Scientific Journal of Impact Factor (SJIF): 4.14
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 12, December -2016

A Study of Main Elements of Word Sense Disambiguation (WSD) for Gujarati Language

Smt Rekha Manish Shah

Computer Engineering Department, Government Polytechnic, Ahmedabad.

Abstract— *WSD is a open problem in the field of natural language processing(NLP). WSD is also considered as an AI-complete problem. The importance of WSD has been widely acknowledged in computational Linguistics.This papers presents a survey on main elements of WSD and various approaches of WSD that are adopted in various research works.*

Keywords— *Word Sense Disambiguation(WSD), context, corpora, supervised WSD, unsupervised WSD*

1. INTRODUCTION

In all the major languages, the task of getting the correct meaning of a word in a particular context, which is very obvious for a human being, is the most difficult problem for the machine. Word sense disambiguation (WSD) is the ability to identify the appropriate meaning of words in context using computational manner. WSD is considered an AI-complete problem. This paper introduces the reader to the main elements of WSD and provides a description of all the elements. It overviews various approaches like supervised and unsupervised, *knowledge-based* (or *dictionary based*) and *corpus-based* approaches. The evaluation measures and baselines employed for evaluation of WSD systems is also discussed. Finally, various real world applications of WSD are discussed at the end of the paper.

2. TASK DESCRIPTION

Many words in Human language are ambiguous, so that these words can be interpreted in multiple ways depending on the context in which they occur. Consider the following sentences:
(a) He lost his *right* leg in the accident.
(b) Your answer is *right*.
The word *right* in the two sentences clearly denote different meanings: the side of the body and true or correct, respectively. Unfortunately, the identification of the specific meaning that a word assumes in context is only apparently simple. While most of the time humans do not even think about the ambiguities of language, machines need to process unstructured textual information and transform them into data structures which must be analyzed in order to determine the underlying meaning. The computational identification of meaning for words in context is called *word sense disambiguation* (*WSD*).

WSD has been described as an AI-complete problem [Mallery 1988], that is, by analogy to NP-completeness in complexity theory, a problem whose difficulty is equivalent to solving central problems of *artificial intelligence* (AI), for example, the Turing Test [Turing 1950]. Its acknowledged difficulty does not originate from a single cause, but rather from a variety of factors.

Word sense disambiguation is the ability to computationally determine which sense of a word is activated by its use in a particular context. If we disregard the punctuation, we can view a text T as a sequence of words (w_1, w_2, . . . , w_n), and we can formally describe WSD as the task of assigning the appropriate sense(s) to all or some of the words in T, that is, to identify a mapping A from words to senses, such that $A(i) \subseteq Senses_D(w_i)$, where $Senses_D(w_i)$ is the set of senses encoded in a dictionary D for word w_i , and $A(i)$ is that subset of the senses of w_i which are appropriate in the context T. The mapping A can assign more than one sense to each word $w_i \in T$, although typically only the most appropriate sense is selected, that is, $|A(i)| = 1$.

WSD can be viewed as a classification task: word senses are the *classes*, and an *automatic classification method* is used to assign each occurrence of a word to one or more classes based on the evidence from the *context* and from *external knowledge sources*.

There are two variants of the generic WSD task:

Lexical sample* or *targeted WSD, where a system is required to disambiguate a restricted set of target words usually occurring one per sentence. Supervised systems are typically employed in this setting, as they can be trained using a number of hand-labeled instances (*training set*) and then applied to classify a set of unlabeled examples (*test set*);

All-words WSD, where systems are expected to disambiguate all open-class words in a text (i.e., nouns, verbs, adjectives, and adverbs). This task requires wide-coverage systems.

Consequently, purely supervised systems can potentially suffer from the problem of *data sparseness*, as it is unlikely that a training set of adequate size is available which covers the full lexicon of the language of interest. On the other hand, other approaches, such as knowledge-lean systems, rely on full-coverage knowledge resources, whose availability must be assured.

3. ELEMENTS OF WSD

There are four main elements of WSD:

A. the selection of word senses (i.e., classes),
B. the use of external knowledge sources,
C. the representation of context, and
D. the selection of an automatic classification method.

A. Selection of Word Senses

A *word sense* is a commonly accepted meaning of a word. For instance, consider the following two sentences:
(1)The chef chopped the vegetables with his ***knife***.
(2) A guard was beaten and cut with a ***knife by the robber***.
The word *knife* is used in the above sentences with two different senses: a tool and a weapon. The two senses are clearly related, as they possibly refer to the same object; however the object's intended uses are different. The examples make it clear that determining the sense inventory of a word is a key problem in word sense disambiguation.

