordNet, such as sense definitions (Kumar et al., 2019; Blevins and Zettlemoyer, 2020) and semantic relations (Bevilacqua and Navigli, 2020; Conia and Navigli, 2021), or with silver data produced via novel generative formulations (Barba et al., 2021b). Sense definitions, in particular, have been shown to significantly improve models' scalability to senses that are underrepresented in the training corpus, and their usage has been thoroughly investigated. Huang et al. (2019) frame WSD as a binary classification problem where, given a word in context and one of its possible definitions in the sense inventory, a model has to determine whether the meaning expressed by the definition provided is suitable for the word considered. Continuing this line of work, Blevins and Zettlemoyer (2020) leverage a bi-encoder that projects both words in context and glosses in a shared vector space; disambiguation is then performed by means of identifying the gloss closest to the target word. Pushing this research trend further, Barba et al. (2021a) propose to frame WSD as a text extraction task where, given a word in context and all its possible glosses, models have to extract the definition that best suits the word under consideration. The authors show that their formulation comes with several benefits, most importantly it allows models to attend to both the input context and all the definitions of the target word together, and does not require large output vocabularies. However, when disambiguating multiple words co-occurring in the same context, all of these approaches process each word independently from the others: neither is a word disambiguated taking into consideration the explicit senses assigned to nearby words, nor does its explicit sense affect their disambiguation.

Conversely, here we clearly point out the limits of this formulation and, standing out from previous research, propose a novel approach, ConSeC, which drops this assumption for the first time in supervised neural WSD. By introducing a feedback loop strategy that iteratively disambiguates the target words in a given context, also conditioning at each step on the sense assignments already performed, we report significant improvements over a wide array of experiments and find that ConSeC outperforms all its alternatives by a large margin.

3 ConSeC

We now describe ConSeC, our proposed approach for WSD. We first introduce our formulation and feedback loop strategy (Section 3.1). Then, we present our Transformer-based architecture (Section 3.2) and, finally, we explain how the disambiguation of words is arranged (Section 3.3).

3.1 Continuous Sense Comprehension

Formally, given a sense inventory S, consider a target word $\hat{w}_i$ occurring in a context $c_{\hat{w}_i}$ and let $D_{\hat{w}_i} = d_{i,1}, \dots, d_{i,k}$ be its k *candidate definitions*. Generally, it is unlikely that $c_{\hat{w}_i}$ is solely composed of $\hat{w}_i$. On the contrary, it is very probable that this context contains several words, among which a number of disambiguation targets $\hat{w}_1, \dots, \hat{w}_i, \dots, \hat{w}_n$ may be present, with $n \geq 1$.

While supervised systems commonly process each $\hat{w}_i$ independently from the others, here we propose the usage of a feedback loop strategy that allows already-made nearby sense assignments to be taken into account. To this end, we first define an ordering function f that sorts the disambiguation targets under consideration according to some criterion, denoting by $\tilde{w}_1, \dots, \tilde{w}_n$ the resulting sorted elements. Then, we frame the disambiguation of each $\tilde{w}_i$ as follows: given $\tilde{w}_i$, its context $c_{\tilde{w}_i}$ and candidates definitions $D_{\tilde{w}_i}$, along with the *context definitions* $\Delta_{\tilde{w}_i} = \delta_1, \dots, \delta_{i-1}$, that is, the definitions of the senses previously assigned to $\tilde{w}_1, \dots, \tilde{w}_{i-1}$, a model has to extract the gloss $d^* \in D_{\tilde{w}_i}$ that best expresses the meaning of $\tilde{w}_i$.

