Paraphrase Identification by Text Canonicalization

Yitao Zhang and **Jon Patrick**
School of Information Technology
University of Sydney
Sydney, Australia, 2006
{yitao, jonpat}@it.usyd.edu.au

Abstract

This paper proposes an approach to sentence-level paraphrase identification by text canonicalization. The source sentence pairs are first converted into surface text that approximates canonical forms. A decision tree learning module which employs simple lexical matching features then takes the output canonicalized texts as its input for a supervised learning process. Experiments on the Microsoft Research (MSR) Paraphrase Corpus give comparable performance to other systems that are equipped with more sophisticated lexical semantic and syntactic matching components, with a Confidence-weighted Score of 0.791. An ancillary experiment using the occurrence of nominalizations suggests that the MSR Paraphrase Corpus might not be a rich source for learning paraphrasing patterns.

1 Introduction

Paraphrase identification is the task of recognizing text fragments with approximately the same meaning within a specific context. It has been recently proposed as an application-independent framework for measuring semantic equivalence in text, which is critical to many natural language systems like Question Answering, Information Extraction, Information Retrieval, Document Summarization, and Machine Translation.

This paper proposes an approach to identifying sentence-level paraphrase pairs by transforming source sentences into more canonicalized text forms. By “canonical form”, we mean a transformed text which is more generic and simpler in someway than the original text, following the idea of restricted languages. For example, the sentence

> Remaining shares will be held by QVC’s management.

is transformed into a more canonicalized form by changing it from the passive to active voice producing

> QVC ’s management will hold Remaining shares.

which is more common in Subject-Verb-Object (SVO) languages like English, while the Passive Voice in the source sentence usually occurs in scientific and business text where a more formal writing style is used.

This approach is consistent with Chomsky’s Transformational Grammar, in which syntactically different, but semantically equivalent sentences can be related by their identical deep semantic structures (Chomsky, 1957). However, it is generally difficult to efficiently analyze any corpus by using the Transformational Grammar due to its high complexity and computational overhead (Hausser, 2001). In our approach, we only attempt to transform parts of the surface structure into a more generic text representation within the context of the paraphrase identification problem. The underlying hypothesis of this approach is that if two sentences are paraphrases of each other, they have a higher chance of being transformed into similar surface texts than a pair of non-paraphrase sentences.

In this paper, only a set of limited canonicalization rules have been applied as a preliminary attempt to evaluate the effectiveness of the methodology. The objective is not to create grammatically correct text sequences from source sentences, but to enable the true paraphrases to share as much surface text, both lexically and syntactically, as possible. Despite this simple model, experiments on the MSR Paraphrase Corpus nevertheless show comparable results to those scores reported in the recent ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment (2005). They also show that this approach increases the Recall rate of the system quite significantly.

Proceedings of the Australasian Language Technology Workshop 2005, pages 160–166,
Sydney, Australia, December 2005.

2 Background

Recent work on sentence-level paraphrasing generally views the problem as one of identifying bidirectional entailment in text pairs. Given an entailment text T and a hypothesis text H, T entails H if H can be inferred from the contents of T (Dagan et al., 2005). A pair of sentences is therefore considered as a paraphrase pair if the entailment relationship holds from both directions.

However, this strict mutual entailment relationship does not hold in most naturally occurred sentence-level paraphrases. Recent attempts on extracting paraphrase pairs from the web, notably the MSR Paraphrase Corpus (Dolan et al., 2004), have shown a large quantity of "more or less semantically equivalent" paraphrase pairs, such as the examples in Table 1. In the paraphrase pair "913945-914112", "Dewhurst" in the first sentence cannot be inferred from the second without giving the specific context knowledge that this person is someone belongs to the "committee". In the pair "420631-420719", the first sentence does not include any information that the minister is Saudi which occurs in the second sentence. Human judges have generally shown little difficulty in identifying these loose semantically equivalent sentence pairs as paraphrases. A surprisingly high inter-rater agreement of 83% was reported in the construction of the MSR Paraphrase Corpus despite the rather vague guideline of identifying sentence-level paraphrases that was used. It suggests that human judges were only interested in the matching of main propositions in sentence pairs, while neglecting the existence of other non-entailed trivial contents.

