RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

Final Report : An In-Depth Exploration of Word Sense Disambiguation in Biomedical Text

CS598 - Topics in AI

By - Ritik Rawat
rr1370

Introduction

Word Sense Disambiguation (WSD) is a natural language processing (NLP) task that involves determining the correct meaning or sense of a word in a given context. Many words in natural language have multiple meanings, and the goal of WSD is to identify the most appropriate sense of a word based on the surrounding words or context.

Consider the word "bat." In isolation, it could be a flying mammal or a piece of sports equipment used in baseball. The sentence "I saw a bat in the cave" would likely refer to the animal, while "He hit the ball with the bat" indicates the sports equipment. Without contextual clues, the intended meaning remains elusive.

Similarly, the word "bank" exhibits diverse meanings. In the sentence "I went to the bank," it could denote a financial institution, but in "I sat on the bank of the river," it takes on the meaning of the riverbank. WSD endeavours to unravel these semantic puzzles, ensuring that machines can accurately interpret language nuances akin to human understanding.

There are three different types of ambiguity-

- **Homonymy** – When the word senses are unrelated.
    - Bat can be an equipment, or a mammal.
    - Right can mean direction, entitlement or correctness.

- **Metaphor** - When one word sense is an analogy.
    - Snake can be a reptile or a person.
    - Spark can mean a particle or intensity of a feeling.

- **Systematic Polysemy** – When the words are related in a way that reflects a coherent structure.
    - Oak can be a type of wood, or a tree.
    - Gold can be the metal or the colour.

Overview

The report introduces Word Sense Disambiguation (WSD) and its challenges, emphasising three types of ambiguity. It explores domain-specific challenges in Biomedical Text WSD and traces the evolution of WSD methods. The MSH WSD dataset is introduced, and our approach, using PubMed articles for WSD, is outlined.

Results show promising disambiguation outcomes, with limitations including reliance on a limited list of Multi-Word Expressions (MWEs) and challenges in interpreting MWE-only clusters. The report concludes with recommendations for future work, addressing limitations and suggesting enhancements for a more comprehensive WSD model.

How is WSD for Biomedical Text different from General Text?

Domain Specificity:

- **Technical Terminology**: The biomedical domain is rife with highly specialized terms and acronyms with meanings specific to the field. General WSD techniques might struggle with unfamiliar vocabulary and nuanced distinctions between these terms.

- **Ontologies and Taxonomies**: Biomedical knowledge is often organized in structured ontologies and taxonomies. Biomedical WSD systems often leverage these resources to understand the relationships between concepts and choose the most relevant sense in a specific context.

Data Scarcity and Annotation Costs:

- **Limited Annotated Data**: Training accurate WSD models requires large amounts of annotated data where the correct meaning of each ambiguous word is labeled. Access to such data is limited and expensive in the biomedical field, making it challenging to train robust models.

- **Domain Expertise**: Annotating biomedical data often requires expertise in the specific field, further increasing the cost and resource constraints.

Sentence Structure and Ambiguity:

- **Complex Sentence Structures**: Biomedical texts often feature complex sentence structures, dense information, and nested dependencies. This can make it harder for WSD systems to accurately identify the relevant context for choosing the correct meaning.

- **Domain-Specific Ambiguity**: Certain words or phrases have specific meanings in the biomedical context that differ from their general usage. For example, "tumor" might refer to a neoplasm in general WSD, but in a clinical context, it might specifically refer to a malignant neoplasm.

Previous Work

Early Forays (1990s - 2000s):

- **Knowledge-based Methods**: These pioneering efforts leveraged dictionaries, thesauruses, and other lexical resources to identify the contextually relevant sense of a word. Systems like SENSEVAL and MEDSCAPE paved the way for exploiting structured knowledge within the biomedical domain.

- **Supervised Learning Approaches**: Early supervised models relied on manually annotated corpora to train algorithms for WSD. However, the scarcity of such annotated data in the biomedical field posed a significant obstacle.