3.2 Model Architecture

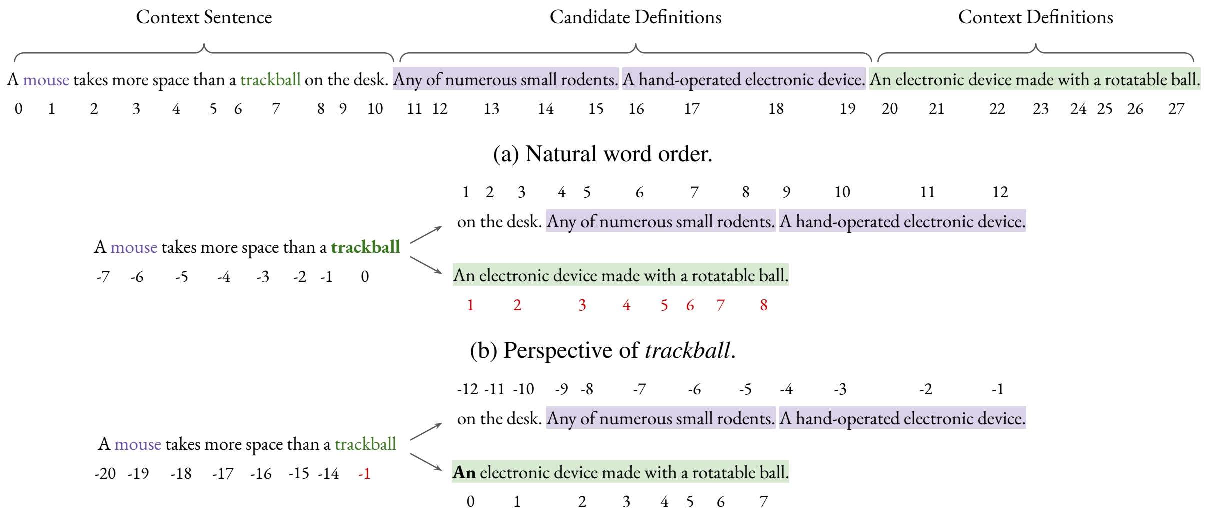
Starting from the text extraction framing of WSD introduced by Barba et al. (2021a), we implement our formulation as follows: given a context $c_{\tilde{w}_i}$ with a target word $\tilde{w}_i$, we first concatenate it with $D_{\tilde{w}_i}$ and $\Delta_{\tilde{w}_i}$, and then feed the resulting string to our model. As a signal that $\tilde{w}_i$ is the disambiguation target under consideration, we *mark* it, that is, surround it with the special tokens <*d*> and </*d*>. Furthermore, we use an additional special token, <*def*>, and prepend it to each candidate definition to denote their beginning.

Given this input, the model computes which <*def*> corresponds to the start of the gloss that best represents $\tilde{w}_i$. As our reference architecture, we use a linear classification head on top of DeBERTa[2] (He et al., 2021), a recently proposed Transformer model that improves over RoBERTa (Liu et al., 2019). The main reason behind this choice is the DeBERTa usage of *relative positions*, an encoding which, differently from its absolute counterparts commonly used in other Transformer architectures, models text positional information via a bi-dimensional matrix that stores the relative distance between each word pair. Leveraging this encoding and inspired by Liu et al. (2020b), we propose an elegant approach to inform the model that, $\forall \tilde{w}_j \in \{\tilde{w}_1, \dots, \tilde{w}_{i-1}\}$, the meaning of $\tilde{w}_j$ is the one expressed by δ_j: we place definitions immediately after the words they refer to, while simultaneously leaving unchanged the word order of $c_{\tilde{w}_i}$. This is achieved by overriding the relative distances in the positional matrix so that $\tilde{w}_j$ perceives next to it both δ_j and its subsequent words in $c_{\tilde{w}_i}$. Formally, we manipulate the relative positions as follows:

- with the exception of the distances between $\tilde{w}_j$ and the words in δ_j, we leave the natural order unchanged, with words in $c_{\tilde{w}_i}$, $D_{\tilde{w}_i}$ and $\Delta_{\tilde{w}_i}$ occurring sequentially one after the other (Figure 1a);
- $\tilde{w}_j$ witnesses $\delta_{\tilde{w}_j}$ after it (Figure 1b);
- symmetrically, $\delta_{\tilde{w}_j}$ perceives $\tilde{w}_j$ as immediately before it (Figure 1c).

This strategy, which lets different words see different word orders, provides a way to communicate the model, *in place*, the meaning of each word that

[2] We use DeBERTa $_{large}$ in all our experiments.


Context Sentence
Candidate Definitions
Context Definitions
A mouse takes more space than a trackball on the desk. Any of numerous small rodents. A hand-operated electronic device. An electronic device made with a rotatable ball.
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
1 2 3 4 5 6 7 8 9 10 11 12
on the desk. Any of numerous small rodents. A hand-operated electronic device.
A mouse takes more space than a trackball
-7 -6 -5 -4 -3 -2 -1 0
An electronic device made with a rotatable ball.
1 2 3 4 5 6 7 8
-12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1
on the desk. Any of numerous small rodents. A hand-operated electronic device.
A mouse takes more space than a trackball
-20 -19 -18 -17 -16 -15 -14 -1
An electronic device made with a rotatable ball.
0 1 2 3 4 5 6 7

(a) Natural word order.

(b) Perspective of *trackball*.

(c) Perspective of the first word of the definition of *trackball*.

