TEXTUAL SIMILARITY DETECTION-A SURVEY

Shital Liladhar Patil*
Deparment Of Computer Engineering
SSBT's COET,Bambhori

Krishnakant P.Adhiya
Deparment Of Computer Engineering
SSBT's COET,Bambhori

Abstract**—** Measuring Textual Similarity ,between words/ terms, sentences, paragraph and document plays an important role in computer science. In natural language processing (NLP), Semantic textual similarity is an important component for many tasks such as document summarization, word sense disambiguation, short answer grading, information retrieval and extraction . The lexical overlapping approach evaluate similarity among sentence and find whether a sentence pair semantically equivalent or not . Existing methods for computing sentence similarity have been adopted from approaches used for long text documents. These methods process sentences in a very high-dimensional space and are consequently inefficient, require human input, and are not adaptable to some application domains. Semantic textual similarity methods improved in two areas ,first is semantic relation between word and second is semantic resources to reduce dimension and overcome disadvantages of existing methods . In this paper, we have given the survey of various techniques and methods for textual similarity detection from sentence.

Keywords**— Natural language processing , Semantic textual similarity, Word similarity, Sentence similarity, Text similarity**

I. INTRODUCTION

Finding the similarity among the sentences in natural language processing (NLP) plays a vital role because a sentence can be expressed in many forms without varying the sentence meaning. Therefore there is a need to identify the semantic similarity among the sentence pair. Measuring and recognizing semantic relation between the pair of sentence is the problem of semantic similarity. The similarity can be measured at different levels of abstraction i.e., between words or sentences or paragraphs or documents and at multi levels such as word to sentence or sentence to paragraph etc., Traditionally, techniques for detecting similarity between documents have centered on analyzing shared words. [1]

Such methods are usually effective when dealing with long texts because similar long texts will usually contain a degree of same words. In short texts the word co-occurrence may be rare. And mainly due to the inherent flexibility of natural language enabling people to express similar meanings using quite different sentences in terms of structure and word content. The information in short texts is very limited and this problem poses a difficult computational challenge.

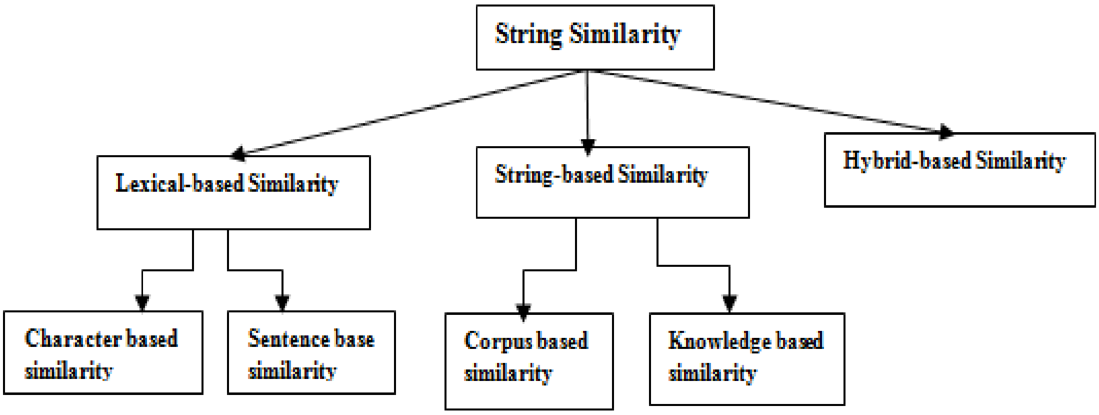

Fig 1 : Sentence Similarity Architecture

1.1 Sentence Similarity Methods:

- Corpus based Methods**:** These methods are based on a corpus features. The first category, traditional information retrieval methods, Term Frequency –Inverse Document Frequency (TF-IDF) methods , assume that documents have common words. However, these methods are not valid for sentences because sentences may have null common words. For example, the sentences “my boy went to school” and “kids learn math” do not have any common word although they are semantically related to education.[2]

