SANDWiCH: Semantical Analysis of Neighbours for Disambiguating Words in Context ad Hoc

Daniel Guzman-Olivares
Bulil Technologies
Autonomous University of Madrid
daniel.guzmano@estudiante.uam.es

Lara Quijano-Sanchez
Autonomous University of Madrid
lara.quijano@uam.es

Federico Liberatore
Cardiff University,
liberatoref@cardiff.ac.uk

Abstract

The rise of generative chat-based Large Language Models (LLMs) over the past two years has spurred a race to develop systems that promise near-human conversational and reasoning experiences. However, recent studies indicate that the language understanding offered by these models remains limited and far from human-like performance, particularly in grasping the contextual meanings of words—an essential aspect of reasoning. In this paper, we present a simple yet computationally efficient framework for multilingual Word Sense Disambiguation (WSD). Our approach reframes the WSD task as a cluster discrimination analysis over a semantic network refined from BabelNet using group algebra. We validate our methodology across multiple WSD benchmarks, achieving a new state of the art for all languages and tasks, as well as in individual assessments by part of speech. Notably, our model significantly surpasses the performance of current alternatives, even in low-resource languages, while reducing the parameter count by 72%.

1 Introduction

In 2022, OpenAI fine-tunned their previously released *GPT-3* (Brown et al., 2020) model using Reinforcement Learning from Human Feedback (RLHF), resulting in the *InstructGPT* model (Ouyang et al., 2022). Using this model and a massive dataset as a base, in November of the same year, OpenAI released a sibling model, the now famous *ChatGPT*. The release served as the starting pistol for the ongoing race of chat-based Large Language Models. During the last two years, we have seen a consistent improvement in capabilities on different tasks (Minaee et al., 2024; Chiang et al., 2024) from bigger and newer models like *LLama* (Touvron et al., 2023), PALM-2 (Google, 2023) *Falcon* (Penedo et al., 2023), *Mistral* (Jiang et al., 2023) or *GPT-4* (OpenAI, 2023). However, recent studies (Kocoń et al., 2023; Qin et al., 2023; Balloccu et al., 2024; Liu et al., 2023) suggest that these models struggle in logic reasoning tasks when the data is out of distribution from their train corpus and fail to match the performance of previously introduced specialized solutions. The difference in performance is particularly noticeable in the tasks requiring to assess the meaning in which words are used in a sentence, where we observe that recent chat-based models lag behind much smaller fine-tuned architectures (Eisenschlos et al., 2023; Kocoń et al., 2023; Sumanathilaka et al., 2024; Qorib et al., 2024).

The Word Sense Disambiguation (WSD) task consists in identifying the sense in which a word is used in some given context from a pool of possible senses (Bevilacqua et al., 2021), e.g., in the sentence "*The crane was lifting a concrete block.*", a *crane* refers to a lifting machine used in construction rather than a large, long-necked bird. Far from the massive chat-based language models that we can find today, the state-of-the-art models for this task are considerably smaller and were introduced during the last five years (e.g. (Barba et al., 2021b,a; Blevins and Zettlemoyer, 2020; Huang et al., 2019; Kumar et al., 2019)) . In general, these models address the WSD problem by assessing the semantic similarity between the target word or its neighbouring context and the candidate definitions, using a fine-tuned encoder-based model (Barba et al., 2021b,a; Blevins and Zettlemoyer, 2020; Huang et al., 2019) or external resources (Blevins and Zettlemoyer, 2020; Kumar et al., 2019). Although these approaches generally provide remarkable results, they show a significant decrease in performance for verbs compared to other parts of speech, rare glosses, and underrepresented languages (Maru et al., 2022; Liu and Liu, 2023; Barba et al., 2021b). These problematic cases suggest that current solutions struggle to accurately model word senses in low-resource settings, as verbs often have multiple possible senses that

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7019–7033
April 29 - May 4, 2025

are unevenly distributed across different contexts, rare glosses are typically endemic to specialized domains, and training data for underrepresented languages is generally scarce. Addressing these deficiencies is crucial for improving generalization across out-of-distribution domains, where the traditional approach of training on large batches of annotated general-domain examples often fails to produce satisfying results (Maru et al., 2022; Navigli et al., 2023) and bridging the WSD performance gap between English and low-resource languages. Ideally, a general solution to the WSD problem should reduce the dependency of performance on the frequency of senses and contexts present in the training data (Kilgarriff, 2004), while addressing the issue at a structural level shared by all languages. Under this premise, we hypothesize that reframing the WSD task as a cluster discrimination task over a semantic network (e.g., BabelNet (Navigli and Ponzetto, 2010)) could address the aforementioned challenges.

In this work, we introduce SANDWiCH[1], a word disambiguation framework that leverages the close relationship between a candidate sense and its neighbors in a semantic network to shift the task from discriminating individual senses to discriminating semantically-close clusters. To this end, SANDWiCH incorporates additional elements into the two-level framework introduced in Barba et al. (2021b), which consists of coarse sense retrieval followed by a fine-tuned encoder-based model. Specifically, we introduce the processing of the semantic network to ensure it is *sense-separated* [**C1**], the inclusion of neighboring key concepts as part of the training data for the encoder-based models [**C2**], the separation of models by part of speech (POS) [**C3**], and the definition of a context-cluster score [**C4**].

Through extensive experimentation on the English all-words WSD task (Raganato et al., 2017a), we establish a new state of the art, achieving a 8% improvement in F1 score across all datasets, consistently outperforming existing solutions in every subset, including those defined by individual datasets and parts of speech. We further evaluate our framework on the more challenging dataset introduced by Maru et al. (2022), achieving an improvement over the previous state of the art ranging between 10-30% depending on the dataset. Additionally, on the multilingual dataset (Pasini et al., 2021), we improve state-of-the-art results for all languages, with particularly notable gains in underrepresented ones.

