RESEARCH REVIEW JOURNALS
Volume-04
Issue-02
February -2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Punjabi to English Translator Using Supervised Learning

*[1]Navjot Kaur, [2]Vijay Dhir & [3]Vijay Rana

[1]Research Scholar, Sant Baba Bhag Singh University, Jalandhar, PB (India)
[2]Director, Sant Baba Bhag Singh University, Jalandhar, PB (India)
[3]Assistant Professor, Sant Baba Bhag Singh University, Jalandhar, PB (India)

ARTICLE DETAILS

Article History
Published Online: 20 February 2019

Keywords
WSD, Relatedness, Supervised learning.

***Corresponding Author**
Email: parmar4007[at]gmail.com

ABSTRACT

WSD has a long history of research and is considered one of the hardest problems in Artificial Intelligence Building of sense tagged data is a main challenge for supervised techniques that achieve promising results in word sense disambiguation, problem, in this paper (preset tense), WSD can be encapsulates follows: given an ambiguous word, such as bank, determine which sense of the word (i.e. a financial institution, the side of a river, a type of basketball shot, etc.) is being used in given context. In this paper we exploit supervised technique and we find the relatedness between two or more words.

1. Introduction

The material age has been described by the advancement as well as convergence of computing, automations and multilingual information systems. This has resulted in the availability of enormous volumes of information in electronic media [11], but whose natural language form, unlike the data presentation formats typical of computer systems in the past, is more suited for human users than computer systems. This has prompted the development of technologies that would solve this problem and support faster and more efficient access to this information.

All essential wording, have words that can mean various things in different contexts. In English, for example, the word bank can refer to a financial institute or side of a river. Human beings are especially good at this. However, it is so challenging for computer machines to grasp the predetermined meaning of a word in a given context [13]. Correctly understanding the meaning of particular instances of a given word requires successfully distinguishing between different senses of that word. It is then generally very convenient if machines used for manipulating language have the capacity to differentiate different senses of a word.

Words can have different senses [12]. Some words have multiple meanings. This is called Polysemy. For example: bank can be a economic convention or a riverbank. Sometimes two completely different words are spelled the same. For example: Can, can be used as a model verb: He can play chess, or as bottle: She brought a can of Pepsi. This is called Homonymy. Division between Polysemy and homonymy is not forever clear. Word sense disambiguation (WSD) is the problem of conclusive in which [14], sense a word having a number of different senses is used in a given context.

One result some analysts have used is to prefer an appropriate dictionary, and just need its set of senses.

2. Theoretical Background

2.1. Word Sense Disambiguation

WSD has a extended history of research and is well thought-out one of the hardest problems in Artificial Intelligence. In the late 1940s, WSD was first contemplation of as a part of Machine Translation, the universal field investigating the use of computer software to translate from one natural language to another. WSD can be encapsulates follows: given an ambiguous word, such as bank, conclude which sense of the word (i.e. a financial institution, the side of a river, a type of basketball shot, etc.) is organism used in given context. WSD rapidly proved an immensely challenging problem.

2.2 Intelligent retrieval Method

All WSD systems, regardless of the consequences of the approach they take, make use of familiarity present in the context and in outside resources. It is based on the explanation that words used collected in text are connected to each other and that the relative can be noticed in the definitions of the words and their senses [18]. Two (or more) words are disambiguated by determining the pair of terminology senses with the maximum word extends beyond in their dictionary translations. For example, when disambiguating the words in "pine cone", the definitions of the suitable senses both includes the words evergreen and tree (at least in one dictionary).The most commonly encountered knowledge sources in WSD are as follows:

2.2.1 Machine-readable dictionaries: (MRDs) are electronic versions of conventional paper dictionaries. Most dictionaries provide at smallest amount the word's part(s) of speech and sense definitions (also called glosses). Some will also provide the articulation, morphology, etymology, valiancy, domain, register, derivations, semantically related terms, illustration sentences, and/or other information for individual words or senses.

2.2.2 Dictionary: Dictionary lexical situation works which group words according to semantic relations-usually synonymy but sometimes also antonym, and rarely other relations. Dictionary used in early NLP research were uncomplicated digitized versions of obsolete editions of Roget's (1852); nowadays more modern offerings are available.

2.2.3 WordNet: is an electronic lexical database for English. This is a suitable but over– simplified explanation of a very complex store. WordNet can be expecting as a large graph or semantic network, where each node of the network represents a real world concept. For example, the perception could be an object like a house, or an entity like a teacher, or an abstract concept like art, and so on.

