Comparing Semantically Related Sentences: The Case of Paraphrase versus Subsumption

Jahna Otterbacher
University of Michigan
jahna@umich.edu

Dragomir Radev
University of Michigan
radev@umich.edu

Abstract

Paraphrases and other semantically related sentences present a challenge to NLP and IR applications such as multi-document summarization and question answering systems. While it is generally agreed that paraphrases contain approximately equivalent ideas, they often differ from one another in subtle, yet non-trivial, ways. In this paper, we examine semantic differences in cases of paraphrase and subsumption, in an effort to understand what makes one sentence significantly more informative than another. Using manually annotated data from the news domain, we concentrate on developing a framework for analyzing and comparing pairs of related sentences.

1 Introduction

News is a domain where paraphrases are frequently found, particularly across topically related documents published by different sources, or even by the same source over time. It has been noted that the prevalence of paraphrases in news is in part due to the way journalists are instructed to write. For example, many agencies extensively use the same newswire sources, with journalists reusing text in the creation of a story (Clough, 2001). Additionally, in following breaking news stories that evolve over time, updates are published at set intervals, in which writers may simply freshen up an article, even if there is little or no new information to report (Mitchell and West, 1996).

Recent work has considered both the automatic detection of paraphrases (Barzilay and McKeown, 2001; Shinyama and Sekine, 2003) as well as paraphrase generation (Barzilay and Lee, 2003; Pang et al., 2003). However, in the current work we contrast the relationship of paraphrase with that of information subsumption. In particular, we focus on developing a framework for comparing pairs of related sentences, in order to determine if they are paraphrases or if one sentence subsumes the other. As we will show, semantic differences between sentences are often subtle yet non-trivial, such that a means to compare them must be established.

2 Motivating Example

In this example, we consider how a Q&A system might go about comparing sentences that contain possible responses to users' questions regarding a major plane crash (Gulf Air flight 072 in August of 2000). Suppose that in its document collection, the system has located eleven news stories that are related to the crash. Consider the following question that might be asked by users interested in the details surrounding the story:

Q1 How many people were on the flight?

1) Bahrain television reported 143 people, including 36 children, were on board.
2) Gulf Air said 135 passengers and eight crew members were on board.
3) There were 135 passengers and eight crew members on board, according to Khaleej Times, a daily newspaper in the emirate.
4) All 143 crew and passengers on board were killed.
5) A Gulf Air Airbus A320 carrying 143 people from Cairo, Egypt to Bahrain crashed today.
6) Victims included 64 Egyptians, 35 Bahrainis, 12 Saudi Arabians, 9 Palestinians, 6 from the UAE, 3 Chinese, two British and one each from the U.S., Canada, Oman, Kuwait, Sudan, Australia, the Philippines, Poland, India, Morocco and Egypt.
7) More than 130 bodies are reported to have been recovered after a Gulf Air jet carrying 143 people crashed into the Gulf off Bahrain on Wednesday.

Figure 1: Extracted sentences with possible answers

Figure 1 lists a subset of the sentences extracted from the source articles that might contain the answers to the users' question. While

Sentence	Q1
1	143 people, 36 children
2	135 passengers and 8 crew members
3	135 passengers and 8 crew members
4	143 crew and passengers
5	143 people
6	unknown
7	143 people

Table 1: Answers to Q1 from extracted sentences

most of the sentences express the number of people on the plane at the time of the crash, there are differences between them with respect to whether or not there were victims and who they were. For example, sentences 1, 2, and 3 involve the predicate "were on board," so although they answer question 1, they cannot confirm to the user whether there were victims and who they were. Likewise, answers 5 and 7 express the number of people carried on the plane. Sentence 7 implies that there were victims since it details the number of bodies found thus far, although it does not provide a firm answer regarding the number of victims. Table 1 shows the answers as found in the seven sentences.