Figure 1: Our positional strategy for “A mouse takes more space than a *trackball* on the desk”, where mouse is the disambiguation target and *trackball* a word already disambiguated whose meaning is *An electronic device made with a rotatable ball*: (a) natural word order (with absolute positions, to ease readability of (b) and (c)); (b) perspective of *trackball*, that is, relative distance of all words with respect to *trackball*; distances in red are different compared to their natural order; (c) perspective of the first word of the definition of *trackball*; subsequent words behave similarly but for their position offset (e.g. the relative position of *trackball* w.r.t. *electronic* is -2). For the sake of conciseness, we omit special symbols from the example reported. Best seen in color.

has already been disambiguated, without resorting to semi-structured formats (e.g. the usage of special linking variables) and, most importantly, while leaving the original context unaltered.

We train the whole architecture with a cross-entropy criterion, considering, however, for the loss computation only the logits corresponding to the candidate definition beginnings. Inspired by common practices in autoregressive models literature (Goodfellow et al., 2016), at training time, we use teacher forcing on the context definitions $\Delta_{\tilde{w}_i}$, that is, $\forall \delta_j \in \Delta_{\tilde{w}_i}$, δ_j is the gold definition assigned to $\tilde{w}_j$ in the training set. Conversely, at prediction time, we use a greedy decoding strategy and let δ_j be the definition the model deemed as the most likely one when disambiguating $\tilde{w}_j$.

3.3 Disambiguation Order

As function f (Section 3.1) defines which context definitions will be available for the disambiguation of $\tilde{w}_i$, providing a better characterization of its meaning, an adequate choice is crucial for ConSeC. Here, we make the common assumption in WSD that the more polysemous a word is, the harder it is to disambiguate, and define a function f that sorts $\hat{w}_1, \ldots, \hat{w}_n$ into $\tilde{w}_1, \ldots, \tilde{w}_n$ by increasing order of polysemy. However, if n is relatively large, as $i \rightarrow n$, $\tilde{w}_i$ will be swarming with context definitions, most of which are likely to be unnecessary and a source of noise for the neural model. For example, imagine $n \gg 20$ and that $\tilde{w}_n$ is some inflection of a very polysemous verb: the model would likely be flooded with trivial definitions and hindered from identifying those that are helpful. Furthermore, having too many definitions results in long encoded sequences and is particularly troublesome for most pre-trained Transformer language models as the complexity of their attention mechanisms scales quadratically.

To cope with this issue, we limit the number of dependencies to a maximum of *max_deps*, prioritizing their selection as follows: we first compute the positive normalized pointwise mutual information[3] (Bouma, 2009) between $\tilde{w}_i$ and each $\tilde{w}_j \in \tilde{w}_1, ..., \tilde{w}_{i-1}$. Then, applying an L_1-normalization, we convert these scores into a probability distribution and select all the highest-scoring $\tilde{w}_j$ needed to reach an α cumulative probability; in order to handle potential fat tail distributions that may occur, we enforce a minimum probability β, discarding $\tilde{w}_j$ if its probability is lower.[4]

[3] Further details in Appendix A.

[4] We treat *max_deps*, α and β as hyperparameters and will discuss them further in Section 4.1.

4 WSD Evaluation

We now assess the effectiveness of ConSeC examining first its applicability to English all-words WSD both in terms of performances (Section 4.1) and via an ablation study of its components (Section 4.2). We then proceed to investigate how ConSeC fares in the cross-lingual setting (Section 4.3).

4.1 English WSD

Data We evaluate ConSeC on English all-words WSD through the framework presented by Raganato et al. (2017a), using SemCor (Miller et al., 1993) as the training corpus. Following established practices in the WSD literature (Raganato et al., 2017b; Huang et al., 2019; Blevins and Zettlemoyer, 2020), we perform model selection on SemEval-2007 (Pradhan et al., 2007, **SE07**), while carrying out testing on Senseval-2 (Edmonds and Cotton, 2001, **SE2**), Senseval-3 (Snyder and Palmer, 2004, **SE3**), SemEval-2013 (Navigli et al., 2013, **SE13**) and SemEval-2015 (Moro and Navigli, 2015, **SE15**). As in previous works, we report the F1 score WSD systems achieve on each of these evaluation datasets and on their concatenation (**ALL**).

In order to have a better picture of models' performances and generalization power, we also consider the five synthetic datasets introduced by Barba et al. (2021a), namely: i) **MFS**, containing all the instances in ALL where the target word is tagged with its most frequent sense[5]; ii) **LFS**, containing all the instances in ALL annotated with a least frequent sense of the target word that appeared at least once in the training corpus; iii) **Unseen Senses**, containing all the instances in ALL tagged with a sense that is not in the training set; iv) **Unseen Words**, containing all the instances in ALL whose lemma and part of speech never co-occurred in the training dataset; v) **Unseen Definitions**, containing all the instances in ALL whose definition never appears in the training dataset.[6]

Finally, a number of recent works have started to use tagged glosses and examples coming from WordNet[7] (**WNGE**) as additional training data. For fair comparability with these systems, we also consider the setting where SemCor is complemented with these supplementary resources and train ConSeC on the resulting corpus.