- **Statistical Techniques**: N-grams, Hidden Markov Models (HMMs), and Conditional Random Fields (CRFs) were employed to capture statistical patterns and dependencies within biomedical text, offering preliminary success in specific contexts.

The Rise of Machine Learning (2010s):

- **Word Embeddings**: The advent of word embeddings like Word2Vec and GloVe reshaped the WSD landscape. These representations captured semantic relationships between words, enhancing the ability to disambiguate based on contextual similarities.

- **Neural Networks**: Deep learning models, particularly Recurrent Neural Networks (RNNs) and their variants, revolutionised WSD. By capturing long-range dependencies and learning complex representations of biomedical text, these models achieved significant performance improvements.

- **Attention Mechanisms**: Attention mechanisms within neural networks allowed focusing on specific parts of the sentence that provide the most relevant clues for WSD. This proved particularly effective in handling complex biomedical sentences with nested and intricate concepts.

Contemporary Landscape (2020s):

- **Hybrid Approaches**: Combining knowledge-based methods with neural networks has emerged as a promising direction. Leveraging domain-specific knowledge sources like ontologies and biomedical databases in conjunction with the learning power of deep learning models facilitates context-aware disambiguation.

- **Transfer Learning**: Pre-trained language models like BERT and BioBERT, originally trained on massive general corpora, are fine-tuned for specific biomedical tasks, including WSD. This transfer of knowledge from pre-trained models further bolsters the accuracy and efficiency of disambiguation methods.

- **Domain-Specific Resources**: Development of specialised corpora and lexical resources tailored to the biomedical domain, such as UMLS and MeSH, provides valuable reference points for WSD models and facilitates domain-specific understanding.

Problems

- Training accurate WSD models requires large amounts of annotated data where the correct meaning of each ambiguous word is labelled. In the biomedical field, access to such data is limited and expensive. This is because annotating biomedical data often requires expertise in the specific field, increasing the cost and resource constraints.

- Unsupervised approaches do not require labelled training examples and often make use of knowledge bases, such as the UMLS Metathesaurus. But the UMLS was not created for this purpose. It has issues with how word senses are distinguished and how they are ordered.

- Biomedical knowledge evolves rapidly with ongoing research. New terms emerge, and the meanings of existing terms can shift. Adapting WSD systems to stay current with these changes is a persistent challenge.

- Biomedical literature is often published in multiple languages. Translating and disambiguating biomedical terms across languages introduce additional complexities, including language-specific nuances and terminology variations. Most of the previous work in the field of biomedical WSD has been focused on only English text.

The MSH WSD Dataset

The MSH Word Sense Disambiguation (WSD) dataset is a collection of ambiguous biomedical terms and their corresponding senses in the Medical Subject Headings (MeSH) vocabulary. It was created to provide a benchmark for evaluating WSD algorithms in the biomedical domain. The dataset consists of 203 ambiguous terms, including 106 abbreviations, 88 full terms, and 9 combinations of both. It is the most popular dataset for training and evaluation of disambiguation models in the biomedical domain.