Once sentences containing possible answers have been extracted, the Q&A system must decide which answer to present to the user. Assuming that users prefer to receive the most specific and informative answers as possible, the system should recognize the differences between the candidate answers. In response to question 1, the system should ideally choose one of the first three answers, since they are more fine-grained than those contained in sentences 5 and 7. In addition to aiding answer selection in a Q&A system, the above information would be useful in the context of a multi-document summarizer. As stated previously, in extractive MDS, one wishes to avoid including sentences that are paraphrases and would thus lead to redundancy in the summary. Therefore, a summarizer should be able to identify similar sentences and then choose the most informative one of the set. In our example, a summary might include the best sentence from table 1. However, including more than one sentence from the set would lead to redundancy.

3 Data

We examined two clusters of news articles from the CST Bank (Radev et al., 2004) in which

	Gulf Air	HK News
Articles	11	8
Sources	7	1
Publication time span	4 days	2.5 years
Annotated sentence pairs	2,242	1,729

Table 2: Characteristics of the corpus of annotated sentence pairs.

Cluster	Paraphrase	Subsumption
Gulf Air	33	41
HK News	64	60
Total	97	101

Table 3: Instances of paraphrase and subsumption.

pairs of sentences were manually annotated as to the relationships between them. Table 2 compares the characteristics of the data clusters.

For each cluster, two hired judges worked independently in labeling each sentence pair for Cross-document Structure Theory (CST) relationships. The labels used, which describe the relationship between a sentence (S1) and a second sentence (S2) include:[1]

- Paraphrase: the two sentences are equivalent with respect to information content
- Subsumption: S1 contains all of the information in S2, plus additional information not conveyed in S2

Table 3 shows the number of sentence pairs labeled as being either a case of paraphrase or subsumption by at least one judge. We analyzed the similarities and differences between the sentences in each pair, quantitatively and qualitatively, in establishing a framework for comparison of related sentences.

4 Properties of Paraphrase and Subsumption

Previously, we developed a classifier for predicting the existence of CST relationships (Zhang et al., 2003), therefore, we currently concentrate on examining the differences between the paraphrase and subsumption relationships. We examined several content-based features of the sentence pairs in our dataset:

[1]The full set of CST relationships, their definitions and examples are available at http://tangra.si.umich.edu/clair/CSTBank.

Measure	Paraphrases	Subsumption
Simple cosine	0.48 (0.24)	0.45 (0.19)
Cosine	0.43 (0.29)	0.38 (0.24)
Token overlap	0.35 (0.25)	0.30 (0.20)
Bigram overlap	0.22 (0.27)	0.17 (0.21)
Norm. LCS	0.38 (0.29)	0.34 (0.24)
Bleu*	0.24 (0.31)	0.16 (0.24)

Table 4: Mean and standard deviation of content measures.

1. Simple cosine: Cosine similarity with a simple binary count (1 if word is shared by S1 and S2, 0 if not shared, regardless of number of occurrences).
2. Cosine: Cosine similarity between S1 and S2.
3. Token overlap: Proportion of shared tokens between S1 and S2.
4. Big. overlap: Bigram overlap between S1 and S2.
5. Norm. LCS: Longest common substring normalized for sentence length.
6. Bleu: A linear combination of n-gram matches between S1 and S2 with a penalty for length differences (Papineni et al., 2002).

Table 4 shows the comparisons of the content based measures between the sentence pairs labeled as paraphrases and those labeled as subsumption. As can be seen, the Bleu metric is the only one that is significantly different between the cases of paraphrase and subsumption (p-value 0.06). Interestingly, the Bleu scores between sentence pairs labeled as having a subsumption relationship tend to be lower than those labeled as cases of paraphrase.

We have attempted to build a classifier to automatically distinguish between cases of subsumption and paraphrase using the content-based metrics, but as the descriptive statistics suggest, these features alone are not able to reliably discriminate between them. Therefore, we have decomposed a subset of the data and used it to come up with a qualitative framework (or taxonomy) of paraphrases and cases of subsumption that might guide our future work. This is described in the next section.