Hyperparameters We train our system with a token batch size of 1536 and 5 steps of gradient accumulation. We use Rectified Adam (Liu et al., 2020a) as the optimizer, fine-tuning the whole architecture with a learning rate of 3^{-6} and a gradient clipping of 1.0, as in He et al. (2021). We limit the number of maximum training steps to 100,000 and evaluate model performance on the validation dataset every 2000 steps, enforcing a patience of 25 evaluation rounds. As regards ConSeC-specific parameters, we use $max\_deps = 9$, $\alpha = 0.7$ and $\beta = 0.1$ in all our experiments.[8] When working with document-level datasets, rather than treating the sentence where $\tilde{w}_i$ occurs as its context, we augment it so that $c_{\tilde{w}_i}$ also includes its preceding and subsequent sentence.

Comparison Systems We compare ConSeC with two common baselines in the WSD literature, namely i) **MFS-SemCor**, where target words are disambiguated by simply emitting their most frequent sense in SemCor, and ii) **BERT-base**, which employs a linear classifier over WordNet senses on top of frozen BERT representations (Devlin et al., 2019; Blevins and Zettlemoyer, 2020). Furthermore, to contextualize ConSeC performances in the current landscape of English WSD, we further consider a number of recent state-of-the-art systems and evaluate our approach against: **SVC** (Vial et al., 2019), which leverages WordNet relations to compress the output vocabulary and compensate for the lack of annotated data; **ARES** (Scarlini et al., 2020), a nearest-neighbor approach based on sense embeddings; **GlossBERT** (Huang et al., 2019), **BEM** (Blevins and Zettlemoyer, 2020), **EWISER** (Bevilacqua and Navigli, 2020)[9], **WMLC** (Conia and Navigli, 2021) and **ESCHER** (Barba et al., 2021a), all of which are supervised systems that exploit sense definitions or relational knowledge to better model the meaning of words. Finally, following the trend of augmenting training data with WNGE, we also consider how ConSeC fares in this setting and evaluate it against SVC, EWISER, WMLC and ESCHER.

[5] We compute sense frequencies from SemCor.

[6] This dataset differs from Unseen Senses as WordNet groups synonymous senses into lexical units called *synsets*, to which definitions are assigned, and therefore different senses may have the same definition.

[7] `https://wordnetcode.princeton.edu/glosstag.shtml`

[8] Further details on this choice in Appendix B.

[9] We note that Bevilacqua and Navigli (2020) use SE15 for model selection, therefore hindering direct comparability.

		Dev Set	**Test Sets**				**Concatenation of all Datasets**				
	Model	SE07	SE2	SE3	SE13	SE15	Nouns	Verbs	Adj.	Adv.	ALL
SemCor	MFS - SemCor	54.5	65.6	66.0	63.8	67.1	67.7	49.8	73.1	80.5	65.5
	BERT$_{base}$	68.6	75.9	74.4	70.6	75.2	75.7	63.7	78.0	85.8	73.7
	SVC	-	-	-	-	-	-	-	-	-	75.6
	SVC - *Ensemble*	69.5	77.5	77.4	76.0	78.3	79.6	65.9	79.5	85.5	76.7
	GlossBERT	72.5	77.7	75.2	76.1	80.4	79.8	67.1	79.6	87.4	77.0
	ARES	71.0	78.0	77.1	78.7	75.0	80.6	68.3	80.5	83.5	77.9
	EWISER	71.0	78.9	78.4	78.9	79.3	81.7	66.3	81.2	85.8	78.3
	WMLC	72.2	78.4	77.8	76.7	78.2	80.1	67.0	80.5	86.2	77.6
	BEM	74.5	79.4	77.4	79.7	81.7	81.4	68.5	83.0	**87.9**	79.0
	ESCHER	76.3	81.7	77.8	82.2	83.2	83.9	69.3	83.8	86.7	80.7
	ConSeC	**77.4**	**82.3**	**79.9**	**83.2**	**85.2**	**85.4**	**70.8**	**84.0**	87.3	**82.0**
+ WNGE	SVC	-	-	-	-	-	-	-	-	-	77.1
	SVC - *Ensemble*	73.4	79.7	77.8	78.7	82.6	81.4	68.7	83.7	85.5	79.0
	EWISER	75.2	80.8	79.0	80.7	81.8	82.9	69.4	83.6	87.3	80.1
	WMLC	76.2	80.4	77.8	81.8	83.3	82.9	70.3	83.4	85.5	80.2
	ESCHER	77.6	82.5	78.5	82.7	85.1	84.6	71.5	83.7	86.7	81.6
	ConSeC	**78.5**	**82.7**	**81.0**	**85.2**	**87.5**	**86.4**	**72.4**	**85.4**	**89.0**	**83.2**

Table 1: Results on the English all-words WSD task, when training on SemCor (top) and when also using WNGE (bottom). We mark in bold best scores per column and section, and underline the highest score on ALL whose difference with ConSeC is statistically significant ($p < 0.01$ according to the McNemar's test (Dietterich, 1998)).