But the dataset is plagued with numerous problems, including-

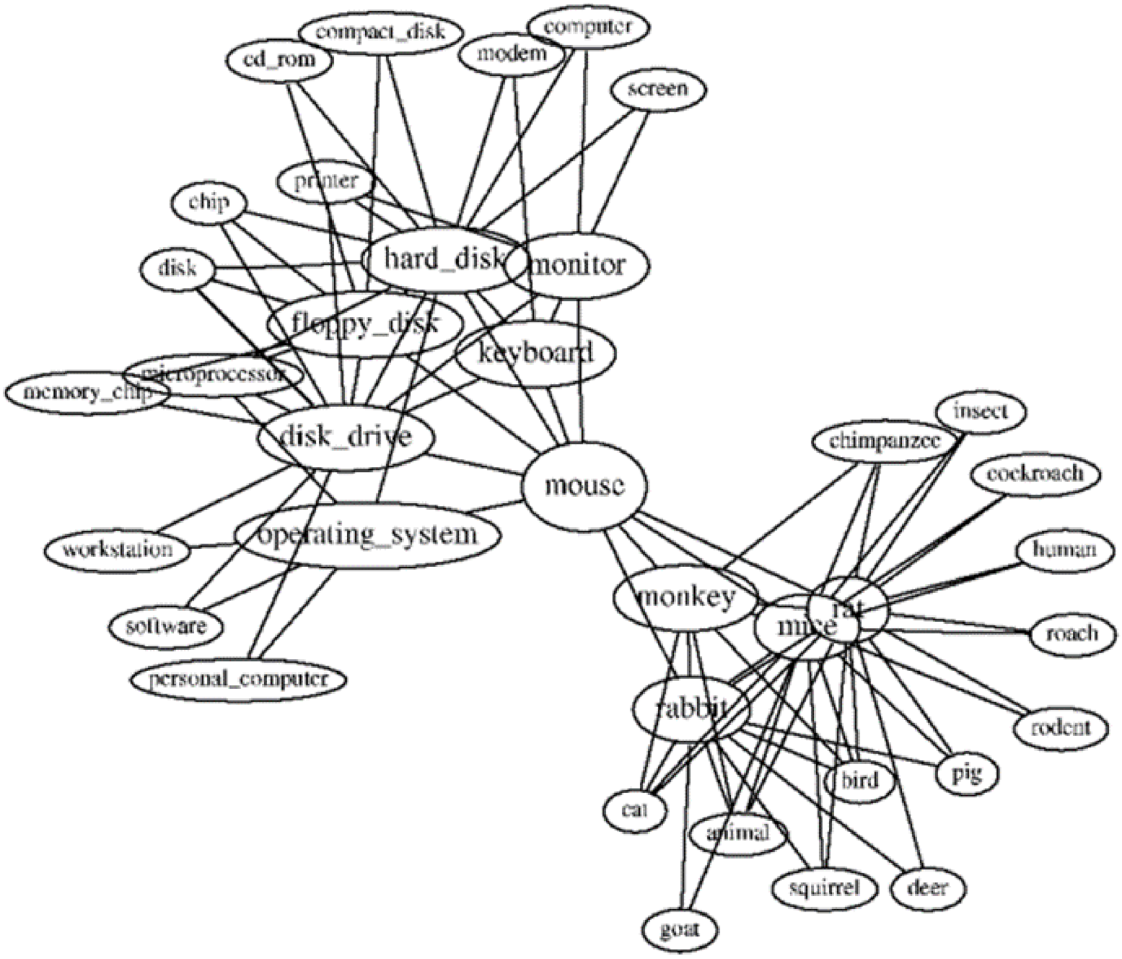
- **Small Size** - The dataset contains only 203 ambiguous terms.
- **Abbreviations** – Nearly half the dataset is just abbreviations.
- **Class Imbalance** – Certain senses have significantly more examples than other senses
- **UMLS** – As it was generated automatically using UMLS, the dataset includes ‘weird’ ambiguity.

Some examples of the aforementioned problems -

- **Diseases** – Cholera, Malaria etc. are also marked as ambiguous terms. One meaning as “Disease” and the other as “Vaccine”
- **Plurals** – Follicle and Follicles both appear in the dataset, with the same senses.
- **Weird** – Veterinary has the senses “Assistants to a veterinarian, biological or biomedical researcher, or other scientist who are engaged in the care and management of animals” and “The medical science concerned with the prevention, diagnosis, and treatment of diseases in animals”.

My Approach

This approach harnesses the information embedded in approximately 16 million PubMed articles by extracting and analysing titles and abstracts to construct a comprehensive co-occurrence network of terms. Employing advanced clustering methodologies (Markov Clustering), we partition the network into distinct clusters, each indicative of a particular semantic sense.



Here, you can clearly see the two different senses of the word “mouse” being represented in different highly connected components of the network.

