Proceedings of the
International Conference on Engineering Research, Innovation and Education 2017
ICERIE 2017, 13–15 January, SUST, Sylhet, Bangladesh

# Review on Homograph Disambiguation: approach for Bangla context

**Abu Hanife Nayem, Jahid Hasan Polash and Md. Saiful Islam**
Department of Computer Science and Engineering, Shahjalal University of Science & Technology, Sylhet, Bangladesh.

**Keywords:**

- Supervised;
- Unsupervised;
- Clustering;
- Induction;
- Dataset;

**Abstract**: Homograph ambiguity is a fundamental issue of NLP. Disambiguation of homographs has various applications including Text-to-Speech (TTS) synthesis, Sentiment analysis etc. Several methodologies have been applied for many languages. In this paper machine learning based homograph disambiguation techniques are reviewed. Best suited approach for Bangla Homograph disambiguation is discussed. The application and usefulness of word sense induction and clustering is reviewed in this paper.

## 1. INTRODUCTION

At the beginning of Machine learning (ML) and Artificial intelligence (AI), study of NLP reveals several problems. One of them is that some words can have multiple meanings in the context. Homograph disambiguation (HGD) as a form of Word sense disambiguation (WSD) means sensing a words meaning in local or global scope.

Homograph as a broader sense Homonyms are words that have similar pronunciation (might be same spelling) but different meaning. In example,

হার (defeat), হার (necklace)
কর (palm of hand), কর (tax)
পাত্র (Bowl), পাত্র (groom)

are homographs in Bangla.

In Bangla as poor-resourced language, identifying homographs is a great challenge. Some popular techniques resulted promising output and scope to apply them in Bangla is discussed in this paper.

* Corresponding author: jahidhasan@student.sust.edu

## 2. HOMOGRAPH DISAMBIGUATION

Considering following two examples, the first sentence depicts the meaning north for উত্তর where in the second sentence it means answer.

1. আমি উত্তর দিকে যেতে চাই।
2. আমি তার প্রশ্নের উত্তর দিতে পারিনি।

To find contextual meaning for the underlined homograph, knowledge-based approaches as well as machine learning approaches can be taken. But as Bangla is a poor resourced language for computational processes knowledge based approach might not be appropriate. Rather homograph word list, shrinks the homograph word searching area in a document.

## 3. LITERATURE REVIEW

Natural language processing works on well-resourced languages (English, French, Italian, German) introduce us to word sense techniques like *Structure*-based methods, *semi*-supervised, *Similarity* based methods etc. Where Chinese, Japanese, Thai uses *minimally* supervised techniques for the lake of word boundary delimiter. Languages as Bangla, Assamese, Marathi are knowledge-poor languages. Thus minimally supervised and unsupervised methodologies have been used for these languages.

However, for Bangla there are very few works have been conducted.

## 4. APPRAOCHES

Machine learning approaches include supervised methods, semi-supervised methods and unsupervised methods based on corpus evidence, training a model by tagged or untagged corpus and probabilistic/statistical models. Supervised and semi-supervised methods tend to apply on knowledge-rich languages where knowledge-poor language processing is more likely to use unsupervised methods.

### 4.1. Supervised HGD

Supervised learning methods for HGD are classifier algorithms based on manually labeled training data with the usefulness of support vector machine (SVM) (Tong, Simon and Koller, 2001). Set of examples (sentences) encoded as vectors are used as training data set whose elements represent features of the example. SVMs' and Memory-based ML algorithms has been proven successful approach. From various knowledge-based (Lesk 1986; Galley and McKeown 2003; Navigli and Velardi 2005) and data-driven (Yarowsky 1995; Ng and Lee 1996; Pedersen 2001) word sense disambiguation methods that have been proposed, supervised systems have been constantly observed as leading to the highest performance. Wikipedia can be a labeled resource for supervised learning data set (Mihalcea and Rada 2007).

### 4.2. Semi-supervised HGD

Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning and so, the data set is divided into labeled and unlabeled data. (Chapelle 2009) Semi supervised methods use two step training on raw corpora. First apply induction sense or categories by a few previously sense annotated

examples. Then functions as an unsupervised clustering approach is performed. For the supervised part of the SSL support vector machine (SVM) is used (Bennett, Kristin, and Ayhan 1999).