- Knowledge based Methods**:** knowledge based methods use semantic dictionary information such word relationships, information content to get word semantic features. proposed a sentence similarity based on the aspects that a human interprets sentences; objects the sentence describes, properties of these objects and behaviour of these objects. Generally, the knowledge based methods are limited because of this, not all words are available in the dictionary and even if a few words exits they usually do not have the required semantic information. As an example, WordNet has a limited number of verbs and adverbs synsets compared to the list of available nouns synsets in the same ontology.[2]

- Hybrid Methods**:** Hybrid methods are a combinations of the previous mentioned methods. A combination of eight knowledge base measures and three corpus based measures is proposed.. The final word similarity measure is the average of all eight measures. [3]The sentence similarity methods are derived using word overlapping over an IDF function of words in related segments. Hybrid approaches shows promising results on standard benchmark datasets.

1.2 Computing Sentence Similarity Approaches:

- Syntactic similarity approach: Syntactical similarity approach in which syntactic means structure of the words and phrases. The similarity of two sentences corresponds to the equivalent relation between the vectors. This is quantified as the cosine of the angle between vectors .This is so-called cosine similarity[5]. The cosine similarity is the given pair of sentences are related to each other .The cosine similarity specify the score based on the words overlapped in the sentences.

- Semantic similarity approach: The two sentences with different symbolic structure and information could convey the same or similar meaning. Semantic similarity of sentences is based on the meanings of the words and the syntax of sentence. Semantic similarity of sentences is based on the meanings of the words and the syntax of sentence. If two sentences are similar, structural relations between words may or may not be similar. [6]Structural relations include relations between words and the distances between words. If the structures of two sentences are similar, they are more possible to convey similar meanings.

1.2.1 Architecture of Textual Similarity Detection From Sentence

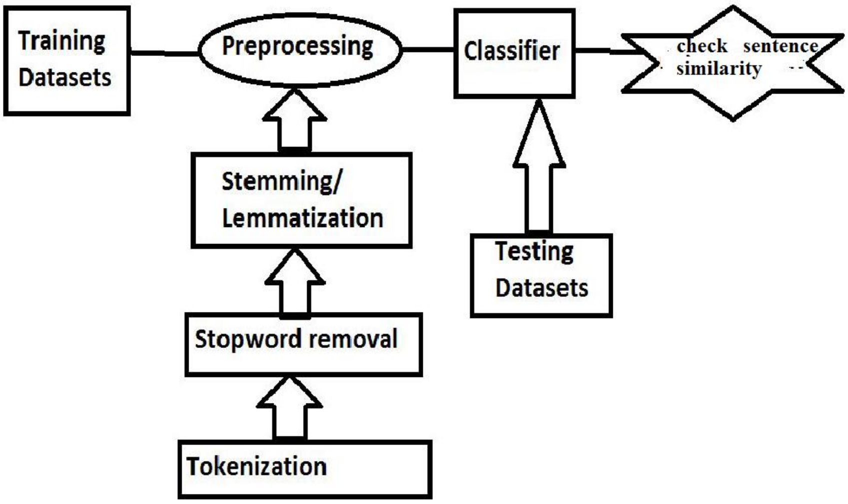

Fig 2:Architecture of Sentence Similarity

1.2.1 Processing Steps : The goal of this phase is to reduce inflectional forms of words to a common base form. In this section the basic preprocessing techniques are discussed.[8]

- Tokenization is the task of chopping up sentences into tokens and throwing away punctuation and other unwanted characters.

- Tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context. In our case we tagged the word to noun and verb.

- Lemmatization is a technique from Natural Language Processing which does full morphological analysis and identifies the base or dictionary form of a word, which is known as the lemma.

- Syntax similarity is a measure of the degree to which the word sets of two given sentences are similar. A similarity of 1 (or 100%) would mean a total overlap between vocabularies, whereas 0 means there are no common words.

- Similarity returns a score denoting how similar two word or sentence senses are, based on some measure that connects the senses in is-a taxonomy.

II . LITERATURE REVIEW

Yuhua Li et.al[1] **,** proposed the novel algorithm for computing similarity between very short texts of sentence length. The author introduced a method that takes account of not only semantic information but also word order information implied in the sentences. The computational method that is able to measure the similarity between very short texts (sentences). This results in a conversational agent knowledge base that is easier to compile, far shorter, more readable and much easier to maintain.