Therefore, the key contributions of this work are as follows:

- **SANDWiCH framework**: We introduce a novel word sense disambiguation framework that shifts the focus from individual sense discrimination to cluster-based sense discrimination, utilizing sense-separated semantic networks and neighboring key concepts to improve performance and robustness.

- **State-of-the-art results on English datasets**: Our system achieves a 8% improvement in F1 score on the English all-words WSD task, consistently surpassing the state of the art across all datasets and parts of speech, including the challenging dataset introduced by Maru et al. (2022).

- **Multilingual generalization**: The proposed framework generalizes effectively to multilingual settings, achieving state-of-the-art results across all languages in the multilingual WSD dataset (Pasini et al., 2021), with significant improvements in underrepresented languages.

2 Related Work

Historically, the WSD problem was introduced in the second half of the twentieth century as part of machine translation efforts (Weaver, 1949/1955; Bar-Hillel, 1960), later evolving into a standalone problem. Early successful approaches primarily relied on rule-based algorithms, statistical methods, and unsupervised techniques (Gale et al., 1992; Yarowsky, 1992; Lesk, 1986; Cowie et al., 1992; Yarowsky, 1995). The development of large-scale structured language resources like Wikipedia and BabelNet (Navigli and Ponzetto, 2010) enabled models to use gloss similarity heuristics and graph proximity metrics to address the WSD problem (Moro et al., 2014; Wang et al., 2015; McCarthy et al., 2016; Jain and Lobiyal, 2015).

The introduction of the first word embedding algorithms (e.g., Word2Vec (Mikolov et al., 2013), fastText (Bojanowski et al., 2017), or GloVe (Pennington et al., 2014)) significantly advanced WSD performance by leveraging seq-to-seq supervised approaches (Kågebäck and Salomonsson, 2016; Taghipour and Ng, 2015; Yuan et al., 2016; Luo

[1]We release all the code for reproducing the paper results in https:www.github.com/danielguzmanolivares/sandwich

et al., 2018). The use of word embeddings as a foundation for neural approaches led to substantial performance gains, which became even more pronounced with the adoption of dynamic embeddings from encoder models (e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), or DeBERTa (He et al., 2020)), derived from the transformer architecture (Vaswani et al., 2017). Fine-tuning an encoder model has since then become the cornerstone of top-performing systems, which can be broadly categorized into two variants based on their approach to the WSD problem.

The first variant comprises purely transformer-based architectures that leverage the representational power of large encoder models. These models often frame the problem by jointly encoding all candidate definitions alongside the given sentence to extract the correct sense of a target word (Scarlini et al., 2020; Huang et al., 2019; Hadiwinoto et al., 2019). Notable examples include ConSec (Barba et al., 2021b), the previous state-of-the-art, which encodes not only the candidate senses of the target word but also non-ambiguous or already disambiguated words in the sentence; BEM (Blevins and Zettlemoyer, 2020), which separates the encoding of glosses and context to compute similarity through a dot product; and ESC (Barba et al., 2021a), which redefines the WSD problem as a span extraction task, analyzing the concatenation of all possible senses to determine the start and end indices of the correct sense.

The second variant includes transformer models that integrate external information, usually from a lexical knowledge base or semantic network (Loureiro and Jorge, 2019; Conia and Navigli, 2021; Song et al., 2021). Notable examples include EWISE (Kumar et al., 2019) and its improved version EWISER (Bevilacqua and Navigli, 2020), which incorporate WordNet information into the neural model; DHFM (Liu and Zeng, 2024), which enriches pretrained embeddings with graph encodings of senses; and Mizuki and Okazaki (2023), which use synonyms and hypernyms from WordNet to train an encoder via a triplet loss over semantically related glosses. Recent approaches have started exploring parallel alternatives, such as Dong and Sifa (2024), which propose using neurosymbolic embeddings that reach 90% F1 for target senses with an explicit class structure (about 70% of the Raganato et al. (2017a) dataset), and Zhang et al. (2023), which use a representation based on superposition states to eliminate dependency on training set size and improve accuracy for rare senses.

Although both variants produce competitive results, to the best of our knowledge, no system has surpassed the 82% F1 score on the unified benchmark (Raganato et al., 2017a). Additionally, Maru et al. (2022) highlighted that generalization to out-of-domain tasks remains a challenge, complicating the ability of current solutions to scale to specialized domains. Moreover, most modern systems rely heavily on encoder models predominantly trained in English, limiting their applicability to under-represented languages (Barba et al., 2021b). To address these challenges, we propose reframing the WSD problem as a semantic cluster discrimination task within a semantic network (BabelNet) and, in the next section, introduce the SANDWiCH framework as a comprehensive solution to the multilingual word sense disambiguation problem.

3 The SANDWiCH framework

3.1 Theoretical motivation

Formally, a written language $\mathcal{L}$, can be defined by the generator $\mathcal{L} :=< \mathcal{V}, \bigoplus >$, where $\mathcal{V}$ is a vocabulary and $\bigoplus$ is the word concatenation operation. Using this notation we denote the dictionary space, that contains the definitions of every word sense as $\mathcal{D} \subset \mathcal{L}$. Naturally, we can define a function $\eta : \mathcal{V} \to \mathcal{P}(\mathcal{D})$, mapping each word to a set of possible definitions. Additionally, given a sentence $s \in \mathcal{L}$, and a target word $w \in s$, we can define a function φ that selects the correct definition from $\eta(w)$. The disambiguation process can be then formalized as:

$$\begin{aligned}\varphi \circ \eta : \mathcal{V} \times \mathcal{L} &\xrightarrow{\eta} \mathcal{P}(D) \times \mathcal{V} \times \mathcal{L} \xrightarrow{\varphi} \mathcal{D} \\ (w, s) &\mapsto \left(\{d_w^i\}_{i=1}^n, w, s\right) \mapsto d_w^k\end{aligned}$$

Where d_w^i is a definition associated with the word w, n is the total possible definitions associated with w, and d_w^k is the correct definition for w in s. Usually η is provided and the WSD task consists in approximating φ.

