AN APPROACH TO WORD SENSE DISAMBIGUATION COMBINING MODIFIED LESK AND BAG-OF-WORDS

Alok Ranjan Pal[1, 3], Anirban Kundu[2, 3], Abhay Singh[1], Raj Shekhar[1], Kunal Sinha[1]

[1]College of Engineering and Management, West Bengal, India 721171
{chhaandasik, abhaysingh3185, rajshekharssp, kunalsameer87}@gmail.com
[2]Shenzhen Key Laboratory of Data Science and Modeling, Kuang-Chi Institute of Advanced Technology, Shenzhen, Guangdong, P. R. China 518057
anirban.kundu@kuang-chi.org
[3]Innovation Research Lab (IRL), West Bengal, India 711103
anik76in@gmail.com

ABSTRACT

In this paper, we are going to propose a technique to find meaning of words using Word Sense Disambiguation using supervised and unsupervised learning. This limitation of information is main flaw of the supervised approach. Our proposed approach focuses to overcome the limitation using learning set which is enriched in dynamic way maintaining new data. We introduce a mixed methodology having "Modified Lesk" approach and "Bag-of-Words" having enriched bags using learning methods.

KEYWORDS

Word Sense Disambiguation (WSD), Modified Lesk (ML), Bag-of-Words (BOW).

1. INTRODUCTION

Word Sense Disambiguation (WSD) [1-2] is the process for identification of probable meaning of ambiguous words based on distinct situations. Words with multiple meaning are ambiguous in nature. The process of identification to decide appropriate meaning of an ambiguous word for a particular context is known as WSD. People decide the meaning of a word based on the characteristic points of a discussion or situation using their own merits. Machines have no ability to decide such an ambiguous situation unless some protocols have been planted into the machines' memory.

In supervised learning, a learning set is considered for the system to predict the meaning of ambiguous words using a few sentences having a specific meaning of particular ambiguous

Jan Zizka (Eds) : CCSIT, SIPP, AISC, PDCTA - 2013
pp. 517–524, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3656

words. Specific learning set is generated as a result for each instance of different meaning. A system finds the probable meaning of an ambiguous word for the particular context based on defined learning set. It shows the result based on information available in database [3-4].

In unsupervised learning, online dictionary is taken as learning set avoiding the inefficiency of supervised learning. “WordNet” is the most widely used online dictionary [5-7] maintaining “words and related meanings” as well as “relations among different words”.

Organization of rest of the paper is as follows: Section 2 is about related activities of our paper, based on the existing methods; Section 3 describes the proposed approach; Section 4 depicts experimental results along with comparison; Section 5 represents conclusion of the paper.

2. RELATED WORKS

Many algorithms have been designed in WSD based on supervised and unsupervised learning. “Lesk” and “Bag-of-Words” are two well-known methods which are discussed in this section as the basis of our proposed approach.

Typical Lesk approach selects a short phrase from the sentence containing an ambiguous word. Then, dictionary definition (gloss) of each of the senses for ambiguous word is compared with glosses of other words in that particular phrase. An ambiguous word is being assigned with the particular sense, whose gloss has highest frequency with the glosses of other words of the phrase.

The Bag-of-Words approach is a model, used in Natural Language Processing (NLP), to find out the actual meaning of a word having different meaning due to different contexts. In this approach, there is a bag for each sense of a keyword (disambiguated word) and all the bags are manually populated. When the meaning of a keyword would be disambiguated, the sentence (containing the keyword) is picked up and the entire sentence would be broken into separate words. Then, each word of the sentence (except “stop words”) would be compared with each word of each “sense” bags searching for the maximum frequency of words in common.

This paper adopts the basic ideas from typical Lesk algorithm and Bag-of-Words algorithm introducing some modifications. In Modified Lesk Approach, gloss of keyword is only considered within specific sentence instead of selection of all words. Number of common words is being calculated between specific sentence and each dictionary based definitions of particular keyword. A list of distinct words from the “Lesk” approach and “Bag-of-Words” approach is prepared based on successful disambiguation of the keyword.

