Paraphrase Substutution for Recognizing Textual Entailment

Wauter Bosma[1] and Chris Callison-Burch[2]

[1] University of Twente,
NL-7500AE Enschede, the Netherlands,
bosmaw@cs.utwente.nl
[2] University of Edinburgh,
2 Buccleuch Place, Edinburgh, EH8 9LW, United Kingdom
callison-burch@ed.ac.uk

Abstract. We describe a method for recognizing textual entailment that uses the length of the longest common subsequence (LCS) between two texts as its decision criterion. Rather than requiring strict word matching in the common subsequences, we perform a flexible match using automatically generated paraphrases. We find that the use of paraphrases over strict word matches represents an average F-measure improvement from 0.22 to 0.36 on the CLEF 2006 Answer Validation Exercise for 7 languages.

1 Introduction

Recognizing textual entailment has recently generated interest from a wide range of Natural Language Processing related research areas, such as automatic summarization, information extraction and question answering. Advances have been made with various techniques, such as aligning syntactic trees and word overlap. While there is still much room for improvement, Vanderwende and Dolan [1] showed that current approaches are close to hitting the boundaries of what is feasible with lexical-syntactic approaches.

Proposed directions to cross this boundary include using logical inference, background knowledge and paraphrasing [2]. We explore the possibility of applying paraphrasing to obtain a more reliable match between a given text and hypothesis for which the presence of an entailment relation is to be determined. For instance, consider the following RTE2 pair.

Text: *Clonaid said, Sunday, that the cloned baby, allegedly born to an American woman, and her family were going to return to the United States Monday, but where they live and further details were not released.*

Hypothesis: *Clonaid announced that mother and daughter would be returning to the US on Monday.*

In this example, text and hypothesis use different words to express the same meaning. Although deep inference is required to recognize that a mother is part

of the family, and that *daughter* and *baby* in this context most likely refer to the same person, most variation occurs on the surface level. For instance, *announced* in the hypothesis could be replaced by *said* without changing its meaning. Similarly, the phrases *the United States* and *the US* would not be matched by a system relying solely on word overlap.

The criterion that our system uses to decide whether a text entails a hypothesis is the length of the longest common subsequence (LCS) between the passages. Rather than identifying the LCS using word matching, our system employs an automatic paraphrasing method that extends matches to synonymous, but non-identical phrases. We automatically generate our paraphrases by extracting them from bilingual parallel corpora.

Whereas many systems use dependency parsers and other linguistic resources that are only available for a limited number of languages, our system employes a method that is comparatively language independent. For this paper we extract paraphrases in Dutch, English, French, German, Italian, Spanish, and Portuguese.

The paraphrase extraction algorithm is described in section 2. Section 3 describes how the entailment score is calculated, and how paraphrases are generated in the entailment detection system. The results of our participation in the CLEF2006 Answer Validation Exercise [3] (henceforth AVE) are described in section 4. We will wrap up with a conclusion and directions for future work in section 5.

2 Paraphrase extraction

Paraphrases are alternative ways of conveying the same information. The automatic generation of paraphrases has been the focus of a significant amount of research lately [4–7]. In this work, we use Bannard and Callison-Burch's method [7], which extracts paraphrases from bilingual parallel corpora.

Bannard and Callison-Burch extract paraphrases from a parallel corpus by equating English phrases which share a common foreign language phrase. English phrases are aligned with their foreign translations using alignment techniques drawn from recent *phrase*-based approaches to statistical machine translation [8]. Paraphrases are identified by pivoting through phrases in another language.

