Paraphrase Identification using Machine Learning Techniques

A.CHITRA[1], C.S.SARAVANA KUMAR[2]
1-Professor, 2-PG Student
Department of Computer Science
PSG College of Technology
Coimbatore-641004
INDIA
csskbaba@yahoo.co.in

Abstract: - Paraphrases are different ways of expressing the same content. Two sentences are said to be paraphrases if they are semantically equivalent. Identification of paraphrases has numerous applications such as Information Extraction, Question Answering, etc. The traditional systems use threshold values to decide whether two sentences are paraphrases. This threshold determination process is independent on the training data and apart may lead to incorrect paraphrase reasoning. In order to avoid the threshold settings, we propose to use machine learning techniques. The advantages of a ML approach is its ability to account for a large mass of information and the possibility to incorporate different information sources like morphologic, syntactic, and semantic among others in a single execution. With the objective to increase the performance of the system and to develop a machine learning approach for paraphrase identification, we scrutinize the influence of the combination of lexical and semantic information, as well as techniques for classifier combination

Key-Words: - Paraphrase, SVM, Natural Language Processing, n-grams, skip grams, cardinal number

1 Introduction

A fundamental phenomenon of natural language is the variability of semantic expression, where the same meaning can be expressed by, or inferred from, different texts. Paraphrases are alternative ways to convey the same information. One can express a single event in thousands of ways in natural language sentences. This greatly adds to the difficulty of natural language processing. Paraphrase identification, the ability to determine whether or not two formally distinct strings are similar in meaning is increasingly recognized as crucial to future applications in multiple fields including Information Retrieval, Question Answering, and Summarization. Measures of text similarity have been used for a long time in applications in natural language processing and related areas. From a practical point of view, diversity in expression presents a major challenge for many NLP applications.

In this paper, we focus on the paraphrase identification. For example the sentences "When fully operational, the facility is expected to employ up to 1,000 people" and "The plant would employ 1,000 people when fully built out, the company said" express the same meaning therefore, they are paraphrases of each other. Our approach is to develop a Support Vector Machine (SVM) Classifier which uses lexical and semantic similarity information to identify the paraphrase sentences. The influences of the combination of lexical and semantic information are scrutinized, as well as techniques for classifier combination. The paper is organized in the following way: Section 2 gives the related work. Section 3 describes about the paraphrase. Section 4 gives the results and Section 5 gives the conclusion.

2 Related Works

Most systems [10] used numerous thresholds to decide definitely whether two sentences are similar and infer the same meaning. This threshold determination process is dependent on the training data and apart may lead to incorrect paraphrase reasoning. In order to avoid the threshold settings, we use machine learning techniques. The advantages of a ML approach consists in the ability to account for a large mass of information and the possibility to incorporate different information sources such as morphologic, syntactic, semantic among others in one single execution. The existing classification includes:

2.1 Rule based Machine Classification

The rule-based machine Classification paradigm includes transfer-based machine Classification, interlingual machine Classification and dictionary-based machine Classification paradigms.

2.2 Example based Machine Classification

Example-based machine Classification (EBMC) approach is often characterized by its use of a bilingual corpus as its main knowledge base, at run-time. It is essentially a Classification by analogy and can be viewed as an implementation of case-based reasoning approach of machine learning

2.3 Statistical Machine Classification:

Statistical machine Classification (SMC) is a machine Classification paradigm where Classifications are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora.

3 Paraphrasing System an Overview

The overall functioning of the system is shown below in Figure 1. In order to detect whether two sentences are semantically equivalent, lexical, semantic and syntactic features can be used [1].

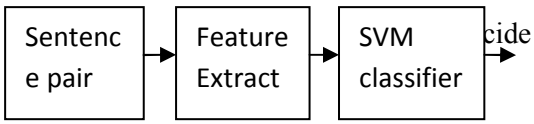

Fig.1 Paraphrase Identification

3.1 Data Set

The Microsoft Research Paraphrase Corpus [17] contains a collection of both positive and negative paraphrase pairs. Sentence pairs from this corpus serve as input to the Feature Extraction module.

3.2 Feature Extraction

The Feature Extraction Module extracts two types of features: Lexical and Semantic.

