Free Paraphrases of Noun Compounds

Vaibhav
Niraj Kumar
Advisor: Dr. Amitabha Mukerjee
Dept. of Computer Science and Engineering
{vaibhavv, nirajkr, amit}@ iitk.ac.in

April 17, 2013

Introduction

We often come across sentences in our daily life containing many nouns which act as a compound noun. The interpretation of these noun compounds may be a very trivial task for us but it is of great value to one who is trying to understand the semantic meaning of any sentence. For a machine to understand the meaning of such compound nouns it should be able to understand the relation between the nouns and then interpret it in that way. Example sea breeze is a compound noun and we will try to paraphrase it suitably to make sense which we automatically do without any effort. Sea breeze can be suitably paraphrased as breeze that comes from sea or breeze that flows from sea, etc.
Noun compounds are so frequent in written text, systems that deal with semantic analysis of text cannot ignore those. And because the meaning of the compound cannot be directly obtained from the nouns, the system should have some kind of way of interpreting it. This clarifies the need and significance of methods that are able to disambiguate and explain the semantics of a compound.

The remainder of this paper will describe a way to paraphrase a given two word noun compound suitably using some algorithms. This is also one of the tasks this year in the **SemEval** competition organized by University of York every year.

Related Works

Many past works have been done to solve this problem. Broadly there are two strategies to tackle this problem one is top-down and other is bottom-up.
In the top-down strategy, the problem of noun compound interpretation is basically converted into a classification problem. Girju et al. (2005) suggested 21 classes of abstract relations.[1]
The second broad strategy to interpret noun compounds is the bottom-up strategy in which noun compounds are being interpreted through paraphrasing those using suitable verb phrases.
It is very clear that the top down approach though is easy but has various drawbacks, like there is an unavoidable loss of information due to the limited classes. Where as in the bottom up approach verbs are infinite and paraphrasing using verbs and prepositions gives a more precise meaning to the compounds. A combination of both these approaches gives the most optimum results.

WordNet :: Similarity

WordNet::Similarity is an open source software package developed at the University of Minnesota. It allows the user to measure the semantic similarity or relatedness between a pair of words. We are using the WS4J, a WordNet similarity API for Java which is a reimplementation of the orginal wordNet:: Similarity developed by Prof. Ted Pedersen's group in University of Minnesota in Duluth.
The system provides six measures of similarity based on the WordNet lexical database [4]. The measures of similarity are based on analysis of the WordNet, which is a lexical Database containing words in synsets in a hierarchy.
The measures of similarity are divided into two groups: path-based and information content-based.
We chose four of the similarity measures in WordNet::Similarity for our project: WUP and LCH as path-based similarity measures, and JCN and LIN as information content-based similarity measures.

- LCH finds the shortest path between nouns
- WUP finds the path length to the root node from the least common subsumer (LCS) of the two word senses that is the most specific word sense they share as an ancestor
- JCN subtracts the information content of the LCS from the sum
- LIN scales the information content of the LCS relative to the sum

Algorithm Used

This method is used to find a similar noun compound from the training noun compounds for any given test noun compound.

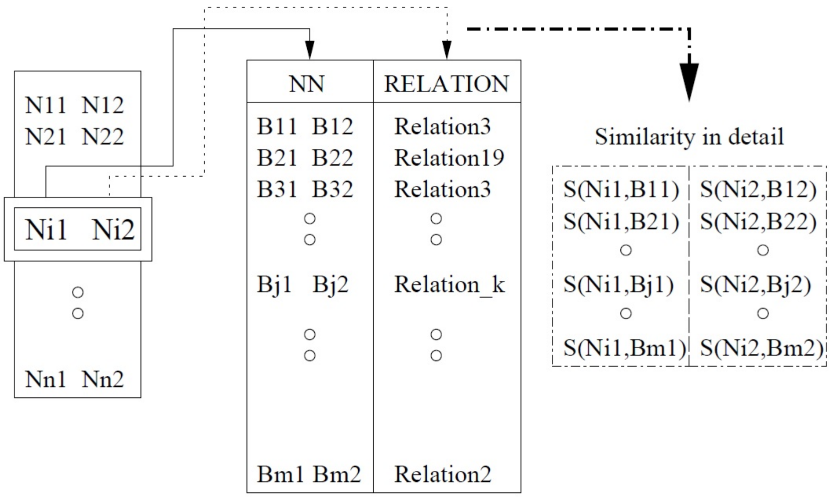

Figure 1: Similarity between the i^{th} NC in the test data and j^{th} NC in the training data [From: 2]