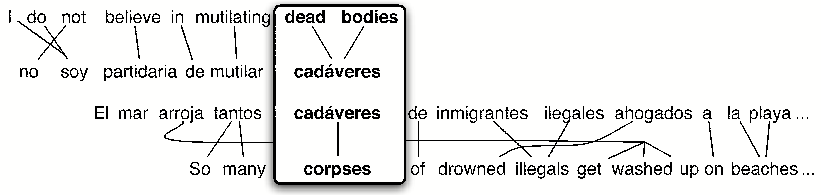

Fig. 1. A bilingual parallel corpus can be used to extract paraphrases

bodies	0.21
dead bodies	0.17
body	0.09
deaths	0.07
dead	0.07
corpses	0.06
bodies of those killed	0.03
the dead	0.02
carcasses	0.02
corpse	0.01

Table 1. Examples paraphrases and probabilities for the phrase *dead bodies*

Candidate paraphrases are found by first identifying all occurrences of the English phrase to be paraphrased, then finding its corresponding foreign language translations of the phrase, finally looking at what other English phrases those foreign languages translate back to.

Figure 1 illustrates how a Spanish phrase can be used as a point of identification for English paraphrases in this way. Often there are many possible paraphrases that can be extracted for a particular phrase. Table 1 shows the paraphrases that were automatically extracted for an English phrase. In order to assign a ranking to a set of possible paraphrases, Bannard and Callison-Burch define a paraphrase probability.

The paraphrase probability $p(e_2|e_1)$ is defined in terms of two translation model probabilities: $p(f|e_1)$, the probability that the original English phrase e_1 translates as a particular phrase f in the other language, and $p(e_2|f)$, the probability that the candidate paraphrase e_2 translates as the foreign language phrase. Since e_1 can translate as multiple foreign language phrases, f is marginalized out:

$$p(e_2|e_1) = \sum_f p(f|e_1)p(e_2|f) \tag{1}$$

The translation model probabilities can be computed using any standard formulation from phrase-based machine translation. For example, $p(e_2|f)$ can be calculated straightforwardly using maximum likelihood estimation by counting how often the phrases e and f were aligned in the parallel corpus:

$$p(e_2|f) \approx \frac{count(e_2, f)}{\sum_e count(e, f)} \tag{2}$$

We extend the definition of the paraphrase probability to include multiple corpora, as follows:

$$p(e_2|e_1) \approx \frac{\sum_{c \in C} \sum_{f \ in \ c} p(f|e_1)p(e_2|f)}{|C|} \tag{3}$$

Pair: 573 (entailment: yes) **Text:** Riding the crest of that wave is Latvia's new currency the handily named lat introduced two years ago. **Hypothesis:** *The* Latvian *currency* is *lat.* (negative judgment; entailment score = 0.60) **Hypothesis paraphrase:** *The Latvia currency lat.* (positive judgment; entailment score = 1.00)
Pair: 2430 (entailment: yes) **Text:** India and Pakistan fought two of their three wars over control of Kashmir and their soldiers still face off across the Siachen Gracier 20 000 feet above sea level in the Himalayas. **Hypothesis:** *India and Pakistan* have *fought two wars* for the possession *of Kashmir.* (negative judgment; entailment score = 0.67) **Hypothesis paraphrase:** *India and Pakistan fought two of their wars* for possession *of Kashmir.* (positive judgment; entailment score = 0.83)
Pair: 8597 (entailment: yes) **Text:** Anthony Busuttil, Professor of Forensic Medicine at Edinburgh University, examined the boy. **Hypothesis:** *Anthony Busuttil* is *professor of Forensic Medicine at* the University of *Edinburgh.* (negative judgment; entailment score = 0.67) **Hypothesis paraphrase:** *Anthony Busuttil professor of Forensic Medicine at Edinburgh University.* (positive judgment; entailment score = 1.00)

Table 2. Examples of text/hypothesis pairs from CLEF AVE for which paraphrasing was required to make a correct assessment. The words in italics are hypothesis words which are aligned with the text sentence.

where c is a parallel corpus from a set of parallel corpora C. Thus multiple corpora may be used by summing over all paraphrase probabilities calculated from a single corpus (as in Equation 1) and normalizing by the number of parallel corpora. We calculate the paraphrase probabilities using the Europarl parallel corpus [9], which contains parallel corpora for Danish, Dutch, English, French, Finnish, German, Greek, Italian, Portuguese, Spanish and Swedish.

