Word Sense Ambiguity in Question Sentence Translation: A Review

Sanjay Kumar Dwivedi and Shweta Vikram(✉)

Babasaheb Bhimrao Ambedkar University, Lucknow, India
skd200@yahoo.com, shwetavikram.2009@rediffmail.com

Abstract. Word Sense Disambiguation has been a major challenge for various linguistic researches. Enough research has been carried in the past four decades. In machine translation, WSD plays a vital role in improving the accuracy of the translation. The automated translation of question papers from English to Hindi is one such key area which requires suitable WSD techniques to resolve ambiguity in a question word. When machine translates question sentences, it faces ambiguity problem that results in ambiguous translation. Identification of question type is important for remove ambiguity in the question paper. In this paper besides discussing WSD its approaches, resources for translation. We have also discussed question classification word sense disambiguation.

Keywords: Machine translation · Types of question · Word sense disambiguation · English language · Hindi language

1 Introduction

Word Sense Disambiguation (WSD) is a very challenging task of machine translation when source and target languages are different in many aspects. WSD is a process which selects the correct sense of the word with respect to context. It is an open problem of Natural Language Processing and it comes under the NP-complete problem. WSD is very helpful in machine translation. The term WSD was first time introduced by Warren Weaver in his famous memorandum on translation in 1949. Machine Translation refers to the use of the computer to automate some part of the task or the entire task of translating between human languages. It is a subfield of computational linguistics that investigates the use of computer software to translate text from one source language (SL) to target language (TL). Many attempts are being made all over the world to develop MT systems for various languages using rule-based, example-based, dictionary based, corpus-based and statistical-based approaches. MT systems can be designed either specifically for two particular languages, called a bilingual system, or for more than a single pair or languages, called a multilingual system. A bilingual system may be either unidirectional, from one source language into one Target Language, or may be bidirectional. Multilingual systems are usually designed to be bidirectional, but most bilingual systems are unidirectional.

S.C. Satapathy and A. Joshi (eds.), *Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 2*, Smart Innovation, Systems and Technologies 84, DOI 10.1007/978-3-319-63645-0_7

2 Literature Review

In India one of the first attempts made to disambiguate the nouns of Hindi language using Hindi WordNet [11]. Corpus is necessary for many purposes such as training purpose, pattern identification. Different types of resources of corpora such as speech, text, and image [12, 13]. Creation of dictionaries, thesauri and corpora started in 1980. Corpora provided a vast amount of knowledge and information on various hands-on language parameters. WSD research has been ongoing for the four decades. Word sense disambiguation is relevant when a word has multiple senses. The first experiments work done on word sense disambiguation with machine translation activities (Hutchins 1999) [5, 25]. In this machine lookup information and translate a word into a target or desired language. Many senseval workshops held on NLP and they gave many unique ideas for WSD approaches. Word Sense Induction (WSI) [7] is a method which is introduced by Navigli, WSI comes under unsupervised techniques which aimed at automatically identifying the set of senses denoted by a word. In this method word senses from the text by clustering word occurrences based on the idea that a given word used in a specific sense which co-occurs with the same neighboring words [2]. Anusaaraka: Machine Translation can be categorizing into rule-based [16], statistically based, example based, and hybrid [1]. Rule-based systems are still higher than statistical systems.

Many works are done and ongoing in machine translation for one natural language to other natural languages. Here we give the brief introduction of work on machine translation in India. Matra[1] English to Hindi Machine Translation systems (MTS) started at CDAC Pune in 2004. Shakti-English to Hindi, Marathi, and Telugu MTS (combines rule-based and statistical approach) IISc Bangalore and IIIT Hyderabad started in 2004. Anglabharti MTS (English-Hindi, Tamil MTS), Angla-Hindi (combines example based approach and AnglaBharti approach) started 1991 and Anubhart Hindi-English MTS; Combines syntax started on 2004 at IIT Kanpur. Anglabharti is a pattern directed rule-based system with a context free grammar like structure for English (source language) that generates a ‘pseudo-target’ applicable to a group of Indian languages (target languages). Siva-English–Hindi translation started on 2004 at IISC Bangalore. Anusaaraka[2] is a unique approach to machine translation based on the theory of information dynamics inspired by the Paninian grammar formalism. MANTRA[3] (MAchiNe assisted TRAnslation tool) translates English text into Hindi in a specified area of personal administration, specifically, gazette notifications, office orders, office memorandums and circulars. MANTRA uses Lexicalized Tree Adjoining Grammar (LTAG) to represent the English as well as the Hindi grammar. Anglabharti uses a pseudo-interlingua approach. Many free translators available on the Internet which is: Google, Babylon, Yahoo Babel Fish, PROMT translation, Microsoft. Every Machine translation based on one or more than one approach.