Ying liu et.al[2]**,** proposed the example-based machine translation approach for chinese sentences are translated into english sentences Firstly, find the most similar examples as the input sentence. Secondly, recombine the translation of the input sentence according to most similar example and bilingual dictionary. Thirdly the translation of the input sentence. The main resources are bilingual dictionary, these resources, the standard template system an bilingual sentence aligned corpora.

LiHong Xu et.al[3] **,** described the text similarity computation method named VSM-Cilin which is based on semantic vector space model and background of radio station. VSM-Cilin improved the traditional VSM in the following areas. First, descibed the semantic relations between words. Second, used semantic resources to reduce dimension. Third, used inverted index to filter out candidate document set. Forth, take the weight of the feature item into consideration when compute the similarity. The experiments show that the accuracy of VSM-Cilin is significantly improved compared with the traditional vector space model and the method of bidirectional mapping based on HITIR-Lab Tongyici Cilin.

Zhao Jingling et.al[6] **,**derived the sentence similarity from semantic and syntactic information contained in the compared sentences. A sentence is considered to be a sequence of words each of which carries useful information. The words, along with their combination structure, make a sentence convey a specific meaning. proposed a new approach to compute the sentence similarity, which is divided into three steps: firstly, obtain word semantic similarity; secondly, obtain semantic similarity between sentences based on the word semantic similarity and the structure of sentences; finally, calculate word order similarity between sentences and combine semantic similarity and word order similarity as the final similarity between sentences.

Jiang et.al[5] ,proposed the new approach for measuring the semantic similarity /distance between words and sentence. It combine the statistical methods and lexico-syntactic patterns so that lexical distance between semantic node in semantic space constructed by taxonomy can be better quantified with computational evidence derived from distributional analysis of corpus data.

Emiliano Giovannetti[7], proposed different methodology to combine two different techniques for semantic relation extraction from texts. On the one hand, generic lexico syntactic patterns are applied to the linguistically analyzed corpus to detect a first set of pairs of co-occurring words, possibly involved in “syntactic” relations. The resulting set of relations can be used to enrich existing ontologies and for semantic annotation of documents or web pages.

Jacob Bank et.al[21] **,**proposed the map reduce programming technique in the analysis of large social networks. There are a number of direct extensions we can see making to this particular work to make it more useful in real world analysis. First, we could add the ability to filter users and pages by certain characteristics. Another potential direction would be to filter the data into time slices to track trends over time. Furthermore, we do not currently weight a co-occurrence that appears many times over a co-occurence that appears only once. All of these would be useful potential expansions on our current work.

Wu et.al[22] ,proposed a new method for semantic representation of verbs and investigated the in sentence on leexical selection problems in machine translation. Wu and Palmer describe semantic similarity measure amongst concepts C1 and C2. Resnik Measure (1995) Similarity depends on the amount of information of two concepts have in common. Lin extended the Resnik(1995) method of the material content (Lin et al., 1998). He has defined three intuitions of similarity and the basic qualitative properties of similarity. Hybrid approach combines the knowledge derived from different sources of information. The major advantage of these approaches is if the knowledge of an information source is insufficient then it may be derived from the alternate information sources.

Table 1. **Summary of word and Sentence similarity approaches**

Similarity methods	**Sentence similarity approaches**		
	Technique	**Advantages**	**Disadvantages**
Corpus based methods	Uses a corpus to get probability or frequency of a word in a corpus	Preprocessed corpus to reduce computations	-Corpus is domain dependent. -Some words might get same similarity. - Semantic vectors are sparse
Knowledge based methods	Uses dictionary information such as WordNet to get similarity	Adoptions of human crafted ontology can increase accuracy	-Limited words. - Some words can get same similarity if they have the same path and depth
Hybrid methods	Uses both corpus and a dictionary information.	Usually performs better	-Additional computations

Issa Atoum et.al[10], proposed a method for comprehensive comparative study of word and sentence similarity measures .The sentence similarity methods can be classified as corpus based, knowledge based and hybrid methods. Hybrid sentence methods are generally better than corpus and knowledge based methods. In the future, it is planned to test more word and sentence methods on other datasets.The word similarity is the foundation of the sentence similarity measures. A Sentence similarity method measures the semantics of group of terms in the text fragments. It has an important role in many applications such as machine translation.