The method is multilingual, since it can be applied to any language which has a parallel corpus. Thus paraphrases can be easily generated for each of the languages in the CLEF AVE task using the Europarl corpus.

3 Recognizing entailment

The longest common subsequence (LCS) is used as a measure of similarity between passages. LCS is also used by the ROUGE [10] summarization evaluation package to measure recall of a system summary with respect to a model summary. We use it not to measure recall but precision, to approximate the ratio of information in the hypothesis which is also in the text. Unlike the longest common *substring*, the longest common subsequence does not require adjacency. A longest common subsequence of a text $T = \langle t_1..t_n \rangle$ and a hypothesis $H = \langle h_1..h_n \rangle$ is defined as a longest possible sequence $Q = \langle q_1..q_n \rangle$ with words in Q also being words in T and H in the same order. $\mathrm{LCS}(T, H)$ is the length of the longest common subsequence:

$$\mathrm{LCS}(T, H) = max\{\ |Q| \mid Q \subseteq T \cup H; (t_i = h_k \in Q \wedge t_j = h_l \in Q \wedge j > i) \rightarrow l > k\} \tag{4}$$

From the LCS, the entailment score $\mathrm{LCS}(T, H)/|H|$ is derived. In order to account for variation in natural language text, the LCS is measured after paraphrasing the hypothesis. The underlying idea is that whenever a paraphrase of

	baseline			LCS			LCS after paraphr.			tree alignment		
lang	Pr.	Rec.	F	Pr.	Rec.	F	Pr.	Rec.	F	Pr.	Rec.	F
DE	.251	1.00	.401	.400[1]	.085[1]	.140[1]	.403	.229	.292			
EN	.158	1.00	.273	.312	.181	.229	.304[2]	.479[2]	.372[2]	.343	.512	.410
ES	.294	1.00	.454	.626	.262	.370	.504[2]	.580[2]	.539[2]	.481[1]	.456[1]	.468[1]
FR	.230	1.00	.374	.463[1]	.052[1]	.094[1]	.495	.210	.295			
IT	.172	1.00	.294	.328[1]	.112[1]	.167[1]	.380	.305	.338			
NL	.104	1.00	.188	.217	.160	.184	.199[2]	.346[2]	.252[2]	.287[1]	.593[1]	.387[1]
PT	.237	1.00	.383	.578[1]	.255[1]	.354[1]	.417	.468	.441			

Table 3. Precesion, recall and F-measure for the baseline (100% YES), LCS, LCS after paraphrasing, and dependency tree alignment.

the hypothesis exists which entails the text, the hypothesis itself also entails the text.

We attempted to extract paraphrases for every phrase in the hypothesis of up to 8 words. Note that by "phrase" we simply mean an (ordered) sequence of words. After generating these candidate mappings we iteratively transform the hypothesis to be closer to the text by substituting in paraphrases. At each iteration, the substitution is made which constitutes the greatest increase of the entailment score. To prevent overgeneration, a word which was introduced in the hypothesis by a paraphrase substitution cannot be substituted itself. The process stops when no more substitutions can be made which positively affect the entailment score. By example, the following paraphrase of the hypothesis from section 1 is obtained by a number of substitutions.

Hypothesis: *Clonaid* announced *that* mother *and* daughter would be returning *to the* US on *Monday*.
Substitutions:
the US → the United States
returning → return
said → announced
would be → is
on Monday → Monday
Paraphrased hypothesis: *Clonaid said that* mother *and* daughter is *return to the United States Monday*.

In this case, paraphrasing caused the length of the LCS to increase from 43% ($\frac{6}{14}$) to 77% ($\frac{10}{13}$). The words in italics are the words which are aligned with the text sentence, i.e. which are part of the longest common subsequence. Table 2 shows a number of CLEF AVE pairs for which paraphrases were used to recognize entailment.