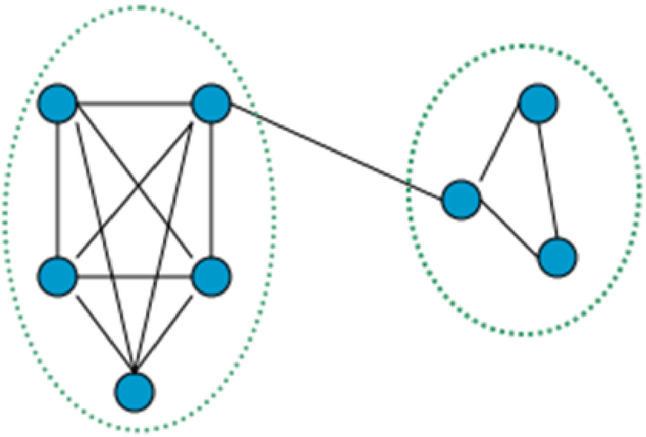
Along with the singletons, we plan to use “Multi Word Expressions” to create the co-occurrence network. The assumption being that MWEs are less ambiguous than singletons and thus, will be good discriminator for disambiguation.

Initially, we planned to create three different types of cluster -

- **Singleton Clusters** – The network is created using only singletons occurring in articles.
- **Mixed Clusters** – The network is created using both MWEs and singletons.
- **Hybrid Clusters** - The network is created using MWEs, singletons are added to the clusters.

The clustering algorithm of choice is “Markov Clustering”. Here is a brief explanation -

The algorithm simulates a process of flow and diffusion within the graph by iteratively inflating and normalising a stochastic matrix, typically representing the adjacency matrix of the graph. The process is akin to random walks, where the matrix operations simulate the probability of transitioning between nodes. Over iterations, densely connected regions of the graph tend to accumulate higher probabilities, while sparser connections diminish.



Problems

The aforementioned approach had some unforeseen issues -

- The PubMed articles provided a list of ~2 billion tokens (singletons). It took a week to complete just the tokenization process using SciSpacy
- Working with this amount of data requires tremendous resources. It was impossible to create a co-occurrence network with 2 billion nodes.
- Markov Clustering doesn’t create overlapping clusters.

To overcome these issued, we changed our approach -

- We Only perform the “Hybrid Clustering” with a subset of 100,000 best MWEs (Based on MRR).
- Use 3 million articles instead of 16 million.
- We made an assumption that MWEs tend to be non-ambiguous, so they will only occur in a single cluster only. Therefore, we can still use Markov Clustering.
- Instead of adding all singleton tokens (~2billion) to the clusters, add the terms with known ambiguity.
    - 96 MSH WSD terms (not including abbreviations)
    - 50 NLM WSD terms
    - List of ~93 ambiguous terms provided by Dr Krovetz.
- We also test the approach with 50,000 most frequent terms.

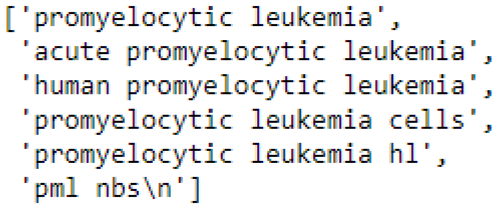
Now the question arises “How are we gonna add these terms to the clusters?”
For each ambiguous term –

- Calculate if it occurs with a MWE more than twice.
- For a cluster, the term is part of that cluster if the term is co - occurring with -
    - A majority of the MWEs in the cluster. (Approach A)
    - More than one MWE in the cluster. (Approach B)
        - The results are not worthwhile.

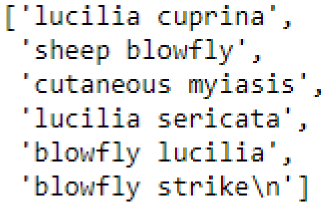
If a term is appearing in more than one cluster, we assume that we are capturing different semantic senses of the term.