### 4.3. Unsupervised HGD

The greatest challenge for HGD researchers is unsupervised approach. Targeting to discover senses automatically based on unlabeled corpora (Boyd, Jordan, Blei, and Zhu 2007; Navigli and Roberto 2009) by word sense induction (WSI) technique and apply them for homograph disambiguation can be considered unsupervised HGD learning.

For Bangla as an under resourced language unsupervised approaches likely to be appropriate.

## 6. WORD SENSE INDUCTION

Words sense induction refers to automatic identification of word senses without manually level data. Not relying on external resources this *unsupervised* technique clusters dataset. Two words are semantically close if they co-occur the same neighboring words (Harris 1954; Curren 2004; Firth 1957).

WSI can be a solution to *knowledge Acquisition Bottleneck* (Wagner 2006) problem. Also WSI can have many more applications other than building sense inventories.

### 6.1. Clustering Approach

The idea of Distribution Hypothesis (Harris 1954; Curren 2004) is that words with similar meaning appear in similar context. Thus in different context words are grouped in different clusters which separates homographic words in contextual domain. Considering a similarity function: *K*-means, *Bisecting K*-means, *Average link*, *Buckshot* are clustering algorithms applied to a test set of word feature vectors (Lin 1998; Pantel and Lin 2002).

Word: হার (defeat), (necklace), (ratio)
Type 1: a) খেলায় হার জিত থাকবেই। b) সবে মিলে করি কাজ,
হারি জিতি নাহি লাজ। c) দলটি ফুটবল ম্যাচে হেরে গেল।
Type 2: a) বাবা আমাকে সোনার হার উপহার দিয়েছেন।
b) হীরের হারটি অপূর্ব। c) কনের গলার হারটি অনেক ভারি।
Type 3: a) বাতাসে জলীয়বাষ্পের হারকে আর্দ্রতা বলে।
b) শিশু মৃত্যুর হার কমানো দরকার। c) বেকারত্বের হার দিন দিন বাড়ছে।

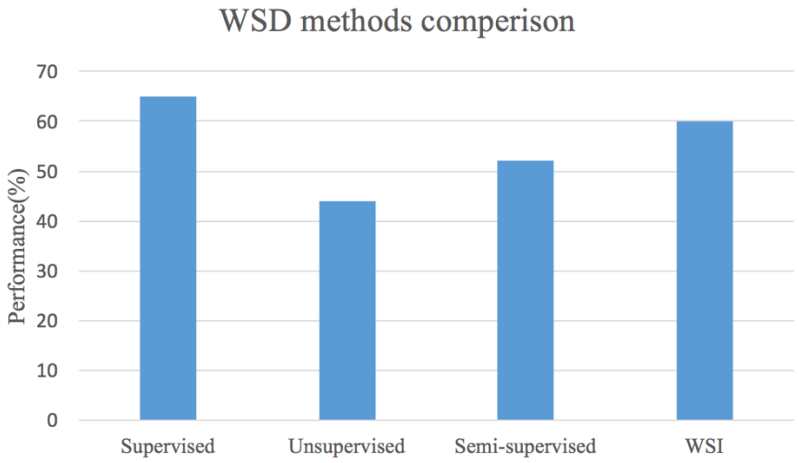
Bi-gram clustering (Schutze 1998), Bayesian contextual framing (Brody and Lapata 2009), Phrasal co-occurrence technique (Dorow and Widdows 2003), clustering using a context window (Ferret 2004) are extended clustering approaches for word sense disambiguation (Nasiruddin 2013) so for Homograph Disambiguation.

### 6.2. Graph-based Approach

Using a sequence of words with their corresponding words(senses), this technique seeks to identify a graph of sense dependencies on which the centrality can be measured, resulting in a set of scores that can be used for sense assignment. By the score of the senses the decision of the sense will be taken (Sinha and Rada 2009; Navigli and Lapata 2010; Dorow , Widdows, Ling, Eckmann, Sergi and Moses 2005).

## 7. APPROACH FOR BANGLA

Supervised approaches for Bangla homograph disambiguation is difficult due to lack of annotated corpora and lexical databases for Bangla. Also sense annotated corpora have less applications and much more difficult to build.




In this case WSI is an attractive alternative. By applying WSI, it is practical to disambiguate particular word instances using the automatically extracted sense inventory even for same spelled words. A particular sense is associated with a particular topic and different senses can be distinguished to their association with particular topic dimensions (Van De Cruys and Apidianaki 2011) shows that the induction step and disambiguation step are based on the same principle.