A *sense inventory* partitions the range of meaning of a word into its senses. Word senses cannot be easily discretized, that is, reduced to a finite discrete set of entries, each encoding a distinct meaning. The main reason for this difficulty stems from the fact that the language is inherently subject to change and interpretation. Also, given a word, it is arguable where one sense ends and the next begins. As a result of such uncertainties, different choices will be made in different dictionaries.

Moreover, the required granularity of sense distinctions might depend on the application. For example, there are cases in machine translation where word ambiguity is preserved across languages (e.g., the word *interest* in English, Italian, and French). As a result, it would be superfluous to enumerate those senses (e.g., the financial vs. the pastime sense), whereas in other applications we might want to distinguish them (e.g., for retrieving documents concerning financial matters rather than pastime activities).

While ambiguity does not usually affect the human understanding of language, WSD aims at making explicit the meaning underlying words in context in a computational manner. Therefore it is generally agreed that, in order to enable an objective evaluation and comparison of WSD systems, senses must be enumerated in a sense inventory (*enumerative approach*). All traditional paper-based and machine-readable dictionaries adopt the enumerative approach.

B. External Knowledge Sources

Knowledge is a fundamental component of WSD. Knowledge sources 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, ontologies, etc. Here is a brief overview of these resources:

Structured resources:

- ***Thesauri***, which provide information about relationships between words, like synonymy, antonymy and, possibly, further relations.
- Machine-***readable dictionaries (MRDs),***
- ***Ontologies,*** which are specifications of conceptualizations of specific domains of interest, usually including taxonomy and a set of semantic relations.

Unstructured resources:

- ***Corpora***, that is, collections of texts used for learning language models. 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.
- ***Raw corpora:*** the Brown Corpus, a million word balanced collection of texts published in the United States in 1961; the British National Corpus (BNC) , a 100 million word collection of written and spoken samples of the English language (often used to collect word frequencies and identify grammatical relations between words); the *Wall Street Journal* (WSJ) corpus, a collection of approximately 30 million words from WSJ; the American National Corpus, which includes 22 million words of written and spoken American English; the Gigaword Corpus, a collection of 2 billion words of newspaper text, etc.

- ***Sense-Annotated Corpora:*** SemCor, the largest and most used sense-tagged corpus, which includes 352 texts tagged with around 234,000 sense annotations; MultiSemCor , an English-Italian parallel corpus annotated with senses from the English and Italian versions of WordNet; the *linehard-serve* corpus containing 4000 sense-tagged examples of these three words (noun, adjective, and verb, respectively); the *interest* corpus with 2369 sense-labeled examples of noun *interest*; the DSO corpus, produced by the Defence Science Organisation (DSO) of Singapore, which includes 192,800 sense-tagged tokens of 191 words from the Brown and *Wall Street Journal* corpora; the Open Mind Word Expert data set, a corpus of sentences whose instances of 288 nouns were semantically annotated by Web users.
- ***Collocation resources***, which register the tendency for words to occur regularly with others: examples include the Word Sketch Engine, JustTheWord, The British National Corpus collocations, the Collins Cobuild Corpus Concordance, etc. A huge dataset of text co-occurrences has been released, which has rapidly gained a large popularity in the WSD community, namely, the Web1T corpus. The corpus contains frequencies for sequences of up to five words in a one trillion word corpus derived from the Web.
- Other resources, such as word frequency lists, *stoplists* (i.e., lists of undiscriminating noncontent words, like *a*, *an*, *the*, and so on), *domain labels*, etc.

C. Representation of Context

As text is an unstructured source of information, to make it a suitable input to an automatic method it is usually transformed into a structured format. A preprocessing of the input text is usually performed, which typically includes the following steps:

- ***tokenization***, splits up the text into a set of tokens (usually words);
- ***part-of-speech tagging***, consisting in the assignment of a grammatical category to each word (e.g., "the/DT boys/NN are/VBD running/JJ," where DT, NN, VBD and JJ are tags for determiners, nouns, verbs, and adjectives, respectively);
- ***lemmatization***, that is, the reduction of morphological variants to their base form (e.g. was → be, boys → boy);
- ***chunking***, which consists of dividing a text in syntactically correlated parts.
- ***parsing***, whose aim is to identify the syntactic structure of a sentence (usually involving the generation of a parse tree of the sentence structure).