Results

- With the 100,000 MWEs, 2648 clusters are created.
  - Used an Inflation factor of 3.5.
  - Average length of a cluster is 37.029 terms per cluster.
  - Skewed by the few largest clusters.
  - The 5 largest clusters created have 9647, 8347, 5985, 3527, 2072 MWEs respectively.
  - The average size without these clusters is 25.9 terms per cluster.
  - There are 24 clusters with a single MWE.



```
['promyelocytic leukemia',
 'acute promyelocytic leukemia',
 'human promyelocytic leukemia',
 'promyelocytic leukemia cells',
 'promyelocytic leukemia hl',
 'pml nbs\n']
```



```
['lucilia cuprina',
 'sheep blowfly',
 'cutaneous myiasis',
 'lucilia sericata',
 'blowfly lucilia',
 'blowfly strike\n']
```

- Total terms to disambiguate – 217
  - 50 from NLM WSD dataset
  - 93 provided by Dr Krovetz with known ambiguity type.
  - 96 from MSH WSD dataset.
  - [' cold ', ' ganglion ', ' plaque ', ' radiation ','pressure'] were common between the datasets

- Out of these 217 terms, 35 either didn't appear in the corpus more than a 100 times or didn't co-occurrence with MWEs more than twice.

- The ambiguity captured (term appearing in more than one cluster) is 85.56% (The aforementioned 35 terms were not taken into consideration)
  - For the NLM WSD dataset we captured 86.36.% ambiguity.
    - 88 out of the 96 terms were taken into account.
  - For the MSH WSD dataset we captured 100% ambiguity.
    - 49 out of the 50 terms were taken into account.
  - For the professor's list we captured 73.33% ambiguity.
    - 60 out of the 93 terms were taken into account.

- Total terms provided by Dr Krovetz with known ambiguity type – 93
    - Only 60 of these terms appeared in the corpus more than a hundred times or co-occurred with any MWE more than twice.

- The total captured ambiguity is 73.33%.
    - For homonymy, we captured 77.77% of the ambiguity.
        - 18 out of 25 terms were taken into account.

    - For metaphors we captured 62.5% of the ambiguity.
        - 24 out of 43 terms were taken into account.

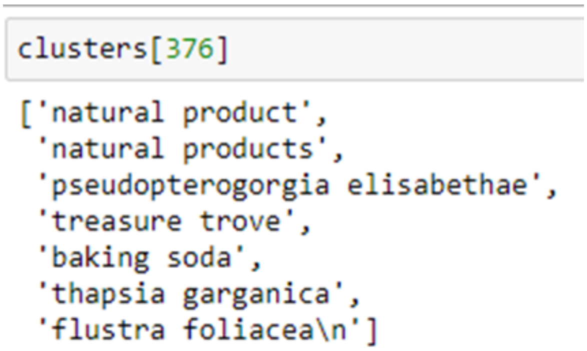
    - For systematic polysemy, we captured 83.33% of the ambiguity.
        - 18 out of 25 terms were taken into account.We repeat this experiment with 50,000 most frequent terms in our dataset.

- 49674 out of the 50,000 terms co-occur with any of the MWE more than twice.
    - Out of which, 51.62% terms occur in more than one cluster.

Limitations

- **Scope Limitation with MWE List:** Our reliance on a specific list of Multi-Word Expressions (MWEs) may limit the system's robustness. Future work should explore methods to expand this list or develop adaptive mechanisms to include additional relevant expressions, ensuring a more comprehensive coverage of biomedical terms.

- **Ambiguous Clusters:** MWE clusters without singletons, can sometimes not make a lot of sense.



```
clusters[376]
```

```
['natural product',
 'natural products',
 'pseudopterogorgia elisabethae',
 'treasure trove',
 'baking soda',
 'thapsia garganica',
 'flustra foliacea\n']
```

- **Dependency on Human-Annotated Terms:** The current evaluation of clusters relies on human-annotated terms. Future work should aim to develop automated evaluation mechanisms, reducing the dependency on manual annotations and enabling more scalable assessments of disambiguation performance.