[1] These runs are submitted.
[2] The submitted runs had a slightly higher precision and lower accuracy, due to a preprocessing error.

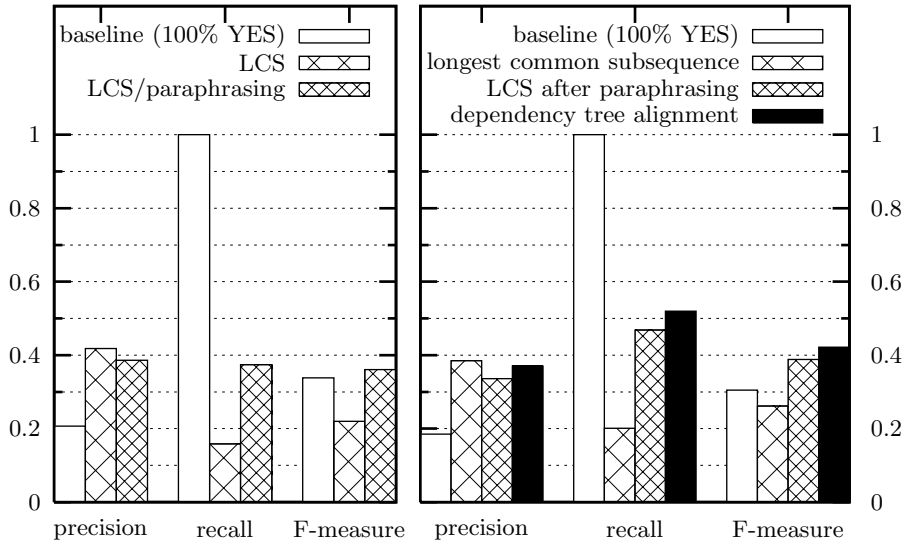

Fig. 2. *Left,* average scores of various metrics for German, English, Spanish, French, Italian, Dutch and Portugese; *right,* average scores, including dependency tree alignment, for English, Spanish and Dutch.

In order to judge whether a hypothesis is entailed by a text we see if the value of the entailment score, $\mathrm{LCS}(T, H)/|H|$, is greater than some threshold value. Support vector machines [11] are used to determine the entailment treshold. Unfortunately, the only suitable training data available was the Question Answering subset of the RTE2 data set [2]. This is a monolingual collection of English passage pairs, with for each pair a boolean annotation of the presence of an entailment relation. Lacking training data for other languages, for our submission we used the RTE2 data to learn the entailment treshold for all languages. The threshold value use used throughout these experiments was 0.75.

4 Results

We compared the performance of the paraphrasing method with two baselines on seven languages within the the CLEF 2006 Answer Validation Exercise. The first baseline is a system which always decides that the hypothesis is entailed. The second baseline is a system which measures the longest common subsequence of text and hypothesis. Table 3 lists the performance of the baselines, the paraphrase-based system and the system which uses dependency trees. Average performance over a number of languages is visualized in Figure 3. Given the fact that paraphrasing is a form of query expansion, we expected that precision drops and recall increases when using paraphrases. Results show show that this

LCS after paraphrasing	dependency tree alignment	entailment relation	
		yes	no
yes	yes	5.9%	8.6%
yes	no	**3.5%**	**7.6%**
no	yes	**3.3%**	**6.5%**
no	no	5.8%	58.8%

Table 4. Entailment assessments by both systems. Percentages of pairs on which the algorithms disagreed are boldfaced. The percentages are averages over Dutch, Spanish and English, compensated for the number of pairs in each language.

is indeed the case, but that the system using paraphrases shows considerably better overall performance, as indicated by the F-measure, compared to plain LCS.

