Supervised Word Sense Disambiguation for Urdu Using Bayesian Classification

Asma Naseer
National University of Computer and Emerging Sciences (Fast)
B Block, Faisal Town, Lahore, Pakistan
asma.naseer@nu.edu.pk

Sarmad Hussain
Al-Khwarzmi
University of Engineering & Technology Lahore
sarmad.hussain@nu.edu.pk

Abstract

In almost all languages words usually have multiple senses or meanings. WSD (Word Sense Disambiguation) is a task of recognizing the correct sense of a word in a particular context. Identifying correct sense of an ambiguous word becomes very vital when a language is needed to be translated into another language or information is needed to be extracted using ambiguous words. There is a massive work in English and other languages for resolving word ambiguity. So far as Urdu is concerned there is not any cited work for resolving ambiguity of words. In this paper a statistical approach i.e. "Bayesian Classification" is applied for resolving this peculiar type of lexical ambiguity, called Word Sense Disambiguation, for some URDU words.

1. Introduction

The process of identifying the correct sense or meaning of a word in a particular context is called Word Sense Disambiguation. When a human being is encountered with a word with multiple senses he easily identifies the exact sense of the word with the help of context without giving a single thought to the other senses. But when the same situation is provided to a computer it is not an easy task to correctly identify the desired sense. WSD process helps in resolving such ambiguity issues. Sometimes a word differs in meaning when its POS is different. For example *butter* can be a verb or a noun as it can be seen in the following example:

Will you spread butter [Noun] on toast?
Don't think you can butter [Verb] me up that easily.

In one sentence butter as a noun means "*a solid yellow food made from milk or cream*" [14], while in the other sentence butter as a verb means "*to say nice things to someone so that they will do what you want*" [14]. As such ambiguities can easily be resolved with the help of POS, WSD does not entertain such words. The word with different meanings having same POS needs some WSD process to conclude the accurate sense. For example *Bank* in English can be "*the rising ground bordering a river or stream etc.*" or "*financial institute*".

The current work focuses on implementing WSD process for Urdu language. Urdu is spoken by more than a 100 million people [1]. It is an Indo-Aryan language and is also the national language of Pakistan and one of the national languages of India. The current work is an initial step to resolve the ambiguity of words in Urdu context. In the current work, four words of Urdu, one noun and three verbs, are focused. The technique that is implemented to resolve ambiguity is Bayesian Classification.

The structure of the rest of the paper is as follows: In section 2 a brief history of related work is described. Section 3 and 4 illustrate corpus and methodology. Results and discussions are presented in sections 5 and 6 respectively. Section 7 describes future enhancement.

2. Previous Research

For sense disambiguation two different approaches exist i.e. supervised and unsupervised [15]. Supervised approach is a classification task while unsupervised approach is a clustering task. The classification approach of WSD makes use of statistical approaches either referring lexicons or using corpus for training. Thesauri, lexicons and corpus are the main source of training in the supervised approach. The unsupervised techniques use different machine learning algorithms like EM (Expectation Maximization), Decision Trees and Neural Networks etc. and resolve ambiguity by making clusters of different senses.

Both types of approaches mentioned above have been used by different researchers for different languages. A brief description of these researches is given below.

In the early efforts of disambiguating senses of words, handmade dictionaries were used [4, 5]. Such efforts performed very well on trained data but on large scale their performance collapsed. In 1986 an online dictionary was used for WSD [6] which was applicable to large context. In this approach senses were identified using sense definition in dictionary. The technique proved very successful in IR. A thesaurus based approach is employed by Walker [7] which decides the sense of a word by identifying the semantic category of the word. For this the semantic category of the context as a whole is decided. The algorithm does not perform well in different domains. Black [9] achieved 50% accuracy on 5 difficult and highly ambiguous words using thesaurus based approach. Another interesting approach is translation of context into another language [10]. Such a translation can give a clear clue to the sense of an ambiguous word. Yarowsky et al. [3] used Bayesian classifier to disambiguate 6 nouns *(duty, drug, land, language, position, and sentences)*. They achieved accuracy around 90%.

Manish Sinah et al. has used Wordnet for Hindi for word sense disambiguation [8]. Their statistical approach identifies the correct sense of nouns in Hindi language. The overlapping of the words in the context and the information in Hindi Wordnet decides the appropriate sense. The application of the process on different domains reveals accuracy varying from 40% to 70%.