5 Qualitative Comparisons of Subsumption Versus Paraphases

In decomposing a pair of sentences to be compared, we first analyzed the basic thematic structure of each sentence by identifying the following elements (Ouhalla, 1999):

- Event: the process or action described in the sentence (predicate)
- Entity: the agents, patients, themes taken by the subcategorization frame of the predicate
- Reason: the cause or motivation behind the predicate of the sentence
- Time: the time of the event described in the sentence
- Location: the physical or contextual setting of the event

Note that although the five elements constitute the basic "who, what, when, where and why" that describe any complete story, only the first two thematic elements are required for a sentence to be grammatical. Therefore, many cases of subsumption can be identified based on one sentence containing only a subset of the elements conveyed in the other. Figure 2 illustrates such a case, in which S2 lacks a reason as compared to S1.

```
(S1) Six French government experts and a
representative of Airbus Industries arrived
Thursday evening to look into the crash.
<event: arrive>
<entity: experts, representative>
<reason: to look into the crash>
<time: Thursday evening>
<location: (none)>

(S2) Six French government experts and an
Airbus Industries representative flew in
Thursday evening.
<event: flew in>
<entity: experts, representative>
<reason: (none)>
<time: Thursday evening>
<location: (none)>
```

Figure 2: Subsumption: S1 has 4 thematic elements while S2 has 3

In other cases, such as that shown in figure 3, the sentences are syntactic transformations of one another. In this example, the second sentence is in the passive voice, so it does not contain information that the government is the agent responsible for the action. Although this fact may have been evident in the original context of the source document, when the sentences are in isolation, it can be argued that S1 subsumes S2.

Finally, in the first step of analysis, complex sentences can also be identified. If one of the sentences making up the complex sentence contains all of the thematic elements present in the

```
(S1) The government announced measures to
improve air quality.
<event: announce>
<entity: government, measures>
<reason: to improve air quality>

(S2) Measures were announced to improve
air quality.
<event: be announced>
<entity: measures>
<reason: to improve air quality>
```

Figure 3: Subsumption in the case of an active vs. passive construction

second, simple sentence, this constitutes a clear case of subsumption. Figure 4 shows such an example.

```
(S1) Action to deal with air pollution
from other sources continues but priority
is being given to street level pollution
from vehicles.
<conj: but>
<event1: continue>
<entity1: action>
<event2: be given>
<entity2: priority>

(S2) Action to tackle street level air
pollution continues.
<event: continue>
<entity: action>
```

Figure 4: Subsumption in the case of a complex vs. simple sentence

The above three examples illustrated cases of sentence pairs in which there are differences with respect to the number of thematic elements present in the sentences in question and where the presence or absence of the elements can be used in determining if one sentence expresses more information than the other, or if they are approximately equivalent. However, if both sentences contain the same elements, the differences between related sentences are typically more complex and require making comparison within each of the thematic elements.

6 Conclusion and Future Work

We have motivated the need for a means to distinguish between sentences that are paraphrases of one another from those pairs in which one sentence subsumes the other. As was shown, simple content based measures of the sentence pairs are not able to make this distinction. Therefore, our further work will focus on using more rich features, such as the syntactic and thematic structure of sentences. To this end, we have developed a framework for comparing sentences with respect to their thematic structure that might guide our future work.

References

Regina Barzilay and Lillian Lee. 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In *Proceedings of NAACL-HLT03*, Edmonton.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the ACL/EACL.*

Paul D. Clough. 2001. Measuring Text Reuse and Document Derivation. Postgraduate transfer report, Department of Computer Science, University of Sheffield, UK.

Catherine C. Mitchell and Mark D. West. 1996. *The News Formula: A Concise Guide to News Writing and Reporting.* St. Martin's Press, New York.

Jamal Ouhalla. 1999. *Transformational Grammar: From Principles and Parameters to Minimalism.* Oxford University Press, New York.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In *Proceedings of HLT/NAACL 2003*, Edmonton, Canada.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Dragomir Radev, Jahna Otterbacher, and Zhu Zhang. 2004. CST Bank: A Corpus for the Study of Cross-document Structural Relationships. In *Proceedings of LREC'04*, Lisbon, May.

Yusuke Shinyama and Satoshi Sekine. 2003. Paraphrase Acquisition for Information Extraction. In *Second International Workshop on Paraphrasing (IWP2003)*, Sapporo, Japan.

Zhu Zhang, Jahna Otterbacher, and Dragomir Radev. 2003. Learning Cross-document Structural Relationships using Boosting. In *Conference on Information and Knowledge Management (CIKM'03)*, New Orleans, November.