2.3 Supervised based approach

Supervised methods are based on the supposition that the context can make available enough confirmation on its have to disambiguate words (hence, world knowledge and reasoning are deemed unnecessary). Possibly every machine learning algorithm administration has been applied to WSD, containing identify techniques such as feature selection, parameter optimization, and collection learning. However, these supervised methods are subject to a new knowledge attainment traffic jam since they rely on substantial amounts of manually sense-tagged corpora for training, which are laborious and expensive to generate.

2.3.1 Support Vector Machines (SVM)
This method is based on the idea of knowledge a linear hyperplane from the training set that isolated positive examples from unhelpful examples. The hyperplane is located in that point of the hyperspace which maximizes the interval to the closest positive and negative examples (called support vectors). In other words, SVMs influence at the same time to minimize the empirical classification error and maximize the arithmetical margin between positive and negative examples. As SVM is a binary classifier, in order to be usable for WSD it must be adapted to multiclass classification (i.e., the senses of a target word).

2.3.2 Memory-based learning
This is a supervised algorithm in which the classification model is built from examples. The model retains examples in memory as points in the feature space and, as new examples are subjected to classification, they are progressively added to the model.MBL has as its defining characteristic that it stores in memory all available instances of a task, and that it extrapolates from the most similar instances in memory to solve problems for which no solution is present in memory.

2.4 Unsupervised approach

Unsupervised approach do not depend on external knowledge sources or sense inventories, machine readable dictionaries or sense-annotated data set. These algorithms naturally do not allow meaning to the words rather they discriminate the word meanings based on information, create in un-annotated corpora.

2.4.1 Context Clustering is based on clustering techniques in which first context vectors are created and then they will be grouped into clusters to identify the meaning of the word. This procedure uses vector space as word space and its extensions are words only. Also in this procedure, a word which is in a corpus will be stand for as vector and how many times it develop will be counted within its context. After that, co-occurrence matrix is created and similarity measures are tested. Then discrimination is performed using any clustering technique.

2.4.2 Word Clustering This technique is similar to context clustering in terms of finding sense but it clusters those words which are semantically identical. For clustering, this approach uses Lin's method. It checks identical words which are similar to target word. And relationship among those words is determined from the elements they are sharing. This can be obtained from the corpus. As words are related they share same kind of dependency in corpus. After that, clustering algorithm is applied to discrimination among senses.

3. Literature Review

Devandra Singh Chaplot et.al [1], 2018 proposed leverage the formalism of topic model to design a WSD system that scales linearly with the number of words in the content. System is able to utilize the whole document as the content for a word to be disambiguated. It also further utilizes the information in the WordNet by assigning a non uniform prior to synset distribution over words and a logistic –normal prior for document distribution over synsets.

GONGQING WU et.al [2], 2018 presented entity linking process to selection query mentions in documents, and then link them to their corresponding entities in a knowledgebase. The application of entity linking involves many fields, such as search Engine retrieval, knowledge fusion, and knowledge base population. It also learns about the Knowledge bases, data sets, and the evaluation criterion and some challenges of entity linking.

Abdulgabbar Saif et al [3], 2018 purposed supervised techniques for sense tagged data to achieved promising results in WSD. It proposes a Knowledge based method for building the Arabic sense tagged corpus from Wikipedia. The advance opening with considerate Arabic WordNet containing Wikipedia to general the Wikipedia article for the proportionate sense in wordnet. In this calculating step, the cross lingual mechanism is used to measure the similarity between features of a Wikipedia article and a wordnet sense independently. For handling the lack of instances of some articles in Wikipedia,

the multiword based technique is purposed to increase a number of instances for each concept.

Alok Ranjan Pal et.al [4] 2017 proposed a knowledge-based Word Sense Disambiguation (WSD) technique for low resource language. This technique has been tested on database in Bengali Language. Bengali Wordnet is used in this task. Lesk algorithm is not worked properly due to less number of overlap. Then an extension of Context Expansion through Synset analysis has been developed.

Sudha Bhingardive et al [5], 2016 purposed WFS baseline for Hindi language by manually ranking the synsets of Hindi Wordnet. A tool of ranking is adopted where human experts can see the frequency of the word senses in the sense-tagged corpus and have been asked to rank the senses of a word using this information also with user intuition. WFS baseline is tested on different datasets.

Satyendr Singh et al [6], 2015 purposed Lesk's dictionary-based algorithm for Hindi WSD.

This enhances opportunity of match between contextual details and extended sense translation. This experiment done with five different cases: Hypernymy and hyponym, hypernymy and homonym, all semantic relations are considered.

