

# Word Sense Disambiguation for a domain-specific lexical sample task

<sup>1</sup>Navjot Kaur, <sup>2</sup>Vijay Dhir & <sup>3</sup>Vijay Rana

<sup>1</sup>Research Scholar, Sant Baba Bhag Singh University, Jalandhar (India)

<sup>2</sup>Director R&D, Sant Baba Bhag Singh University, Jalandhar (India)

<sup>3</sup>Assistant Professor, Sant Baba Bhag Singh University, Jalandhar (India)

---

## ARTICLE DETAILS

### Article History

Published Online: 20 February 2019

### Keywords

NLP, WSD, WordNet.

### Corresponding Author:

Email: [parmar4007@gmail.com](mailto:parmar4007@gmail.com)

---

## ABSTRACT

Another limitation of already available supervised based word sense disambiguation (WSD) systems is that they only represent a word as a disorganized entity. However, representation of words using Vector Space Model can offer important and valuable information and hence improve overall accuracy. There are various methods that are already developed in the past. With some minor disadvantages, proposed relatedness system measures some properties for the upcoming generation of the Word Sense Disambiguation applications: sense inventory, domain understandability and universality. This paper investigates a novel approach in word sense disambiguation by exploiting every word in space and evaluates this approach on a domain-specific lexical sample task.

---

## 1. Introduction

Word Sense Disambiguation is one of the crucial tasks in natural language processing. There has been an extensive past information available on WSD and has measured many of the hardest troubles in Artificial Intelligence. Since 1940, WSD has been deliberately used as a part of Machine Translation for performing many prominent tasks. In computer software, it is generally investigating the use of computer software to give details of things from one natural language to another.

We can perform WSD task generally in two dissimilar ways. A model of vocabulary in a lexical sample task, is preferred from the constructor database and based on a small given context the intended words are disambiguated. It is bogus that the expression to be disambiguated has a recognized set of senses in the sense catalog, where the sense inventory contains the measuring of language and their dissimilar senses. Given this supposition, WSD can be considered as an organization problem where, specified a word in a context and its potential senses in a lexicon, the task is to organize the happening of the statement in one or more intelligence classes. Nevertheless, lexical illustration disambiguation is defined we can disambiguate only a few words at a time. In order to lead it in a realistic future this ambiguity should be treated in a more prominent way. The task based on all-words serves that idea.

WSD can be categorized into three forms: supervised, unsupervised and reinforcement techniques. The word contains many meanings and all senses of a word are defined to one or more of its probable senses that depend on the word hierarchy.

Each word has imitation which has a unique and identical set of meanings from a WordNet, a vocabulary knowledge base, or an ontology which means showing the properties and relationship between the various contexts. It can also be done with the help of Application-specific inventories. To exemplify, in a structure of machine translation (MT), a word sense can be treated as word translation, which helps in enhancing more and more realistic use of huge amount of multi-lingual

analogous collection of written text that serve as data for preparation.

## 2. Applications of WSD

The ambiguity of a word in the context is removed with the help of Word Sense Disambiguation which is very important for various applications of NLP like:

- **Information Retrieval:** - Word Sense Disambiguation supports in developing a term basis in retrieval of information. Word senses enhance revival of presentation and the senses of Word Sense Disambiguation are controlled like catalog terms. In this manner, credentials would not be ordered that are based on words separate, the documents should be supposed to be ranked on the basis of word senses, or on a grouping of words and words senses. To illustrate: Using dissimilar the indexes for the keyword "Java" such as "programming language", like "type of coffee", and as similar to "location" will recover precision of an Information Retrieval system. Instead of indexing, WSD is also serviced in the query development. The short queries are prolonged using words that are in the right place to same synsets. Amplified queries provide better results than inventive queries in case of recovery. Thus, Word Sense Disambiguation mainly improves the accuracy of Information Retrieval by eliminating the irrelevant hits. There are two common processes in IR. The first one is to dismiss from the documents general, and as a result meaningless words, like "the," and second process is to delete the enduring letters from their roots. The word "running," for illustration, would be changed to "run." These processes are also known as stop wording and arising, correspondingly.
- **Machine Translation:** - The mechanical classification of the exact transformation of the words in the framework is known as machine translation (MT) [2], which is a very complex effort. WSD facilitates in