[1] http://cdacmumbai.in/matra/.

[2] http://anusaaraka.iiit.ac.in/.

[3] http://en.wikipedia.org/wiki/MANTRA-Rajbhasha.

3 Approaches to WSD

Many approaches are used to disambiguate the senses:

- *Supervised WSD*: In this approach needs of supervision. These approaches used trained data set of machine-learning techniques [4, 8, 18].
- *Semi-supervised WSD*: Semi-supervised or minimally supervised approaches do not need to a tagged corpus. Semi-supervised WSD used labeled and unlabeled data [10].
- *Unsupervised WSD*: It is based on unlabeled corpora, and does not courage any manually sense-tagged corpus to provide a sense choice for a word in context. Clustering comes under this approach [23, 24].
- Other WSD: Example based (or Instance based), Bootstrapping, AI based, Hybrid WSD (Combination of all WSD approaches or some approaches is known as hybrid WSD approaches).

Gathering of information needed for word sense disambiguation and every machine translation based on a different approach. Every WSD approaches do not do anything without knowledge [6]. Comparison of the Word Sense Disambiguation approaches with some specific criteria is shown in Table 1 [6, 9, 14, 15, 20].

Table 1. Comparison of supervised, semi-supervised and unsupervised WSD approach

Specifications	Supervised WSD	Semi-supervised WSD	Unsupervised WSD
Type of data used in WSD	Secondary	Primary and secondary	Primary
Main data	Labeled data	Small set of labeled data	Unlabeled data
Time	It is time-consuming approach	It takes less time to supervised WSD	It takes less time
Output	It gives relevant output	Sometimes gives relevant output	No guarantee for relevant output
Representation	Tagged data in text form	Small set of tagged data in text form	Untagged data in text form
Cost nature	Expansive	In between Supervised and Unsupervised WSD	Cheap
Algorithm	Naive Bayes (NB), K-nearest neighbor (K-NN), Support Vector Machine (SVM), Neural Network (NN)	Decision list	Agglomerative, Divisive, K-means, Bisecting K-means
Requirement	Collection of very large data set	Medium size of data set	Small data set

4 External Knowledge Resources

The basic component of WSD is Knowledge. Knowledge resources provide data which are essential to associate senses with words. They can vary from corpora of texts, either unlabeled or annotated with word senses, to machine-readable dictionaries, thesauri, glossaries, ontology's, etc. [4]. External knowledge resource can be two types one is structure and other is unstructured [8].

4.1 Structured Resources

Arrangement of data in some definite or fixed order in resources is known as structure data. For example, Dictionary follows an alphabetical order.

- *Machine-readable dictionaries (MRDs)*, which have become a popular source of knowledge for natural language processing since the 1980s, when the first dictionaries were made available in electronic format: among these, we cite the Collins English Dictionary, the Oxford Advanced Learner's Dictionary of Current English, the Oxford Dictionary of English [26], and the Longman Dictionary of Contemporary English (LDOCE) [27].
- *Thesauri* provided information about the relationships between words, like synonymy antonym and, possibly, further relations.
- *Ontology* is the specifications of conceptualizations of specific domains of interest [28], usually including a taxonomy and a set of semantic relations.
- *Glossary* is an alphabetical list of a particular domain; it appears at the end of the book.

4.2 Unstructured Resources

Lacking a definite structure of data in the unstructured resource. Wikipedia and corpus do not have any definite structure.

- *World Wide Web:* It is the huge collection of online data.
  - *Wikipedia:* It is a huge collection of articles and this article written by many different Indian and another language. Wikipedia, mainly differentiates in three articles which are Feature, Good, and normal articles [24].
  - *Search Engine Result:* It has collection of searched data.
- *Corpora:* Corpora can be sense-annotated or raw (i.e., unlabeled). Both kinds of resources are used in WSD and are most useful in supervised and unsupervised approaches, respectively [22]. WordNet: WordNet [29, 30] is a computational lexicon of English based on psycholinguistic principles, created and maintained at Princeton University. It encodes concepts in terms of sets of synonyms its latest version, WordNet 3.1, so the English WordNet is a collection of English synsets. WordNet as a graph whose nodes represented by synset and whose edges represented the semantic relation between synset [19, 21].

- *Raw Corpus:* It is a collection of data in text format. It can be anything, such as articles, stories, and poems.
- *Sense-Annotated Corpora:* It is a set of data, but data have sensed. This corpus is very helpful for supervised learning.
- *Collocation resources*, which register the tendency for words to occur regularly with others.