Bar-Haim et al. (2005) decomposed the entailment task into two sub-levels, namely, lexical and lexical-syntactic. At the lexical level, for each word or phrase h in Hypothesis H, if h can be matched with a corresponding item t in Text T using either lexical matching, or a sequence of lexical transformations, then H and T are tagged as a true entailment pair. Lexical transformation rules include morphological derivations like nominalization (example "913945-914112" in Table 1, "proposal => propose"), ontological relations like synonym and hypernym, or world knowledge such as "Taliban => organization". At the lexical-syntactic level, entailment between H and T holds if both the lexical and syntactic relations in H are also found in T. The relations evaluated at the lexical-syntactic level include syntactic movement triggered by morphological derivation of words, passive to active voice transformation of verbs, co-reference in text, and the syntactic level paraphrases like "X was born in Y <=> X is Y man by birth". In an empirical analysis of the PASCAL Recognising Textual Entailment Challenge (RTE) corpus (Dagan et al., 2005), 240 sentence pairs were randomly chosen and tagged by human annotators based on the above criteria for semantic entailment. What they have found is that working on the lexical-syntactic level outperforms on the lexical level by a significant increase of the Precision score, namely, from 59% to 86%. However, the Recall rate shows only 6% improvement by switching from lexical to a lexical-syntactic level.

In a similar effort to evaluate the contribution of syntactic knowledge in the entailment task, Vanderwende et al. (2005) found that 37% of the RTE Entailment Corpus examples could be handled by syntax alone, assuming the existence of an ideal parser. With additional help from a thesaurus, this figure can be increased to 49%.

Corley and Mihalcea (2005) proposed a bag-of-words model for identifying entailments and paraphrases by measuring the semantic similarity of two texts. In their model, the semantic similarity of two text segments T_i and T_j is defined as a score function that combines the semantic similarities of nouns and verbs, the lexical similarities of other open class words, together with word specificities measured by the inverse document frequency metric derived from the British National Corpus. Experimental results on the MSR Paraphrase Corpus showed a 4.4% increase of system accuracy by incorporating semantic knowledge.

Inversion Transduction Grammars (ITG), which is previously proposed as a framework for machine translation, has also been applied in the context of the paraphrase and entailment task by Wu (2005). Without consulting any thesaurus, the Bracketing ITG model worked mainly on a syntactic matching level and achieved a Confidence-weighted Score of 0.761, which is 10% higher than the random baseline.

3 The Dataset

The Microsoft Research Paraphrase Corpus has been used throughout our experiments. It is the result of a recent effort to construct a large scale

Table 1: Examples of MSR Paraphrase Corpus

ID	Text1	Text2	Description
913945-914112	Dewhurst’s proposal calls for an abrupt end to the controversial ”Robin Hood” plan for school finance.	The committee would propose a replacement for the ”Robin Hood” school finance system.	Nominalization
2484044-2483683	The tour plans to make stops in 103 cities before rallying in Washington on Oct. 1-2, and in New York City on Oct. 3-4.	The tour will stop in 103 cities before rallying in Washington on Oct. 1 and 2, and New York on Oct. 3 and 4.	Nominalization + Future Tense
420631-420719	Those reports were denied by the interior minister, Prince Nayef.	However, the Saudi interior minister, Prince Nayef, denied the reports.	Passive/Active Voice

paraphrase corpus for generic purposes (Dolan et al., 2004). It consists of 5,801 sentence pairs extracted from online newswire text, in which 3,900 are tagged as true paraphrases by human judges. This high proportion of occurrences of paraphrase pairs can be explained by the methodology used to create the corpus. In the construction of the corpus, edit distance is used as the only metric to filter out lexically unsimilar sentence pairs, which means the remaining instances have large lexical overlaps. As a consequence, although the MSR Paraphrase Corpus is rich in the number of paraphrase pairs, it is not enriched with a good variety of lexical and syntactic patterns. Weeds et al. (2005) argue that this ”high overlap in words” makes it a poor source for studying the distributional similarity of syntactic paraphrasing patterns.

In an effort to substantiate this claim, we made an evaluation of the occurrence of nominalization, which is a classical linguistic device for paraphrasing, in both the MSR Paraphrase Corpus and the RTE Entailment Corpus. We used a semi-automatic method to calculate the occurrence of nominalizations. First we pos-tagged sentence pairs in the corpus and lemmatized all the verbs and nouns. If there was an exact string match between a lemmatized verb and a lemmatized noun in a sentence pair, we marked it as a candidate of nominalization, and asked human judges to verify it at a later stage. A walk-through example of finding nominalization is shown in Table 2.