enhanced consideration of input language and production of the sentences in the desired language. It is also concerned with the lexical variety of words which depends upon the procedure of the context. Word Sense Disambiguation is essential for the choice of lexical in Machine Translation for the word that provides various translations for dissimilar senses which are potentially indistinct within a provided field. To exemplify, English-French economic news translator uses the English noun revolutionize to translate either changement or monnaie. The Machine Translation [3] directly describes the vocabulary as the senses in the destination language. Nonetheless, most Machine Translation has models which do not use unambiguous Word Sense Disambiguation. The glossary is previously disambiguated for the specified rule known as hand-crafted rules.

- **Information Extraction (IE) and text mining:** - Word Sense Disambiguation is an appropriate for a specific examination of text in various applications. To cite an example, an intelligence get-together system may involve the deteriorating and all the references to the various illegal drugs. Bioinformatics research wants the relationships between the connecting genes and the genes products that are formulated from the wide systematic literature; nonetheless, the same name is given to the genes and their proteins. Generally, an automatic explanation of documents is required by the Semantic Web according to the orientation of ontology. Co-reference purpose, named-entity categorization, and acronym expansion may also be casted as Word Sense Disambiguation [4] problems that are suited for names. Word Sense Disambiguation is only useful in the beginning of these areas.
- **Lexicography:** - WSD and lexicography [5] (i.e., the professional writing of dictionaries) can definitely benefit from each additional: WSD can help supply practical sense groupings and statistically important gauges of framework for new or accessible senses. Moreover, WSD can help create semantic networks out of machine-readable dictionaries. Modern lexicography is corpus-based, thus WSD and lexicography can exertion in a loop, with WSD providing uneven empirical sense grouping and statistically significant contextual indicators of sense to lexicographers, who construct better sense inventories and sense-annotated corpora to WSD. Moreover, intelligent dictionaries and thesauri might one day make available us with a semantically-cross-referenced dictionary as well as better appropriate look-up services.
- **Text Processing:** - Text to Speech conversion i.e., when vocabulary are prominent in more than one way depending on their denotation.
- **Speech Recognition and Part of Speech tagging:** - Speech recognition i.e. when dealing out homophones words which are spelled in a different way but prominent the same way

- **The Semantic Web:-** Finally, the semantic Web [7] visualization can potentially benefit starting most of the above-mentioned applications, as it essentially needs domain-oriented and unlimited sense disambiguation to collection with the semantics of (Web) documents, and facilitate interoperability between systems, ontologies, and users. WSD has been used in semantic Web-related research fields, like ontology learning [6], to make domain taxonomies and develop significant semantic networks.

### 3. Literature Review

**Marine CARPUAT et al [8]** introducing a new model that incorporates many word phrasal Word Sense Disambiguation models into the phrase based Statistical Machine Translation (SMT), that helps to performs fully phrasal multi-word disambiguation. SMT techniques applied on the widest range of possible class of the models so Phrase based SMT model is choose as the Baseline. Their work redefines the Word Sense Disambiguation effort to counterpart the same phrasal translation disambiguation work faced by the phrase based Statistical Machine Translation systems.

**Alessandro Raganato et al [9]** have reviewed a classification problem at the word level, and framed this problem using Neural Sequence learning. Sequence learning approaches receive a individual model in one pass from the practice data, and next disambiguate together all destination words within an absorption text. Sequence learning implements more functional all-words models that consistently lead to state of the art.

**Marine CARPUAT and Dekai WU et al [10]** have reviewed to use standard Word Sense Disambiguation models to improve the deficiencies of statistical Machine Translation systems. The authors presented empirically based analysis and its major advantage is that the SMT system is quite accurate within the training domain. Word Sense Disambiguation model is the only model which is able to enhance the prediction of the lexical choice in a relatively small amount of cases. When the dedicated WSD model makes better predictions, recently used statistical MT models are not able to exploit this.

**Annette Rios et al [11]** have presents two methods which enhance the word sense disambiguation in the field of Neural Machine Translation that works by integrating the sense embeddings. Firstly, sense embeddings is passed as an additional input to a neural machine translation system. Secondly, sense embeddings are based on extract lexical chains taken from the various document and integrated this information in the Neural Machine Translation model.

