Comparison between Bag of Words and Word Sense Disambiguation

Ayoub Mohamed H. Elyasir and Kalaiarasi Sonai Muthu Anbananthen

Faculty of Information Science and Technology, Multimedia University (Melaka Campus), Ayer Keroh Lama, Bukit Berung 75450, Malaysia

ayoub_it@msn.com, kalaiarasi@mmu.edu.my

Abstract - Bag of Words (BoW) and Word Sense Disambiguation (WSD) are the main approaches utilized in almost every data mining project for classification and data processing. The two approaches are extensively used in constructing various classifiers including supervised, unsupervised and semi-supervised classifiers. In this paper, we introduce new method of defining and comparing between BoW and WSD based on three categories. First, introduce and explain the approaches through the human brain analogy to simplify the overall concept. Secondly, sort their classifiers, methodologies and algorithms in the data mining field. Finally, introduce our developed cognitive miner to illustrate the practical functionality of these two approaches.

Index Terms - Data Mining, Bag of Words, Word Sense Disambiguation, Classifier

1. Introduction

In this paper we compare the two approaches, Bag of Words (BoW) and Word Sense Disambiguation (WSD), relative to data mining field specifically mining opinions as unstructured text. Processing the unstructured text is no trivial task and requires extensive amount of research and sophisticated algorithms to produce accurate results in terms of polarity classification. BoW is commonly used in processing the raw text as it has faster and light processing mechanism, because it relies mainly on the statistical and counting techniques. However, WSD utilizes external resources and more cognitive knowledge to perform such opinion mining task.

As far as our knowledge, none of the existing papers devotes itself in comparing between these two approaches especially in the comprehensive way that we are proposing. Fig 1 shows that this paper provides concise benchmark to compare between BoW and WSD within three categories: First is using the human brain analogy to explain and define both approaches in a simple manner, second comparison criteria include methodology, classifiers and algorithms, and then produce summarized result for their pros and cons. Third is developing cognitive miner utilizing both approaches to understand their practical functionality and applications.

Human brains are the extreme machines for comprehending structured and unstructured text, and produce nearly 100% accurate polarity classification. The process of comprehension involves lots of cognitive functions, all mental processes including problem solving, memory recalls, language understanding and concentration, as well as cooperation among the brain regions and hemispheres just to read and understand a corresponding piece of text. In this category of comparison, we attempt to simplify the concept of the two approaches and their difference using the brain comprehension analogy. WSD tends to have closer working mechanism to the brain comprehension technique, in which it attempts to interpret the unstructured text based on the given grammatical structure and syntactic meaning while BoW disregards the words' associations and relationships.

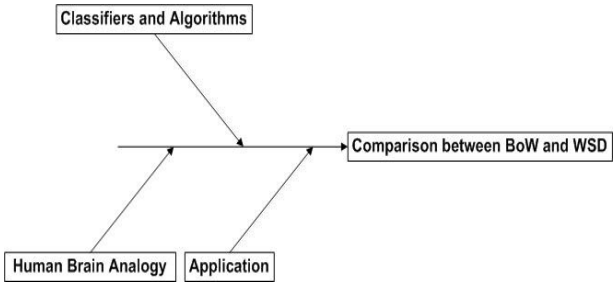

Fig 1: Three categories to compare between BoW and WSD

Second category is to distinguish between each one through their methodology, classifiers and algorithms. The classifiers range from supervised to unsupervised learning either with external resource like thesauri and dictionaries or without resources. At the end of this comparison category, we provide summarized pros and cons to depict briefly their strengths and weaknesses.

For better understanding on both approaches, we attempt to discover their functionalities by embedding them into cognitive mining application. The cognitive miner is capable of finding an answer for any structured question by crawling through the web using focused web crawler and then process the unstructured text using both concepts BoW and WSD. Our developed cognitive miner is a hybrid of both Natural Language Processing (NLP) approaches to produce an accurate answer with confidence precision.

2. Compare Using Human Brain Analogy

Brains are sophisticated organs and perform complicated functions in which is almost impossible to compare their performance against computers. However, we attempt to explain the meaning of BoW and WSD using the brain analogy and how it learns to read text. Reading ability is more ambiguous than speaking as it is more likely involves grammatical structures and difficult phrase combination. This composition of grammar and phrase sequence is known as

© 2013. The authors - Published by Atlantis Press

syntactic structure where the full text comprehension is required to understand the entire sentence.

Before learning to read, children are able to comprehend a meaning of text through the experience of speaking and listening, because they develop the sense of correct syntax and how the sentence should be made. For example, the correct phrase "I want to play" is twisted first to "I play want" or "I to want play" before the brain recognizes the appropriate pattern. However, they might face difficulty understanding certain words due to the word's independent meaning but some still able to figure it out through the sentence syntactic. Hence, we conclude that a word is interpreted using dependency and the syntactic properties as shown in Fig 2.