Besides supervised word sense disambiguation, unsupervised WSD (clustering) has also been used in different languages. Such type of an approach is used in [11]. In this technique the record for the context of an ambiguous word is kept. For this a weighted list of distributed similar words, based on the syntactic context of the ambiguous word, is built. The precision attained with this approach is 69.86%.

The results of different approaches of WSD reveal that the accuracy of supervised WSD techniques using dictionaries and thesauri is not much elevated. The usage of Hindi Wordnet [8] achieves maximum 70% accuracy while the usage of thesauri by Black [9] results in only 50% accuracy. Even the performance of unsupervised algorithms for WSD does not reveal high accuracy [11]. Among all these techniques the one with better performance is Bayesian Classification which can achieve accuracy up till 90% [3].

In this paper Bayesian Classifier is used for resolving lexical ambiguity of words because of its accuracy and consistency in performance. This method has been used for different classification tasks but for WSD it was first used in 1992 by Gale et al.

3. Corpus

The corpus used for training and testing is taken from [13]. It consists of 18 million words and encompasses different domains like sports, news, finance, culture, consumer information and personal communication etc. For WSD process four words were selected based on high and medium frequency from the respective corpus. Table 1 describes the number of sentences tagged for each word.

Table 1: Number of sense tagged sentences per word

Words	# of Tagged Sentences
کھلانا *khalana* Verb	80
بجانا *bajana* Verb	1,115
بھاگنا *bhagna* Verb	1,194
ملک *mlk* Noun	21,528

4. Methodology

For WSD process four words, three verbs and one noun, are selected form the corpus (Table 2 contains information about these words). As Urdu is an agglutinative language it has got different morphological forms of verbs. The forms that a verb may contain are more than 40. For example the word بھاگنا *bhagna* (to run) can have different morphological forms such as:

بھاگ *bhag* (ran)
بھاگی *bhagi* (ran –female singular)
بھاگا *bhaga* (ran –male singular)
بھاگے *bhagay* (ran –male plural)
بھاگیں *bhagin* (ran –female plural)

Among all these forms only the base form is considered for all the three verbs (for verbs base form is the one that ends with نا *na*, and for noun its male singular).

For each of the four words all of the sentences from the corpus containing these words are fetched and manually sense tagged. From theses tagged sentences 80% sentences are used for training and remaining 20% are utilized for testing. For training and testing Bayesian Classifier is used. The following algorithm describes the classifier:

4.1. Algorithm

Training
Look at the words around an ambiguous word in a context window.
for all senses s_k of w do
  for all v_j in vocabulary do
    $P(v_j|s_k) = C(v_j,s_k)/C(s_k)$
  end
end
for all senses s_k of w do
  $P(s_k) = C(s_k)/C(w)$
End

Disambiguation
for all senses s_k of w do
  $score(s_k) = P(s_k)$
  for all v_j in context window c do
    $score(s_k) = score(s_k) + P(v_j|s_k)$
  end
end
choose $argmax_{sk}$ score (s_k) [15]

4.2. Training

For WSD process four words were selected, based on high and medium frequencies, from the corpus [13]. Among the four words one is Noun and the other three are Verbs (the detail of these words and their senses is illustrated in Table 2).

Table 2: Words and their senses

Words	Sense 1	Sense 2	Sense 3	Sense 4	Sense 5
کھلانا *khalana* Verb	To make one eat	To make one play	-	-	-
بجانا *bajana* Verb	Play (music)	Knock	Striking of Clock	-	-
بھاگنا *bhagna* Verb	Run	Elope	Escape	Avoid	Rush
ملک *mlk* Noun	Country	Angel	Milk	Name of a Cast	-

For training all the sentences of a particular word are stored separately and then sense tagged. With the help of these tagged sentences bag of words are created for each sense of each word using different window sizes i.e. 3x3, 5x5, 7x7. For all the words which occur in the bag of words of a particular sense, frequencies are computed and stored against these words of each sense.

Each word in the context of the ambiguous word contributes potentially useful information about that sense of the ambiguous word which is most likely to occur. So for training these content words are chosen and their frequencies for each sense are kept in record. As the frequencies of different senses of a word may have a significant difference, it was decided that the context beyond the sentence boundary should not be considered.