**Adam Poliak et al [12]** have used sentence representations from pre-trained 1neural Machine translation encoders to train classifiers for Natural Language Inference (NLI). There is four distinct datasets are used to train NLI classifier: Multi NLI and three recast NLI datasets. Neural Machine Translation encoders might learn the most about semantic proto-roles.

**Gurpreet Singh Josan et al [13]** have developed Punjabi to Hindi Machine Translation system which is based on Direct Machine Translation architectures and language collection. The system is resolved in ASP.net at front end along with MSAccess at back end. Statistical and rule based approach maintain a huge efficiency and simple result. Another function enables the user to translate a webpage in Punjabi to Hindi on the fly. Another interface implements the users to write E-mail in Punjabi.

**Sam Henry et al [14]** have analyzed various methods to develop distributional lexicon vectors of multi-words for the effort of the semantic relatedness. Create perception vectors with the MetaMap tool. His model also compares the efficiency of multi-word term with the help of aggregation methods of the component word vectors, creating multi-word term vectors with the help of the compoundify tool. The highest sum of the correlations is achieved with the help of vectors concept across all the datasets. The multi-word aggregation method has no statistical importance between all the dimensionally reduction techniques.

**Costanza Conforti et al [15]** a WSD system has been developed for the Low Resource Language (LRL) venation, which is a Gallo-Italic language spoken in Italy. There is two methods are used to develop a system. First, select and annotate a corpus for venation, considering two words and using two levels of annotation. Second, supervised WSD performed with support vector Machine on this corpus for proof-of-concept experiment.

**Tommaso Pasini et al [16]** have presents two knowledge based methods to correct distribution of senses within an input corpus which can potentially boost the performance of WSD. Fully automatic, Language independent, Knowledge based methods are EnDI and DaD for computing the correct senses given a raw input corpus of sentences without relying on manual training data. These methods perform well on Intrinsic and extrinsic evaluations and the strongest baselines for the most frequent sense in multiple languages.

**Yee Seng Chan et al [17]** have shows that the word sense disambiguation system helps to improve the performance of statistical Machine Translation Systems. Using a Hierarchical Phrase based model (Hiero) for statistical MT, based on weighted synchronous context-free grammar (CFG). And the added features of WSD into Hiero to rerun the experiment. Hiero, which achieves enhanced results on paraphrase presentation relatively than other approaches.

**David Yarowsky et al [18]** have purposed an unsupervised learning technique algorithm that can accurately disambiguate word senses in a large, completely untagged corpus. The algorithm does not need the expensive hand-tagged training dataset, to use two powerful properties of human language: one sense per collocation and one sense per discourse. The algorithm performs well using only local collocation information, treating each token of the word independently. Indeed, this procedure is robust and self-correcting and exhibits much strength of supervised techniques.

**Domagoj Alagic et al [19]** have developed algorithm for describing lexical ambiguity and resolving it. Using Word Sense Induction (WSI) that induces word senses from unannotated corpora, classically by performing clustering. Lexical substitutes can improve on context-based representations for Clustering based WSI. This method works as first computing the similarity of word instances based on presets and contexts and after that performs clustering using a standard clustering technique, also explored the trade-off between automatically predicted and human-provided substitutes for WSI on a small hand annotated dataset.