For Dutch, Spanish and English, we made a syntactic analysis of each sentence using the parsers of [12], [13] and [14] respectively. As a fourth entailment recognition system, we measured the largest common subtree of the dependency trees of the text and hypothesis. The algorithm of Marsi et al. [15] was used to align dependency trees. Interestingly, as shown in Figure 3 (right), the dependency tree alignment system performs comparably to the largest common subsequence after paraphrasing, while the first uses syntactic information and the latter uses paraphrase generation. The fact that both systems disagree on 37 percent of all pairs with positive entailment (see Table 4) indicates that performance can be further increased when employing both types of information in an integrated system.

5 Conclusion

We evaluated the effect of paraphrasing on a longest common subsequence-based system for recognizing textual entailment. In our CLEF experiments on 7 languages, the system using paraphrases outperformed the system relying on merely the longest common subsequence. Our method is applicable to a wide range of languages, since no language specific natural language analysis or background knowledge is used other than paraphrases automatically extracted from bilingual parallel corpora. Although our system performs similarly to a syntax based system, we showed that there is relatively little overlap between the sets of correctly recognized pairs of both systems. This indicates that information conveyed by paraphrases and syntax are largely complementary for the task of recognizing entailment. In the future we plan to investigate if our system can be improved by using a combination of syntax-based and paraphrase-based approaches to

entailment recognition. Also, we plan to improve methods for determining the entailment threshold.

Acknowledgments

This work is funded by the Interactive Multimodal Information Extraction (IMIX) program of the Netherlands Organization for Scientific Research (NWO).

References

1. Vanderwende, L., Dolan, W.B.: What syntax can contribute in the entailment task. In: PASCAL Challenges Workshop on Recognizing Textual Entailment, Southampton, United Kingdom, Springer-Verlag (2005) 205–216
2. Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., Szpektor, I.: The second PASCAL Recognising Textual Entailment Challenge. In Magnini, B., Dagan, I., eds.: Proceedings of the Second PASCAL Recognising Textual Entailment Challenge, Trento, Italy (April 2006)
3. Peñas, A., Rodrigo, Á., Sama, V., Verdejo, F.: Overview of the answer validation exercise 2006. In: Evaluation of Multilingual and Multi-modal Information Retrieval — Seventh CLEF Workshop. LNCS, Alicante, Spain (September 2006)
4. Barzilay, R., McKeown, K.: Extracting paraphrases from a parallel corpus. In: ACL-2001. (2001)
5. Pang, B., Knight, K., Marcu, D.: Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In: Proceedings of HLT/NAACL. (2003)
6. Barzilay, R., Lee, L.: Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In: Proceedings of HLT/NAACL. (2003)
7. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: ACL-2005. (2005)
8. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of HLT/NAACL. (2003)
9. Koehn, P.: A parallel corpus for statistical machine translation. In: Proceedings of MT-Summit. (2005)
10. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Proceedings of ACL 2004 Workshop Text Summarization Branches Out, Barcelona, Spain (2004)
11. Vapnik, V.N.: The nature of statistical learning theory. 2nd edn. Springer (November 1999)
12. Bouma, G., van Noord, G., Malouf, R.: Alpino: wide-coverage computational analysis of Dutch. In: Proceedings of CLIN. (2000)
13. Carreras, X., Chao, I., Padró, L., Padró, M.: FreeLing: an open-source suite of language analyzers. In: Proceedings of the 4th international Language Resources and Evaluation Conference, Lisbon, Portugal (2004)
14. Lin, D.: Dependency-based evaluation of MiniPar. In: Proceedings of LREC Workshop on the Evaluation of Parsing Systems, Granada, Spain (1998)
15. Marsi, E., Krahmer, E., Bosma, W., Theune, M.: Normalized alignment of dependency trees for detecting textual entailment. In Magnini, B., Dagan, I., eds.: Second PASCAL Recognising Textual Entailment Challenge, Venice, Italy (April 2006) 56–61