Here is an example of the processing flow:

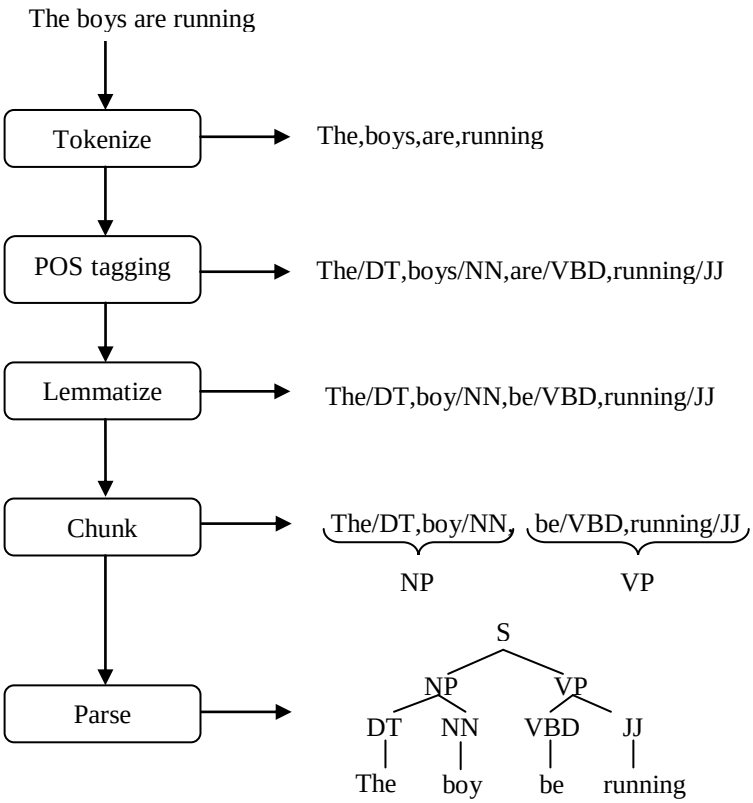

Figure 1 Processing Flow Of Input Text

As a result of the preprocessing phase of a portion of text (e.g., a sentence, a paragraph, a full document, etc.), each word can be represented as a vector of features of different kinds or in more structured ways, for example, as a tree or a graph of the relations between words. The representation of a word in context is the main support, together with additional knowledge resources, for allowing automatic methods to choose the appropriate sense from a reference inventory.

A set of features is chosen to represent the context. These include information resulting from the above-mentioned preprocessing steps. We can group these features as follows:

- ***local features***, which represent the local context of a word usage, that is, features of a small number of words surrounding the target word, including part-of-speech tags, word forms, positions with respect to the target word, etc.;
- ***topical features***, which—in contrast to local features—define the general topic of a text or discourse, thus representing more general contexts (e.g., a window of words, a sentence, a phrase, a paragraph, etc.), usually as bags of words;
- ***syntactic features***, representing syntactic cues and argument-head relations between the target word and other words within the same sentence (note that these words might be outside the local context);
- ***semantic features***, representing semantic information, such as previously established senses of words in context, domain indicators, etc.

Based on this set of features, each word occurrence (usually within a sentence) can be converted to a feature vector.

It must be noted that choosing the appropriate size of context (both in structured and unstructured representations) is an important factor in the development of a WSD algorithm, as it is known to affect the disambiguation performance.

D. Choice of a Classification Method

The final step is the choice of a classification method. Most of the approaches to the resolution of word ambiguity stem from the field of machine learning, ranging from methods with strong supervision, to syntactic and structural pattern recognition approaches. We can broadly distinguish two main approaches to WSD:

- ***supervised WSD***: these approaches use machine-learning techniques to learn a classifier from labeled training sets, that is, sets of examples encoded in terms of a number of features together with their appropriate sense label (or class);
- ***unsupervised WSD*:** these methods are based on unlabeled corpora, and do not exploit any manually sense-tagged corpus to provide a sense choice for a word in context.

There is further distinction between *knowledge-based* (or *knowledge-rich*, or *dictionary based*) and *corpus-based* (or *knowledge-poor*) approaches. The former rely on the use of external lexical resources, such as machine-readable dictionaries, thesauri, ontologies, etc., whereas the latter do not make use of any of these resources for disambiguation.