#### 4. Motivation

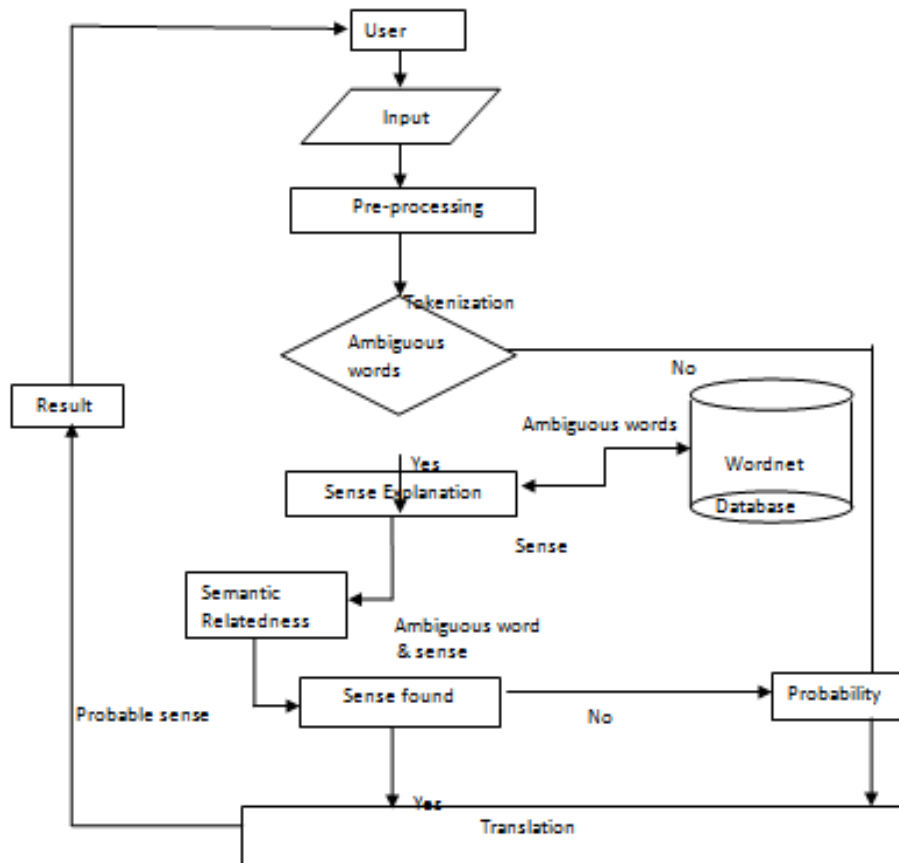
In the process of material on internet is propagating exponentially so are the actual information retrieval system [20]. It is long for that these structures shall simplify interoperability and also address the problems concerning semantic relatedness. For instance, major work that has been done by the various automatic spell corrections and checking in the English language, for the Indian language mistake alteration, has shown more imperfections than any other English because of the Indian Language characteristics. There are a few works on developing an appropriated representation of the words applying context, an inappropriate over the fitting of the words that have more than one meaning is having an issue which deteriorates the learning the terminology when these are used and needs to be fixed. Most of the existing search systems are not able to retrieve the desired results with the intended meaning and this is primarily due to the fact that these systems have not been designed with the intention of extracting wisdom from the web. Therefore, the need for research activities in information retrieval system [63] supporting the relatedness infrastructure is apparent. To accomplish this vision this chapter proposes Punjabi Information Retrieval System (PIRS) that shall push retrieval of meaningful results.

#### 5. Methodology

Punjabi Information Retrieval System (PIRS) is proposed with the idea to give an interactive & a novel platform that shall execute tasks in parallel in discovering useful information and knowledge. In fact, PIRS is a knowledge based method that measures the amount of semantic similarity numerically and applicability between the dissimilar words depending on the exploration of lexical resources such as WordNet [21]. The proposed method will justify itself as a breakthrough in the field of an information retrieval which makes information more meaningful. In contrast to the existing systems which focus on finding non-ambiguous words? PIRS focuses on initially finding the words which has highest ambiguity and later computing the applicability between the remaining query and the ambiguous word searched during initial phases. PIRS comprises of four modules namely Query Extraction Module (QEM), Maximum word Explanation Module (MWEM), Sense Relatedness Module (SRM), Translation Module(TM). Figure 4.1 presents the abstract view of interaction of all four modules where QEM gets activated when a user inputs a query and outputs the ambiguous keywords. The second module MWEM identifies the set of maximum meanings for such ambiguous keywords. In semantic relatedness [22] module system find the relatedness between two words on the basis of their semantic

hierarchy. In translation module translates those words, which creates the ambiguity in the sentence. Working of each module

along with its algorithm which is described in detail in upcoming sections.



**Query Extraction Module (QEM)**

In general, the computer based machine does not have the complex capabilities of vocabulary that human beings have. The proposed system uses QEM technique to handle this kind of situation.

Algorithm 1 describes the working of QEM module. The algorithm returns the most ambiguous word.

