DETECTION OF MULTIWORD FROM A WORDNET IS COMPLEX

Md J. Abedin[1], B. S. Purkayastha[2]
Department of Computer Science, Affiliated to Assam University, Silchar.
Email: jaynal84@gmail.com; bipul_sh@hotmail.com

Abstract

Multiword detection is very difficult task in Language Processing. There has been a great change in the field of Natural Language Processing. Manual encoding of linguistic information is being challenged by automated corpus based learning methodologies for NLP with linguistic Knowledge. Although, Corpus based approaches have been successful in many different areas of Natural Language Processing. Multiword Detection using human evaluation method and machine evaluation method is a matter of discussion in present NLP research areas. Languages are not use not only to gather information but for communication as well. The problem is to take the information provided by the outside the world and translate the information into precise internal representation .This internal representation is called semantic representation. In this paper, we deals with how MWEs are generalized to optimized expensive evaluation of Multiword detection from a WordNet.

Keywords: *MWEs, Wordnet, StopWrod, lexeme.*

----------***----------

1. INTRODUCTION

Languages are made up of words that are interpreted in the form of phrases and sentences. MWEs can happened in nominal, verbal and adverbials form. Formal definition of Multiword Expression define by [1] as: Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into multiple lexemes, and (b) display lexical, syntactic, semantic, pragmatic or statistical idiomaticity. In English language, Decomposability of lexemes is that MWEs must be made up of multiple white space delimited words. For example, ATM Card, is potentially a MWE which is made up of two Lexemes ATM and Card, While fused word such as Whitehouse is not considered to be a perfect MWE. However, Decomposition of an expression into multiple lexemes is still applicable.

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept [2]. The basic building block of WordNet is a synset. Each synset has a unique number associated with it called as synset identity or synset id.

These synsets are created manually by lexicographers and they also manually create links between synsets representing semantic and lexical relations, [3].

Multiword deals with Information Retrieval (IR) System, Which is an application area of computer technology for acquisition, organization, storage, retrieval, and distribution of information. Multiword detection can be applied to detect multiword available from a different repositories. Our research discipline is concerned with both the theoretical underpinnings and the practical improvement of Query Retrieval System including the construction and maintenance of large information repositories through Web interfaces using hypertext and multimedia databases.

Natural language Interface to Database (NLIDB) can acts as an alternative interface for finding structured information from database particularly on a small hand held devices because writing questions in natural language is much easier for a casual user than to implement practically because it is complicated and time consuming, Navigation required in the traditional database interfaces.

The summaries of a word or a sentence using query from an information retrieval can be classified as abstractive and extractive summarization .The task of disambiguation requires that ambiguity is controlled at each level of the analysis and that plausible solutions surface as an output while implausible ones are discarded.

In Multiword detection, we deal with Multiword expression, text summarization, sentence selection, sentence weighting and term weighting. All these are key idea of words selecting from a documents or web search results. When a query is given by a user either word mode or sentence mode, it will be appeared with several meaning from the index of the words or sentences, finding the real idea of the query is a matter of discussion in Natural Language processing that cause multiword detection searching techniques to be boarded in NPL research areas.

Our research Multiword detection basically based upon MWEs .The major NLP tasks relating to MWEs are: (1) identifying and extracting MWEs from corpus data, and disambiguating their internal syntax, and (2) interpreting MWEs. Increasingly, these tasks are being pipelined with parsers and applications such as machine translation [4].

2. MULTIWORD DETECTION TECHNIQUE

In multiword detection individual terms (word) are analyse in both syntax and semantic form [5]. Various algorithms can be applied for multiword detection techniques which are:

a) Graph Algorithms
b) Clustering Algorithms

c) Generic Algorithms etc which facilitate text summarization task [6].

While considering multiword detection [5] broadly consider the following steps in details at the initial stage:

1. Term Selection: Individual term.
2. Term Weighting: It is a process of estimating the sum of usefulness weights of individual term of which the sentence consists.
3. Sentence Selection: Selecting appropriate sentences to assign some numerical measure of usefulness of a sentence for the summary and then select the best ones.
4. Sentence weighting: It is a process of assigning usefulness weight of a sentence.