Future Work

- **Integration of Singletons and MWEs:** Currently, our approach relies on a list of Multi-Word Expressions (MWEs). Future work should explore the integration of both singletons and MWEs, creating a more comprehensive and robust disambiguation model. This expanded scope could enhance the system's ability to handle a broader range of terms and contexts.

- **Ambiguity Type Classification:** The current approach focuses on capturing ambiguity without classifying the type (homonymy, metaphor, systematic polysemy). In future we can classify the terms in the NLM WSD and MSH WSD datasets to have a better understanding of our model’s capability.

- **Scaling Up Computational Resources:** Given the resource-intensive nature of working with large datasets, future work should explore the use of more powerful computational resources. Scaling up the system's capacity can potentially lead to the inclusion of a larger set of MWEs, improving the overall performance and coverage of the disambiguation model.

- **Dynamic Adaptation to Biomedical Knowledge Evolution:** Biomedical knowledge evolves rapidly with ongoing research. Future work should focus on developing mechanisms to dynamically adapt WSD systems to stay current with changes in biomedical terminology. This involves regularly updating the system with the latest terms and meanings as they emerge from new research findings.

- **Multilingual Biomedical WSD:** Biomedical literature is often published in multiple languages. Future research could extend the scope of WSD to handle multilingual biomedical texts. Addressing challenges related to language-specific nuances and terminology variations would contribute to the applicability of biomedical WSD in a global context.

- **Overlapping Clusters:** Explore the use of overlapping clustering algorithms such as Overlapping Partitioning Cluster and Markov Chain and Link Clustering.

- Use both biomedical and General text together to create a more robust model.

Summary

We delve into Word Sense Disambiguation (WSD), a crucial task in Natural Language Processing (NLP), aiming to discern the correct meaning of words in context. We then explore three types of ambiguity – homonymy, metaphor, and systematic polysemy – illustrating challenges in interpreting words with multiple meanings. We then highlight distinctions between WSD in general and biomedical text, emphasising domain-specific complexities like technical terminology, ontologies, and data scarcity.

The historical overview traces WSD evolution from knowledge-based methods to machine learning approaches, focusing on word embeddings, neural networks, and attention mechanisms. The contemporary landscape discusses hybrid approaches and transfer learning, presenting the challenges and advancements in the field.

We introduce the MSH WSD dataset, a benchmark for biomedical WSD, addressing its limitations, including class imbalance and issues stemming from automatic generation using UMLS. We then discuss our approach of using PubMed articles to construct a co-occurrence network and using clustering technique for WSD. We outline challenges faced during implementation and adjustments made to address resource constraints.

Results indicate promising disambiguation outcomes, capturing a significant percentage of ambiguity in various datasets. The limitations include reliance on a limited list of Multi-Word Expressions (MWEs) and challenges in interpreting MWE-only clusters. The need for human-annotated terms for evaluation and future recommendations for comprehensive models incorporating both singletons and MWEs are discussed.

In summary, this report provides a comprehensive exploration of WSD, highlighting challenges in the biomedical domain, proposing an innovative approach, and presenting initial outcomes along with avenues for future research and improvement.

Appendix

This section lists prompts given to ChatGPT to help write this report by sections.

- Introduction-
  - “Provide examples of ambiguous words.”
- How is WSD for Biomedical Text different from General Text? -
  - “How is wsd on biomedical text different from general text based on the following criterions - Data availability, Domain specific jargon”
- Problems -
  - “What are the problems with existing methods based on following - lack of annotated data, pace of research, multiple languages.”
- My Approach -
  - “Give a brief explanation of markov clustering.”
- Limitations -
  - “Rephrase the following – “We are limited in scope by a list of MWEs.” “
  - “Rephrase the following - “We are dependent on a small set of human annotated terms for evaluation.””
- Future Work -
  - “Rephrase the following statements in terms of limitations -“ We can use singletons and MWE together.””
  - “We can classify the terms in the NLM and WSD datasets into known ambiguity types”
  - “We can use more computational resources.”
- Summary -
  - “Write a summary for the following report.”