**Algorithm 1: Query Extraction Module**

```

    Step 1: Import Punjabi WordNet
    Step 2: Input Query
    Step 3: Results = Aambiguous words
    Step 4: Tokenization words = w.split(Query)
    Step 5: Part of speech
    Step 6: To detect polysemy words
    m = data_text('words')
    wrd = for m ∈ String.split()
    for m ∈ wrd
    if(len(PB wordnet.synset(m) > 1))
    append(m)
    end if
    end for
    end for
    
```

It is started with to import Punjabi WordNet. A set of words do inclined at the time absorption to Query Extraction which detects ambiguous words that outline the user's target words demand. In the existing translators that only provide the results as Polysemy words while neglect the semantics of user

compulsion. QEM different algorithm that target on finding unclear queries whatever rest more search to detect a suitable meaning to summarize the user requirements. The algorithm tries to handle the ambiguity challenge. This algorithm works as eliminating those words, which have little value in the sentence and find the ambiguous words in given query. To achieve this objective processing phase performs some set of techniques like, POS, tokenization and Polysemy words finding etc. WordNet is treated elected the biggest important resource applicable to analyst in computational linguistics, text analysis and many related fields. Natural Language Processing is necessary as long as trade accurately along the broad volume of text today usable online. NLP [23] is used into retrieve information and text summarization in MT Systems. There are various relations are used in WordNet like antinomy, synonymy etc.

**Sense Exploration Module (SEM)**

Sense Exploration module executes in two steps. It first considers the set of linked words and later resolves the perfect meaning of each occurrence. There are two common approaches for WordNet manufacture and these are merge approach and expansion approach. Algorithm 2 describes the working of SEM module. The algorithm represents the maximal applicable text for the similar words.

**Algorithm 2: Sense Exploration Module:**

```

    Input: Wordnet wrdnt; Ambiguous Keywords: data
    Output: Set of similar sense: {s1, s2 ... Sn}
    
```

```

Step 1: Import Punjabi Wordnet as wrdnt
Step 2: wrdnt.synset('data')
Step 3: Extraction and similarity computation
if(wrdnt.synset('data') < 1)
  then return('data')
  else
    W × POS → 2 SYNSETS
Step 4: Set of Similar senses as output
return(SS{s1, s2 ... ..., sn});

```

In merge approach, various senses in which a word perhaps used are firstly registered. After that, the lexicographer formulates a synset individually of the sense, successively the three basis of minimality, coverage and replace ability for synset creation. And the second is expansion approach in which the synsets of a WordNet of the provided language are given. Then, meaning of each synset can be carefully studied. WordNet is like a vocabulary in that it supplies words and meanings. WordNet has arranged verbs, nouns, adverbs and adjectives in synsets. One and the same words are arrange together to design equivalent sets, or synsets. All such synset therefore produce a single specific sense of or concepts.

Different types of words form in WordNet are:-

- Polysemous: words that have different sense are known as Polysemous. In WordNet, any word appears in as many synsets as it receive senses.
- Monosemous: words which can have only one sense are known as Monosemous.

The various sense numbers are assigned in WordNet according to their frequency. They are occurred in the corpus in such a way that the first sense of word is more common than the second word. In the word explanation module all the possible senses of given Polysemy words are obtained from WordNet.

Algorithm 3 describes the working of SDM module. The algorithm finds the relatedness between two words.

#### Algorithm 3: Sense discovery module

```

Step: 1 Take input from MWE
Step2: Find the relatedness
2.1 if(relatedness == si1, si2, ..., simi)
  R ← sij × skl
  then return correct sense
  2.2 else
    if(relatedness! = words)
      then return(minsynset);
    end step 2

```

Measures of similarity or relatedness have various applications. Some of them are automatic indexing, information retrieval, word sense disambiguation [1] and an automatic text

correction. Relatedness is a special case of semantic similarity and works only in hyponymy/hypernymy relations. It measures combined relationships that exist between the various words which depends on the content or their significance. There are various semantic similarity methods that have been developed. These methods can be categorized into two categories. First one is information content method and second one is edge counting method. The existing semantic similarity [24] methods based on lexical and path based algorithm. Information content method is also known as corpus based method that measures the difference between the information content of the two words as the function of the probability of an occurrence in the corpus.

