

Paraphrase Substitution for Recognizing Textual Entailment

Wauter Bosma
University of Twente
bosmaw@cs.utwente.nl

Chris Callison-Burch
University of Edinburgh
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.

Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing—*Language parsing and understanding*

General Terms

Experimentation, Languages, Reliability

Keywords

Recognizing textual entailment, Paraphrase generation, Question answering

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 [13] 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.*

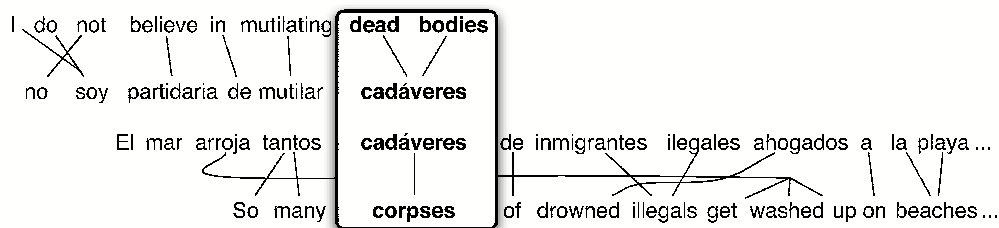


Figure 1: A bilingual parallel corpus can be used to extract paraphrases

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*

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 employs 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 (henceforth AVE) are described in section 4. We will wrap up with a conclusion and suggestions 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, 12, 3, 1]. In this work, we use Bannard and Callison-Burch’s method [1], which extracts paraphrases from bilingual parallel corpora.

Clonaid	announced	that	mother	and	daughter	would	be	returning	to	the	US	on	Monday
announced 0.31	that 0.36	mother 0.59	and 0.78	daughter 0.70	would 0.19	be 0.33	return 0.28	to 0.32	the 0.50	us 0.35	on 0.25	monday 0.62	
declared 0.03	the 0.09	mothers 0.11	, 0.10	daughters 0.05	will 0.07	is 0.08	returning 0.15	of 0.08	of 0.04	united states 0.06	of 0.08	on monday 0.16	
announce 0.02	this 0.08	parent 0.03	or 0.03	girl 0.02	would be 0.04	are 0.03	back 0.04	that 0.07	a 0.03	\$ 0.02	to 0.05	friday 0.02	
declare 0.02	it 0.04	stem 0.03	; 0.01	girls 0.02	be 0.02	a 0.02	returned 0.03	the 0.05	that 0.03	usd 0.02	in 0.04	thursday 0.02	
stated 0.02	, 0.03	maternal 0.02		her daughter 0.01	is 0.02	was 0.02	returnees 0.02	for 0.04	it 0.03	dollars 0.01	about 0.04		
said 0.02	said 0.02	the mother 0.01		development 0.01	could 0.02	that 0.01	again 0.01	in 0.04	of the 0.02	usa 0.01	for 0.03		
announced that 0.01	which 0.01	a mother 0.01		female 0.01	should 0.02	become 0.01	repatriation 0.01	on 0.04	to 0.01	us dollars 0.01	the 0.03		
proclaimed 0.01	to 0.01	home 0.01			will be 0.01	been 0.01	returns 0.01	at 0.02	by 0.01	us \$ 0.01	with regard 0.02		
	:	0.01					come back 0.01	a 0.02			on the 0.02		
announced that	mother and	would be	returning to	the US	on Monday								
announced that 0.33	mother and 0.58	would be 0.19	returning to 0.17	the united states 0.26	on monday 0.55								
announced 0.10	mother 0.07	would 0.08	return to 0.12	united states 0.19	monday 0.20								
said that 0.03	mothers and 0.06	is 0.06	return 0.11	the usa 0.12	last monday 0.04								
declared 0.03	woman and 0.03	will 0.05	returning 0.03	the us 0.09	next monday 0.01								
stated that 0.03	mother, 0.03	be 0.04	to return to 0.02	us 0.03	this monday 0.01								
proclaimed 0.02	mothers 0.02	will be 0.04	back to 0.02	america 0.02	monday, 0.01								
has announced that 0.02	parent 0.01	should be 0.02	to 0.02	states 0.01	on a monday 0.01								
announce that 0.02	and 0.01	it would be 0.02	returned 0.02	united states of america 0.01	of monday 0.01								
declare 0.02	maternal and 0.01	are 0.01	a return to 0.01	american 0.01									
indicated that 0.02	parents 0.01	would have 0.01	go back to 0.01	\$ 0.01									
				to the US									
				to the us 0.15									
				united states 0.10									
				to the united states 0.09									
				the united states 0.08									
				to the usa 0.03									
				the americans 0.02									
				us 0.02									
				the usa 0.02									
				the american 0.02									
				to the american 0.02									