## 8. CONCLUSION AND DISCUSSION

The techniques described in this paper are commonly used in various languages for Homograph disambiguation. For Bangla we might need to be a little selective. The advantages of unsupervised learning over supervised and semi-supervised methods for Bangla are described in this paper. WSI is introduced which is a promising solution implements clustering and graph-based clustering algorithms.

As supervised techniques illustrates most correctly predicted results for NLP, building sense annotated corpus can be a matter of great interest for researchers. Then Homograph disambiguation algorithms will be much easier and accurate.

## REFERENCES

Tong, Simon, and Daphne Koller. "Support vector machine active learning with applications to text classification." *Journal of machine learning research*2.Nov (2001): 45-66.

M.E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the SIGDOC Conference 1986, Toronto, June.
M. Galley and K. McKeown. 2003. Improving word sense disambiguation in lexical chaining. In Proceedings of IJCAI 2003, Acapulco, Mexico.
R. Navigli and P. Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 27.
D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of ACL 1995, Cambridge.
H.T. Ng and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach. In Proceedings of ACL 1996, New Mexico.
T. Pedersen. 2001. A decision tree of bigrams is an accurate predictor of word sense. In Proceedings of NAACL 2001, Pittsburgh.
Mihalcea, Rada. "Using Wikipedia for Automatic Word Sense Disambiguation." *HLT-NAACL*. 2007.

Chapelle, Olivier, Bernhard Scholkopf, and Alexander Zien. "Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006)[Book reviews]." *IEEE Transactions on Neural Networks* 20.3 (2009): 542-542.
Bennett, Kristin, and Ayhan Demiriz. "Semi-supervised support vector machines." *Advances in Neural Information processing systems* (1999): 368-374.
Navigli, Roberto. "Word sense disambiguation: A survey." *ACM Computing Surveys (CSUR)* 41.2 (2009): 10.

Boyd-Graber, Jordan L., David M. Blei, and Xiaojin Zhu. "A Topic Model for Word Sense Disambiguation." *EMNLP-CoNLL*. 2007.
Harris, Z. (1954). Distributional Structure. In Papers in Structural and Transformational Linguistics, pp. 775–794

Curren, J. R. (2004). PhD Thesis: From distributional to semantic similarity. University of Edinburgh. Edinburg, U

Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis (Oxford: Philological Society), pp. 1–32. Reprinted in Palmer, F.R., (ed.) (1968). Selected Papers of J.R. Firth 1952-1959. London: Longman.
Wagner, C. (2006). Breaking the knowledge acquisition bottleneck through conversational knowledge management. In Information Resources Management Journal (IRMJ) 19(1), pp. 70–83. IGI Global.
Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In Proceedings of the 17th Int'l Conference on Computational Linguistics, pp. 768–774. Quebec, Canada.
Pantel, P. and Lin, D. (2002). Discovering Word Senses from Text. In Proceedings of the 8th Int'l Conference on Knowledge Discovery and Data Mining, pp. 613–619. Canada.
Schutze, H. (1998). Automatic Word Sense Discrimination. In Computational Linguistics 24(1), pp. 97–124. MIT Press.
Brody, S. and Lapata, M. (2009). Bayesian Word Sense Induction. In Proceedings of the 12th Conference of the EACL 2009, pp. 103–111. Athens, Greece.
Dorow, B. and Widdows, D. (2003). Discovering Corpus-Specific Word Senses. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003), pp. 79–82. Budapest, Hungary.
Ferret, O. (2004). Discovering Word Senses from a Network of Lexical Cooccurrences. In Proceedings of the 20th Int'l Conference on Computational Linguistics, pp. 1326–1332.
Nasiruddin, Mohammad. "A state of the art of word sense induction: A way towards word sense disambiguation for under-resourced languages." *arXiv preprint arXiv:1310.1425* (2013).
Sinha, Ravi, and Rada Mihalcea. "Unsupervised graph-based word sense disambiguation." *Recent Advances in Natural Language Processing V: Selected Papers from RANLP* (2009).
R. Navigli and M. Lapata, "An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 678-692, April 2010.
Dorow, B., Widdows, D., Ling, K., Eckmann, J., Sergi, D. and Moses, E. (2005). Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination. In Proceedings of the MEANING-2005 Workshop.

Van De Cruys, T. and Apidianaki, M. (2011). Latent Semantic Word Sense Induction and Disambiguation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 1476– 1485. Oregon, USA.