Courtney Corley et.al[8] **,** described a method that combines word to word similarity metrics into a text to text metric. This method outperforms the previous text similarity metrics based on lexical matching.Measures of text similarity have been used for a long time in applications in natural language processing and related areas. In Text similarity has been also used for relevance feedback and text classification, word sense disambiguation , and more recently for extractive summarization , and methods for automatic evaluation of machine translation or text summarization.

Yuhua Li et.al[11] , proposed the novel algorithm for computing similarity between very short texts of sentence length. The proposed method introduced that takes account of not only semantic information but also word order information implied in the sentences. The similarity between two sentences is obtained the information from a structured lexical database . The word order similarity is computed from the position of word appearance in the sentence and the sentence similarity is computed as a combination of semantic similarity and word order similarity.

Table 2. Survey Table For Sentence Similarity

Authors	**Contribution**	**Approach**	**Strength**	**Limitation**
Jay J. Jiang et.al[5]	Semantic similarity between documents using hybrid approach	Corpus statistics and lexical taxonomy	Useful in word sense disambiguation	-Corpus is domain dependent. -Some words might get same similarity
Y. Liu et.al[2]	Computing similarity for Chinese-English language.	Machine Translation	It is easy to adapt to new languages and new domains.	The accuracy rate is low when some chunks are complex phrase's or sentences

Yuhua Li et.al[1]	Computing similarity between very short texts of sentence length.	Knowledge based	The short text message require a less space.	For short texts,word co-occurrence may be rare or not present.
Emiliano Giovannetti et.al [7]	Semantic relations extraction from text using hybrid approach	Statistical methods and lexico-syntactic patterns	-Improved accuracy -Usually performs better	-Additional computations -Data sparseness of the corpus
Jacob B, et. al [21]	Computing the sentence similarity from text	Jaccard similarity coefficient	Calculation of sentence similarity involves fewer computations	Word co-occurrence may be null. Considers only the surface similarity which is not reliable
Zhao Jingling et.al[6]	Measuring the sentence similarity of short text sentence	Use semantic vector model	Shows the higher Accuracy than other method	Some words can get same similarity if they have the same path and depth
LiHong Xu et.al[3]	Computing the text similarity from the sentence.	Vector space model	Allows computing a continuous degree of similarity between queries and documents.	The order in which the terms appear in the document is lost in the vector representation.
Wu et al.[22]	Semantic similarity representation of verb..	Syntactic similarity approach	knowledge of an information source is insufficient then it may be derived from the alternate information sources.	The accuracy of similarity is low.
Rada Mihalcea et.al[8]	Tex similarity detection from document	-Knowledge Based -Coroups Based	-Adoptions of human crafted ontology can increase accuracy -Preprocessed corpus to reduce computations	-Corpus is domain dependent -Limited words.

Eneko Agirre et.al[4] ,proposed the method for Smantic Textual Similarity (STS) detection from sentence .Semantic textual similarity measures the degree of semantic equivalence between two texts. This paper presents the results of the STS pilot task in Semeval. Machine translation evaluation resources and previously existing paraphrase datasets are contained 2000 pairs of sentences

Maurer, et al[12] ,focused on textual plagiarism rather than plagiarism in music, paintings, pictures, maps, technical drawings, etc., firstly they discussed the complex general setting, then report on some results of plagiarism detection software. They believed that this type of papers have a value to all researchers, educators and students and should be considered as influential work that optimistically will support many still deeper analyses. Finally they claimed that the improvement of existing plagiarism techniques and algorithms are highly needed due to increasing digitizing documents day after day.

Metzler et.al[13] , described that text similarity spans a spectrum, with broad topical similarity near one extreme and document identity at the other Intermediate levels of similarity resulting from summarization, paraphrasing, copying, and stronger forms of topical relevance are useful for applications such as information low analysis and question-answering tasks. Proposed explore mechanisms for measuring such intermediate kinds of similarity focusing on the task of identifying where a particular piece of information originated. The proposed method consider both sentence-to-sentence and document-to-document comparison, and have incorporated these algorithms into a prototype information low analysis tool.