The SANDWiCH framework assumes that we are additionally given a graph structure $G := (\mathcal{D}, \mathcal{E})$ over the definitions space $\mathcal{D}$, in which an edge $(d_i, d_j) \in \mathcal{E} \subset \mathcal{D} \times \mathcal{D}$ connecting the definitions $d_i, d_j \in \mathcal{D}$ exists if d_i is semantically related with d_j (e.g. *apple (Fruit.)* $\sim$ *fruit (The ripened reproductive body of a seed plant.))*. Using this notation, we can define the sense neighbourhood of $d_i \in \mathcal{D}$ as $\mathcal{N}(d_i) := \{d_j : (d_i, d_j) \in \mathcal{E}\}$,

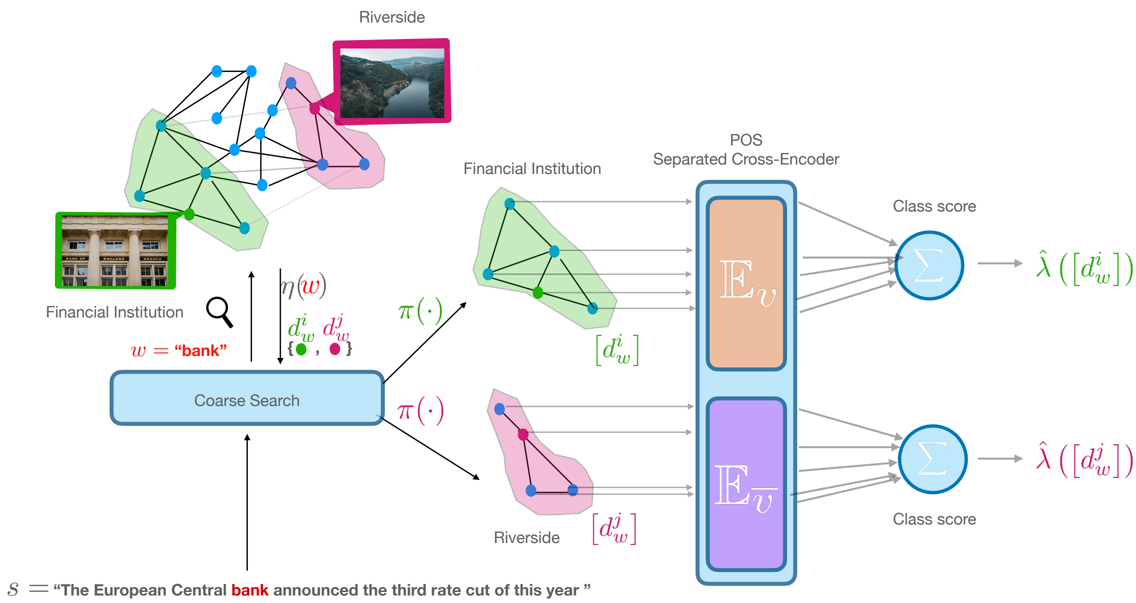

Figure 1: Illustration of the SANDWiCH architecture in the processing of the word *bank* in context.

and the union set of all neighbourhoods of candidate definitions given a word w as $\mathcal{D}_{\eta(w)} := \bigcup_{d_w^i \in \eta(w)} \{\mathcal{N}(d_w^i)\}$.

The graph that SANDWiCH uses is a modified version of the semantic graph to ensure that the graph is *sense-separable*:

Sense-Separability Condition: *Given the semantic graph $G := (\mathcal{D}, \mathcal{E})$, defined as above, we say that the graph is sense-separable if and only if for any given word w we have that*:

$$\forall d_w^i, d_w^j, \in \eta(w) : \mathcal{N}(d_w^i) \cap \mathcal{N}(d_w^j) = \emptyset$$

In practice, its enough with eliminating the edges connecting neighborhoods in the graph G. Given a word w if the sense-separability condition holds, and we consider the subgraph defined by $\mathcal{D}_{\eta(w)}$ and its connecting edges $\mathcal{E}_{\eta(w)}$, the relation $\sim_w$: $d_i \sim_w d_j \iff \exists k : d_i, d_j \in \mathcal{N}(d_w^k \subset \mathcal{D}_{\eta(w)})$ defines an equivalence relation over $\mathcal{D}_{\eta(w)}$ and therefore we have that:

$$\begin{tikzcd}
\mathcal{P}(\mathcal{D}) \times \mathcal{V} \times \mathcal{L} \arrow[rr, "\varphi"] & & \eta(w) \\
\mathcal{D}_{\eta(w)} \times \mathcal{V} \times \mathcal{L} \arrow[u, hook, "\cup"'] \arrow[urr, "\varphi^*"] \arrow[r, "\pi"] & (\mathcal{D}_{\eta(w)} / \sim_w \times \mathcal{V} \times \mathcal{L}) \arrow[ur, "\lambda"'] &
\end{tikzcd}$$