Results We show in Table 1 the F1 scores ConSeC and its alternatives achieve on the evaluation datasets when training on SemCor (top).

Arguably the most interesting finding we report is the improvement ConSeC shows over ESCHER. Indeed, both systems use Transformer-based architectures with an almost identical number of parameters, and build upon text extraction formulations. The major difference lies in the usage of the feedback loop strategy we are proposing for ConSeC. The improvement of 1.3 points, which is statistically significant, clearly highlights the effectiveness of our proposal and the inherent limitations of performing WSD independently. Taking a broader look at the board, we can see that ConSeC surpasses all its comparison systems on all evaluation datasets except the POS-specific partition of ALL containing only adverbs, thus setting a new state of the art in English WSD.

Furthermore, we evaluate ConSeC when feeding both SemCor and WNGE to the learning procedure (Table 1, bottom). As Barba et al. (2021a) did not report on this setting, we run this experiment ourselves and train ESCHER on this augmented data.[10] Overall, we witness a similar trend compared to when training on SemCor only, with ConSeC surpassing all its alternatives. Scores are however much higher and the additional data allow ConSeC to achieve a completely unprecedented improvement, attaining 83.2 F1 on ALL. These results back our claim that the additional information coming from the glosses of already disambiguated instances does indeed help in identifying the correct meaning of words in context.

Frequency-Specific Evaluation We now carry out a coarse-grained error analysis, using the five datasets presented in Barba et al. (2021a) to examine how the model effectiveness changes when considering different frequency classes for target words and senses. We show the behavior of ConSeC in terms of F1 score in Table 2, comparing it with BEM and ESCHER on the *SemCor-only* training setting. Overall, our formulation achieves higher performances on 3 out of 5 datasets; the only exceptions are Unseen Definitions, where it behaves on par with ESCHER, and Unseen Words, where, instead, it falls short compared to it. Among these findings, the improvement on the LFS subset is particularly interesting. Indeed, as word senses follow the Zipfian distribution (Kilgarriff, 2004), supervised WSD systems tend to have a strong bias towards the MFS, making improvements in this setting hard to achieve. Thus, the 1.7 points over ESCHER, which was the prior best system, suggest that ConSeC better counterbalances the strong bias in senses distribution, while experiencing no side effect on MFS performances which, in fact, rise. This finding further highlights the posi-

[10] We used the authors' code released at `https://github.com/SapienzaNLP/esc`.

Model	MFS	LFS	U-Words	U-Senses	U-Defs
BEM	94.7	52.1	91.2	67.1	68.2
ESCHER	93.7	55.7	**95.1**	75.0	**76.8**
ConSec	**95.3**	**57.4**	92.8	**75.3**	**76.8**

Table 2: F1 scores of BEM, ESCHER and ConSec on MFS, LFS and Unseen* datasets. Results underlined or in bold have the same meaning as in Table 1.

Model	Stat. Sign.	ALL
ESCHER	-	80.7
+DeBerta	✗	80.6
+More Context	✗	81.0
+Context Defs	✓	82.0

Table 3: Ablation study over ConSec components. The *Stat. Sign.* column denotes whether the difference w.r.t. the row before is statistically significant. The line of +*Context Defs* is equivalent to ConSec.

Language	SyntagRank	EWISER	XLMR	ConSec
English	70.0	73.3	76.3	**79.0**
Dutch	56.0	57.5	59.2	**63.3**
Estonian	56.3	66.0	66.1	**69.8**
French	70.0	80.9	83.9	**84.4**
German	76.0	80.9	83.1	**84.2**
Italian	69.6	74.6	77.6	**79.3**
Japanese	57.5	55.8	61.9	**63.0**
Spanish	68.6	71.9	75.9	**77.4**

Table 4: Cross-Lingual Word Sense Disambiguation results on Pasini et al. (2021). Results underlined or in bold have the same meaning as in Table 1.

tive impact of our formulation and the benefits of introducing context definitions into the disambiguation process.