Disambiguation probability would be increased based on enrichment of the bag. It means that learning method is tried to introduce within the typical concept of bags. If the bag grows infinitely, then disambiguation accuracy would be near to 100% in a typical way. Actual growth of the bag is limited depending on real-time memory management.

3. PROPOSED APPROACH

Design of our approach is presented in form of flow chart in this section. This approach is designed to achieve a disambiguated result with higher precision values.

In our approach, “stop words” like ‘a’, ‘an’, ‘the’, etc. are being discarded from input texts as these words are meaningless to derive the “sense” of the particular sentence. Then, the text containing meaningful words (excluding the stop words) is passed through “Bag-of-Words” and “Modified Lesk” algorithms in a parallel fashion. “Bag-of-Words” algorithm is considered as “Module 1”; and, “Modified Lesk” is considered as “Module 2”. These two algorithms are responsible to find the actual sense of ambiguous words in the particular context. The unmatched words in these algorithms are being stored in a temporary database for further usage. After that, results of “Module 1” and “Module 2” have been being analysed to formulate the particular sense depending on the context of the sentence in “Module 3”. If at least either of the algorithms (using “Module 1” or “Module 2”) find the sense applying logical “OR” operation on the projected results, then particular sense is assigned to the unmatched words in the temporary database. Correctness of results based on the implemented algorithms is checked in “Module 4”. If both algorithms derive same result obtained by applying “AND” operation on two results of “Module 1” and “Module 2”, then the sense is considered as disambiguated sense. Therefore, unmatched words (kept in a temporary database) has to be moved to related sense bag as per the “Bag-of-Words” algorithm in “Module 1” to participate in disambiguation method now onwards. Otherwise, the derived senses are considered as the probable senses and unmatched words are being moved to an anticipated database in “Module 5”. Figure 1 shows the modular division of our proposed approach. If the occurrence of a word in the anticipated database with a particular sense crosses specified threshold, the word is considered to be used for decision making and is moved to the related sense bag of the “Bag-of-Words” algorithm in “Module 1” to participate in disambiguation.

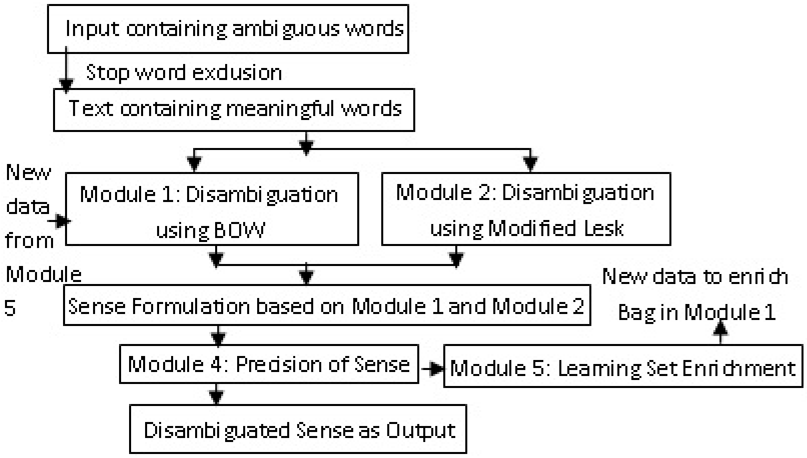
Input containing ambiguous words
Stop word exclusion
Text containing meaningful words
New data from Module 5
Module 1: Disambiguation using BOW
Module 2: Disambiguation using Modified Lesk
Sense Formulation based on Module 1 and Module 2
New data to enrich Bag in Module 1
Module 4: Precision of Sense
Module 5: Learning Set Enrichment
Disambiguated Sense as Output

Fig. 1. Modular Division of Proposed Design.