Where φ^* is the natural extension of φ to the $\mathcal{D}_{\eta(w)}$ domain (i.e. $\varphi^*(d_w^i) = d_w^i$ if $\varphi\left((d_w^{(1)}, \ldots, d_w^{(n)}), w, s\right) = d_w^i$, otherwise $\varphi^*(d_w^i) = \emptyset$); π is the canonical projection of the equivalence relation and λ maps an equivalence class $[d_w^i]$ to the correct definition. Since φ is class invariant under $\sim_w$, then $\varphi \cong \lambda \circ \pi$, and there exists a unique λ satisfying this relation (Mac Lane and Birkhoff, 1967). This means that we can approximate the disambiguation process given by φ at the definition level by the equivalence class of all semantically related definitions in the graph $G_{\eta(w)} = (\mathcal{D}_{\eta(w)}, \mathcal{E}_{\eta(w)})$. To this end, we use the following approximation:

$$\hat{\lambda} = \arg \max_{[d_w^i] \in \frac{\mathcal{D}_{\eta(w)}}{\sim_w}} \sum_{d_w^j \in [d_w^i]} \left(\mathbb{E}_v(d_w^j) + \mathbb{E}_{\overline{v}}(d_w^j)\right) \delta_j$$

Where $\mathbb{E}_{\overline{v}}(\cdot)$ is an encoder-based model fine-tuned using data including every POS except verbs to predict the probability of a definition d_w^j being semantically relevant given the word w in a given context. Analogously, we can define $\mathbb{E}_v(\cdot)$ for an encoder-based model using data including nouns and verbs only. Finally, the weight scores δ_j are defined as

$$\delta_j = \frac{e^{2|\mathbb{E}_v(d_w^j) + \mathbb{E}_{\overline{v}}(d_w^j) - 1|}}{\sum_{d_w^j \in [d_w^i]} e^{2|\mathbb{E}_v(d_w^j) + \mathbb{E}_{\overline{v}}(d_w^j) - 1|}}$$

3.2 Implementation Details

From a theoretical perspective, the Word Sense Disambiguation (WSD) task consists of two main components. The first is an information retrieval step, where the goal is to estimate η, the top-K sense candidates associated with a given word. For certain parts of speech, such as adjectives and adverbs, this retrieval may not be necessary due to their limited number of possible senses. However, for nouns and especially verbs, which often have a wide range of senses, this step is critical to narrowing down the candidate definitions and ensuring a manageable input size for the disambiguation process.

In the SANDWiCH framework, this retrieval is managed by a coarse sense retrieval module (see Figure 1), which fine-tunes a DeBERTa-v3-xsmall model as a cross-encoder (Reimers and Gurevych, 2019) to estimate the relevance of a candidate definition given the sentence and target word. We train this model using the SemCor corpus (Miller et al., 1993) combined with datasets from Raganato et al. (2017a), following established methods to classify candidate definitions as relevant or not. The top-K candidate senses are ranked by probability, and we set K=30 as per Barba et al. (2021b), achieving a recall of 98-99% across all datasets.

The second step selects the most appropriate definition from the retrieved candidates. Instead of directly estimating φ, we focus on composing equivalence classes through $\lambda \circ \pi$, reducing reliance on specific word-level data by estimating at the equivalence class level. For this approach to work, the sense-separability condition must hold, meaning the semantic clusters in the graph must be disjoint. We extract the sense graph from BabelNet (Navigli and Ponzetto, 2010) and remove edges connecting senses of the same word to ensure clean separability. The equivalence classes are defined as the immediate neighborhoods of the target word's senses.

Initial experiments revealed a significant performance boost by partitioning the training data into two groups: one for nouns and verbs, and another for nouns, adjectives, and adverbs. Training separate cross-encoders for these groups further enhanced performance, even beyond a standard ensemble of models, as discussed in Section 4.3.

For training, we generate positive and negative examples by sampling from the neighborhoods of correct and incorrect senses. Unlike the coarse retrieval step, all elements within a neighborhood share the same label. The input consists of a concatenated sentence-definition pair (s, d_w^i), where the word w in sentence s is marked with special tokens [d]. We use DeBERTa-v3-small as the backbone model for the cross-encoders, training with a batch size of 64, 10 epochs, a learning rate of $2e^{-5}$, and gradient clipping at 1. A cosine annealing scheduler (Loshchilov and Hutter, 2017) and binary cross-entropy with logits are used as optimization methods.

After training, the class score is computed using the formula outlined in Section 3.1. The δ_{ij} weights are derived from the softmax of the absolute difference between the model's predictions for relevance and non-relevance, which represents its confidence in assigning the correct sense cluster. Additional training details can be found in Appendix A

4 Experimentation

In this section, we present and discuss the results of our experiments to evaluate the SANDWiCH framework against existing alternatives. In Section 4.1, we first assess our model's performance on the English all-words benchmark (Raganato et al., 2017a), breaking down results by individual datasets and parts of speech. Following this, in Section 4.2, we examine how well SANDWiCH generalizes to previously unseen domains and rare senses, using the more challenging dataset from Maru et al. (2022), and compare it to the current state of the art. We then perform an ablation study in Section 4.3 to evaluate the individual contribution of each system component. In Section 4.4, we investigate the framework's adaptability to other languages. Finally, in Section 4.5, we explore alternative backbone models for the cross-encoders and analyze the trade-off between model size and performance.