3.2.1 Lexical Features

Lexical features quantify the word level overlap found between the two input sentences. The lexical features used are:

Skip-grams:

These look for non-consecutive sequences of words that may have gaps in between, compared to all combinations of words that can appear in the sentences. Given that T1 and T2 are the two sentences the skip-gram measures used are:

Skip_gramT1 = (skip_gram(T1, T2)) / C(n,skip_gram(T1,T2)) (1)

Skip_gramT2 = (skip_gram(T1, T2)) / C(m,skip_gram(T1,T2)) (2)

where n and m are the number of words in T1 and T2; skip_gram(T1,T2) is the number of common skip-grams found in T1, T2 with the maximum length of the skip-grams set to four and C(n,skip_gram(T1,T2)) is a combinatorial function. [9]

Longest Common Subsequence:

One long common sub-sequence of words between two sentences is found. Its length is divided by the number of words present in T1 or T2. [19].

Example:

Gandhi was in London for thirty years

Gandhi was in London less than 10 years.

Here the Least Common Subsequence is four

3.2.2 Semantic Features

Semantic features attempt to detect the similarity in meaning between the candidate sentence pairs. Before measuring the semantic features, the parts-of-speech tags for the sentences are fixed using the Tree Tagger toolkit [20]. The Semantic features used are:

Noun/Verb Similarity measure:

In order to detect this measure the WordNet corpora is used [16]. The Lin similarity measure has been used. For a pair of words w1, w2:

Sim_{lin}(w1, w2)= 2*IC(Least Common Subsumer(w1,w2)/IC(w1)+ IC(w2) where IC refers to the Information Content of a word. IC(w1) = log(P(w1)) where P(w1) is the probability of w1 assessed from WordNet. Least Common Subsumer refers to the lowest super-ordinate of w1, w2 in the WordNet hierarchy. Here for the pair of sentences T1 and T2 [13], similarity is calculated by adding the Sim_{lin} measure between all pairs of nouns/verbs from the two sentences and

dividing it by the total number of noun/verb pairs.

Cardinal Number attribute:
Numbers are matched against text associated with numbers. For example '20' matches with 'more than 15'. The number of such matches is assessed. Also writing as "thirty" is transformed automatically into "30", and then is lexically matched with the corresponding number.
For Example:
Gandhi was in London for thirty years
Gandhi was in London less than 10 years.
Here the output is 0 because it is not a paraphrase.

Proper Name Attribute:
This assesses the match between proper name attributes. When there are no proper names or on matches the value is 0. When the proper names match the value is 1.
For Example:
Gandhi was in London for thirty years
Gandhi was in London less than 10 years
Here the proper name will be given as two (because we have Gandhi and London)

3.3 Support Vector Machine Classifier
Support Vector Machines are based on the notion of a "margin" either side of a hyperplane that separates two data classes. SVMs perform well for two class problems. It is important to choose a suitable kernel and SVM type for the task.
Based upon the type of operation the kernel type is selected:
Linear:
It is the default type of the kernel which involves linear operations.
Polynomial:
When polynomial operations are involved then the kernel to be involved is Polynomial
Radial Basis:
When more complex data are involved then the best choice for the kernel type would be Radial Basis
Sigmoid:
When sigmoid functions are involved then the kernel type can be selected as Sigmoid.

3.4 SVM Prediction:

SVM Prediction is made by inputting the sample data and the model created (by extracting the features) are used to test the sample and the prediction is made based on the model created.

Implementation

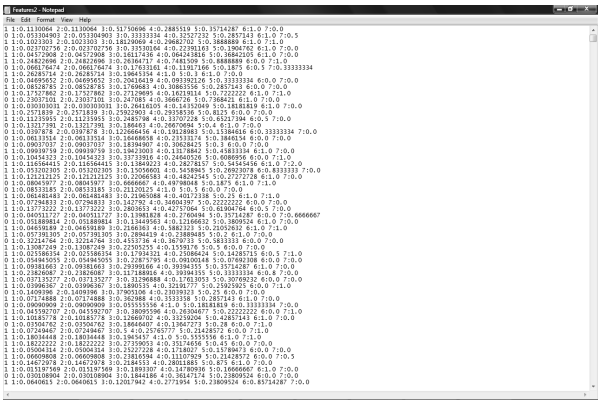

Fig.2: Feature Extraction

Fig.3: Prediction of the test data

The figure 2 depicts the feature extraction. By extracting the feature values it is organised for every pair of sentences and by giving the values as input and specifying the SVM type and kernel type the classification is made. The prediction is made by specifying the test data and it is tested with the model created by classification as shown in figure 3.