An external knowledge resource is a vital part of machine translation similarly, the structure identification of source and target language is also important.

5 Question Classification

Question classification [3] is a process to identify the types of question sentences (such as wh- question, key word specific etc.) and it helps in machine translation. Interrogative sentences always have question marks at the end of sentences, but all question sentences do not follow the same pattern. Classification of question sentences helpful to identify the structure of question sentences such as Wh- question, Keyword specific question, and anomalous verb related question. Question classification is the vital part of WSD.

A question which requires reasoning and thus long explanations are identified by keyword like WH-word, explain, discuss, justify etc.

- Question related to quantity
- Question related to time.
- Question-related to person or place.
- Questions regarding to different passages like kaun-kaun, kya-kya, vibhinn.
- Sentences that ask question are called interrogative sentences. Many types of question in exam question papers: Yes/no Interrogative question. Example: Did Ram go the game Friday night?
- Alternative question - Example: Did Ram go Patna or Lucknow on Friday night?
- Interrogative wh- type question- Example: What is Ram doing?
- Tag question: Tag question is questions attached or tagged onto the ending of a declarative statement. They transform a declarative sentence into an interrogative sentence. Declarative sentences become question: Example: Ram live in the city, don’t Ram?

The computer is not working?

- Subjective questions: fill in the blank- Example: Delhi is a capital of ______
- Objective questions: Many different types of question come under this category such as fill in the blank (Different types of notation such as —, …., , ( ), ____), matching, passage, true false. Example: Who amongst the following is considered to be the Father of ‘Local Self-Government’ in India? (a) Lord Dalhousie, (b) Lord Canning, (c) Lord Curzon, (d) Lord Ripon
- Keyword specific questions: In this types of the question have many different keywords. We collected some keyword from different resources such as exam question papers, online available data, and exercises of the book. Some keywords

are- Explain, discuss, justify, solve, find out, perform. Example: Describe the potential method for amortized analysis of an algorithm with a suitable example.

- Note type Question: Example: Write short notes on following:
  Cyber law in Indian context,
  Miller–Rabin method.

In modern English, only the anomalous verbs are normally inverted with the subject to form the interrogative [17].

6 Discussion on Question Sentence Ambiguity

This paper analyzes many types of question sentences in the English language which is discussed in Question Classification Section.

1. *Question sentences should be in text format:* Text format provides the uniformity to machine translation.
2. *Subject or area of the question sentences should be known:* It is very helpful in disambiguation because the meaning of the word also varies from subject to subject or context. For example word ring has a simple meaning in general language is a finger ring (Anguthi) but in physics, it has circulated (chhalla). Part of speech also changes the meaning of the word. Shabdkosh English to Hindi Dictionary shows the meaning of word "ring".
3. *Declare of nontranslate item (formula, abbreviation):* Non translates item helpful for understanding the meaning of context for example formula, abbreviation etc.
4. *Declare multiword expression:* Multiword expression (ME) or idiom is also change the meaning of the context. A collection of two or more words is known as the multiword expression. It is a pain in the natural language processing (NLP) for example "kick the bucket" means "to die (mar jaanaa)" in ME but in the word to word translation is "Balti ko laat maarnaa".
5. *Identification of the question type (WH question or key word specific):* Classification of question sentences helpful to identify the structure of question sentences such as Wh- question, Keyword specific question, and anomalous verb related question. Question classification is the vital part of WSD.

7 Conclusion

Machine translation is easy to disambiguate the word in simple translation because it identifies the pattern of sentences, but exam, question sentences does not have a pattern so the translation is more difficult. This paper reviews many types of question sentences and try to identify their pattern and discuss how to disambiguate their meaning. Questionable translation is very helpful for the student which is not friendly in English language.