Fig. 1 describes the overall procedure for the disambiguation of words.

Fig. 2 is based on Module 1 and it shows to find the sense of an ambiguous word using Bag-of-Words approach.

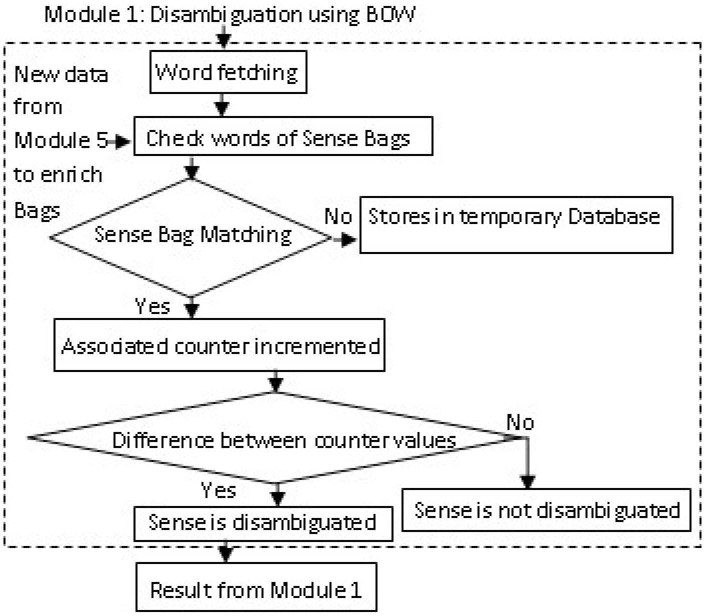
Module 1: Disambiguation using BOW
New data from Module 5 to enrich Bags
Word fetching
Check words of Sense Bags
Sense Bag Matching
No
Stores in temporary Database
Yes
Associated counter incremented
Difference between counter values
No
Yes
Sense is disambiguated
Sense is not disambiguated
Result from Module 1

Fig. 2. Flowchart of Bag-of-Words approach.

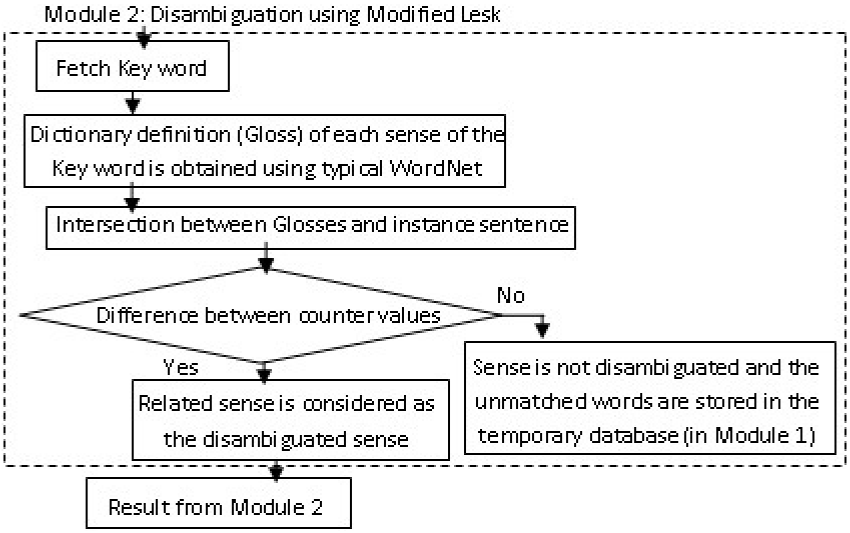
Module 2: Disambiguation using Modified Lesk
Fetch Key word
Dictionary definition (Gloss) of each sense of the Key word is obtained using typical WordNet
Intersection between Glosses and instance sentence
Difference between counter values
No
Yes
Related sense is considered as the disambiguated sense
Sense is not disambiguated and the unmatched words are stored in the temporary database (in Module 1)
Result from Module 2

Fig. 3. Flowchart of Modified Lesk approach.