For POS tagging a short context can be sufficient but so far as sense disambiguation is concerned a broader context may be required. So the experiment is made with different window sizes i.e. 3x3, 5x5 and 7x7 (n words from the left of the ambiguous word and n words from the right, where n = {3, 5, 7}). For all the words which occur in the bag of words of a particular sense, frequencies are computed and stored against these context words of each sense. These frequencies are looked up and utilized during testing. As the frequencies of different senses of a word have a significant difference, it was decided that the context beyond the sentence boundary should not be considered. For example the sense country of the word ملک *mlk* occurred 17,845 times. On the other hand the sense Angel of the same word occurred only for 4 times. If the context beyond the boundary of the sentence is considered then the highly frequently occurring senses can bias the bag of words of low frequency senses.

The sense with the maximum probability among all the candidate senses is declared as the sense of the word. To calculate the probabilities Equation 4.1 through 4.5 are used [15]. The sense with the maximum probability among all the candidate senses is declared as the sense of the word. To calculate the probabilities Equation 4.1 through 4.5 are used [15].

$$s' = argmax_{sk}\, P(s_k|c) \quad \text{(4.1)}$$

$$s' = argmax_{sk}\, \frac{P(c|s_k)}{P(c)}\, P(sk) \quad \text{(4.2)}$$

s' is the sense with the maximum probability which is determined while considering the probabilities of all the other senses as well. *C* is the context of the given ambiguous word while s_k is k^{th} sense of the word.

As the classifier depends on the assumption that all the feature variables (words in the context) are conditionally independent, the Equation 4.3 is used which implements the said assumption.

$$P(c|s_k) = P(\{v_j|\ v_j\ in\ c\}|\ s_k) = \prod_{vj}\ in\ c\ P(v_j\ |s_k)$$ ___ (4.3)

v_j is the j^{th} word in the context *C*.

$$P(\ v_j\ |s_k) = \frac{C(v_j,s_k)}{C(s_k)}$$ ___ (4.4)

Probability of j^{th} word in the context of k^{th} sense of the word is the count of the number of times v_j occurs in the context of k^{th} sense divided by $C(s_k)$ which is the total count of the occurrences of the k^{th} sense.

$$P(c_k) = \frac{C(s_k)}{C(w)}$$ ___ (4.5)

Equation 4.5 uses the count of two features i.e. count of the occurrences of k^{th} sense and the count of the total occurrences of the ambiguous word *w*.

The context words are used as bag of words by taking the assumption of independence as already described (Equation 4.3). After calculating all the probabilities for contextual words the calculations are stored in files. Files are maintained for each sense of the four ambiguous words. Besides contextual words' frequencies the frequency of sense itself is also stored. For different contextual window size (3x3, 5x5, 7x6) the bag of words were different so files were also maintained separately for each window size.

As for all the words frequencies are computed and stored against each context word of the sense so during testing these frequencies are just looked up and implanted in the formulae during testing.

4.3. Testing (Disambiguation)

As all the frequencies already got calculated during training phase, in testing (disambiguation) phase only look up and basic arithmetic operations were required.

Testing phase got completed in three stages, each for a different window size. In each testing stage bags of words based on respective window size were looked up, and the frequencies stored against them were employed in equation 4.2 to calculate argmax among all the sense of the respective word.

Like many other applications of Natural Language Processing WSD also faces the problem of sparseness. To resolve the issue a smoothing technique presented in [2] is used. Equation 4.6 describes the respective formula for smoothing.

$$P(v_j|s_k) = \frac{C(v_j,s_k)+ \alpha*\#Senses*P(s_k)}{C(s_k)+ \alpha*\#Senses*P(s_k)*|Vocabulary|}$$ ___ (4.6)

α is a minor value in fraction that is added to the probability of each v_j in the context. For normalizing the probabilities α is |*vocabulary*| times added in the denominator.

Testing was made on the remaining 20% of the data which was not used in training phase. This data was also manually sense tagged for the verification of the process but the tagging was kept separate and only sentences were fed to the system for annotation. Thus the output of WSD process is verified with the remaining 20% manually sense tagged data.

5. Results

For POS tagging a short context can be sufficient but so far as sense disambiguation is concerned a broader context may be required. So the experiment is made with different window sizes i.e. 3x3, 5x5 and 7x7 (n words from the left of the ambiguous word and n words from the right, where n = {3, 5, 7}).

The algorithm outperformed for the word ملک *mlk* as it has got very high frequency in the corpus, but the words with less occurrences revealed comparatively less accuracy.

The increase in window size also resulted in a better performance. Maximum precision (98.35%) and recall (92.17%) are resulted for 7x7 window. The system is evaluated on the basis of precision, recall and F-Measure with α = 0.5. The overall performance of the system with different window sizes is described in Table 3.