This method gives a reliable lower bound on the occurrence of nominalizations in the corpora. The results are shown in Table 3. Notice that in the MSR training dataset only 60 true nominalizations exist in over 4,000 sentence pairs, compared to the number of 44 over 800 in the RTE testing dataset. This result suggests that the distribution of paraphrasing patterns in the MSR Paraphrase Corpus is likely to be below the normal distribution in natural text, or at least not that rich compared to a human constructed and balanced corpus. Therefore, it might not be a rich resource for studying the real distribution of features of naturally occurring paraphrases and Weeds et al.’s comments are justified.

Table 3: Occurrence of Nominalizations

	True Nominalizations	Corpus Size(sentence pairs)
RTE	44	800
MSR	60	4076

Despite these innate problems of the corpus, it is still by far the largest sample dataset of paraphrasing phenomenon, which provides a solid base for system testing. Therefore, we decided to focus our research on this corpus as the first stage of our experiments.

4 Experiments

This section describes the details of the two modules in the system, namely the text canonicalization module and the supervised learning module.

4.1 Text Canonicalization

The function of the text canonicalization module is to constrain the language choices, both at lexical and syntactic level, of any text that carries meanings. In this paper, only a set of limited canonicalization rules has been applied.

Number Entities Number entities include dates, times, monetary values, and other quan-

Table 2: An Example of Finding Nominalizations

ID	913945	914112
	Dewhurst/NNP ’s/POS proposal/NN calls/VBZ for/IN an/DT abrupt/JJ end/NN to/TO the/DT controversial/JJ ”/NNP Robin/NNP Hood/NNP ”/NNP plan/NN for/IN school/NN finance/NN ./.	The/DT committee/NN would/MD propose/VB a/DT replacement/NN for/IN the/DT ”/NNP Robin/NNP Hood/NNP ”/NNP school/NN finance/NN system/NN ./.
Nouns	proposal=>propos, end, Robin, Hood, plan, school, finance=>financ	committee=>committe, replacement=>replac, Robin, Hood, school, finance=>financ, system
Verbs	calls=>call	propose=>propos
Candidate Nominalizations: (proposal, propose)		

Table 4: Passive to Active Voice

	id = 420631	id = 420719
Before transformation	Those reports were denied by the interior minister, Prince Nayef.	However, the Saudi interior minister, Prince Nayef, denied the reports.
After transformation	the interior minister, Prince Nayef denied Those reports.	unchanged

tities like percentages. In the experiments, the system will replace those number entities with generic tags in the text.

Passive/Active Voice In the passive to active voice transformation, the system first consults Minipar (Lin, 1998), which is a principle-base English parser, to get the parsed dependency tree structure of the text. Then it finds all the verbs in passive voice, together with their grammatical subjects and the objects. Finally, the system swaps the child nodes of the subjects and the objects of each verb. The canonicalized text is then created from the transformed syntactic tree. An example of passive to active voice transformation is shown in Table 4.

Future Tense The expression of future tense in text has also been canonicalized to constrain the lexical choices which refer to future action and willingness. An example of future tense usage in the MSR Paraphrase Corpus is given by the text pair ”2484044-2483683” in Table 1. In the sentence, “plans to” and “will” both refer to the future actions the subject will be taking. They have to be canonicalized into the same surface text to create higher probabilities to be matched at a later stage. In the experiments, we compile a list of common words and phrase structures(like “plan to” and “be expected to”) to be substituted by a single word “will”, which the system defines as the generic expression of future actions.

4.2 Supervised Learning

At the supervised learning stage, the decision tree learning module of Weka (Witten and Frank, 1999) was used. The training dataset and the test dataset used in the experiments are the corresponding training and test dataset in MSR Paraphrase Corpus as described in Section 3.

Lexical Matching Features The features used in the supervised learning stage are

- *Longest Common Substring* measures the length of the longest common strings shared by two sentences. It is a consecutive sequence of words.

- *Longest Common Subsequence* measures the length of the longest common sequence of strings shared by two sentences. It does not require this sequence to be consecutive in the original text.