Fig. 3 is based on Module 2 and it finds the sense of an ambiguous word using Modified Lesk.
Fig. 4 is designed based on Module 3. It formulates actual sense of the ambiguous word using results from previous two modules. If at least one of the two approaches can derive the sense, that is considered as the disambiguated sense.

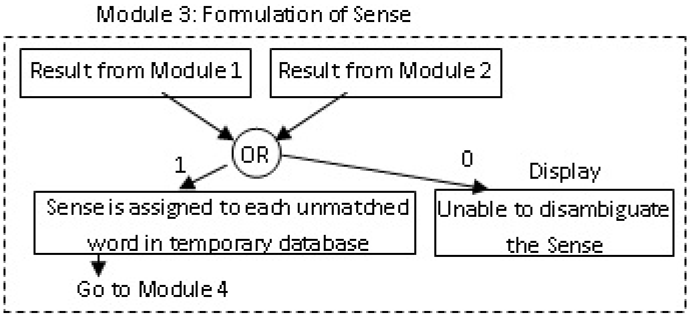
Module 3: Formulation of Sense
Result from Module 1
Result from Module 2
OR
1
0
Display
Sense is assigned to each unmatched word in temporary database
Unable to disambiguate the Sense
Go to Module 4

Fig. 4. Flowchart to Formulate Sense.

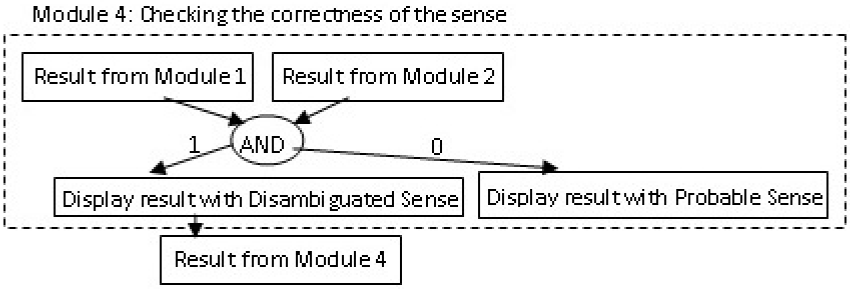
Module 4: Checking the correctness of the sense
Result from Module 1
Result from Module 2
AND
1
0
Display result with Disambiguated Sense
Display result with Probable Sense
Result from Module 4

Fig. 5. Flowchart for checking correctness of sense.

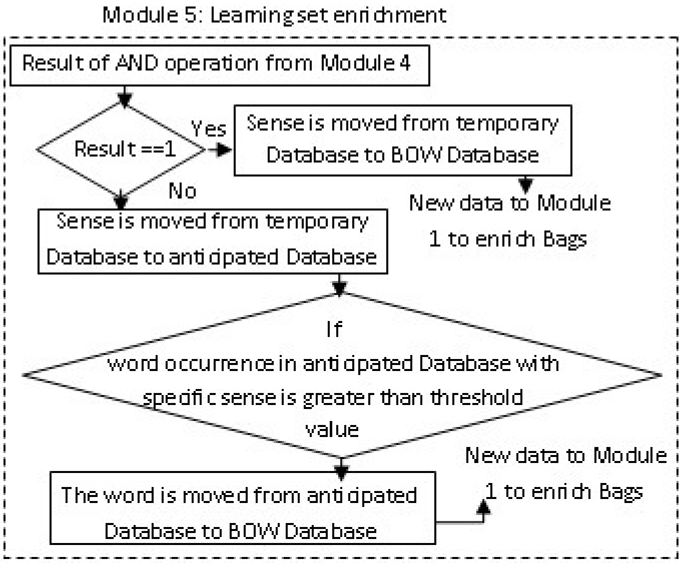
Module 5: Learning set enrichment
Result of AND operation from Module 4
Result ==1
Yes
No
Sense is moved from temporary Database to BOW Database
New data to Module 1 to enrich Bags
Sense is moved from temporary Database to anticipated Database
If word occurrence in anticipated Database with specific sense is greater than threshold value
New data to Module 1 to enrich Bags
The word is moved from anticipated Database to BOW Database

Fig. 6. Flowchart for Learning Set Enrichment.