References

1. Chaudhury, S., et al.: Anusaaraka: an expert system based machine translation system. In: 2010 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), Beijing, pp. 1–6. IEEE. Print ISBN 978-1-4244-6896-6. (2010). doi:10.1109/NLPKE.2010.5587789
2. Navigli, R.: Meaningful clustering of senses helps boost word sense disambiguation performance. In: Proceeding of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp. 105–112, Sydney. Association for Computational Linguistics (2006)
3. Kumar, P., et al.: A Hindi question answering system for e-learning documents. In: ICISIP, Third International Conference on Intelligent Sensing and Information Processing, 2005, Bangalore, pp. 80–85. IEEE. Print ISBN 0-7803-9588-3. (2005). doi:10.1109/ICISIP.2005.1619914
4. Navigli, R.: Word sense disambiguation: a survey. ACM Comput. Surv. **41**(2), Article 10 (2009)
5. VidhuBhaha, R.V., Abirami, S.: Trends in Word Sense Disambiguation. Springer, Berlin (2012)
6. Ponzetto, S.P., Navigli R.: Knowledge-rich word sense disambiguation rivaling supervised systems. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 1522–1531. Association for Computational Linguistics (2010)
7. Navigli, R., Crisafulli, G.: Inducing word senses to improve web search result clustering. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 116–126. Association for Computational Linguistics (2010)
8. Faralli, S., Navigli, R.: A new minimally-supervised framework. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, 12–14 July 2012, pp. 1411–1422. Association for Computational Linguistics (2012)
9. Yadav, R.K., Gupta, D.: Annotation guidelines for Hindi–English word alignment. In: 2010 International Conference on Asian Language Processing, 978-0-7695-4288-1/10. IEEE (2010). doi:10.1109/IALP.2010.58
10. Mishra, N., Mishra, A.: Part of speech tagging for Hindi corpus. In: 2011 International Conference on Communication Systems and Network Technologies, 978-0-7695-4437-3/11. IEEE (2011). doi:10.1109/CSNT.2011.118
11. Agarwal, M., Bajpai, J.: Correlation-based word sense disambiguation. In: 2014 Seventh International Conference on Contemporary Computing (IC3), pp. 382–386. IEEE Conference Publications. (2014). doi:10.1109/IC3.2014.6897204
12. Ahmed, P., Dev, A., Agrawal, S.S.: Hindi speech corpora: a review. In: Oriental COCOSDA held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference, pp. 1–6. IEEE (2013)
13. Bansal, S., et al.: Determination of linguistic differences and statistical analysis of large corpora of Indian languages. In: Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference, pp. 1–5. IEEE (2013)
14. Jurafsky, D., Martin, J.H.: Word sense disambiguation and Information Retrieval. In: Speech and Language Processing an Introduction to Natural Language Processing Computational Linguistics, and Speech Recognition. Pearson Education, Inc. ISBN 678-81-317-1672-4 (2000)

15. Pool, D., Mackworth, A., Goebel, R.: Computational Intelligence—A Logical Approach. Oxford University Press Inc. (1998). Fourth Impression (2012). ISBN 13:978-0-19-568572-5, ISBN 10:0-19-568572-5
16. Bharti, A., Chaitanya, V., Sangal, R.: Hindi grammar. In: Natural Language Processing—A Panninial Perspective. India Prentice-Hall (2010)
17. Wren, P.C., Martin, H., Prasada Rao, N.D.V.: High School English Grammar & Composition. S. Chand & Company Ltd., New Delhi (2010). ISBN 81-219-2197-X
18. Fulmari, A., et al.: A survey on supervised learning for word sense disambiguation. Int. J. Adv. Res. Comput. Commun. Eng. **2**(12), (2013). (ISSN (Print) 2319–5940, ISSN (Online) 2278–1021)
19. Navigli, R., Lapata: Graph Connectivity Measures for Unsupervised Word Sense Disambiguation. In: IJCAI 2007, pp. 1683–1688 (2007)
20. Mante, R., et al.: A review of literature on word sense disambiguation. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) **5**(2), 1475–1477 (2014). (ISSN 0975-9646)
21. Kolte, S.G., et al.: WordNet: a knowledge source for word sense disambiguation. Int. J. Recent Trends Eng. **2**(4), 213–217 (2009)
22. Montoyo, A., et al.: Combining knowledge- and corpus-based word-sense-disambiguation methods. J. Artif. Intell. Res. **23**, 299–330 (2005).
23. Navigli, R., Lapata, M.: An experimental study of graph connectivity for unsupervised word sense disambiguation. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 678–692 (2010)
24. Turdakov, D.Y.: Word sense disambiguation methods. In: Programming and Computer Software, vol. 36, no. 6, pp. 309–326 (2010). ISSN 0361_7688 (Pleiades Publishing, Ltd., Original Russian Text ©, published in Programmirovanie)
25. Hutchins, W.J.: The development and use of machine translation systems and computer-based translation tools. International Conference on Machine Translation & Computer Language Information Processing. Proceedings of the conference, 26–28 June 1999, Beijing, China, ed. Chen Zhaoxiong, pp. 1–16 (1999)
26. Soane, C., Stevenson, A.: Oxford English Dictionary (2003)
27. Procter, P.: Longman Dictionary of Contemporary English Dictionary of Contemporary English. Cartermills Publishing (1978)
28. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquisition **5**(2), 199–220 (1993)
29. Fellbaum, C.: WorldNet. John Wiley & Sons, Inc. (1998)
30. Miller, et al.: Introduction to WordNet: an on-line lexical database. Int. J. Lexicography **3**(4), 235–244 (1990)