4.1 All-words English WSD

Introduced in 2017, the all-words English WSD benchmark is the most widely used standard for evaluating WSD systems. It comprises five datasets: Senseval-2002 (**SE2**) (Edmonds and Cotton, 2001), Senseval-2003 (**SE3**) (Snyder and Palmer, 2004), Semeval-2007 (**SE7**) (Pradhan et al., 2007), Semeval-2013 (**SE13**) (Navigli et al., 2013), and Semeval-2015 (**SE15**) (Moro and Navigli, 2015). Following prior work (Raganato et al., 2017b; Huang et al., 2019; Blevins and Zettlemoyer, 2020; Barba et al., 2021b), we use Semeval-2007 as the development set and train on the SemCor corpus. Our results, reported by individual dataset and POS, are summarized in Table 1.

The SANDWiCH framework significantly outperforms previous state-of-the-art methods across all datasets and parts of speech, improving the overall F1 score by seven points. Notably, noun disambiguation sees an eight-point increase, highlighting the effectiveness of the equivalence class approximation for the WSD task.

4.2 A More Challenging Dataset

In this experiment, we reproduce and evaluate the performance of the SANDWiCH framework on the rare senses benchmark introduced by Maru et al. (2022). This benchmark consists of four parts: a dataset designed to test WSD systems on rare and out-of-domain senses (**42D**); a collection of the most frequent errors made by state-of-the-art models on the all-words English WSD benchmark (**hardEN**); a WSD task similar in nature to those found in the all-words English WSD benchmark (**S10**); and (**softEN**), which is the opposite of the hardEN dataset.

We present the results of previous reported systems alongside the ConSeC model, which represented the state of the art in the all-words English WSD benchmark, in Table 2. Notably, SANDWiCH significantly outperforms all models, achieving improvements of over 10 F1 points in S10, 22 F1 points in 42D, 45 F1 points in hardEN, and two F1 points in softEN.

4.3 Ablation Study

In this section, we assess the individual contributions of each component within the SANDWiCH framework to better understand their interrelations. To do this, we first explore the practical and theoretical contributions that enable the architecture to function effectively. As introduced in Section 1, the three main pillars supporting the SANDWiCH framework are: the use of equivalence classes instead of single senses, the sense-separability condition in the semantic graph, and the part-of-speech (POS) separation of the cross-encoders for computing class scores.

In this ablation study, we first analyze the effect of using equivalence classes instead of senses directly, observing the expected decrease in performance. This decline occurs because the system loses robustness against the frequency bias in the training data (Maru et al., 2022; Navigli et al., 2023), making it overly dependent on the training distribution and limiting its ability to generalize beyond the training domain.

If we maintain the use of classes but cannot ensure the semantic graph is sense-separable, we introduce noise into the training set, particularly with word-definition pairs labeled both positively and negatively for the same sentence.

Finally, we find that employing POS separation in the cross-encoders leads to a considerable performance increase compared to using a standard ensemble of two cross-encoders trained on the entire dataset. This gain may stem from the differing disambiguation strategies for each POS: verbs typically rely on objects, subjects, actions, and tense information (Hashimoto and Tsuruoka, 2015; Wagner, 2009), while nouns, adjectives, and adverbs focus more on their interrelations (Rosso et al., 2005). All results are reported in Table 3 in which we evaluate the performance in the all-words English dataset in the aforementioned cases.

4.4 Multilingual WSD

In this setting, we explore the adaptability of the SANDWiCH pipeline to other languages. Specifically, we assess the performance of our solution across nine languages in the XL-WSD dataset (Pasini et al., 2021). Since the DeBERTa-v3 model is trained exclusively in English, we use mBART-50 (Liu et al., 2020) as the backbone model. For each language, we also adapt the BabelNet semantic network to ensure it meets the sense-separability assumption.

In this context, SANDWiCH outperforms the current state of the art in every tested language. The improvements are consistent across all language groups, with gains exceeding nine F1 points in Germanic languages (English, German, and Dutch), eight F1 points in Romance languages (Spanish, Italian, and French), 18 F1 points in Finno-Ugric languages (Estonian), and 25 F1 points in Japonic languages (Japanese).

Additionally, we evaluated our system in Croatian, representing a low-resource language in the Slavic language group, achieving a competitive performance of 84.1 F1 points. For the other low-resource languages (Estonian, Dutch, and Japanese) in our tests, SANDWiCH's reduced dependency on individual senses resulted in the most significant improvements, surpassing the next best approach by over 20 F1 points.

4.5 Backbone model efficiency analysis

As mentioned in Section 3.2, we use DeBERTa-v3-small as the backbone system in the SANDWiCH architecture. To evaluate the trade-off between parameter count and performance, we modify the backbone model and analyze its im-

Model	SE07	SE2	SE3	SE13	SE15	Nouns	Verbs	Adj.	Adv.	ALL
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 - 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
QR-WSD	74.5	80.6	79.1	80.0	84.7	83.7	71.4	82.8	86.7	80.5
GPT4o	-	76.3	73.2	79.7	83.7	81.2	66.3	79.0	71.3	77.4
GPT4	-	74.3	70.0	77.4	79.5	78.6	59.7	79.5	74.0	74.6
GPT-3.5	-	63.1	59.2	63.8	70.5	68.1	46.7	66.6	64.8	63.3
SANDWiCH	**81.2**	**88.5**	**84.9**	**92.5**	**91.7**	**94.0**	**74.6**	**86.8**	**91.6**	**89.0**

Table 1: Performance (F1 score) of various models, broken down by task and POS in the all-words English WSD benchmark. The best results are highlighted in bold. The compared systems include MFS, which selects the most common sense from SemCor, BERT base (Devlin et al., 2019), SVC (Vial et al., 2019), GlossBERT (Huang et al., 2019), ARES (Scarlini et al., 2020), EWISER (Bevilacqua and Navigli, 2020), WMLC (Conia and Navigli, 2021), BEM (Blevins and Zettlemoyer, 2020), ESCHER (Barba et al., 2021a), ConSeC (Barba et al., 2021b), QR-WSD (Zhang et al., 2023), GPT4o, GPT4, and GPT-3.5. are the most recent versions of the ChatGPT model available at the time of writing.