For multiword detection we need to select a particular domain for which a corpus of data needs to be collected by conducting survey through different databases using different languages or web interface interaction by different user. This corpus is then analyzed to find the pattern of resulting output provided by the detection process. Based on these patterns a context free grammar is designed to represent syntax of input questions. Representing syntax is an important step as in deep analysis method is required and the subsequent steps are dependent on it. In multiword detection, use of extensive knowledge from outside the domain is not required.

2.1 N-Grams In Multiword Detection

An *n*-gram model predicts x_i based on $x_{i-(n-1)}, \cdots, x_{i-1}$. It is one of the text representation model used in consecutive elements that contain the term. These elements can be words or characters. For example, if *n* equals 2, the defined term that will contain 2 words or characters, namely bigrams. Multiword detections basically deals with bigram sequence of n-gram model. We also need to concentrate stop words during word detection from a sentence.

Stop Words: Stop words are words which are filtered out prior to, or after processing of natural language data (text). It is controlled by human input and not automated. There is not one definite list of stop words which all tools use, if even used. We can generate n grams in two ways [7] :

Literal queries use the quoted n-gram directly as a search term for the search engine (e.g., the bigram history changes expands to the query "history changes").

Inflected queries are obtained by expanding an n-gram into all its morphological forms. These forms are then submitted as literal queries, and the resulting hits are summed up (e.g., history changes expands to "history change", "history changes", "history changed", "histories change", "histories changed").

2.2. Sentence Parsing In Multiword Detection

Consider a sentence: '***He withdraw money from ATM Machine using the ATM Card'***

Parse tree for the sentence is outline in fig1 as:

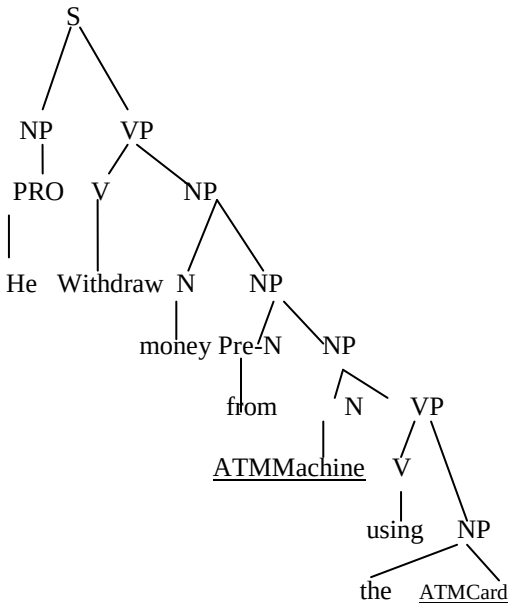

Fig.1: Parsing for the sentence "He withdraw money from ATM Machine using the ATM Card "

In the above sentence lexicons are there ,where each word contains syntactic, semantic information .It is also seen that two multiword such as ATM Machine and ATM Card are considered based on domain specific knowledge. Based on these lexicons information system will detect as an output. It will also reduced Word Sense Disambiguation (WSD) from the sentence.

2.3 Classification Of Multiword Expressions

For Developing a lexicon of MWEs, we need to develop a classification that will capture general properties of MWEs. In section, we present a commonly used high level classification based on the syntactic and semantic properties of MWEs outlined in Fig. 2 [1].

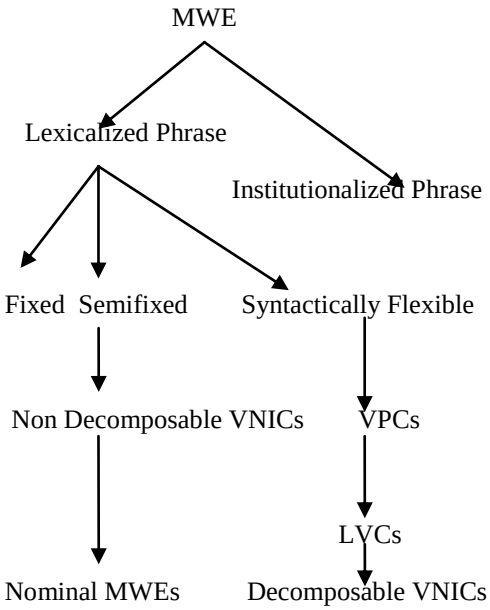

Fig.2: Classifications of MWEs

In the Fig.2, VPCs meant for Verb Particle Constructions, PVC meant for particle Verbs Construction, LVCs meant for Light Verb Construction and VNICs meant Verb Noun Idiomatic Combinations.