Implementation

For a given noun compound, firstly the verb phrases are extracted from the training dataset. Then web based validation is used to rank the paraphrases and remove senseless paraphrases.

Verb Phrase Extraction
Using the wordNet:: Similarity API we calculate the similarity between any pair of words. According to the algorithm described above, we calculate similarities between every noun compound in the training dataset and the given test noun compound. In every noun compound there is a head noun and a modifier noun, so to overcome this dilemma we calculate the similarities for the two possible pair of the test and training noun compound. For example the test noun compound is *milk shake* and one of the training noun compound is *bus stand*, then here, the similarities will be calculated for {*milk, bus*}, {*milk, stand*}, {*shake, bus*}, {*shake, stand*}. Then corresponding to each entry in the training dataset there will be two similarity index, one will be the product of the similarity of {*milk, bus*}, {*shake, stand*} and the other will be the the product of the similarity between {*milk, stand*}, {*shake, bus*}. Then we select that pair of noun compound from the training dataset which has the highest similarity quotient. We use the verb phrases of this noun compound to paraphrase the test noun compound. So a list of paraphrases is generated and they are then validated on the web.

Validation
The purpose of validation is to produce a ranked list of verb phrases and to eliminate those that are less likely to paraphrase the compound. We have used a web based validation technique, in which we are querying the generated paraphrases for the test noun compound, and then from the xml file which we get as the result, we have extracted the time latency for each query, which is then used as a count for ranking the paraphrases. The more is the time latency, the better is the paraphrase. For the web based validation we have used a free web search API, named *FAROO*.

Results

We were able to complete the SemEval task. Here are few top paraphrases which were generated for one of the noun compound *world economy*, in specific.

Paraphrase	**Frequency**	**Score**
economy for world production	412	0.4444444444444444
economy in the world industry	323	0.4000000000000000
economy involved in world manufacture	122	0.11428571428571428
economy that sells and buys world	108	0.020000000000000004
economy that deals with world	74	0.20000000000000000
economy trading in world	62	0.120000000000000000
economy that is always selling world	54	0.11428571428571428
economy is of world	48	0.80000000000000
economy which is doing business in world	47	0.08857142857142858
economy by which world are manufactured	41	0.11428571428571428
economy is for world	40	0.11657142857142858

According to the scorer given in the semEval task, the score for each paraphrase is shown, and also the cumulative score for this particular noun compound are as follows-

Actual score: 10.127489863534963;
Maximum score: 28.834343434343417;
Relative score: 0.35123011857702024

Evaluation is based on the Golstandard reference set and the scorer.java file provided in the semEval task. Two types of scores are assigned with the paraphrases of any noun compound, i.e Isomorphic and Non- isomorphic.
The order of test paraphrases is important in the isomorphic scoring mode, in which each of the test paraphrases is matched to the closest remaining reference paraphrases that has not yet been matched to one of the other test paraphrases. They will be matched in the order in which they are listed.
The order of the test paraphrases is not important in non-isomorphic scoring mode. In this mode, each of the test paraphrases is matched to the closest matching reference paraphrase, and several of the test paraphrases may match the same reference paraphrase.
Isomorphic mapping rewards both precision and recall, where as non-isomorphic mapping just rewards precision.