- *Edit Distance* describes how many edit operations (add, delete, or replace of a word token at a time) are required to convert a source text into a target text. The fewer edit operations needed, the less edit distance and the more lexical overlap of the two text segments.

- *Modified N-gram Precision* is also an important metric adopted from the BLEU algorithm for evaluating machine translations (Papineni et al., 2001). It was originally proposed to capture both the accuracy and the fluency of a translated text with reference to a set of candidate translations. In the context of paraphrases, we try to calculate the modified n-gram precision from both directions of a sentence pair. For example, given the following sentence pair:

T_1: the the the the the the the.
T_2: The cat is on the mat.

we first define the modified count of an n-gram t in T_1 as the minimum between the occurrence of t in T_1 and the maximum occurrence of t in T_2. For instance, $Count_{\text{modified}}(\text{“the”})$ is 2 because the unigram “the” occurs only twice in the second sentence. The directional modified n-gram precisions from T_1 to T_2 is defined in Equation 1, in which m is the order of n-gram (up to trigram m=3 was used in our experiment), and $Count(k)$ simply counts the number of k in the source sentence T_1. We also calculated the directional modified n-gram precision score from T_2 to T_1, and used the average of the two directional precision as the modified n-gram precision of the sentence pair by Equation 2.

Moreover, our calculation of the above features is solely based on word token level. For instance, we use word n-gram instead of letter n-gram in calculating the modified n-gram precision.

$$mnp_{T_1} = \frac{1}{m}\sum_{i=1}^{m} -\log\left(\frac{\sum_{t\in\text{n-gram}_i} Count_{\text{modified}}(t)}{\sum_{k\in\text{n-gram}_i} Count(k)}\right) \tag{1}$$

$$mnp(T_1, T_2) \;=\; \frac{mnp_{T_1} + mnp_{T_2}}{2} \tag{2}$$

Evaluation Metrics To assess the system performance, we adopt the Confidence-weighted Score(CWS) as the main figure for our evaluations. CWS is defined in Equation 3

$$cws \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{\text{\#correct-up-to-i}}{i} \tag{3}$$

in which #correct-up-to-i is the number of correct tagging instances up to the current position i, and the test data samples are first ranked in decreasing order according to their confidence level of tagging judgments. The CWS metric generally rewards a system that assigns higher confidence values to correct tagging decisions than to those wrong ones (Dagan et al., 2005). Meanwhile, traditional machine learning metrics like accuracy, precision, recall, and F_1 values are also reported for better understanding of the system.

The Baselines Two baselines have been provided for the task. The first baseline system uniformly predicts true for paraphrase pairs. The second baseline system uses the lexical matching features in Section 4.2 on the original text pairs for the supervised learning stage.

5 Results

The experiment results are shown in Table 5. For comparison, scores of Wu (2005), and Corley and Mihalcea (2005)'s systems are also included in the table.[1]

For the two baseline systems, B2, which employs pure lexical matching features on the source text, outperforms B1, the system that uniformly predicts paraphrases, both in Accuracy by 6%, and in CWS by 12%. The B2 system also shows comparable results with respect to Wu, and Corley and Mihalcea's systems and sets a high standard as a baseline system. This further reveals the main characteristic of the MSR Paraphrase Corpus: paraphrase text pairs in the corpus share more lexical overlaps than non-paraphrase pairs.

Compared with B2, systems using canonicalized text, namely S1 - S7, generally suffer a slightly poorer performance in the Accuracy score. However, the Recall rate rises significantly in all systems except in S3 and S6. Interestingly, S3 and S6 also show the highest CWS score and the Precision score at the same time. This suggests that the canonicalization of future tense helps systems to make more precise and reliable tagging decisions. Canonicalization on Passive/Active voice (S2) also increases the Recall rate by almost 10% compared with B2. This suggests that a pure lexical matching system could be further improved by even some preliminary syntactic transformations. Number entity canonicalization helps to increase the Recall rate of the system. This could be explained

[1] Wu only reported the CWS score on MSR corpus in his paper, while Corley and Mihalcea did not report any CWS score in their paper.