4.2 Ablation

As both ESCHER and ConSec leverage text extraction formulations, with similar underlying architectures, and yet ConSec significantly outperforms ESCHER, we ablate here the differences between the two systems, namely i) the usage of DeBerta, ii) having $c_{\tilde{w}_i}$ also include the previous and subsequent sentence[11], and iii) the introduction of context definitions. We show iteratively how performances change, in terms of F1 score, as we move from ESCHER to ConSec in Table 3.

As a first result, we note that, as expected, changing the underlying model of ESCHER from BART (Lewis et al., 2020) to DeBerta does not cause any significant difference in performances. Indeed, the two systems feature an almost identical number of parameters and attain similar scores on text extraction tasks such as SQuAD (Lewis et al., 2020; He et al., 2021), with DeBerta behaving slightly better. Once we include more surrounding context in $c_{\tilde{w}_i}$, performances rise to 81.0, suggesting that the additional context provides valuable information to the neural model.[12] However, this system, which differs from ConSec only in what pertains context definitions, achieves 1 F1 point less than ConSec; once we include this component, performances rise back to 82.0, showing the benefits of introducing our feedback loop strategy.

4.3 Cross-Lingual WSD

We now examine whether our approach can scale to different languages, evaluating ConSec against the cross-lingual framework made available by Pasini et al. (2021). Within the scope of this work, we limit ourselves to considering only the following language-specific setting for ConSec: for each language in the test bed, we train a dedicated monolingual model, using both datasets and sense definitions in that language. We defer exploring zero-shot and multilingual settings to future work.

Training is performed using the silver monolingual datasets the authors release within the framework. As our cross-lingual framing expects sense definitions to be in the same language as that of its datasets, we translate English glosses towards each language using the multilingual translation model released by Tang et al. (2020). Since a multilingual version of DeBerta is not available, we replace it with mBART (Liu et al., 2020c);[13] however, as mBART does not support relative positions, to inject the knowledge that $\tilde{w}_j$ *means* δ_j, we first prepend $\tilde{w}_j$ to each $\delta_j \in \Delta_{\tilde{w}_i}$, and then concatenate them right after the candidate definitions.

We compare ConSeC against three systems included in Pasini et al. (2021), namely EWISER, XLMR-Large[14] and SyntagRank (Scozzafava et al., 2020); EWISER and XLMR-Large are supervised systems, while SyntagRank is an unsupervised knowledge-based approach that builds upon syntagmatic relations (Maru et al., 2019). For both EWISER and XLMR-Large, we only consider the zero-shot scenario the authors illustrate, as it is the

[11] ESCHER treats plain sentences as its contexts.

[12] This is especially true for short contexts, where the additional data might solve otherwise unsolvable ambiguities.

[13] Details on the parameters used in Appendix B.

[14] We consider the XLMR-Large architecture as it achieved the best average results on all the languages of the framework.

Test	ALL	Δ
No Context Definitions	80.7	-1.3
Adversarial	80.0	-2.0
Teacher Forcing	82.5	+0.5

Table 5: Behavioral tests on ConSeC. The Δ column reports the relative difference w.r.t. ConSeC trained on SemCor and evaluated on ALL.

only setting where EWISER is available and, besides, the one where XLMR-Large fares the best.[15]

Table 4 reports the F1 scores on 8 different languages[16]. As the results show, ConSeC establishes a marked new state of the art on all the languages under examination; in particular, our approach outperforms all its competitors even on low-resource languages (e.g. Estonian and Japanese). The formulation we propose is therefore a viable option for cross-lingual WSD, especially when requiring high-performing systems.

5 Analysis

To get better insights into the impact context definitions have on ConSeC performance, we devise three behavioral tests. First of all, we examine to what extent performances degrade at prediction time when the feedback loop is disabled, that is, when we no longer provide context definitions to ConSeC. As shown in Table 5 (first row), while performances certainly drop, they are still on par with the previous state of the art and suggest that the feedback loop can be treated as a pluggable component: it can be disabled for fast processing, allowing batched independent disambiguation, or, vice versa, enabled when higher accuracy is crucial.

Second, we perform an adversarial attack where we provide ConSeC with wrong context definitions in the feedback loop: instead of using the most likely definition of $\tilde{w}_j$ as δ_j, we let δ_j be the less likely one. As shown in Table 5 (second row), performances drop by only 0.7 points compared to when context definitions are missing rather than purposely wrong. This result seems to suggest that the model learns to ignore noisy or irrelevant definitions at training time and we hypothesize this may be caused by our choice of using a heuristic non-differentiable objective as function f: indeed, as the neural model cannot tune the selection strategy, it may learn instead to ignore unrelated definitions. Thus, the less likely gloss does not necessarily have a negative impact and a more significant test may be performed involving human annotators to wisely select adversarial definitions that could most interfere with the disambiguation.