Finally, WSD approaches can be categorized as *token-based* and *type-based*. Token based approaches associate a specific meaning with each occurrence of a word depending on the context in which it appears. In contrast, type-based disambiguation is based on the assumption that a word is consensually referred with the same sense within a single text. Consequently, these methods tend to infer a sense (called the *predominant sense*) for a word from the analysis of the entire text and possibly assign it to each occurrence within the text.

4. EVALUATION METHODOLOGY

However, one of the real objectives of WSD is to demonstrate that it improves the performance of applications such as information retrieval, machine translation, etc. Here is described the evaluation measures and baselines employed for evaluation of WSD systems.

A. Evaluation Measures

The assessment of word sense disambiguation systems is usually performed in terms of evaluation measures borrowed from the field of information retrieval.

*The **coverage C** is defined* as the percentage of items in the test set for which the system provided a sense assignment that is:

$$C = \frac{\#\text{answers provided}}{\#\text{total answers to provide}}$$

The ***precision P*** of a system is computed as the percentage of correct answers given by the automatic system, that is:

$$P = \frac{\#\text{correct answers provided}}{\#\text{ answers provided}}$$

Recall R is defined as the number of correct answers given by the automatic system over the total number of answers to be given:

$$R = \frac{\#\text{correct answers provided}}{\#\text{total answers to provide}}$$

According to the above definitions $R \leq P$. When coverage is 100%, $P = R$. In the WSD literature, recall is also referred to as *accuracy*.

B. Baselines

A baseline is a standard method to which the performance of different approaches is compared. Here two basic baselines, the random baseline and the first sense baseline, are presented.

- ***The Random Baseline:*** Let D be the reference dictionary and $T = (w_1, w_2, \ldots, w_n)$ be a test set such that word w_i ($i \in \{1, \ldots, n\}$) is a content word in the corpus. The *chance* or *random baseline* consists in the random choice of a sense from those available for each word w_i . Under the uniform distribution, for each word w_i the probability of success of such a choice is $(1 / |Senses_D(w_i)|)$.

The accuracy of the random baseline is obtained by averaging over all the content words in the test set T:

$$Acc_{Chance} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Senses_D(w_i)|}$$

- ***The First Sense Baseline:*** The *first sense baseline* (or *most frequent sense baseline*) is based on a ranking of word senses. This baseline consists in choosing the first sense for each word in a corpus, independent of its context. For instance, in WordNet, senses of the same word are ranked based on the frequency of occurrence of each sense in the SemCor corpus.

C. Lower and Upper Bounds:

Lower and upper bounds are performance figures that indicate the range within which the performance of any system should fall. Specifically, a lower bound usually measures a performance obtained with an extremely simple method and which any system should be able to exceed. A typical lower bound is the random baseline. This baseline poses serious difficulties to WSD systems as it is often hard to beat.

An *upper bound* specifies the highest performance reasonably attainable. In WSD, a typical upper bound is the *interannotator agreement* or *intertagger agreement* (ITA), that is, the percentage of words tagged with the same sense by two or more human annotators. The interannotator agreement on coarse-grained, possibly binary, sense inventories is calculated around 90%, whereas on fine-grained, WordNet-style sense inventories the inter-annotator agreement is estimated between 67% and 80%.

D. Evaluation: The Senseval/Semeval Competitions

Comparing and evaluating different WSD systems is extremely difficult, because of the different test sets, sense inventories, and knowledge resources adopted. Before the organization of specific evaluation campaigns, most systems were assessed on in-house, often small-scale, data sets. Therefore, most of the pre-Senseval results are not comparable with subsequent approaches in the field.

Senseval (now renamed *Semeval*) is an international word sense disambiguation competition, held every three years since 1998. The objective of the competition is to perform a comparative evaluation of WSD systems in several kinds of tasks, including all-words and lexical sample WSD for different languages, and, more recently, new tasks such as semantic role labeling, gloss WSD, lexical substitution, etc.

The systems submitted for evaluation to these competitions usually integrate different techniques and often combine supervised and knowledge-based methods (especially for avoiding bad performance in lack of training examples). The Senseval workshops are the best reference to study the recent trends of WSD and the future research directions in the field. Moreover, they lead to the periodic release of data sets of high value for the research community. (http://alt.qcri.org/semeval2016/)

5. APPLICATIONS

There are numerous real-world applications which might get benefit from WSD and can improve their performance.