3. WE MENTION SOME OF THE AREAS OF NLP WHILE CONSIDERING MULTIWORD DETECTION

Word Sense Disambiguation (WSD) [8]**:** WSD solves what sense has a given word, generally based on its context. This task is very important because of its successful resolution, depends the correctness of other applications such as Machine Translation, Question Answering, etc.

Information Retrieval (IR)[9]**:** It consists of finding documents of an Unstructured nature that satisfies an information need from within large collections of documents usually on local computer or on the internet. This area overtakes traditional database searching, becoming the dominant form of information access. Now hundreds of millions of people use IR systems every day when they use a web search engine or search their emails.

Machine Translation [10]: It is a machine-assisted system responsible for translation from one language to another. This application is very useful for bussiness and scientific purposes by the reason that the international collaboration grows exponentially.

Question Answering (QA) [11]: It is a complex task that combines techniques from NLP, IR and machine learning. The main aim of QA is to localize the correct answer to a question written in natural language in a nonstructural collection of documents. Systems of QA look like a search engine.

4. CONCLUSION

Detection of Multiword using human evaluation method is expensive and time consuming therefore we need to generalized method for quick, easy, in an inexpensive, and language independent way. Parsing a sentence reduces the time complexity to search a Multiword from a wordnet. MWEs are a key issue and a current weakness for natural language parsing and generation, as well as real-life applications depending on language technologies, such as machine translation, POS etc.

REFERENCE

[1] A. Copestake, F. Lambeau, A. Villavicencio, F. Bond, T. Baldwin, I. A. Sag, and D.Flickinger, 2002.Multiword expressions: Linguistic precision and reusability. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages1941–7, Las Palmas, Canary Islands.

[2] Fellbaum, Christine, ed.: 1998, *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press.

[3] Jackendoff, Ray: 1997, *The Architecture of the Language Faculty*, Cambridge, MA: MIT Press.

[4] S. Venkatapathy and A. Joshi (2006). Using information about multi-word expressions for the word-alignment task. In Proceedings of the COLING/ ACL 2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Sydney, Australia, pp. 53–60.

[5] M.en C.Yulia Nikolaevna Ledeneva, Dr.Alexander (November, 2008), Automatic Language – Independent of Multiword Description for Text Summarization. Bolinger, Dwight, ed.: 1972, *Degree Words*, the Hague: Mouton.

[6] Bolinger, Dwight, ed.: 1972, *Degree Words*, the Hague: Mouton.

[7] L.Mirella and K.Frank,2005. ACM Transactions on Speech and Language Processing, Vol. 2, No. 1.

[8] A.Gelbukh, G. Sidorov, H.Y.Sang. Evolutionary Approach to Natural Language Word Sense Disambiguation through Global CoherenceOptimization.WSEAS Transactions on Communications, ISSN 1109-2742, Issue 1 Vol. 2, pp. 11–19, 2003.

[9] C. Manning, An Introduction to Information Retrieval. Cambridge University Press, 2007.

[10] A. Gelbukh, I. Bolshakov, Internet, a true friend of translator. International Journal of Translation, ISSN 0970-9819, Vol. 15, No. 2, pp. 31–50, 2003.

[10] I. Bolshakov,A.Gelbukh . Computational Linguistics: Models, Resources, Applications.
IPN-UNAM-FCE, ISBN 970-36-0147-2, 2004.

[11] R. A.Peréz ,M.G.y Montes y G,L.P. Villaseñor. Enhancing Cross-Language Question Answering by Combining Multiple Question Translations. Lecture Notes in Computer Science, Springer-Verlag, vol. 4394, pp. 485–493, 2007.