#Dataset	ARES	BEM	ESC	EWS	GEN	GBT	SYN	CSC	SandWiCH
S10	77.9	77.1	78.0	76.1	72.3	75.8	64.0	77.5	**87.5**
42D	41.8	53.2	58.9	43.9	50.2	45.7	32.8	56.6	**77.1**
softEN	78.7	80.3	83.7	79.2	76.4	77.1	63.4	87.7	**89.4**
hardEN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.35	**53.4**

Table 2: F1 performance metrics on the Maru et al. (2022) benchmark. The compared models are ARES (Scarlini et al., 2020), BEM (Blevins and Zettlemoyer, 2020), ESC (Barba et al., 2021a), EWS (Bevilacqua and Navigli, 2020), GEN (Bevilacqua et al., 2020), GBT (Huang et al., 2019), SYN (Scozzafava et al., 2020), and CSC(Barba et al., 2021b). Best scores are highlighted in bold.

Active Components	Score ALL
Classess	53.5
Classes + Encoders	66.6
Classes + Separability	79.5
Encoders + Separability	57.5
Classes + Encoders (No sep.)	55.1
Whole Pipeline	89.0

Table 3: Ablation study on different components of the SANDWiCH pipeline, *Classes* denotes using the equivalence class structure instead of senses directly, *Encoders* refers to the splitting of the cross-encoders by POS as described in Section 3.1, *Separability* is whether the separability condition holds or not, and *Encoders (No sep.)* is an ensemble of cross-encoder non-separted by POS. The F1 ALL score refers to the score in the all-words English dataset.

pact on the all-words English WSD dataset. In Table 5, we compare several models: BERT-base, BERT-large, BART-large (Lewis et al., 2020), RoBERTa-base, RoBERTa-large, DeBERTa-xsmall, DeBERTa-small, DeBERTa-base, and DeBERTa-large. Our results indicate that the DeBERTa family offers the highest overall performance, with performance gains diminishing as model size increases. For instance, the leap from DeBERTa-xsmall to DeBERTa-large (with a 1300% increase in parameter count) yields substantial improvements (4.5 F1 points), but moving from DeBERTa-small to DeBERTa-large results in only a 0.1 F1 point gain. This suggests that DeBERTa-small provides the optimal balance between parameter count and performance, outperforming the

Language	SyntagRank	EWISER	XLMR	ConSeC	SANDWiCH
English	70.0	73.3	76.3	79.0	**88.9**
Dutch	56.0	57.5	59.2	63.3	**83.7**
Estonian	56.3	66.0	66.1	69.8	**89.5**
French	70.0	80.9	83.9	84.4	**92.8**
German	76.0	80.9	83.1	84.2	**93.2**
Italian	69.6	74.6	77.6	79.3	**86.6**
Japanese	57.5	55.8	61.9	63.0	**85.7**
Spanish	68.6	71.9	75.9	77.4	**84.0**
Croatian	-	-	-	-	**84.1**

Table 4: Comparison of F1 scores across different languages in the XL-WSD (Pasini et al., 2021) for SyntagRank (Scozzafava et al., 2020), EWISER (Bevilacqua and Navigli, 2020), XLMR (Pasini et al., 2021), ConSeC (Barba et al., 2021b), and SANDWiCH.

Model	N° Params.	F1 Score
DeBERTa v3 xsmall	22M	84.6
DeBERTa v3 small	44M	89.0
DeBERTa v3 base	86M	89.0
DeBERTa v3 large	304M	89.1
BERT base	110M	78.7
BERT large	340M	80.2
BART large	406M	83.3
RoBERTa base	125M	80.5
RoBERTa large	355M	81.9

Table 5: Performance of the SANDWiCH framework in the all-words English dataset (Raganato et al., 2017a) changing the backbone model.

ConSec model by six F1 points while using just 28% of its parameters.

5 Results Analysis

The performance of the proposed SANDWiCH framework across existing datasets demonstrates that reframing the WSD problem as a discrimination task over semantically related clusters effectively addresses the limitations of current solutions, confirming our initial hypothesis. Specifically, in the all-words English benchmark, we surpass the previous state of the art across each dataset and in the combined total (ALL). This improvement extends to rare senses and out-of-domain data, as shown by results on the Maru et al. (2022) dataset, where SANDWiCH significantly outperforms prior solutions. This success includes cases where words have a large number of possible senses (see Appendix B), indicating that our approach mitigates challenges in such scenarios.

Additionally, we analyze the individual contributions of each component in the architecture, concluding that the key to SANDWiCH's success is the creation of separable clusters over the semantic network. Furthermore, separating cross-encoders by POS leads to considerable performance gains. We also extend the framework to multiple languages, outperforming all existing alternatives and making significant strides in low-resource language disambiguation. Notably, we demonstrate that SANDWiCH achieves these results with only 28% of the parameters used by the previous state-of-the-art, proving the robustness of the sense-cluster approach.

6 Conclusion

In this paper, we introduced the SANDWiCH framework, a novel approach to the WSD problem, which arose from the hypothesis that reframing the disambiguation task as sense cluster discrimination over a semantic network could address the challenges faced by previous state-of-the-art solutions when generalizing to low-resource languages and domains. Through extensive experimentation, we confirmed our hypothesis, surpassing the state of the art across all benchmarks, including rare senses and multiple languages. Furthermore, we evaluated various alternatives for the backbone model and demonstrated the efficiency of our architecture, achieving a 72% reduction in model size while still surpassing the state of the art.