Kaveh Thagipur et al [7] 2014 purposed Semi-supervised approach which incorporates Knowledge from unlabeled datasets by using word embeddings. This method uses distributed word representations and improves the accuracy. Semi-supervised learning approaches are: co-training and self-training.

Hamed Valizadegan et al [8] 2013 reviewed a new approach multi-expert learning method to combine class-label information obtained from multiple experts and that these experts may differ in their class label assessments. To combine labels from different human experts there are two models are used, a *consensus* model was representing the classification model and *individual expert* models representing the class label decisions exhibited by individual experts.

David Vickrey et al [9] have presented a word translation model to determine the correct translation of a word from context. Used corresponding language entirely as a large supply of partially labelled data for this task. The algorithms are presented for solving the word translation problem and demonstrate a significant improvement over a baseline system. The word translation model could be improved in a variety of ways, drawing upon the large body of work on WSD. The model introduces the novel blank filling task, which decouples the impact of word translation from other factors, such as syntactic correctness. Also, the model could be extended handle phrases.

Edouard Grave et al [10] 2018 have explored models to trained such high quality word representations for 157 languages. Word representations are also known as word vectors, which have been widely used in natural language processing. There are two sources of data are used to train these models: First, is a common source of data to learn word vectors, available in number of languages, the online *encyclopaedia Wikipedia,* which provides high quality data for comparable across languages. The size of Wikipedia is relatively small and not enough to learn high quality word vectors with wide range. Second is, an alternative source of large scale of data is the web and resources such as crawl. The authors also reviewed three new word analogy datasets to evaluate these word vectors, for Hindi, French as well as Polish. And evaluate pre-trained word representations on 10 languages, which showing strong performance.

4. Problem Definition

A critical look at the above literature highlights the fact that there are numerous debatable issues which still need to be researched upon. The Word Sense Disambiguation (WSD) has been the vision for the next generation of the intelligent system, where information is desired to be useful not only for the people but also for the computers. The major obstacle in implementing WSD is that machines don't have the kind of vocabulary that people have. Different methods have been proposed in the past; however, current relatedness measures lack some desirable properties for a new generation of Word Sense Disambiguation applications: sense inventory, domain understand ability and universality.

Natural languages are notoriously ambiguous on various levels. Semantically, a single word can have more than one meaning, with the two readings belonging either to the same (i.e., *bank*). This problem is the ambiguous nature of English words, particularly polysemy words. The ambiguous word creates an enormous problem in machine translation. E. g

Consider the sentences:

P: ਜਿੱਥੇ ਚਾਹ, ਉੱਥੇ ਰਾਹ

E: Where tea, the way there

P: ਦਰ ਵਿੱਚ ਨਾ ਬੈਠੇ

T: Do not sit in the rate

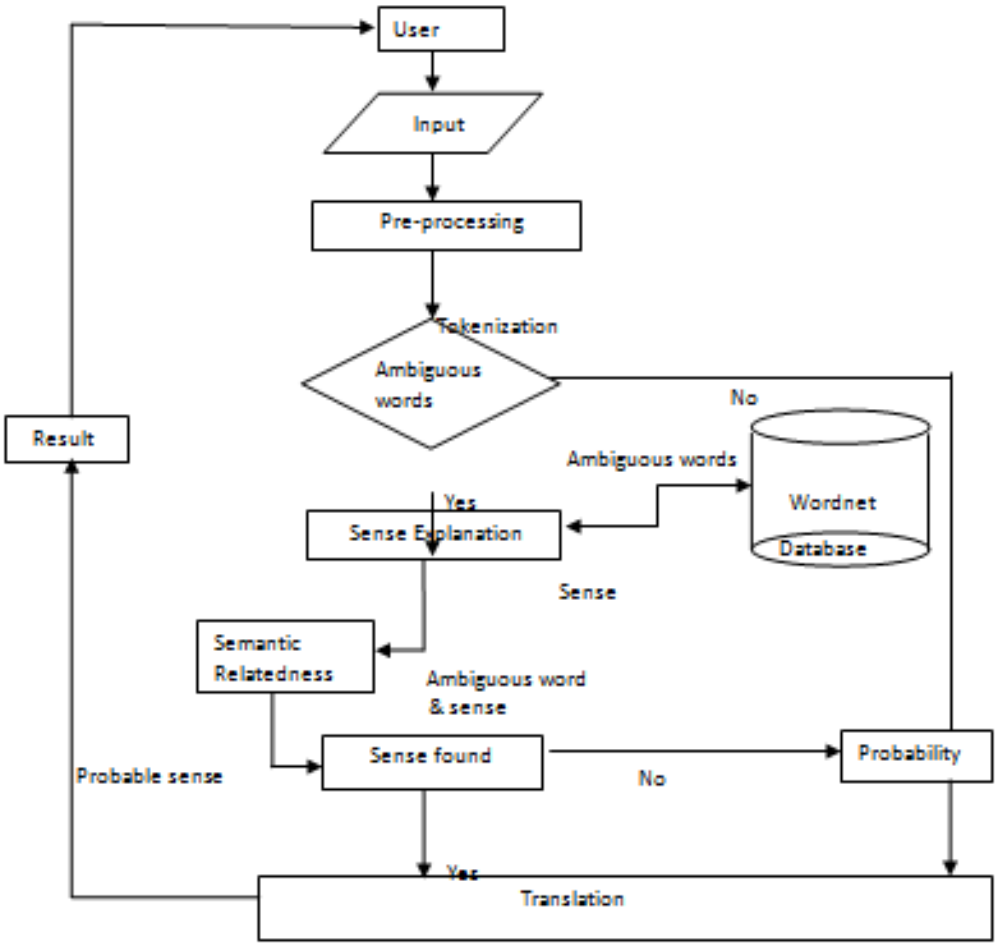
In the above sentence, ਚਾਹ (*chaah*) is a ambiguous word, ਦਰ (*dar*) is also a ambiguous word. Both the ambiguous words present, in the system increases complexity, such type of words are not present in the database or word in the sentence not understand by the system.