Meanwhile, BoW uses the same concept of interpreting the word based on its independent meaning regardless of its syntactic structure. Whereby a document or article is represented as bag of words each one is independent from the other counting their appearance frequency and candidacy strength. BoW utilizes statistical techniques to count the frequency and elect the candidate word dependent on the corresponding application or domain. However, WSD cares about the word's relationship or association with the surrounding words that is an interpretation based on the syntactic meaning. Thus, multiple senses are important in which the word may appear as a verb in one sentence and noun in the other as depicted in Fig 2.

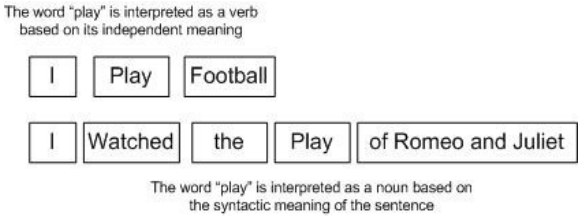

Fig 2: Word's meaning interpretation

Multiple senses and syntactic meanings are not trivial to analyze even for the brain. Thus, the brain attempts to break down and simplify the comprehension techniques using the following elements:

- Word Sequential Order
- Analysis of the grammatical structure:
  - Negation
  - Passive Voice
  - Word Encapsulation
  - Conjunctions

On the other hand, the BoW approach is much easier and does not require comprehensive interpretation for the syntactic structure of a sentence. However, WSD relies on various elements to interpret the word dependent on the syntactic and associations among the words. Therefore some WSD concepts utilize the same elements for grammatical analysis and word sequential order.

3. Comparison Based on Classifiers and Algorithms

In this category of comparison, we attempt to explain the difference and distinguish between BoW and WSD using their methodology, classifiers and algorithms. This comparison serves as an introductory for what we are implementing in the last category which explains how these two approaches are the underneath mechanism of many commercial and research applications.

A. The Bag of Words methodology

BoW is the simple approach in opinion mining where it processes the unstructured text regardless of its syntactic or grammatical structures. The words or phrases are treated independently without consideration to the association or relationship among the words. Old opinion mining architecture made use of BoW to represent the documents as bag of words to count the appearance frequency and discards the sequential information provided by the word order and their associations.

B. Bag of Words Classifiers

BoW relies on statistical and probabilistic methods to tokenize the unstructured text and produce polarity classification, by counting the word frequencies and degrees to filter out the candidates. Classification algorithms in BoW attempt to label the word given the training example, data instance, in the case of supervised learning, while unsupervised learning classifiers tend to label based on the internal distances among the data instances.

1) Supervised Learning

The example in the supervised learning consists of a pair of an input value and the pre-defined output, whereby analysis are committed over the data instance to produce a classifier:

Naïve Bayes: Simple probabilistic classifier based on the most likelihood strategy and can be trained effectively using supervised learning settings. It is oversimplified classifier but provides reasonable accuracy and only requires small set of data instance for training which makes it one of the most popular classifiers in data mining in general.

Decision Tree: Predictive model that generates binary like structure output. Two types of decision trees are there; classification and regression tress, in which some classifiers utilize them to improve their accuracy rate.

Maximum Entropy: Alternative for Naïve Bayes classifiers because no assumption is made over the statistical independence but requires larger set of training data.

2) Unsupervised Learning

There is no label for the example or training data in the unsupervised learning, instead clustering techniques are in use which distinguishes it from the supervised learning:

K-means: Is one of the few unsupervised learning algorithms that solve the renowned clustering problem, it does that by calculating the gap from the centers of the clusters (J MacQueen 1967).

N-grams: Probabilistic model that can be used for parsing the unstructured text by generating sequence of characters

such that the size one is known as unigram, size two is bigram, size three is trigram and the larger is n-gram.

Stemming: Technically considered as a technique rather than an algorithm, in which attempts to transform the word back to its root and eliminate any extras such as, the word "playing" is transformed to "play" using any stemming technique. It does this transformation without training data, instead relies on dictionaries and word mapping.

WordNet and SentiWordNet are also unsupervised and heavily utilized in the field of natural language processing and opinion mining respectively. We also utilize SentiWordNet in our opinion mining architecture to evaluate the word or sentence polarity and classify the result accordingly.

C. Word Sense Disambiguation Methodology

It is the ability to recognize the word multiple meanings across various sentences in different domains. The word or phrase may have multiple senses (meanings) dependent on the domain that the word belongs to; therefore BoW is unable to digest this concept and process over it. WSD is not a replacement but a complement for what BoW is unable to accomplish, better to think of it as in the cons of first is covered by the pros of second and vice versa.