4 Results
A Radial Basis kernel using nu-SVC has been found to work well for Paraphrase Identification using the libSVM.

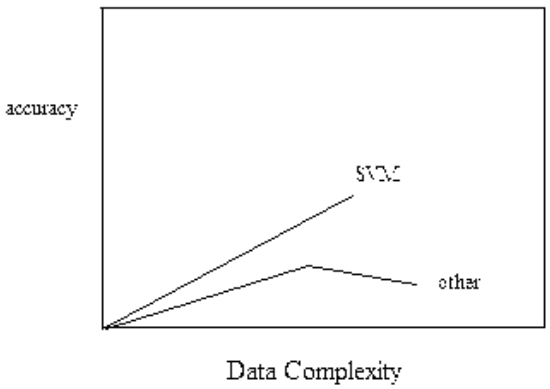

Fig.4 Comparison of SVM with Others

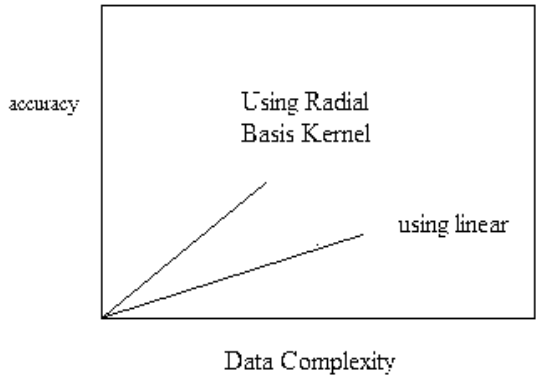

Fig.5 Comparison of Radial Basis Kernel type with the standard Linear type

5 Conclusion

The machine learning approach is used for sentence level paraphrase identification. The features set extraction include lexical and semantic features. The model built is based on the feature set extracted and the prediction is made by predicting the test sample with the model built. Several experiments were conducted and the obtained results were compared to a baseline and already existing systems. The experiments revealed that the features relying on common consecutive or insequence matches can resolve correctly 70% of the paraphrases. For all experiment, the best performance is obtained with SVM. With the analysis of the results, it has been determined that this is due to the ability of SVM to work with high dimensional attribute spaces.

In the future, incorporating a Named Entity Recognizer will improve the performance of the proper name attribute. As paraphrases act on different representation levels lexical, semantic, and syntactic or even a combination among them all, the incorporation of syntactic information would improve the recognition rate.

References:

[1] Zornitsa Kozareva and Andr´es Montoyo, Paraphrase Identification on the Basis of Supervised Machine Learning Techniques Departamento de Lenguajes y Sistemas Inform´aticos Universidad de Alicante, 2006.

[2] Regina Barzilay and Lillian Lee. Learning to paraphrase: An Unsupervised Approach using multiple-sequence alignment. *In HLT-NAACL 2003: Main Proceedings,* pages 16–23, 2003.

[3] Regina Barzilay and Kathleen McKeown. Extracting paraphrases from a Parallel corpus. *In 39th Annual Meeting of the Association for Computational Linguistics,* pages 50–57, 2001,

[4] Chris Brockett and William B. Dolan. Support vector machines for Paraphrase identification and corpus construction. *In Second International Joint Conference on Natural Language Processing*, 2005

[5] Jacob Cohen. A coefficient of agreement for nominal scales. Educ. Psychol. Meas, 1960.

[6] Ronan Collobert and Samy Bengio. Svmtorch: Support vector machines for Largescale regression problems. *Journal of Machine Learning Research, 1,* issn 1533-7928:143–160, 2001.

[7] Courtney Corley and Rada Mihalcea. Measures of text semantic similarity. *In Proceedings of the ACL workshop on Empirical Modeling of Semantic Equivalence,* July 2005.

[8] Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van Den Bosch.Timbl: Tilburg memory-based learner. Technical Report ILK 010, Tilburg University, November 2003.

[9] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using ngram co-occurrence statistics. In *Conference*

of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 71–78, 2003.

[10] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. *In Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2006.

[11] William B. Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. *In International Conference on Computational Linguistics, COLING, 2005.*

[12] Zornitsa Kozareva and Andr´es Montoyo. The role and resolution of textual entailment in natural language processing applications. *In 11th International Conference on Applications of Natural Language to Information Systems* (NLDB 2006), 2006.

[13] Dekang Lin. An information-theoretic definition of similarity. In *Proceddings of 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA, 1998.

[14] R. Barzilay and M. Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the ACLWorkshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain, August.

[15] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer.1993. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2):263–311.

[16] Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*.

[17] Chris Quirk, Chris Brockett, and William B. Dolan. Monolingual machine translation for paraphrase generation,. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[18] Ted Pedersen. Assessing system agreement and instance difficulty in the lexical sample tasks of senseval-2. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

[19] Ido Dagan and Oren Glickman. Probabilistic textual entailment: Generic applied modeling of language variability. In *PASCAL Workshop on Learning Methods for Text Understanding and Mining*.

[20] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, 1994.