Yusuke Shinyama et.al[14] ,proposed the method for paraphrase acquisition from news article . Article derived from the different newspaper can contain the paraphrases if they can report same event on same day. Proposed the named entity recognition.

Allan J.et.al[15] , proposed the novel approach for measuring the text similarity from sentences.Previous research in novelty detection has focused on the task of finding novel material, given a set or stream of documents on a certain topic. This study investigates the more difficult two-part task defined by the TREC 2002 novelty track: given a topic and a group of documents relevant to that topic, 1) and the relevant sentences from the documents, and 2) and the novel sentences from the collection of relevant sentences. Our research shows that the former step appears to be the more difficult part of this task, and that the performance of novelty measures is very sensitive to the presence of non-relevant sentences.

Osman, et al[16] , presented a professional study as they classified most techniques in text plagiarism into seven categories and explained the advantages and limitation each of them. Moreover they argued man important issues regarding plagiarism detection like tasks and processes of the current plagiarism detection. Finally they explained the weaknesses of some techniques which are lacking for detecting some types of plagiarized text.

Bin-Habtoor, et al[17] ,classified their survey into four categories which are plagiarism in (documents, code, techniques and algorithms). They stated that plagiarism detection for information is a big concern in universities and for teachers, policy-makers and students. Hence they proposed a system that is able to detect many plagiarism tries in deferent fields (E-Learning, E-Business, and E-Journals) and can be used to check programs, papers with images included.

Eisa et al[18] ,described and identified the state-of-the-art plagiarism techniques in terms of their attributes, limitations, processes and taxonomies. They revealed that the existing techniques are incapable to perform an intelligent detection efficiently for plagiarized ideas, figures, tables, formulas and scanned documents therefore they recommended that the integration of structural features and contextual information with semantic similarity methods can help to detect these types of plagiarism. They also stated that Turnitin is the most accurate in detection and steadiest tool among the existing seven tools, after analyzing their performance. Furthermore they discovered areas where further improvements are required in existing techniques and the current trends in plagiarism detection.

Hatzivassiloglou et al[19] ,described the knowledge-based approach and corpus-based approach for text similarity detection from sentence. Because of the growing demand from applications, this study is concerned with the development of a method by investigating the underlying information that contributes to the meaning of a sentence. The sentence similarity compute using semantic knowledge from a lexical database and statistical information from a corpus dataset. The impact of syntactic information is also considered in the calculation of similarity. The proposed algorithm differs from existing methods in two aspects. Firstly, we strictly consider text in sentence units, so the surface information is very limited compared to that in document units. Secondly we investigate a method to incorporate word order information in the detection of syntactic similarity.

Mihai Lintean et.al[20] ,described the greedy method for problem of measuring semantic similarity between short texts. Proposed method is based on the principle of compositionality which states that the overall meaning of a sentence can be captured by summing up the meaning of its parts, i.e. the meanings of words in our case. Proposed method extend wordto- word semantic similarity metrics to quantify the semantic similarity at sentence level. The results using several word-to-word semantic similarity metrics, based on word knowledge or vectorial representations of meaning. Our approach performs better than similar approaches on the tasks of paraphrase identification and recognizing textual entailment, which are two illustrative semantic similarity tasks.

III. CONCLUSION

Sentence similarity is considered the basis of many natural language tasks such as information retrieval, question answering and text summarization. The set of word and sentence similarity measures using knowledge based, corpus based , hybrid based method . The survey shows that word similarity is not enough to select a good sentence similarity measure. Hybrid sentence methods are generally better than corpus and knowledge based methods.

ACKNOWLEDGMENT

First and foremost, I would especially like to thank my adviser and guide Prof. Dr.K.P.Adhiya, for being a wonderful instructor and always being in bright spirits. I am incredibly grateful to Principal K. S. Wani , Head of the department Prof. Dr. Girish K. Patnaik and the Professors in the M.E. Computer Engineering Program at SSBT's COET,Jalgaon. Without their time and support, I would not have been able to complete this work. Last, but not least, I like to thank my family and friends who also supported me throughout my studies..

REFERENCES

1. Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering, 18(8):1138–1150, August.

2. Y. Liu and C.Q. Zong, "Example-Based Chinese-English MT", Proc.2004 IEEE Int'l Conf Systems, Man, and Cybernetics, Vol.1-7, 2004, pp.6093-6096.

