Word Sense Disambiguation Models Emerging Trends: A Comparative Analysis

Nemika Tyagi¹, Dr. Sudeshna Chakraborty¹, Jyotsna¹, Aditya Kumar¹, and Nzanzu Katasohire Romeo¹
¹School of Engineering and Technology, Sharda University, Uttar Pradesh, India
E-mail: nemikatyagi@gmail.com

Abstract. Word Sense Disambiguation (WSD) arises due to the presence of ambiguity in the text during the semantic analysis of natural languages. It is a major unsolved problem in the area of Natural Language Processing (NLP) and its applications. This paper explores and reviews WSD algorithms that have contributed to, or created state-of-art solutions in recent years. Moreover, this paper also aims to analyze the recent technological trends in the domain of WSD which can give us leverage to identify the possible future trajectory of the search for better WSD solutions.

Keywords: Word Sense Disambiguation, Natural Language Processing, Semantic analysis, Computing Methodologies, Context disambiguation, Machine-Readable Dictionaries.

1. Introduction
Words having multiple senses are called polysemous words and most words in natural languages are polysemous in nature [1]. A problem arises when computers attempt to understand the specific sense of these words and try to comprehend their context. For example, a homonym like 'park' can refer to the act of parking a vehicle in the sentence “She struggled to parallel park her new car” whereas it can refer to a large public garden in the sentence "She took her dog out for a walk in the park". Not just computers but humans too come across such a problem in their conversations but they can associate the correct context with ambiguous words with much better accuracies. It is because of the vast knowledge humans have gained through their years of experiences and interactions in the real world.

Programmed machines, on the other hand, do not possess the same cognition, critical thinking, and adaptive intelligence as a human to resolve this issue which is regarded as Word Sense Disambiguation. WSD has remained one of the prolonged unsolved problems in NLP for decades. Various methods such as knowledge-based, supervised, semi-supervised, unsupervised, etc. have been used to solve this problem previously.

1.1. Transformation of computing methodologies
The knowledge-based or supervised approaches like Lesk-based algorithm [1], Statistical Models, and Conceptual Density-based methods among many others relied on the accessibility to a large corpus of annotated text. This data could be in the form of Machine-Readable Dictionary (MRD), thesaurus, or WordNet-based corpora. Supervised algorithms appear to be the most efficient over time but suffer from the lack of expansive annotated corpus. In contrast to these models, we have been moving towards semi-supervised or unsupervised approaches in the past decade. Despite these advances of newer techniques the state-of-art still mostly comprises of supervised algorithms. There has also been a revival of knowledge-based techniques with the advancement of graph-based methods like PageRank [2].
1.2. Motivation to resolve WSD
Solving the complex issue of word sense disambiguation is helpful when dealing with other Natural Language Processing tasks. It is also required in various domains of linguistic researches like Machine Translation from one language to another, Retrieval and Extraction of Information, and Text mining [3]. In this survey, we explore some prominent models that attempt to resolve Word Sense Disambiguation using progressive approaches in more recent times.

2. Evaluation of WSD models
To evaluate the models that attempt to solve WSD, an organized and systematic way was introduced for the first time in 1998 as 'Senseval' conferences. Initially, these conferences were held once every three years beginning with Senseval-1 held in 1998, followed by Senseval-2 in 2001, and then Senseval-3 organized in 2004. These initial conferences were predominantly concentrated on WSD only. In the successive years, Senseval evolved into 'SemEval' which refers to Semantic Evaluation. Besides WSD, SemEval conferences also include other aspects and domains of semantic analysis that are beyond WSD. SemEval is a series of evaluation conferences that have been ongoing for almost two decades now.

Since 2012 the SemEval conferences have been held yearly and the evaluation is done on the various tasks that are designed by the committee for the given year. Models are evaluated for tasks related to specific languages like English, Chinese, French, etc. as well as for multilingual and cross-lingual tasks. A total number of 12 tasks were evaluated upon in the SemEval-2020 workshop ranging from Lexical semantics, Extraction of Knowledge, Common Sense Knowledge, and Reasoning to the Societal Applications of Natural Language Processing.

SemEval workshops and conferences are organized to assess WSD and other semantic analysis models to reach a higher level of precision and create state-of-art solutions. These workshops have also reevaluated their standards over the years to better suit the design and demands of upcoming technological solutions and adapt to the trends to prevent any insight overlook.

3. Approaches to resolve WSD
Despite WSD being a problem that has not yet been completely resolved, tremendous progress has still been made over the years to tackle this issue. The most significant change in the computational methodologies used in recent times as compared to previous decades is the utilization of unsupervised or minimally supervised models besides the highly supervised models. Machine learning and the advancement of Artificial Intelligence-based algorithms have greatly influenced this shift. Discussed below are some of the prominent approaches undertaken by researchers in the course of approximately the entire last decade (2012-2021). This study gives us an overview of the evolution of WSD methodologies undertaken during this period. Table 1. gives a comparative analysis of all these approaches more comprehensively.