In future work, it would be valuable to explore alternative methods for creating sense clusters, extend our approach to additional languages, and investigate whether SANDWiCH's disambiguation capabilities can serve as a baseline or be combined with existing solutions for translation into low-resource languages and specialized text analysis.

Limitations

As detailed in Section 3.2, the implementation of the SANDWiCH framework requires of a previously given semantic network. However depending on the language, this might be a complicated resource to get or not as complete as needed for ensuring a reasonable accuracy (e.g. for low-resource languages). Our architecture also depends on the performance of the cross-encoders used to calculate the score of the equivalence classes, even if we manage to greatly improve the performance for some underrepresented languages, the backbone models used are not available for every language and that can limit the usability of our proposed solution.

Acknowledgments

We sincerely thank the anonymous reviewers for their thorough reviewing and valuable suggestions. The research of Guzman-Olivares was conducted with the support of Bulil Technologies S.L., that provided the hardware for the development of this project, and the financial support of the Spanish Ministry of Science and Innovation, grant PID2022-139131NB-I00. The research of Quijano-Sanchez was conducted with financial support from the Spanish Ministry of Science and Innovation, grants PID2022-139131NB-I00 & PID2021-122677NB-I00.

References

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Yehoshua Bar-Hillel. 1960. The present status of automatic translation of languages. *Adv. Comput.*, 1:91–163.

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, and Roberto Navigli. 2021b. 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.

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, 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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

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.

Jim Cowie, Joe Guthrie, and Louise Guthrie. 1992. Lexical disambiguation using simulated annealing. In *COLING 1992 Volume 1: The 14th International Conference on 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.

Tiansi Dong and Rafet Sifa. 2024. Word sense disambiguation as a game of neurosymbolic darts. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024*, pages 22–32, Torino, Italia. ELRA and ICCL.

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.

Julian Martin Eisenschlos, Jeremy R. Cole, Fangyu Liu, and William W. Cohen. 2023. WinoDict: Probing language models for in-context word acquisition. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102, Dubrovnik, Croatia. Association for Computational Linguistics.

William A. Gale, Kenneth Ward Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

Google. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

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.

Kazuma Hashimoto and Yoshimasa Tsuruoka. 2015. Learning embeddings for transitive verb disambiguation by implicit tensor factorization. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 1–11, Beijing, China. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

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.

Amita Jain and D. K. Lobiyal. 2015. Fuzzy hindi wordnet and word sense disambiguation using fuzzy graph connectivity measures. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 15(2).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mikael Kågebäck and Hans Salomonsson. 2016. Word sense disambiguation using a bidirectional lstm. In *CogALex@COLING*.

Adam Kilgarriff. 2004. How dominant is the commonest sense of a word? volume 3206, pages 103–112.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

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.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, page 24–26, New York, NY, USA. Association for Computing Machinery.

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.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *Preprint*, arXiv:2304.03439.

Jiaheng Liu and Haonan Zeng. 2024. Achieving fine-grained word sense disambiguation with context hypergraph and sememe hypergraph. In *Proceedings of the 4th International Conference on Artificial Intelligence and Computer Engineering*, ICAICE '23, page 383–388, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. 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. *Preprint*, arXiv:1907.11692.

Zhu Liu and Ying Liu. 2023. Ambiguity meets uncertainty: Investigating uncertainty estimation for word sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3963–3977, Toronto, Canada. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts. *Preprint*, arXiv:1608.03983.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018. Incorporating glosses into neural word sense disambiguation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2473–2482, Melbourne, Australia. Association for Computational Linguistics.

Saunders Mac Lane and Garrett Birkhoff. 1967. *Algebra*. Macmillan, New York. MR:0214415. Zbl:0153.32401.

Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the hard core of Word Sense Disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737, Dublin, Ireland. Association for Computational Linguistics.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics*, 42(2):245–275.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

George Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. pages 303–308.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *Preprint*, arXiv:2402.06196.

Sakae Mizuki and Naoaki Okazaki. 2023. Semantic specialization for knowledge-based word sense disambiguation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3457–3470, Dubrovnik, Croatia. 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.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

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 Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

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.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems*, volume 36, pages 79155–79172. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Muhammad Qorib, Geonsik Moon, and Hwee Tou Ng. 2024. Are decoder-only language models better than encoder-only language models in understanding word meaning? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16339–16347, Bangkok, Thailand and virtual meeting. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Paolo Rosso, Manuel Montes-y Gómez, Davide Buscaldi, Aarón Pancardo-Rodríguez, and Luis Villaseñor Pineda. 2005. Two web-based approaches for noun sense disambiguation. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'05, page 267–279, Berlin, Heidelberg. Springer-Verlag.

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.

Yang Song, Xin Cai Ong, Hwee Tou Ng, and Qian Lin. 2021. Improved word sense disambiguation with enhanced sense representations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4311–4320, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deshan Sumanathilaka, Nicholas Micallef, and Julian Hough. 2024. Assessing gpt's potential for word sense disambiguation: A quantitative evaluation on prompt engineering techniques. In *2024 IEEE 15th Control and System Graduate Research Colloquium (ICSGRC)*, pages 204–209.

Kaveh Taghipour and Hwee Tou Ng. 2015. Semi-supervised word sense disambiguation using word embeddings in general and specific domains. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–323, Denver, Colorado. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

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*, volume 30. 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.

Wiebke Wagner. 2009. Verb sense disambiguation using a predicate-argument-clustering model.