3. L. Xu, D. Wang, and M. Huang, "Improved Sentence Similarity Algorithm based on VSM and its application in Question Answering System." Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on IEEE, 2010, pp. 368 - 371.

4. Eneko Agirre, Daniel Cer, Mona Diab and Gonzalez-Agirre Aitore, 2012. SemEval-2012 task6: A pilot on Semantic textual Similarity. In Proc. 6th International Workshop on Semantic Evaluation (SemEval2012), First joint conference on Lexical and Computational Semantics, Montreal, Canada.

5. Jiang, Jay J., and David W. Conrath. "Semantic similarity based on corpus statistics and lexical taxonomy", arXiv preprint cmp-lg/9709008, 1997.

6. Z.Jingling , Z. Huiyun , Cui .Baojiang ." Sentence Similarity Based on Semantic Vector Model" Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing 2014

7. Emiliano Giovannetti, Simone Marchi, and Simonetta Montemagni, "Combining Statistical Techniques and Lexico -syntactic Patterns for Semantic Relations Extraction from Text", SWAP 2008

8. R. M. Courtney Corley, \Measuring the semantic similarity of texts," in ACL workshop on Empirical Modeling of semantic Equivalence and Entailment (EMSEE),2013. IEEE,2005, pp. 13-18.

9. E. Blanco and D. Moldovan, "A Semantic Logic-Based Approach to Determine Textual Similarity," Audio, Speech, Lang. Process. IEEE/ACM Trans., vol. 23, no. 4, pp. 683–693, Apr. 2015.

10. Issa Atoum, Ahmed Otoom and Narayanan Kulathuramaiyer. Article: A Comprehensive Comparative Study of Word and Sentence Similarity Measures. *International Journal of Computer Applications* 135(1):10-17, February 2016. Published by Foundation of Computer Science (FCS), NY, USA.

11. Y. Li, D. McLean, Z. A. Bandar, J. D. OShea, and K. Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering, 18(8):11381150, August.

12. A. H. Osman, *et al.*, "Survey of text plagiarism detection," *Computer Engineering and Applications Journal (ComEngApp)*, vol. 1, pp. 37-45, 2012.

13. Metzler, D., Bernstein, Y., Croft, W., Moffat, A., and Zobel, J. "Similarity measures for tracking information flow" Proceedings of CIKM, 517– 524, 2005 .

14. Shinyama, Y., Sekine, S., Sudo, K., Grishman, R.: Automatic paraphrase acquisition from news articles. In: Proceedings of the Human Language Technology Conference (HLT-2002), San Diego, California, pp. 40–46 (2002)

15. Allan, J., Bolivar, A., and Wade, C. "Retrieval and novelty detection at the sentence level" In Proceedings of SIGIR'03, 314–321, 2003.

16. A. H. Osman, *et al.*, "Survey of text plagiarism detection," *Computer Engineering and Applications Journal (ComEngApp)*, vol. 1, pp. 37-45, 2012.

17. A. Bin-Habtoor and M. Zaher, "A survey on plagiarism detection systems," *International Journal of Computer Theory and Engineering*, vol. 4, p. 185, 2012.

18. T. A. E. Eisa, *et al.*, "Existing plagiarism detection techniques: A systematic mapping of the scholarly literature," *Online Information Review*, vol. 39, pp. 383-400, 2015

19. Burgess, C., Livesay, K., and Lund, K. 1998. Explorations in Context Space: Words, Sentences, Discourse. *Discourse Processes* 25: 211-257

20. Mihai Lintean and Vasile Rus (2005)," Measuring Semantic Similarity in Short Texts through Greedy Pairing andWord Semantics ".Proceedings of the Twenty Fifth International Florida Artificial Intelligence Research Society Conference

21. Jacob B, Benjamin C(2008) “Calculating the Jaccard Similarity Coefficient with Map Reduce for Entity Pairs in Wikipedia”, http://www.infosci.cornell.edu/weblab/papers/Bank2008.pdf

22. Z. Wu and M. Palmer, "Verb semantics and lexical selection" ,in Proceedings of 32nd annual Meeting of the Association for Computational Linguistics, (1994) June. IEEE,2005, pp. 27-30.