Figure 2: All paraphrases for each subphrase within the example hypothesis

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. 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:

$$p(e_2|e_1) = \sum_f p(f|e_1)p(e_2|f) \quad (1)$$

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:

$$p(e_2|f) \approx \frac{\text{count}(e_2, f)}{\sum_{e_2} \text{count}(e_2, f)} \quad (2)$$

<p>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: <i>The Latvian currency is lat.</i> (negative judgment; entailment score = 0.60) Hypothesis paraphrase: <i>The Latvia currency lat.</i> (positive judgment; entailment score = 1.00)</p>
<p>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 Glacier 20 000 feet above sea level in the Himalayas. Hypothesis: <i>India and Pakistan have fought two wars</i> for the possession of <i>Kashmir.</i> (negative judgment; entailment score = 0.67) Hypothesis paraphrase: <i>India and Pakistan fought two of their wars</i> for possession of <i>Kashmir.</i> (positive judgment; entailment score = 0.83)</p>
<p>Pair: 8597 (entailment: yes) Text: Anthony Busuttil, Professor of Forensic Medicine at Edinburgh University, examined the boy. Hypothesis: <i>Anthony Busuttil is professor of Forensic Medicine at the University of Edinburgh.</i> (negative judgment; entailment score = 0.67) Hypothesis paraphrase: <i>Anthony Busuttil professor of Forensic Medicine at Edinburgh University.</i> (positive judgment; entailment score = 1.00)</p>

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.

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

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

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 [7], 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 [9] 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. $LCS(T, H)$ is the length of the longest common subsequence:

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

From the LCS, the entailment score $LCS(T, H)/|H|$ is derived. In order to account for variation in natural language text, the LCS is measured after paraphrasing the hypothesis. The underlying

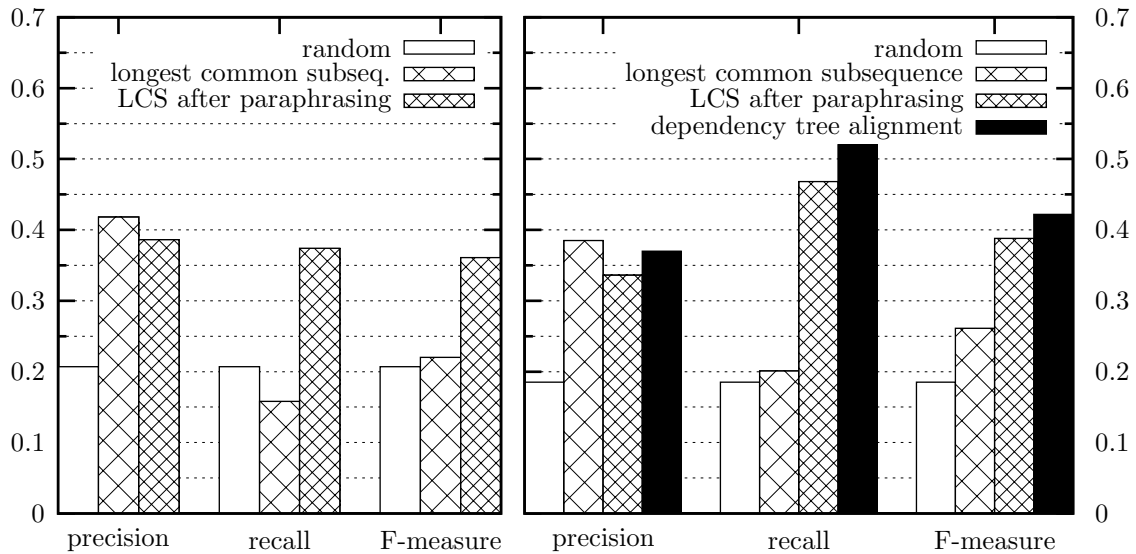


Figure 3: *Left*, average scores of various metrics for German, English, Spanish, French, Italian, Dutch and Portuguese; *right*, average scores, including dependency tree alignment, for English, Spanish and Dutch.