3.1. A Minimally-Supervised Framework - 2012 [4]
This model proposes an unsupervised approach for Domain WSD. The creators of this model make use of domain glossaries obtained by doing repetitive bootstrapping instead of using an annotated corpora with slight constrain on the distinctiveness of selected relations. It gives solutions on dual standards of sense granularity: an extremely fine-grained level and a more coarse-grained level. They select the best-suited gloss with the use of Personalized PageRank (PPR) and domain-boosted PPR algorithms. Both give a high performance of 80% and 69% F1-scores respectively.

3.2. WSD Algorithm based on Lesk - 2014 [5]
This is an Unsupervised WSD algorithm that enhances the gloss-context overlapping algorithms from the knowledge-based WSD approach. Babel-Net, which is a combination of Word-Net and Wikipedia is the annotated corpora here. This method makes use of finding the highest similarity rather than simple overlapping by building semantic vectors on the Distributional Semantic Space of expanded gloss and
computing their similarity as the cosine. After adjustments, this approach reaches a performance of over 70% of the F1-score measure.

3.3. Semi-supervised WSD with Neural Models - 2016 [6]
This publication introduces the use of two different neural models based on LSTM, that is, Long Short-Term Memory and the other, a semi-supervised algorithm that makes use of Label Propagation (LP). The first algorithm is trained on some unlabelled text and predicts the context of the word by computing similarity among the words in the corpus that form the context. The other algorithm starts with labeled sentences and augments them to unlabelled sentences and uses cosine similarity to predict contexts from the vertices representing all sentences in a graphical approach. The LSTM approach achieved the highest all-words F1 scores except for tasks in Sem-Eval 2013. The LSTM LP classifier which was coupled with an LSTM language model attained the best scores on nouns and adverbs.

3.4. Neural sequence learning models for WSD - 2017 [7]
This learning model introduces a supervised WSD algorithm that treats the disambiguation of word senses as a sequence learning problem. On the contrary, most models treat WSD as a classification problem for a single word at a time. Three progressive algorithms are employed starting from a Bidirectional LSTM (BLSTM) architecture to an Attentive Bi-directional architecture that has an additional attentive layer, and finally, a Recurrent Neural Network based Sequence-to-Sequence (Seq2Seq) Model that separates the algorithm into encoder and decoder. The results obtained by both BLSTM and Seq2Seq are better or equivalent to best performing supervised models.

3.5. Knowledge-based models for learning sense distribution - 2018 [8]
This research work advocates that to improve supervised WSD models which constitute the state of the art, systems must be able to automatically understand the way word's senses are distributed. This sense distribution at the sentence level is implemented in two separate stages: first is the computation of a Semantic vector and second is the actual word sense distribution at Sentence-level. This distribution is then utilized by two fully automatic methods that are also language-independent. They are Entropy-Based Distribution (EnDi) Learning and Domain-Aware Distribution (DaD) Learning. Both EnDi and DaD outperform several alternative methods.

3.6. WSD model using pre-trained contextualized word representations - 2019 [9]
Here the authors integrate pre-trained contextualized word representations using Bi-directional Encoder Representations from Transformers (BERT) for WSD. BERT is firstly trained on the semantic task called masked word prediction and then exploited for WSD by two techniques: first uses the Nearest Neighbour Matching algorithm and the second one uses Linear Projection of hidden layers. These BERT-based models can be used to perform WSD tasks for a particular language at a time. Both these approaches either perform on par with the best approaches or set state-of-art F1 scores.

3.7. GlossBERT - 2020 [10]
This model focuses on leveraging gloss information using a supervised BERT-based neural WSD system. WSD task is treated as a problem of sentence-pair classification in this method. Firstly, a context-gloss matchup is generated and then GlossBERT is employed in 3 different techniques: GlossBERT (Token-CLS), GlossBERT (Sent-CLS), and GlossBERT (Sent-CLS-WS), where CLS denotes the initial token and WS denotes Weak Supervision. All the models give state-of-art results and GlossBERT (Sent-CLS-WS) especially sets new and higher F1-scores against most SemEval tasks.

3.8. SensEmBERT - 2020 [11]
SensEmBERT is a WSD approach that uses knowledge-based methodology and utilizes sense embedding in multiple languages. It extracts the knowledge from Wikipedia, BabelNet, NASARI (i.e., Novel Approach to Semantically-Aware Representation of Items), and BERT. Mainly it retrieves
context from Wikipedia for a synset and then picks sentences from BERT that suit them best. Finally, a vector is created by concatenating context and sense gloss. A nearest neighbor-based approach and a supervised SensEmBERT approach are used. The latter outperforms all the contemporary models with an increment of 2.1 points on F1-scores.