When we turn from human to machine processing of translation, various problems are occurred at different - different levels. Commonly, there are two types: Polysemy- is a specific statement from acquires more than one sense. Similar- numerous vocabulary obtains the equivalent meaning are one and the other relevant problems in natural language processing (NLP) or artificial intelligence (AI). To develop a Translation Model for Punjabi-English pair of languages, where Punjabi is source language and English is a destination language. Machine translation can help to overcome the technological barriers. Algorithm 4 describes the working of TTM module. The algorithm finds translates words.

#### Algorithm 4: Text Translation module (TTM)

```

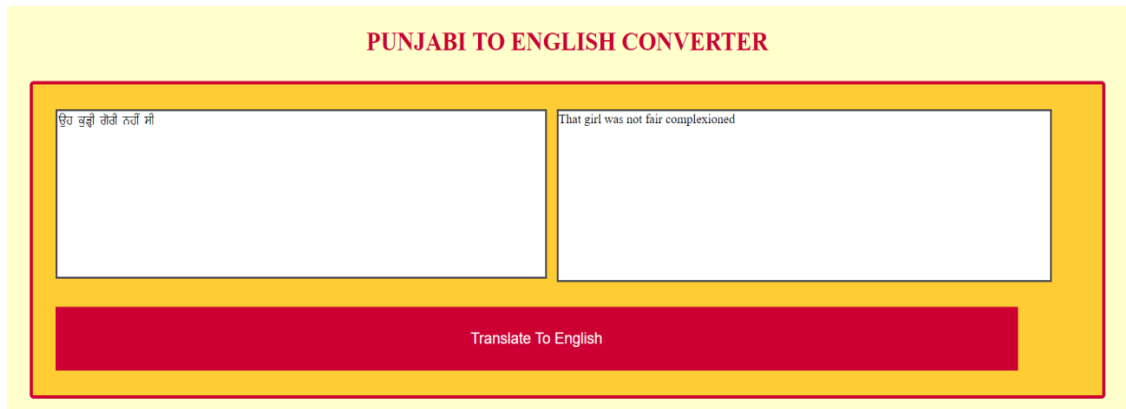
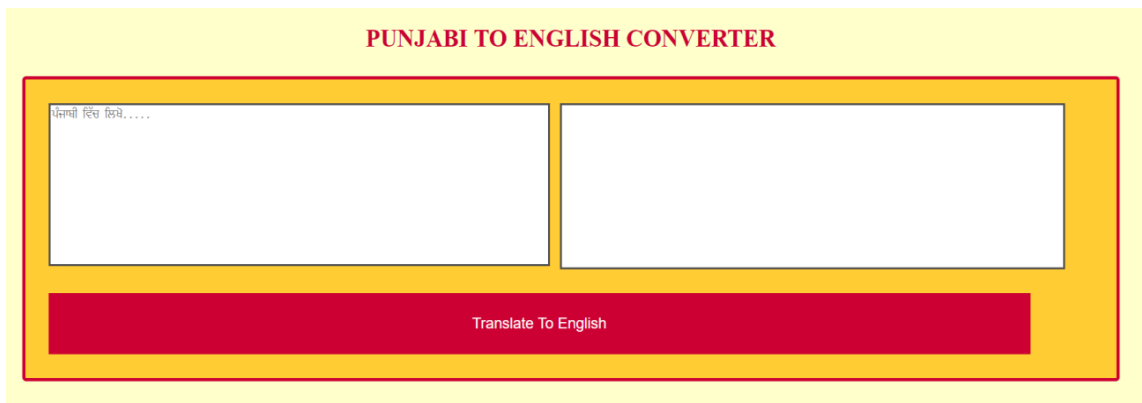
Step 1: import grammar rules,
Step 2: input SRM
Step 3: GR = np(noun phrase), vp(verb phrase),
          n(noun), prep(preposition),
          adj(adjective), s(sentence)
Step 4: Results

```

## 6. Results and discussion

This part comprises the results that are creating during the results. The result is determined with using quality framework.

This work is executed in PHP with a motivation behind WampServer. It is capable of translating a sentence into another language (whereas Punjabi language is source language and destination language is English language). Various sentences are adding in the translation system using the technique of supervised learning. PHP Myadmin is used to storing data (sentences) provide by WampServer. Firstly take the sentences or words which shows the ambiguities in the destination language. And then chooses which Programming language is used for write the coding. There is PHP is used which is server side scripting language. And then make the algorithm for making Punjabi to English translator. After that select the Server (Wamp, Xamp Server) which is used to run the coding, programs and provide services like PHP Myadmin for storing the values in Database. Following are the figure shows the result for an input query in more subtle elements.