Finally, we investigate the discrepancy between teacher forcing and greedy decoding at inference time, and test the model when, instead, setting δ_j to the gold definition of $\tilde{w}_j$ in the evaluation datasets. As reported in Table 5 (third row), the two techniques do not differ significantly. While surprising at a first glance, a number of factors might actually explain this result. First, ConSeC performances are above 80%, implying that many of the definitions chosen by greedy decoding are correct. Furthermore, as instance polysemy and error rate strongly correlate,[17] our function f has the following implications: i) early instances are less likely to be wrongly disambiguated and become inaccurate hints to subsequent words, and ii) while late instances have a higher probability of receiving wrong context definitions, they also have access to more data and can therefore counterbalance possible mistakes. Finally, even when mistakes do occur and wrong hints are generated, due to WordNet sense granularity, the misclassified sense may not be that dissimilar from the correct one or, in fact, even appear as a close alternative (Erk and McCarthy, 2009), thus still acting as a valuable hint to the neural model.

6 Conclusion

In this work we presented ConSeC, a novel extractive approach to WSD that allows the disambiguation of words to be conditioned not only on their context and possible meanings but, for the first time in neural WSD literature, also on the explicit senses assigned to nearby words. By explicitly embedding their definitions within the model input, we report significant results on both English and cross-lingual WSD, establishing a new state of the art in both settings. Most notably, our system reaches the unprecedented performances of 82 F1 on the

[15] Pasini et al. (2021) also explore training XLMR-Large on their silver monolingual datasets, which are the same as the ones we use here. Interestingly, they report inferior performances compared to plain zero-shot, suggesting possible quality concerns regarding the automatic silver creation procedure.

[16] We chose to consider ConSeC only on Dutch, English, Estonian, French, German, Italian, Japanese and Spanish as this language set is the result of the intersection between the languages supported by mBART and the ones available in the multilingual framework.

[17] The two variables correlate with a 0.94 Spearman correlation, further analysis in Appendix C.

standard English framework when trained on SemCor only, and 83.2 when leveraging the WordNet tagged glosses and examples as additional training data. We perform several experiments investigating different aspects of our new formulation and, in particular, addressing its potential speed concerns, we demonstrate that the usage of context definitions can be treated as a pluggable component in our system, to be activated when higher accuracy is required, or seamlessly removed when speed is of primary importance. As future work, we plan to make the disambiguation ordering and the dependencies choice fully differentiable, while at the same time expanding on ConSec applicability to the cross-lingual setting.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487.

This work was partially supported by the MIUR under the grant “Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of Sapienza University.

References

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-21*, pages 3779–3785. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

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.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Comet.ML. 2021. Comet.ML home page.

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.

Montse Cuadros and German Rigau. 2008. KnowNet: Building a large net of knowledge from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 161–168, Manchester, UK. Coling 2008 Organizing Committee.

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.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

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.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 440–449. ACL.

William Falcon, Jirka Borovec, Adrian Wälchli, Nic Eggert, Justus Schock, Jeremy Jordan, Nicki Skafte, Ir1dXD, Vadim Bereznyuk, Ethan Harris, Tullie Murrell, Peter Yu, Sebastian Præsius, Travis Addair, Jacob Zhong, Dmitry Lipin, So Uchida, Shreyas Bapat, Hendrik Schröter, Boris Dayma, Alexey Karnachev, Akshay Kulkarni, Shunta Komatsu, Martin.B, Jean-Baptiste SCHIRATTI, Hadrien Mary, Donal Byrne, Cristobal Eyzaguirre, cinjon, and Anton Bakhtin. 2020. Pytorch Lightning.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. 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 5297–5306, Hong Kong, China. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021*.

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.

Mikael Kågebäck and Hans Salomonsson. 2016. Word sense disambiguation using a bidirectional LSTM. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 51–56, Osaka, Japan. The COLING 2016 Organizing Committee.

Adam Kilgarriff. 2004. How dominant is the commonest sense of a word? In *Text, Speech and Dialogue*, pages 103–111, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

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.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020a. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020b. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2901–2908.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020c. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.