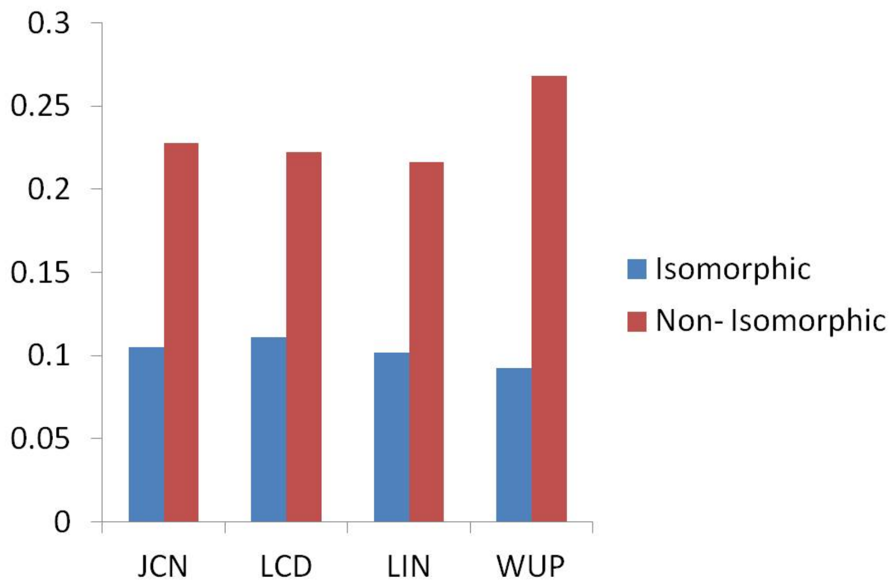

Figure 2: Scores for 20 test noun compounds for various algorithms of wordNet:: Similarity

Finally, the score was computed for the complete test dataset containing 180 test noun compounds (using LCD, which had the highest isomorphic score among all) and it was found that the overall isomorphic and non-isomorphic scores were 0.13 and 0.21 repectively.

Future Work

Many things can be done in future for further improvement in the accuracy of the results.
Selection of a more appropriate traning noun compound for a given test noun compound. Every noun compound can be classified in some semantic category like time, possessions, etc. So while calculating the similarity for each pair of test and training noun compound one can assign some weights to the more important of the noun compound pair and give it more weightage to come up with a single similarity.
One can use a more powerful web based validation technique, by using some standard web search API like Google, Yahoo, Bing, etc. After the results from web based validation, results can be further narrowed by aplying n-gram models.
A completely different approach could be to take a huge corpus consisiting of natural english sentences like BNC, and then find the suitable paraphrases for every test noun compounds.

Given the noun compound (n_1 n_2), if we can find an occurrence of n_1 with a verb phrase v_p such that it is the verbs object, and an occurrence of n_2 with the same verb phrase v_p such that it is the verbs subject, then, v_p might be suitable for paraphrasing the compound in the format: n_2 that v_p n_1. This approach is described in [3].

Acknowledgement

We thank Prof. Amitabha Mukerjee for his valuable support throughout the project, guiding us from time to time and looking into the project when it was needed. We have used the GoldStandard Reference set for calculating the final score, which was provided to us by Nitesh Surtani, IIITH. We acknowledge him for that.

Datasets

- Semeval 2013: Task 4 - Training Dataset
- Semeval 2013: Task 4 - Test Input

$$From : http : //www.cs.york.ac.uk/semeval - 2013/task4/index.php?id = data$$

- Semeval 2013: Task 4 - GoldStandard reference set for evaluation [Refer acknowledgement]
- Fellbaum, C. (1998). WordNet: An electronic lexical database, MIT press Cambridge, MA.[A]

Bibliography

[**1**] Butnariu, C., Kim, S., Nakov, P., O Seaghdha, D., Szpakowicz, S. & Veale, T. (2009). SemEval-2010 Task 9: *The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions.*, In Proceedings of the SemEval-2010 Workshop.

[**2**] Su Nam Kim and Timothy Baldwin (2005). *Automatic Interpretation of Noun Compounds using wordNet similarity.*

[**3**] Lilit Darbinyan supervised by Stephen Pulman (2010). *Interpretation of noun compounds*, Oxford University.