3.9. *CluBERT - 2020* [12]
It is a multilingual cluster-based methodology that prompts the distribution of word senses into a random input corpus. It does so by exploiting the contextual information obtained from BERT and the knowledge about lexical semantics available in BabelNet. Sense distribution is attained by sentence clustering, cluster disambiguation, and distribution extraction. CluBERT is not only able to scale well on multiple languages but also achieves state-of-art outcomes on both the intrinsic and the extrinsic evaluations.

3.10. *Interpretability in WSD using Tsetlin Machine - 2021* [13]
This paper tries to resolve the drawback of supervised neural models like Deep Neural Networks (DNN) for WSD. DNN models have no interpretability, and this problem is settled by making use of the newly introduced Tsetlin Machine (TM). Sense categorization is done by a TM-based sense classifier by using conjunctive clauses to demonstrate a particular feature of each category. It achieves almost the same result as the state-of-art and is human-interpretable at the same time.

| Reference | Year | Overview | Results | Limitations |
|-----------|------|----------|---------|-------------|
| [9]       | 2012 | Uses Bootstrapping to attain domain glossaries and employs 2 algorithms: PPR and domain-boosted PPR. | Domain PPR is more robust than PRR and achieves 80% and 69% F1 respectively. Applicable on multilingual gloss. | Made use of 30 domains only and can still be extended. |
| [10]      | 2014 | Uses DSS to build vectors of gloss expansion from BabelNet. The highest similar sense is chosen. | Best performance w.r.t other participants at SemEval 2013 Task 12 obtaining over 70% of F1-score measure. | Doesn't work as well for Italian or other non-English languages. |
| [11]      | 2016 | Consists of two approaches, based on: LSTM (Long Short-Term Memory) and LSTM-LP (Label Propagation). | LSTM approach scores highest F1 scores except SemEval 2013. LSTM LP attained highest score on nouns and adverbs tasks. | Assessment of non-English languages and idiomatic phrases have not been performed. |
| [12]      | 2017 | WSD is treated as a Sequence learning model. Makes use of BLSTM and Seq2Seq algorithms. | BLSTM outperforms Seq2Seq on F-scores of 83% to 82%. Both perform better or equivalent with state-of-art models. | Global coherence is an issue. Encoder-Decoder approach may not be optimal. |
| [13]      | 2018 | Automated learning of distribution of word senses through 2 knowledge-based methods: EnDi and DaD | EnDi outperforms some models. DaD surpasses performance of all the multilingual WSD models. | In intrinsic evaluation both EnDi and DaD have shortcomings. |
| [14]      | 2019 | Incorporates BERT for WSD in a novel approach through | The LP model of BERT outperforms all of prior | The NN model lacks in performance |
| References | Year | Methodology | Neural WSD Models | Compared to LP Model |
|------------|------|-------------|-------------------|---------------------|
| [15]       | 2020 | Uses context-gloss pair, trained on BERT using 3 different GlossBert methods: Token-CLS, Sent-CLS, and Sent-CLS-WS. | All the models collectively set new high F1 scores across all SemEval tasks. GlossBert (Sent-CLS-WS) performs best. | GlossBERT (Token-CLS) and GlossBERT (Sent-CLS) do not perform as good as GlossBert (Sent-CLS-WS). |
| [16]       | 2020 | Knowledge-based, sense embedding WSD approach. Uses an NN model and a supervised model for implementation. | Supervised SensEmBERT appears to be the best system among the available models outperforming the existing state-of-art. | NN SensEmBERT ranks 2nd to the LMMS (Language Modelling Makes Sense) approach of WSD. |
| [17]       | 2020 | An automatic cluster-based approach that exploits both BERT and BabelNet. | CluBERT outperforms the Most Frequent Sense standards of WSD tasks in intrinsic, extrinsic, and multilingual tasks. | It falls back on domains that are poorly connected in BabelNet. |
| [18]       | 2021 | Resolves the issue of interpretability in neural models by using TM based WSD classifier. | TM surpasses a lot of models close to state-of-art and achieves high accuracies. | It falls short of the BERT models that comprise the state-of-art. |

**Figure 1.** Work-flow in a basic BERT-based WSD model.

4. **The recent shift in Methodology**

On assessing the various techniques that are used to solve WSD it can be concluded that the current state-of-art WSD models rely a lot on Bidirectional Encoder Representations from Transformers or BERT [14] technology. BERT gathers deep bidirectional representations from a corpus of unannotated and unlabelled text and jointly conditions on both sides’ context in all layers [14]. It pertains to left and
right contexts at the same time. Combining BERT with machine learning algorithms has given more precision to the word sense disambiguation algorithms. A BERT model that is used for pre-training, requires to be fine-tuned according to the desired requirement of an NLP task. Using a neural network or a clustering algorithm gives BERT-based models significantly more precise results. But some of these techniques still lack in terms of limited sense inventory and the unavailability of a theoretically complete corpus for multi-lingual senses.

