Automatic Paraphrase Discovery based on Context and Keywords between NE Pairs

Satoshi Sekine
New York University
715 Broadway, 7th floor
New York, NY 10003 USA
sekine@cs.nyu.edu

Abstract

Automatic paraphrase discovery is an important but challenging task. We propose an unsupervised method to discover paraphrases from a large untagged corpus, without requiring any seed phrase or other cue. We focus on phrases which connect two Named Entities (NEs), and proceed in two stages. The first stage identifies a keyword in each phrase and joins phrases with the same keyword into sets. The second stage links sets which involve the same pairs of individual NEs. A total of 13,976 phrases were grouped. The accuracy of the sets in representing paraphrase ranged from 73% to 99%, depending on the NE categories and set sizes; the accuracy of the links for two evaluated domains was 73% and 86%.

1 Introduction

One of the difficulties in Natural Language Processing is the fact that there are many ways to express the same thing or event. If the expression is a word or a short phrase (like "corporation" and "company"), it is called a "synonym". There has been a lot of research on such lexical relations, along with the creation of resources such as WordNet. If the expression is longer or complicated (like "A buys B" and "A's purchase of B"), it is called "paraphrase", i.e. a set of phrases which express the same thing or event. Recently, this topic has been getting more attention, as is evident from the Paraphrase Workshops in 2003 and 2004, driven by the needs of various NLP applications. For example, in Information Retrieval (IR), we have to match a user's query to the expressions in the desired documents, while in Question Answering (QA), we have to find the answer to the user's question even if the formulation of the answer in the document is different from the question. Also, in Information Extraction (IE), in which the system tries to extract elements of some events (e.g. date and company names of a corporate merger event), several event instances from different news articles have to be aligned even if these are expressed differently.

We realize the importance of paraphrase; however, the major obstacle is the construction of paraphrase knowledge. For example, we can easily imagine that the number of paraphrases for "A buys B" is enormous and it is not possible to create a comprehensive inventory by hand. Also, we don't know how many such paraphrase sets are necessary to cover even some everyday things or events. Up to now, most IE researchers have been creating paraphrase knowledge (or IE patterns) by hand and for specific tasks. So, there is a limitation that IE can only be performed for a pre-defined task, like "corporate mergers" or "management succession". In order to create an IE system for a new domain, one has to spend a long time to create the knowledge. So, it is too costly to make IE technology "open-domain" or "on-demand" like IR or QA.

In this paper, we will propose an unsupervised method to discover paraphrases from a large untagged corpus. We are focusing on phrases which have two Named Entities (NEs), as those types of phrases are very important for IE applications. After tagging a large corpus with an automatic NE tagger, the method tries to find sets of paraphrases automatically without being given a seed phrase or any kinds of cue.

2 Algorithm

2.1 Overview

Before explaining our method in detail, we present a brief overview in this subsection.

First, from a large corpus, we extract all the NE instance pairs. Here, an NE instance pair is any pair of NEs separated by at most 4 syntactic chunks; for example, "IBM plans to acquire Lotus". For each pair we also record the context, i.e. the phrase between the two NEs (Step1). Next, for each pair of NE *categories*, we collect all the contexts and find the keywords which are topical for that NE category pair. We use a simple TF/IDF method to measure the topicality of words. Hereafter, each pair of NE categories will be called a *domain*; e.g. the "Company – Company" domain, which we will call CC-domain (Step 2). For each domain, phrases which contain the same keyword are gathered to build a set of phrases (Step 3). Finally, we find links between sets of phrases, based on the NE instance pair data (for example, different phrases which link "IBM" and "Lotus") (Step 4). As we shall see, most of the linked sets are paraphrases.

This overview is illustrated in Figure 1.

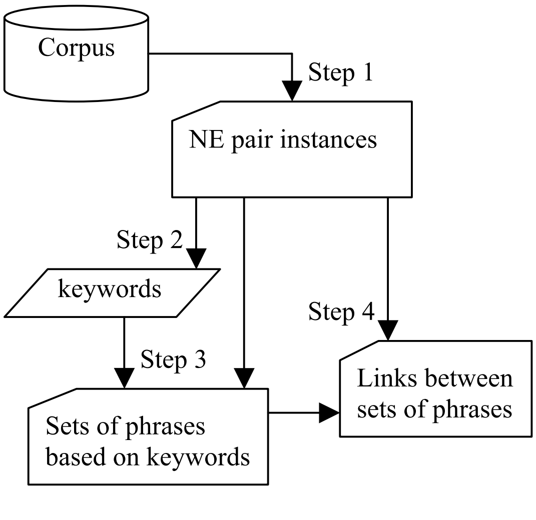
Corpus
Step 1
NE pair instances
Step 2
keywords
Step 4
Step 3
Links between sets of phrases
Sets of phrases based on keywords

Figure 1. Overview of the method

2.2 Step by Step Algorithm

In this section, we will explain the algorithm step by step with examples. Because of their size, the examples (Figures 2 to 4) appear at the end of the paper.