Fig. 5 is designed based on Module 4. It finds the correctness of disambiguated sense using "AND" operation, derived by Module 1 and Module 2. If both approaches derive same sense, the result of "AND" operation is '1'. Otherwise, for all other cases, the result is '0'.

Fig. 6 is designed based on Module 5 activities. It enriches the learning set by populating with words from temporary database.

Key feature of this approach is based on auto enrichment property of the learning set. Initially, if any word is not present in the learning set, it could not be able for participation in case of disambiguation. Though, its probable meaning would be stored in the database. When the number of occurrences of the particular word with a particular sense crosses specific threshold value, the word is inserted in the learning set to take part in disambiguation procedure. Therefore, the efficiency of the disambiguation process is increased by this auto increment property of the learning set.

4. EXPERIMENTAL RESULTS

Typical word sense disambiguation based approaches examine efficiency based on three parameters such as "Precision", "Recall", and "F-measure". Precision (P) is the ratio of "matched target words based on human decision" and "number of instances responded by the system based on the particular words". Recall value (R) is the ratio of "number of target words for which the answer matches with the human decided answer" and "total number of target words in the dataset". F-Measure is evaluated as "(2*P*R / (P+R))" based on the calculation of Precision and Recall value. Different types of datasets are being considered in our experimentation to exhibit the superiority of our proposed design.

Testing has been performed on huge datasets among which a sample is considered for showing the comparison results between typical approaches and our proposed approach. In Table 1, "Plant" and "Bank" have considered as target words. Main focus is the precision value as it is the most dependable parameter in this type of disambiguation tests. Comparison among three algorithms has been depicted in Table 1.

Sample Data for Test 1:

This is SBI bank. He goes to bank. Ram is a good boy. Smoke is coming out of cement plant. He deposited Rs. 10,000 in SBI bank account. Are you near the bank of river? He is sitting on bank of river. We must plant flowers and trees. To maintain environment green, all must plant flowers and trees in our locality. The police made a plan with a motive to catch thieves with evidence.

Target Words: Bank, Plant.

Table 1. F-Measure Comparison in Test 1.

Algorithms	Precision	Recall Value	F-Measure
Modified Lesk	1.0	0.3	0.5
Bag-Of-Words	1.0	0.67	0.80
Proposed Approach	1.0	0.88	0.94

Sample Data for Test 2:

We live in an era where bank plays an important role in life. Bank provides social security. Money is an object which makes 90% human beings greedy but still people deposit money in bank without fear. Reason for above activity is trust. The bank which creates maximum trust in the hearts of people is considered to be most successful bank. Few such trustful names in India are SBI, PNB and RBI. RBI is such a big name that people can bank upon it. Here is a small story, one day a boy found a one rupee coin near the bank of the river. He wanted to keep that money safe. But he could not found any one upon whom he can bank upon. He thought to deposit the money under a tree, in the ground, near the bank of river. Moral of the story kids find earth as the safest bank. Here is another story about a beggar. A beggar deposited lot of money in her hut which was near the bank of Ganga. One day other beggars found her asset and they planned to loot that money. When the beggar came to know about the plan she shouted for help. Nobody but a bank came to rescue and they helped the 80 year old to open an account and keep her money safe.

Target Word: Bank.

Table 2. F-Measure Comparison in Test 2.