Jing Wang, Mohit Bansal, Kevin Gimpel, Brian D. Ziebart, and Clement T. Yu. 2015. A sense-topic model for word sense induction with unsupervised data enrichment. *Transactions of the Association for Computational Linguistics*, 3:59–71.

Warren Weaver. 1949/1955. Translation. In William N. Locke and A. Donald Boothe, editors, *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA. Reprinted from a memorandum written by Weaver in 1949.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *International Conference on Computational Linguistics*.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Annual Meeting of the Association for Computational Linguistics*.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1374–1385, Osaka, Japan. The COLING 2016 Organizing Committee.

Junwei Zhang, Ruifang He, and Fengyu Guo. 2023. Quantum-inspired representation for long-tail senses of word sense disambiguation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

A Additional Implementation details

A.1 Hardware Specifications

All experiments were performed in a machine with the technical capabilities reported in Appendix A.1.

CPU	AMD Ryzen Threadripper 3975WX
RAM	256 GB
Cores	64
GPU	2x Nvidia A100 160GB

Table 6: Specifications of the machine in which the experiments were executed.

A.2 Training Hyperparameters

The full table of hyperparameters used in the training of the system can be found in Table 7. Different options for the settings of the system appear between curly braces, while the selected ones appear in bold. The only hyperparameter endemic to the SANDWiCH system is the number K of candidate senses returned in the coarse search (see Figure 1).

A.3 Transforming from WordNet Synset to BabelNet Synsets

Our system uses a dump of BabelNet 5.0 as its information source. The graph we employ is a post-processed version that is restricted to a specific language. Given that we work with the version from the Raganato et al. (2017b) dataset, implemented by Pasini et al. (2021), for the all-English WSD task, we had to adapt our comparison across models to accommodate BabelNet. This involves mapping the WordNet-based results of some systems (like ConSec) to BabelNet synsets. Since WordNet differentiates synsets at a finer level, we adjust the predictions from WordNet-based systems by associating all related BabelNet synsets to the predicted WordNet synset. These are then treated as a single unit when compared to the gold standard group. If there is any overlap between the predicted and the gold standard synsets, the prediction is considered correct. To ensure the accuracy of our comparison method, we reproduced all results reported in the ConSec paper, validating the correctness of our

Parameter	Value
Optimizer	AdamW
Learning Rate	$\{10^{-7}, 10^{-6}, \mathbf{10^{-5}}, 10^{-4}\}$
Gradient Accumulation Steps	{1,**5**,10}
Maximum Gradient Norm	{**1**, 5, 10, 50, 100}
Batch Size	{4, 16, 32, **64**, 128}
Epochs	1, 5, **10**, 15, 20
Evaluation Steps	1000
Scheduler	{**Cosine Annealing**, Linear}
Weight Decay	0.01
Maximum Gradient Norm	**1**, 5, 10
Loss Function	Cross-Entropy with logits
Max Tokens	512
K (Top K retrieval SANDWiCH)	5, 10, 15, 20, 25, **30**, 35, 40

Table 7: Training hyperparameters for the proposed system. Between curly braces are all values tested during optimization, the one selected are marked in bold.

mapping methodology and ensuring that no system has an unfair advantage.

B Polysemic Words accuracy comparison

In this section, we compare the performance of ConSec, the previous state-of-the-art model, with the SANDWiCH framework on the all-words English WSD dataset, focusing on polysemic words grouped by their number of possible meanings (see Figure 2). SANDWiCH consistently reduces the error across all polysemic words, with this reduction becoming more pronounced as the number of possible senses increases. This suggests that the clustering approach employed by SANDWiCH is more effective in managing words with multiple senses and is less dependent on the frequency with which a particular sense appears in the training data.

C Licensing and BabelNet derived data

BabelNet is covered under a license that does not permit the usage of the resource or any derived products from it for other purposes than scientific research. For this reason, following the terms stated in BabelNet's license, we explicity prohibit the usage of the derived sense networks or the model trained with them for any usage different than scientific research.

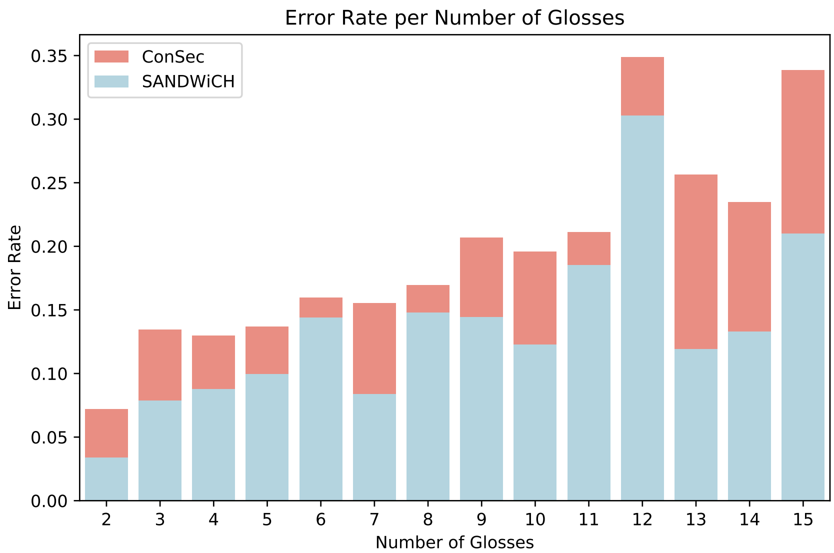

Figure 2: Error rate difference between ConSec (in salmon) and SANDWiCH (in light blue) for words with different number of glosses.