Table 3: Results for different window sizes

Window	Precision	Recall	$F_{\alpha=0.5}$
3x3	92.38 %	85.58 %	88.85 %
5x5	96.20 %	89.79 %	92.88 %
7x7	98.35 %	92.17 %	95.15 %

6. Discussion

Two factors affected the accuracy of the system i.e. window size and overlapping of the context of different senses of a particular word. Table 3 reveals gradual improvement in the accuracy of the system, as the window size increases from 3x3 to 7x7. As more and more context is entertained while determining the desired sense, the accuracy of classifying the correct sense got increased. The precision of the system increased from 92.38% to 98.35%, when window size increased from 3x3 to 7x7. Same is true for recall, it increased from 85.58% to 92.17%.

Although it seems very attractive to enhance the window size and get more accuracy but due to the sparseness there is always some upper limit for the size of window. The experiments made us to fix upper limit at 7.

The over lapping of the context also immensely affected the accuracy. For example the word ملک *mlk* has got the highest frequency for the sense country amongst all the four senses. Table 4 describes a huge variation in the frequency of different senses. Although sense *angel* got very low frequency but as the context of other senses did not overlap the context of the sense *angel* so its precision and recall both are 100%. On the other hand the context of *cast's name* and *country* has got highly overlapping context. Among these two senses the sense *cast's name* is the lower one with respect to the frequency. Thus the overlapping of context resulted in poor performance for the sense with lower frequency. Maximum precision and recall gained for this sense is 78.46% and 86.02 % respectively.

Table 4: Results for the word ملک *mlk*

ملک *mlk*			Window 3x3	
Senses	# of Sentences	Sense Frequency	Precision	Recall
Angel	4	0.00019	100%	100 %
Milk	32	0.0015	100%	100 %
Cast	2075	0.104	83.73%	42.1 %
Country	17575	0.89	93.5 %	99.03 %

7. Future Enhancement

In the current work only Bayesian Classifier is used for WSD. There are a lot of other supervised and unsupervised statistical techniques which can be explored and applied in the context of Urdu. As Urdu is an agglutinative language the morphological features of Urdu can help in disambiguating a word. Using an Urdu stemmer can also address sparseness issue up to some extent.

8. References

[1] http://crl.nmsu.edu/Resources/lang_res/urdu.htm

[2] Ido Milstein, "*Bayesian Word Sense Disambiguation and Spelling Correction*", April 12, 2000.

[3] David Yarowsky, "*Word Sense Disambiguation using Statistical Model of Rogert's categories trained on Large Corpora*", In Proceedings of the 14th International Conference on Computational Linguistics (COLING-92), pages 454-460, Nantes, France, 1992.

[4] Weiss SF, "*Learning to Disambiguate*", Information Storage and Retrieval, 1973; 9:33-41.

[5] Kelly E, Stone P., "*Computer Recognition of English Word Senses*", North-Holland Publishing Co., Amsterdam, 1975.

[6] http://crulp.org/oud

[7] Walker, D.E., Amsler, R.A. "*The Use of Machine-Readable Dictionaries in Sublanguage Analysis"*, In R. GRISHMAN and R. KITTEDGE (Eds.). Analyzing Language in restricted domains, 1986.

[8] Manish Sinha et.al., "Hindi Word Sense Disambiguation", In Proceedings of the 3rd Global Wordnet Conference (GWC 05), 2006.

[9] Black, Ezra (1987), "*Towards Computational Discrimination of English Word Senses*", Ph. D. thesis, City University of New York.

[10] Dagan, Ido, Alon Itai, and Ulrike Schwall, "*Two Languages are more Informative than One*", Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, pp 130-137, 1991.

[11] Javier Tejada-Cárcamo et al., "*Improving Unsupervised WSD with Dynamic Thesauri*", In Proceedings of the 11th International Conference on Text, Speech and Dialogue, 2008.

[12] Hussain, S. "*Resources for Urdu Language Processing*", In the Proceedings of the 6th Workshop on Asian Language Resources, IJCNLP'08, IIIT Hyderabad, India, 2008.

[13] Ijaz, M and Hussain, S. "*Corpus Based Urdu Lexicon Development*", In Proceedings of Conference on Language Technology (CLT07), University of Peshawar, Pakistan, 2007.

[14] Compaq Oxford Dictionary and Thesaurus.

[15] Christopher D. Manning, Hinrich Schuetze, "Foundations of Statistical Natural Language Processing", 1999.