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, 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 and Paola Velardi. 2004. Structural semantic interconnection: a knowledge-based approach to word sense disambiguation. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic*

Analysis of Text, pages 179–182, Barcelona, Spain. Association for Computational Linguistics.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13648–13656.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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, Claudio Delli Bovi, and Roberto Navigli. 2017b. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167, Copenhagen, Denmark. 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.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, page 6000–6010. Curran Associates, Inc.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2019. Sense vocabulary compression through the semantic knowledge of WordNet for neural word sense disambiguation. In *Proceedings of the 10th Global Wordnet Conference*, pages 108–117, Wroclaw, Poland. Global Wordnet Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Positive Normalized Pointwise Mutual Information

To choose the most important context definitions for each disambiguation instance, we use the positive normalized pointwise mutual information measure (*pnpmi*), that we compute using a document collection as follows. For each pair of words (w_i, w_j), we first calculate the pointwise mutual information (*pmi*) between w_i and w_j:

$$pmi(w_i, w_j) = \log_2 \frac{p(w_i, w_j)}{p(w_i) \cdot p(w_j)}$$

where $p(w_i, w_j)$ is the probability for w_i and w_j to co-occur in a document of the collection while $p(w_i)$ ($p(w_j)$) is the probability for w_i (w_j) to occur in a document. Then we compute:

$$npmi(w_i, w_j) = \frac{pmi(w_i, w_j)}{h(w_i, w_j)}$$

Hyperparameter	**Values**
Neural	
Optimizer	RAdam
Learning Rate	$3e^{-6}$
Gradient Accumulation Steps	[1.0, **5.0**, 10.0]
Weight Decay*	0.01
Tokens Batch Size	1536
Patience	25
Validation Check Interval	2000
ConSec Specific	
Dependencies Cumulative Probability (α)	[0.5, **0.7**, 0.9]
Minimum Normalized Pmi (β)	[0.0, **0.1**, 0.2]
Maximum Number of Dependencies	[5, **9**, 20]

Table 6: Explored hyperparameters ranges via grid search. When multiple values for a hyperparameter have been evaluated, we report in bold the best performing one. Top: standard hyperparameters involved in training neural architectures. Bottom: ConSec-specific hyperparameters. * We do not apply weight decay on neither the bias weights of Linear layers nor the weights of Layer Norm layers.

with $h(w_i, w_j)$ being $-\log_2 p(w_i, w_j)$, and finally:

$$pnpmi(w_i, w_j) = max(0, npmi(w_i, w_j))$$

We use Wikipedia as our document collection, and compute the words frequencies on a sample of 100,000 documents for each language.

B Training Details

Hyperparameters We report in Table 6 the hyperparameters with which we trained ConSec on the English WSD task. With the exception of the number of gradient accumulation steps, all hyperparameters come from previous literature, especially from He et al. (2021). Conversely, this parameter, together with ConSec-specific hyperparameters, has been tuned on the standard WSD framework (Raganato et al., 2017a) via a grid search approach.

For what concerns cross-lingual WSD, we follow the training hyperparameters used in Barba et al. (2021a) for their BART-based model while, for ConSec-specific parameters, we rely on the values that performed best in the English setting.

Implementation Our work is implemented in PyTorch (Paszke et al., 2019), using PyTorch Lightning (Falcon et al., 2020) as the underlying framework. We retrieve the pretrained models for DeBERTa-Large and mBART from HuggingFace Transformers (Wolf et al., 2020); we note the two models have $406M$ and $610M$ parameters respectively, on top of which we only add a linear classifier with shape $l \times 1$, where l represents the size of the final hidden states and amounts to 1024 for both architectures. To track and optimize our experiments, we used Comet.ML (2021).

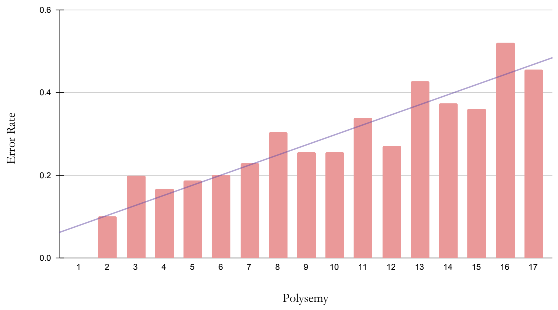


Figure 2: Polysemy to Error Classification Rate that ConSec trained on SemCor shows on the ALL dataset.

Hardware and Runtime We trained every model on a GeForce RTX 3090 graphic card with 24 gigabytes of VRAM. All trainings on SemCor lasted between 7 and 20 hours, while those on SemCor and WNGE between 25 and 45 hours.

C Polysemy to Error Rate Correlation

To back our claims that more polysemous words are more difficult to disambiguate, we computed the Spearman's correlation coefficient between the polysemy of senses and their classification error rate by our best system trained on SemCor and tested on ALL. We find that the correlation coefficient is very high, amounting to 0.94 with a *p-value* $\ll 0.01$. We show in Figure 2 the classification error rate for words with up to 17 possible senses.