Algorithms	Precision	Recall Value	F-Measure
Modified Lesk	0.83	0.45	0.58
Bag-Of-Words	0.71	0.45	0.55
Proposed Approach	0.77	0.6	0.68

In Table 2, the result is below our expectations as initial database is small for "Bag-of-Words" approach. "Modified Lesk" (unsupervised) has shown better results than "Bag-of-Words" (supervised).

Sample Data for Test 3:

This is PNB bank. He goes to bank. He was in PNB bank for money transfer. He deposited Rs 10,000 in PNB bank account. Are you near the bank of river? He is sitting on bank of river. He was in PNB bank for money transfer. We must plant flowers and trees. He was in PNB bank for money transfer. This is PNB bank. This is PNB bank. This is PNB bank. He was in PNB bank for money transfer. He was in PNB bank for money transfer. He was in PNB bank for money transfer. He was in PNB bank for money transfer. This is PNB bank. This is PNB bank. This is PNB bank. This is PNB bank. This is his SBI bank.

Target Words: Bank.

Table 3. F-Measure Comparison in Test 3.

Algorithms	Precision	Recall Value	F-Measure
Modified Lesk	1.0	0.15	0.26
Bag-Of-Words	1.0	0.45	0.62
Proposed Approach	1.0	0.85	0.92

In Table 3, the text is long enough to give combined approach more chances to show its efficiency. Few lines are repeated in order to overcome the threshold value.

Table 4. Average of Test Results.

Algorithm	Precision	Recall Value	F-Measure
Modified Lesk	0.94	0.3	0.45
Bag-of-Words	0.90	0.52	0.66
Proposed Approach	0.90	0.78	0.85

Table 4 shows average values of all the tests performed. Efficiency of an algorithm based on fixed size learning set is improved in this paper enriching datasets. "Bag-of-Words" and "Modified

Lesk" approaches individually exhibit the "F-Measure" as 0.66 and 0.45 respectively; whereas proposed approach shows "F-Measure" as 0.85 since learning set is dynamically enriched with new context sensitive definitions.

5. CONCLUSION

Our approach has established better performance in enhanced WSD based on learning sets. Disambiguation accuracy is improved using enriched datasets. Higher precision value, recall value, and F-Measure have achieved.

ACKNOWLEDGMENT

Flow cytometry data analysis system based on GPUs platform (No. JC201005280651A)

REFERENCES

[1] Nameh , M., S., Fakhrahmad, M., Jahromi, M. Z.: A New Approach to Word Sense Disambiguation Based on Context Similarity, Proceedings of the World Congress on Engineering, Vol. I, 2011

[2] Navigli, R.: Word Sense Disambiguation: a Survey, ACM Computing Surveys, Vol. 41, No. 2, 2009, ACM Press, pp. 1-69

[3] Kolte, S. G., Bhirud, S. G.: Word Sense Disambiguation Using WordNet Domains, First International Conference on Digital Object Identifier, 2008, pp. 1187-1191

[4] Xiaojie, W., Matsumoto, Y.: Chinese word sense disambiguation by combining pseudo training data, Proceedings of The International Conference on Natural Language Processing and Knowledge Engineering, 2003, pp. 138-143

[5] Liu, Y., Scheuermann, P., Li, X., Zhu, X.: Using WordNet to Disambiguate Word Senses for Text Classification, Proceedings of the 7th International Conference on Computational Science, Springer-Verlag, 2007, pp. 781 - 789

[6] Cañas ,A. J., Valerio,A., Lalinde-Pulido,J., Carvalho,M., Arguedas,M.: Using WordNet for Word Sense Disambiguation to Support Concept Map Construction, String Processing and Information Retrieval, 2003, pp. 350-359

[7] Seo, H., Chung, H., Rim, H., Myaeng, S. H., Kim, S.: Unsupervised word sense disambiguation using WordNet relatives, Computer Speech and Language, Vol. 18, No. 3, 2004, pp. 253-273