Using Dependency Relations to Decide Paraphrasing

Mihai Lintean and Vasile Rus and Arthur C. Graesser
Department of Computer Science
Department of Psychology
Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152, USA
{M.Lintean,vrus,a-graesser@memphis.edu}

Abstract

In this paper we investigate the role of dependency relations, i.e. explicit syntactic relations among words in a sentence, to decide whether a sentence is a paraphrase of another sentence. The proposed paraphrase algorithm basically quantifies the degree of overlap, at dependency relations level, between two sentences. We evaluated the algorithm on the industry standard data set, the Microsoft Paraphrase Corpus. The experiments revealed promising results. Another contribution of the paper is the comparison of two different methods to extract the dependency relations.

1 Introduction

Paraphrase is a text-to-text relation between two non-identical text fragments that express the same meaning in different ways. We focus on sentential paraphrases in this paper.

Paraphrases are important in a number of applications. In natural language generation, paraphrases are a method to increase diversity of generated text [1]. Paraphrases are useful in Intelligent Tutoring Systems (ITSs) with natural language input to assess whether student articulated answers to deep questions, e.g. conceptual physics questions, are similar/paraphrases-of ideal answers [2].

Paraphrase identification was previously explored, most notably by Corley and Mihalcea [4] and Qiu and colleagues [3]. Corley and Mihalcea [4] proposed an algorithm that combines word-to-word similarity metrics into a text-to-text semantic similarity metric. To get the semantic similarity between words they used the WordNet similarity package. Qiu and colleagues [3] proposed a two-phase architecture. A classifier first tries to pair nuggets (individual semantic content units) shared by two sentences, and then it classifies the rest of unpaired nuggets as significant or not for paraphrasing. If the sentences do not contain significant unpaired nuggets, then the sentences are considered paraphrases.

2 Approach

In this paper we explore the role of dependency relations in paraphrase identification. A dependency relationship [5] is an asymmetric relationship between two words in a sentence, a **head**, or modifee, and a **modifier**. A sentence can be represented by a set of dependency relations. As an example of a dependency, we consider the *subject* relation between *John* and *drives* in the sentence *John drives a car.* Such a dependency can be viewed as a triple *subj(John, drives)*.

Our approach to paraphrase identification is based on the idea that two sentences express the same meaning if they are lexically and syntactically similar. Because the lexical and syntactic aspects of a sentence can be captured as sets of dependencies, our approach relies on dependencies to identify paraphrases. Thus, two sentences are in a paraphrase relationship if their sets of dependencies are similar. To check whether two sets of dependencies are similar, we take each dependency from the first sentence and find a match in the second sentence. When matching two dependencies, the label and the base form (or stem form) of the heads and modifiers, which are the elements of a dependency, must be identical. To decide whether two sentences are in a paraphrase relation, a score is computed as the number of common dependency relations between the two sentences divided by the average number of relations in the two sentences. The result is a normalized score which should have higher values if the sentences are in a paraphrase relation and lower values otherwise. Given a score, a threshold is used to decide whether the sentences are in a paraphrase relation. The threshold was empirically established to 0.19 as it provided optimal results on the training/development data. Enforcing that the stem of the words in the dependency relation be identical may be too restrictive. We plan to explore in future work, matching dependencies even when the corresponding words are similar but not necessarily identical. To find similarity among words we could use word-to-word similarity metrics.

To obtain the dependency relations for a given sentence, we used two different parsers. The first parser is Minipar, an efficient broad-coverage dependency parser for the English language. The parser was trained on the Susanne corpus, a dependency-annotated collection of texts. The second parser we used is the Stanford Parser, which is a a phrase-based parser trained on Penn Treebank, a corpus annotated with phrase-based parse trees. Stanford parser was extended to extract dependency relations from the syntactic parse trees it generates. Because the dependency identification module is just an add-on to the standard phrase-based Stanford parser, the Stanford parser it is much slower than Minipar when it comes to extracting the dependency relations. For instance, Minipar can parse 1725 pairs of sentences, i.e. 3450 sentences, in 48 seconds while Stanford parser takes 1926 seconds, i.e. 32 minutes and 6 seconds.

3 Results

We conducted experiments on the Microsoft Research Paraphrase Corpus (MSR; [6]) which consists of 5801 newswire sentence pairs, 3900 of which were labeled as paraphrases by human annotators. The MSR corpus is divided into a training set (4076 sentence pairs) and a test set (1725 pairs). The MSR is the largest publicly available paraphrase annotated corpus that we are aware of.

We report performances using accuracy, precision, and recall. Accuracy is the proportion of sentence pairs correctly predicted. Precision is the proportion of correctly predicted paraphrases. Recall is the proportion of correct paraphrases that were predicted.

We achieved an accuracy of approximately 69.33% on the MSR test data set (using Minipar, see below) which is significantly better (Fisher's exact test yields a p=0.00005) than the accuracy of 66.49% of the naive majority classifier of guessing the majority class of paraphrase in the test set (1141 pairs out of 1725 are paraphrases). When compared to the SimFinder approach [3], which achieved the best performance of 72.90% on the MSR corpus, we could say that our approach is promising. In our tests the dependency relations computed by the Minipar parser (69.33% accuracy) seemed to give better results than the Stanford parser (68.23% accuracy). The difference is statistically significant (p=0.04 based on same Fisher's exact test). Precision is 70.64% for Stanford and 82.43% for Minipar, while Recall is 89.36% for Stanford and 90.76% for Minipar.

4 Future Work

We plan to test our current approach and future improved approaches on new paraphrase data sets from Intelligent Tutoring Systems.

References

[1] L. Iordanskaja, R. Kittredge, and A. Polgere. 1991. Natural Language Generation in Artificial Intelligence and Computational Linguistics. *Lexical selection and paraphrase in a meaning-text generation model*, Kluwer Academic.

[2] P.M. McCarthy, V. Rus, S.A. Crossley, S.C. Bigham, A.C. Graesser, and D.S. McNamara. 2007. Assessing entailer with a corpus of natural language. in *Proceeding of FLAIRS*, Menlo Park, CA. AAAI Press.

[3] Qiu, L.; Kan M.Y; and Chua T.S. 2006. Paraphrase Recognition via Dissimilarity Significance Classification In *Proceeding of EMNLP*, Sydney 2006.

[4] Corley, C., and Mihalcea, R. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. Ann Arbor, USA.

[5] Hays, D. 1964. Dependency Theory: A Formalism and Some Observations. *Languages*, 40:511–525.

[6] Dolan, B., Quirk, C., and Brockett, C. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of COLING*, Geneva, Switzerland.