The natural language is very complex and paradoxical in many ways that make it very dependent on the domain or the context. WSD can spot this dependency by working out the problem of polysemy, having multiple meanings for the same word. For example, the word "play" in "The boy plays football" is different from "I am watching the play of Romeo and Juliet". The process involves two main steps:

- Determine all the different meanings for each word relative to the corresponding domain or context:
  - Extract the meanings that are found in dictionaries of the same language and other languages
  - Extract the associated features such as synonyms in the thesaurus
- Assign each occurrence of a word to the suitable meaning:
  - Using the linguistic data available within the unstructured text to extract information relative to the context
  - Using the external sources such as lexical resources to draw extra information regarding the word context and domain

WSD is all about matching the context of the word to the word itself using either external data sources like dictionaries and thesaurus or deriving helpful information from within the context and its relation with the words. It is a classification task using two methods: Machine Learning and Thesaurus based approaches.

1) Machine Learning Approaches

During this approach, the initial input is processed for a disambiguated word using Part of Speech Tagging (POS) where each word is attached a label according to its syntactic function. Number of methods has emerged to tackle the problem of WSD in the machine learning process including but not limited to: Artificial Intelligence (AI), Symbolic Methods, and Machine-readable Dictionaries

Word Sense in AI is done for larger systems including those full language understanding. WSD early history is deeply connected to the AI as a subtask of sematic interpretation but then started to draw another path toward the lexical resources as it is a rule-based and hand-coded. The process of word disambiguation remained dictionary based until the utilization of statistical techniques which allows the usage of supervised learning in the early nineteen.

2) Thesauri

It contains information on the relationship and the syntactic association among the words. These relationships are represented in synonyms and antonyms of the given word which is different from the normal dictionary that only contain definitions and pronunciations. However, a thesaurus is mainly used in information retrieval where the data is drawn based on the suggested relations in the thesaurus; the process of Information Retrieval (IR) involves indexing and tagging. Thesaurus may have different terms across various fields such as thesaurus in the field of AI is known as ontology and in the IR is a semantic database.

D. Word Sense Disambiguation Algorithms

Most of WSD algorithms work by statistically analyzing (n) words that surround the words to be disambiguated. The old tradition of WSD algorithms is to use Naïve Bayes and Decision Trees to train and disambiguate the words, but after the proposal of more accurate kernel based algorithms on the top is SVM in the supervised learning.

Graph based algorithms have boosted the performance of lexical knowledge base to overcome the weaknesses found in the supervised learning techniques. Other methods like the knowledge transfer from online resources to the WordNet have shown good accuracy in sensing the words that might outperform the supervised techniques in certain domains.

Most of the supervised algorithms have been applied to WSD where it depends on the context to provide sufficient information to disambiguate the words. Generally supervised techniques and methods provide better performance and accuracy than unsupervised learning in the WSD, therefore some researchers prefer to utilize both of them and make a hybrid system known as semi-supervised learning that allow both labeled and unlabeled examples.

4. Summary for Comparison Between Bow and WSD

Although BoW is more suitable to work on a domain with lots of features for extraction like education domain, we still combine both approaches BoW and WSD to complement each other. Therefore we utilize BoW for crawling and text processing .i.e., HTML parsing, sentence segmentation, stemming and stop words removal. While WSD provides superior performance in POS tagging and Named Entity Recognition (NER) which classifies the words based on their context.

Table 1: Comparison between BoW and WSD

Bag of Words (BoW)	Word Sense Disambiguation (WSD)
Document is represented as bag of words. i.e. “Study Hard” is the same as “Hard Study”	Relationships and associations among the words are important for the analysis
Uses statistical methods to do opinion mining	Variety of methods are proposed but all of them are concerned about the word meaning in the corresponding domain
Counts the word frequencies and degrees	Assigns the different senses to the word from the same context
Discards the word sequential order	Maintains the word sequential order
BoW is much faster in opinion mining as the sequential order is neglected	WSD presents better accuracy than BoW in certain domain where the word syntactic meaning is crucial
Higher performance during the object categorization as it provides vector representation especially with the SVM classifier	Slower and incapable for vector representation, therefore object categorization is made for BoW than WSD
BoW leads to accurate classification when it comes to positive opinion mining	It depends on the domain and the opinion context
Less applications including: part of speech tagging, natural language processing, correlation and regression, word statistics and semantic information	Has more applications including: Information Retrieval, indexing, parsing, intelligent natural language processing, machine translation where it helps in understanding the generation of sentences in a target language, part of speech tagging, spelling correction and annotation
BoW serves as the pre-processing stage in opinion mining such as object categorization	More to the final stages of opinion mining specially in the classification part

A. BoW and WSD in the Action

Both approaches have variety of applications in different field across data mining and machine learning. BoW is involved in the picture processing where pixels are treated and represented as bags and then processed according to certain parameters. It has some usefulness in the voice recognition as well; where sound pitches are classified using BoW supervised learning techniques. WSD applications are mostly devoted to natural language processing where the concern on the syntactic meaning is needed.