idea is that whenever a paraphrase of 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 sequence of words. Figure 2 shows all paraphrases that we were able to extract for the example hypothesis. 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 are were to recognize entailment.

In order to judge whether a hypothesis is entailed by a text we see if the value of the entailment score, $LCS(T, H)/|H|$, is greater than some threshold value. Support vector machines [14] are used to determine the entailment treshold. Unfortunately, the only suitable training data available was

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 3: 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.

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 threshold for all languages. The threshold value 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 decides at random with a probability distribution reflecting the corpus distribution of entailment values. The second baseline is a system which measures the longest common subsequence of text and hypothesis. Given the fact that paraphrasing is a form of query expansion, we expected that precision drops and recall increases when using paraphrases. Figure 3 (left) shows that this 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 [5], [6] and [10] 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. [11] was used to align dependency trees. Interestingly, as shown in Figure 3 (right), the dependency tree alignment system performs slightly better but 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 3) indicates that performance can be further increased when employing both types of information in an integrated system.

5 Conclusion

We showed that paraphrases can boost the performance of an entailment recognition system relying on the longest common subsequence to decide for entailment. 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 three things: a combination of syntax-based and paraphrase-based approaches to entailment recognition, improved methods for determining the entailment threshold, and further enhancements to paraphrase extraction techniques, such as taking syntax into account.

Acknowledgments

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

References

- [1] Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *ACL-2005*, 2005.
- [2] Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second PASCAL Recognising Textual Entailment Challenge. In Bernardo Magnini and Ido Dagan, editors, *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge*, Trento, Italy, April 2006.
- [3] Regina Barzilay and Lillian Lee. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL*, 2003.
- [4] Regina Barzilay and Kathleen McKeown. Extracting paraphrases from a parallel corpus. In *ACL-2001*, 2001.
- [5] Gosse Bouma, Gertjan van Noord, and Robert Malouf. Alpino: wide-coverage computational analysis of Dutch. In *Proceedings of CLIN*, 2000.
- [6] Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. FreeLing: an open-source suite of language analyzers. In *Proceedings of the 4th international Language Resources and Evaluation Conference*, Lisbon, Portugal, 2004.
- [7] Philipp Koehn. A parallel corpus for statistical machine translation. In *Proceedings of MT-Summit*, 2005.
- [8] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*, 2003.
- [9] Chin-Yew Lin. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of ACL 2004 Workshop Text Summarization Branches Out*, Barcelona, Spain, 2004.
- [10] Dekang Lin. Dependency-based evaluation of MiniPar. In *Proceedings of LREC Workshop on the Evaluation of Parsing Systems*, Granada, Spain, 1998.
- [11] Erwin Marsi, Emiel Krahmer, Wauter Bosma, and Mariët Theune. Normalized alignment of dependency trees for detecting textual entailment. In Bernardo Magnini and Ido Dagan, editors, *Second PASCAL Recognising Textual Entailment Challenge*, pages 56–61, Venice, Italy, April 2006. PASCAL.
- [12] Bo Pang, Kevin Knight, and Daniel Marcu. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of HLT/NAACL*, 2003.
- [13] Lucy Vanderwende and William B. Dolan. What syntax can contribute in the entailment task. In *PASCAL Challenges Workshop on Recognizing Textual Entailment*, pages 205–216, Southampton, United Kingdom, 2005. Springer-Verlag.
- [14] Vladimir Naumovich Vapnik. *The nature of statistical learning theory*. Springer, 2nd edition, November 1999.