**7. Conclusion**

In this manuscript, we implemented a translator which translates/converts the text from Punjabi to English language. A translator is basically an application program used to transform sentences written in one language and converts to the target language. The Punjabi-English machine translation system is based on supervised learning technique. We ensure that the input sentence contains ambiguous word with multiple

English translations. The system performs translations at sentence level. In the process of future work, we examine the appropriateness algorithms for Punjabi word sense disambiguation. We train the data as much as maximum. Therefore, our proposed system of disambiguation senses can be examined to be useful and suitable for other research efforts in natural language processing.

**References**

1. Sussna, M. (1993, December). Word sense disambiguation for free-text indexing using a massive semantic network. In Proceedings of the second international conference on Information and knowledge management (pp. 67-74). ACM.
2. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
3. Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M., & Liu, Y. (2015). Minimum risk training for neural machine translation. arXiv preprint arXiv:1512.02433.
4. Sugawara, H., Takamura, H., Sasano, R., & Okumura, M. (2015, May). Context representation with word embeddings for wsd. In International Conference of the Pacific Association for Computational Linguistics (pp. 108-119). Springer, Singapore.
5. Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational linguistics, 16(1), 22-29.
6. Zhou, L. (2007). Ontology learning: state of the art and open issues. Information Technology and Management, 8(3), 241-252.
7. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 34-43.
8. Carpuat, M., & Wu, D. (2007). Improving statistical machine translation using word sense disambiguation. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).
9. Raganato, A., Bovi, C. D., & Navigli, R. (2017). Neural sequence learning models for word sense disambiguation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1156-1167).
10. Carpuat, M., & Wu, D. (2005, June). Word sense disambiguation vs. statistical machine translation. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 387-394). Association for Computational Linguistics.
11. Rios, A., Mascarell, L., & Sennrich, R. (2017). Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings. In Proceedings of the 2nd Conference on Machine Translation, Copenhagen, Denmark.
12. Poliak, A., Belinkov, Y., Glass, J., & Van Durme, B. (2018). On the Evaluation of Semantic Phenomena in Neural Machine Translation Using Natural Language Inference. ArXiv preprint arXiv: 1804.09779.

13. Josan, G. S., & Lehal, G. S. (2008, August). A Punjabi to Hindi machine translation system. In 22nd International Conference on Computational Linguistics: Demonstration Papers (pp. 157-160). Association for Computational Linguistics.
14. Henry, S., Cuffy, C., & McInnes, B. T. (2018). Vector representations of multi-word terms for semantic relatedness. *Journal of biomedical informatics*, 77, 111-119.
15. Conforti, C., & Fraser, A. (2017, May). Supervised Word Sense Disambiguation for Venetan: A Proof-of-Concept Experiment. In The Thirtieth International Flairs Conference.
16. Pasini, T., & Navigli, R. (2018). Two knowledge-based methods for high-performance sense distribution learning. *Proceedings of AAAI, New Orleans, United States*.
17. Chan, Y. S., Ng, H. T., & Chiang, D. (2007). Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 33-40).
18. Yarowsky, D. (1995, June). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* (pp. 189-196). Association for Computational Linguistics.
19. Alagić, D., Šnajder, J., & Padó, S. (2018, January). Leveraging Lexical Substitutes for Unsupervised Word Sense Induction. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.
20. Clarke, I., Sandberg, O., Wiley, B., & Hong, T. W. (2001). Freenet: A distributed anonymous information storage and retrieval system. In *Designing privacy enhancing technologies* (pp. 46-66). Springer, Berlin, Heidelberg.
21. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3(4), 235-244.
22. Clarke, I., Sandberg, O., Wiley, B., & Hong, T. W. (2001). Freenet: A distributed anonymous information storage and retrieval system. In *Designing privacy enhancing technologies* (pp. 46-66). Springer, Berlin, Heidelberg.
23. Sugawara, H., Takamura, H., Sasano, R., & Okumura, M. (2015, May). Context representation with word embeddings for wsd. In *International Conference of the Pacific Association for Computational Linguistics* (pp. 108-119). Springer, Singapore.
24. Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*