5. Objectives

This work during the research period aims to address few of the listed issues by achieving the following objectives:

- To study the structure of ambiguity words.
- To identify the most probable word as per the semantic [16] relatedness.
- To develop a model for translating the Punjabi sentence into English language without any ambiguity.
- Performance Evaluation of proposed framework by comparing the performance of existing interfaces.
- Performance Evaluation of proposed framework by comparing the performance of existing interfaces.

6. Methodology



6.2.1 Query Extraction Module (QEM): It is process of eliminating those words, which have little value in the sentence and find the ambiguous words in given query. To achieve this objective pre-processing phase performs some set of techniques like POS, tokenization and polysemy words finding, etc. After getting a polysemy words the next phase is to find all the possible senses of polysemy word.

6.2.2 Maximum Word Explanation Module (MWEM): In sense explanation module all the possible senses of given polysemy word are obtained from WordNet. In this phase development and clarification of senses is also achieved. After getting all possible senses of polysemy word the next step is to find the semantics relatedness between word and its senses.

6.2.3 Semantic Relatedness Module (SRM): In semantic relatedness module system find the relatedness between two words on the basis of their semantic hierarchy. Semantic relatedness measures quantify the degree in which some words or concepts are related, considering not only similarity but any possible semantic relationship among them. If the relatedness between two words is not found then system will use the existing probability measure to compute the probable result.

6.2.4 Translation Module: - In translation module translates those words, which creates the ambiguity in the sentence. Various phases are executed in this process after these entire procedures translation module is executed. Translated words are provided to the users.

7. Results

This translator is divided into two sections. Firstly, we provide the input section for inputting the Punjabi Sentences into Punjabi language. Secondly, give the output section for showing the destination language. Afterthat provide the button for translating the text or sentence into destination language. The user can write the sentence here or copy the text in punjabi from another site and paste here for translating the text into english language, if the text is copied by user stored in the database then it will be translated otherwise it is not converted. First stored the values of text in database and it will translate the text.