BabelNet [15] is preferably used as the sense inventory for training the data model as it combines the two most exhaustive lexicographical datasets i.e., Wikipedia to WordNet along with other additional word-sense databases like Wiktionary, Wikidata, VerbAtlas, Geonames, and many more. BabelNet is in the form of a multilingual encyclopedia dictionary and gives synsets for senses just like WordNet but for 500 languages instead. BabelNet is currently the best available database that is used for training NLP applications. The corpus generated after training is utilized by the given lexeme to fetch its specific context and resolve the ambiguity of senses. The basic architecture of these BERT-based methodologies is shown in figure 1.

4.1. The architecture of BERT and its input representation.
BERT’s model implementation was originally explained in Vaswani et al. (2017) [16]. Its architecture is a bidirectional transformer encoder with multiple layers. It has a pre-training phase and a fine-tuning phase in its constituent framework. In its first phase, BERT is trained on unlabelled textual data spanning different pre-training semantic tasks. Then during the finetuning step, the model is initialized with attributes obtained after the first phase. These parameter attributes are lastly fine-tuned with the help of labeled data that we get from the diverse downstream tasks [14].

The input of BERT is represented in such a way that it can exemplify both an individual sentence and a couple of sentences in a single sequence of tokens without any ambiguity. It uses different segment embeddings for different sentences. The labels used for the start, separation, and end of sentences are CLS and SEP. The entire representation of input sentences is a combination of the corresponding token, its segment embedding, and its specific position embedding (ranges from 0 to n, where n is the total number of tokens in the input). For example, a pair of sentences: “It rained. She got wet.”, can be represented as given in figure 2.

![Figure 2. BERT’s Input Representation [14]](image)

5. Conclusion
Word Sense Disambiguation is a problem that has not yet been completely resolved despite decades of research. This work aims to throw light on some of the methods that have been undertaken to resolve the issue of WSD in the recent past. It can be concluded from the review that the efficiency of our
semantic analysis systems tackling WSD has progressively improved during this period and it has subsequently benefited various other applications of NLP. Since the creation of a systematic and architecture-based BERT [14] model, the performance of WSD algorithms has increased tremendously and the current state-of-art supervised machine learning models are established on the same. A better understanding of the principles and utility of BERT helps us acknowledge the present best performing models for resolving WSD and will also help in refining the concepts further to reach even better results.

References
[1] Banerjee et al., 2002, February. An adapted Lesk algorithm for word sense disambiguation using WordNet. In International conference on intelligent text processing and computational linguistics (pp. 136-145). Springer, Berlin, Heidelberg. Another reference
[2] Taher H. Haveliwala., 2002. Topic-sensitive PageRank. In Proceedings of 11th International Conference on World Wide Web, WWW 2002, pages 517–526, Honolulu, Hawaii.
[3] Pal et al., 2015. Word sense disambiguation: A survey. arXiv preprint arXiv:1508.01346.
[4] Faralli et al., 2012, July. A new minimally-supervised framework for domain word sense disambiguation. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 1411-1422).
[5] Basile et al., 2014, August. An enhanced leks word sense disambiguation algorithm through a distributional semantic model. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (pp. 1591-1600).
[6] Yuan et al., 2016. Semi-supervised word sense disambiguation with neural models. arXiv preprint arXiv:1603.07012.
[7] Raganato et al., 2017, September. Neural sequence learning models for word sense disambiguation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1156-1167).
[8] Pasini et al., 2018, April. Two knowledge-based methods for high-performance sense distribution learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).
[9] Hadiwinoto et al., 2019. Improved word sense disambiguation using pre-trained contextualized word representations. arXiv preprint arXiv:1910.00194.
[10] Huang et al., 2020. GlossBERT: BERT for word sense disambiguation with gloss knowledge. arXiv preprint arXiv:1908.07245.
[11] Scarlini et al., 2020, April. SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 8758-8765).
[12] Pasini et al., 2020, July. CluBERT: A cluster-based approach for learning sense distributions in multiple languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4008-4018).
[13] Yadav et al., 2021. Interpretability in Word Sense Disambiguation using Tsetlin Machine. In ICAART (2) (pp. 402-409).
[14] Devlin et al., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[15] Navigli et al., 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193, pp.217-250.

[16] Vaswani et al., 2017. Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).