Machine Translation (MT)
The automatic identification of the correct translation of a word in context, that is, machine translation (MT), is a very difficult task. Word sense disambiguation has been considered as the main task to be solved in order to enable machine translation, based on the idea that the disambiguation of texts should help translation systems choose better candidates. In fact, depending on the context, words can have completely different translations. For instance, the English word kite can be translated in Gujarati as પતંગ, સમડી, કાઈટ ચલચિત્ર etc. Unfortunately, WSD has been much harder than expected, as we know after years of comparative evaluations.

Information Retrieval (IR)
It is very important to resolve ambiguity in a query before retrieving information. As for example, a word "depression" in a query may has different meanings like illness,
weather systems, or economics. So, finding the exact sense of an ambiguous word in a particular question before finding its answer is the most vital issue in this area.

Text Mining
In specific domains it is interesting to distinguish between specific instances of concepts: for example, in the medical domain we might be interested in identifying all kinds of drugs across a text, whereas in bioinformatics we would like to solve the ambiguities in naming genes and proteins. Tasks like named-entity recognition (NER), acronym expansion (e.g., MP as member of parliament or military police), etc., can all be cast as disambiguation problems.

Word Processing
Word processing is a relevant application of natural language processing, whose importance has been recognized for a long time. Word sense disambiguation can aid in correcting the spelling of a word, for case change, or to determine when anuswara should be inserted for Indian languages (e.g., in Gujarati, for changing મા (= mother) to માં(= into), or મદ(=proud) to મંદ(=slow), based on semantic evidence in context about the correct spelling). Given the increasing interest in regional languages in NLP, WSD might play an increasingly relevant role in the determination and correction of words.

Content Analysis
WSD can be applied to analyze the context of a text in terms of its area, subject, ideas etc. For instance, the classification of blogs has recently been gaining more and more interest within the Internet community: as blogs grow at an exponential pace, we need a simple yet effective way to classify them, determine their main topics, and identify relevant (possibly semantic) connections between blogs and even between single blog posts. A second related area of research is that of (semantic) social network analysis, which is becoming more and more active with the recent evolutions of the Web.

6. CONCLUSIONS

This paper summarized the main elements of WSD and provided a description of all the elements. It also classified existing WSD algorithms according to their techniques, like supervised and unsupervised, *knowledge-based* (or *dictionary based*) and *corpus-based* approaches. The evaluation measures and baselines employed for evaluation of WSD systems is also discussed. Finally, various real world applications of WSD are discussed at the end of the paper.

7. ACKNOWLEDGMENT

I am very much thankful to all the esteemed authors in a reference list, for their hard work that helped me to continue my research in a better shape and in a right direction.

REFERENCES

[1] Xiaohua Zhou and Hyoil Han, College of Information Science & Technology, Drexel University, "Survey of Word Sense Disambiguation Approaches"

[2] Alok Ranjan Pal and Diganta Saha, International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.5, No.3, July 2015, " Word Sense Disambiguation: A Survey"

[3] Roberto Navigli, Universit `a di Roma La Sapienza, "Word Sense Disambiguation: A Survey"

[4] J.Sreedhar , Dr.S.Viswanadha Raju, Dr.A.Vinaya Babu, International Journal of Scientific & Engineering Research, Volume 5, Issue 6, June-2014 555 ISSN 2229-5518, "A Study of Critical Approaches in WSD for Telugu Language Nouns: Current State of the Art"

[5] Massimiliano Ciaramita, A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Cognitive and Linguistic Sciences at Brown University May 2005, "Broad-Coverage Hierarchical Word Sense Disambiguation"

[6] Aparna Konduri, A Thesis Presented to The Graduate Faculty of the University of Akron, "Clustering Of Web Services Based On Semantic Similarity"

[7] Jonas Ekedahl, Engineering Physics, Lund Univ, Koraljka Golub KnowLib, Dept. of IT, Lund Univ.Sweden, "Word sense disambiguation using WordNet and the Lesk algorithm"

[8] Salini M T, department of computer science, cochin university of science and technology ,kochi, "Word Sense Disambiguation"

[9] Mitesh M. Khapra, "Word Sense Disambiguation"

[10] Rada Mihalcea and Ted Pedersen, 'Slides from the AAAI 2005 Tutorial – Advances in Word Sense Disambiguation' http://www.d.umn.edu/~tpederse/WSDTutorial.html

[11] Eneko Agirre & Philip Edmonds, Word Sense Disambiguation: Algorithms and Applications

[12] Yarowsky, David. 2000. "Word sense disambiguation. *Handbook of Natural Language Processing"*, ed. by Dale et al., 629-654. New York: Marcel Dekker.