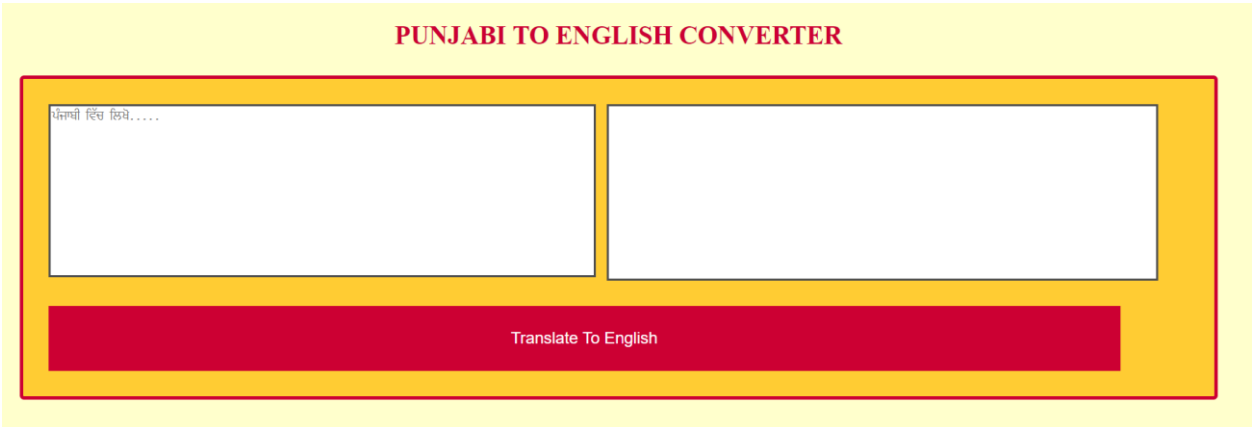


Figure1: Image of Punjabi to English Translator/Converter

Figure2: Translating Text

Now select the Punjabi editor for write the text into input language. The translator removes the ambiguities of Punjabi language and provides the right output to the user. The translator also checks the relationship or semantic relatedness between the Punjabi texts. It first takes the input and Tokenization and translates into English.

8. 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. We have implemented the translation system for the translation of text from Punjabi to English using PHP server side language and run on the wamp server PHP my admin. We present an approach for solving the ambiguity of words in Punjabi language. We evaluate our approach through an experiment using the Punjabi-English parallel corpus aligned at sentence level. We ensured that the input sentence contained ambiguous word with multiple English translations. It is complex to understand the approach, but yields result in short duration of time. It takes less time to develop translation system using PHP, but it is not easy to achieve accuracy. This system use very big data for storing translation sentences.

References

1. Chaplot, D. S., & Salakhutdinov, R. (2018). Knowledge-based word sense disambiguation using topic models. *arXiv preprint arXiv:1801.01900*.
2. Wu, G., He, Y., & Hu, X. (2018). Entity Linking: An Issue to Extract Corresponding Entity With Knowledge Base. *IEEE Access*, *6*, 6220-6231.
3. Saif, A., Omar, N., Zainodin, U. Z., & Ab Aziz, M. J. (2018). Building Sense Tagged Corpus Using Wikipedia for Supervised Word Sense Disambiguation. *Procedia Computer Science*, *123*, 403-412.
4. Pal, A. R., Saha, D., & Pal, A. (2017). A Knowledge based Methodology for Word Sense Disambiguation for Low Resource Language. *Advances in Computational Sciences and Technology*, *10*(2), 267-283.
5. Bhingardive, S., Shukla, R., Saraswati, J., Kashyap, L., Singh, D., & Bhattacharyya, P. (2016). Synset Ranking of Hindi WordNet. In *LREC*.
6. Singh, S., & Siddiqui, T. J. (2015). Role of Semantic Relations in Hindi Word Sense Disambiguation. *Procedia Computer Science*, *46*, 240-248.
7. Taghipour, K., & Ng, H. T. (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*(pp. 314-323).

8. Valizadegan, H., Nguyen, Q., & Hauskrecht, M. (2013). Learning classification models from multiple experts. *Journal of biomedical informatics*, *46*(6), 1125-1135.
9. Vickrey, D., Biewald, L., Teyssier, M., & Koller, D. (2005, October). Word-sense disambiguation for machine translation. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 771-778). Association for Computational Linguistics.
10. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.
11. Zhi-Jun, P., Jian-Ling, P., & Yao-Xi, J. (2015, January). Large-Scale Bilingual Corpus Knowledge Based on Web Mining. In *Proceedings of the 2015 Sixth International Conference on Digital Manufacturing and Automation* (pp. 163-166). IEEE Computer Society.
12. Dhungana, U. R., Shakya, S., Baral, K., & Sharma, B. (2015, February). Word Sense Disambiguation using WSD specific WordNet of polysemy words. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)* (pp. 148-152). IEEE.
13. Tripodi, R., & Pelillo, M. (2017). A game-theoretic approach to word sense disambiguation. *Computational Linguistics*, *43*(1), 31-70.
14. Pelevina, M., Arefyev, N., Biemann, C., & Panchenko, A. (2017). Making sense of word embeddings. *arXiv preprint arXiv:1708.03390*.
15. Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., ... & Weikum, G. (2011, July). Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 782-792). Association for Computational Linguistics.
16. Rana, V. (2012). Blueprint of an ant-based control of semantic web. *IJoAT*, *2*, 603-612.
17. Thukral, S., & Rana, V. (2019). Versatility of fuzzy logic in chronic diseases: A review. *Medical hypotheses*, *122*, 150-156.
18. Sharma, S., & Rana, V. (2017). Web personalization through semantic annotation system. *Advances in Computational Sciences and Technology*, *10*(6), 1683-1690.