Table 5: Experiment results on MSR Paraphrase Corpus

	CWS	Acc	Pre	Rec	F_1
Systems using Canonicalized Text					
S1: (a)number entities	0.740	0.692	0.713	0.898	0.795
S2: (b)passive/active	0.742	0.719	0.743	0.882	0.807
S3: (c)future tense	0.791	0.708	0.784	0.775	0.779
S4: (a)+(b)	0.739	0.697	0.716	0.900	0.798
S5: (a)+(c)	0.731	0.701	0.732	0.869	0.794
S6: (b)+(c)	0.791	0.709	0.784	0.776	0.780
S7: (a)+(b)+(c)	0.723	0.703	0.734	0.867	0.795
Baselines					
B1: Uniform	0.664	0.664	0.664	1	0.798
B2: LexicalMatch	0.783	0.723	0.788	0.798	0.793
Other Systems with Reported Scores					
Wu (2005)	0.761				
Corley and Mihalcea (2005)		0.715	0.723	0.925	0.812

by how the MSR Paraphrase Corpus was constructed. During the tagging process, source sentences were already pre-processed by replacing number entities with generic tags. Human judges then made their decisions based on the canonicalized text. While the dataset revealed to the public, the source text is provided instead of the data used by human judges.

In general, systems S1-S7 show competitive performance with respect to Wu, and Corley and Mihalcea's systems. Corley and Mihalcea's system gives a better Recall rate, which suggests the importance of introducing lexical semantics features in the system. Our approach currently does not model synonyms into any canonicalized form, therefore loses the possibility of capturing this lexical variance. On the other hand, neither Wu, nor Corley and Mihalcea's system outperforms the lexical matching system B2 in terms of CWS and Accuracy. This again suggests that the nature of the paraphrases in the corpus is that they share more lexical overlaps than non-paraphrases, rather than employing sophisticated syntactic paraphrasing patterns.

6 Conclusion

This paper proposes a text canonicalization approach to the paraphrase identification task. Our approach tries to tackle the problem on both the lexical and the grammatical level, as distinct from existing research which has concentrated on lexical analyses. Despite the simple transformation rules applied, this approach has shown competitive figures of system performance on the MSR Paraphrase Corpus with that reported in current state-of-the-art systems. Moreover, this method reports a significant increase in the recall rate of paraphrases compared with a system using non-canonicalized text. It clearly encourages the use of more conceptualized and more canonical syntax which tries to approximate the deeper semantic information of the original text.

However, further research is required to reveal how many transformation rules are needed for the task. It would also be interesting to develop an effective engineering method for managing the expanding canonicalization rule set. In the future, more work has also to be done to equip the system with lexical semantic knowledge from either manually constructed lexical databases like WordNet (Fellbaum, 1998), or other resources that automatically learned from corpora like VerbOcean (Chklovski and Pantel, 2004).

Acknowledgments

This paper was supported by the School of Information Technologies, University of Sydney. We would also like to thank the reviewers for their insightful comments, and all members of the Sydney Language Technology Group for their support.

References

Roy Bar-Haim, Idan Szpecktor, and Oren Glickman. 2005. Definition and analysis of intermediate entailment levels. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 55–60, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, Barcelona, Spain.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Ido Dagan, Bernardo Magnini, and Oren Glickman. 2005. The PASCAL Recognising Textual Entailment Challenge. In *PASCAL Proceedings of the First Challenge Workshop, Recognizing Textual Entailment*, Southampton, U.K., April.

Bill Dolan and Ido Dagan, editors. 2005. *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. Association for Computational Linguistics, Ann Arbor, Michigan, June.

William B. Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of COLING 2004*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Roland Hausser. 2001. *Foundations of Computational Linguistics: Human-Computer Communication in Natural Language*. Springer.

Dekang Lin. 1998. Dependency-based Evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*, Granada, Spain, May.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation.

Lucy Vanderwende, Deborah Coughlin, and Bill Dolan. 2005. What Syntax can Contribute in Entailment Task. In *PASCAL Proceedings of the First Challenge Workshop, Recognizing Textual Entailment*, Southampton, U.K.

Julie Weeds, David Weir, and Bill Keller. 2005. The distributional similarity of sub-parses. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 7–12, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Ian H. Witten and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

Dekai Wu. 2005. Recognizing paraphrases and textual entailment using inversion transduction grammars. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 25–30, Ann Arbor, Michigan, June. Association for Computational Linguistics.