In this category, we introduce our developed cognitive miner to search and crawl the internet for a user inserted search queries in a form of structured question. The two approaches are integrated together to provide the optimum performance of text processing, whereby BoW is utilized extensively in the cognitive analysis, n-gram processing and transformation of the unstructured text into structured format. WSD resolves the answer types, POS tagging and cognitive interpretation.

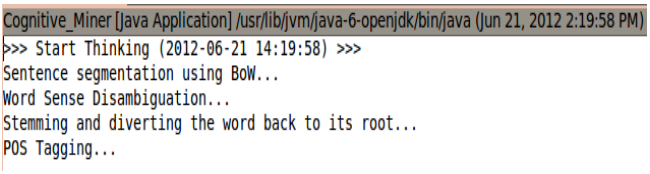

Fig 3: Loading BoW and WSD in the cognitive miner

Fig 3 above depicts the loading of BoW and WSD modules in the cognitive miner including some pre-processing tasks such as, stemming and stop words removal. The timestamp is attached with all processes for its usefulness in file logging.

Fig 4 shows the attributes and content models in which WSD depends on for normalization and resolving the answer types. The attributes are created based on WordNet 3.0 entries and knowledge transfer from the internet.

Fig 5 illustrates the crucial phase in our cognitive miner. Whereby the user inserts structured question and waits for graphical representation answer if the sufficient data is available, otherwise unstructured text answer is provided. BoW performs cognitive analysis to transform the structured question into bag of words to treat them respectively. Then, WSD resolves the answer types based on the entries found in the WordNet and internet knowledge transfer. Cognitive interpretation in the WSD sorts the extracted answer types into their corresponding class to produce fine n-gram results and simplify the process of crawling. N-gram processing in the BoW is helpful to generate cognitive predications to approach the best filtered results for the internet crawling and finally present the result as shown in Fig 6.

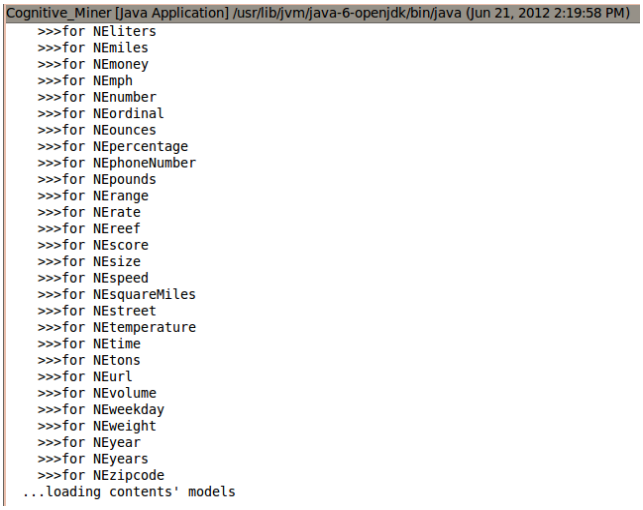

Fig 4: Loading Attributes for WSD

```
         what is the best university in Malaysia ?

>>> Cognitive Analysis (2012-06-21 14:21:07) >>>
Normalizing: what be the best university in malaysia

Answer types:
NEproperName->NEorganization->NEeducationalInstitution

Cognitive Interpretations:
Object Property: NAME
Cognitive Target: best university
Topic of Interest: Malaysia

Cognitive Predications:


>>> N-gram processing (2012-06-21 14:21:20) >>>
Query strings:
best university Malaysia
best university (Malaysia OR Malaya)
"best university" "Malaysia" best university Malaysia
"is the best university in Malaysia "
"the best university in Malaysia  is"

>>> Crawling (2012-06-21 14:21:20) >>>
```

Fig 5: Processing User Structured Question

```
[1]     Unity College
        Score: 3.8350094E-4
        Document: http://www.e2malaysia.com/universities.html
```

Fig 6: Result for the User Inserted Structured Question

5. Conclusion

We provided thorough comparison between the two main approaches in data mining and the unstructured text processing in specific, Bag of Words (BoW) and Word Sense Disambiguation (WSD), using three different categories: distinguish using human brain analogy; depict the details utilizing methodologies, classifiers and algorithms; finally, demonstrate a cognitive miner system to understand the practical functionality of BoW and WSD.

Acknowledgments

This work was supported by Fundamental Research Grant Scheme of Malaysia 2011.

References

[1] J MacQueen. "Some methods for classification and analysis of multivariate observations" university of California, Los Angeles, (1967)