Step 1. Extract NE instance pairs with contexts

First, we extract NE pair instances with their context from the corpus. The sentences in the corpus were tagged by a transformation-based chunker and an NE tagger. The NE tagger is a rule-based system with 140 NE categories [Sekine et al. 2004]. These 140 NE categories are designed by extending MUC's 7 NE categories with finer sub-categories (such as Company, Institute, and Political Party for Organization; and Country, Province, and City for Location) and adding some new types of NE categories (Position Title, Product, Event, and Natural Object). All the NE pair instances which co-occur separated by at most 4 chunks are collected along with information about their NE types and the phrase between the NEs (the 'context'). Figure 2 shows examples of extracted NE pair instances and their contexts. The data is sorted based on the frequency of the context ("a unit of" appeared 314 times in the corpus) and the NE pair instances appearing with that context are shown with their frequency (e.g. "NBC" and "General Electric Co." appeared 10 times with the context "a unit of").

Step 2. Find keywords for each NE pair

When we look at the contexts for each domain, we noticed that there is one or a few important words which indicate the relation between the NEs (for example, the word "unit" for the phrase "a unit of"). Once we figure out the important word (e.g. keyword), we believe we can capture the meaning of the phrase by the keyword. We used the TF/ITF metric to identify keywords.

All the contexts collected for a given domain are gathered in a bag and the TF/ITF scores are calculated for all the words except stopwords in the bag. Here, the term frequency (TF) is the frequency of a word in the bag and the inverse term frequency (ITF) is the inverse of the log of the frequency in the entire corpus. Figure 3 shows some keywords with their scores.

Step 3. Gather phrases using keywords

Next, we select a keyword for each phrase – the top-ranked word based on the TF/IDF metric. (If the TF/IDF score of that word is below a threshold, the phrase is discarded.) We then gather all phrases with the same keyword. Figure 4 shows some such phrase sets based on keywords in the CC-domain.

Step 4. Cluster phrases based on Links

We now have a set of phrases which share a keyword. However, there are phrases which express the same meanings even though they do not share the same keyword. For example, in Figure 3, we can see that the phrases in the "buy", "acquire" and "purchase" sets are mostly paraphrases. At this step, we will try to link those sets, and put them into a single cluster. Our clue is the NE instance pairs. If the same pair of NE instances is used with different phrases, these phrases are likely to be paraphrases. For example, the two NEs "Eastern Group Plc" and "Hanson Plc" have the following contexts. Here, "EG" represents "Eastern Group Plc". and "H" represents "Hanson Plc".

- EG, has agreed to be bought by H
- EG, now owned by H
- H to acquire EG
- H's agreement to buy EG

Three of those phrases are actually paraphrases, but sometime there could be some noise; such as the second phrase above. So, we set a threshold that at least two examples are required to build a link. More examples are shown in Figure 5.

Notice that the CC-domain is a special case. As the two NE categories are the same, we can't differentiate phrases with different orders of participants – whether the buying company or the to-be-bought company comes first. The links can solve the problem. As can be seen in the example, the first two phrases have a different order of NE names from the last two, so we can determine that the last two phrases represent a reversed relation. In figure 4, reverse relations are indicated by `*' next to the frequency.

Now we have sets of phrases which share a keyword and we have links between those sets.

3 Experiments

3.1 Corpora

For the experiments, we used four newswire corpora, the Los Angeles Times/Washington Post, The New York Times, Reuters and the Wall Street Journal, all published in 1995. They contain about 200M words (25M, 110M, 40M and 19M words, respectively). All the sentences have been analyzed by our chunker and NE tagger. The procedure using the tagged sentences to discover paraphrases takes about one hour on a 2GHz Pentium 4 PC with 1GB of memory.

3.2 Results

In this subsection, we will report the results of the experiment, in terms of the number of words, phrases or clusters. We will report the evaluation results in the next subsection.

Step 1. Extract NE pair instances with contexts

From the four years of newspaper corpus, we extracted 1.9 million pairs of NE instances. The most frequent NE category pairs are "Person - Person (209,236), followed by "Country - Country" (95,123) and "Person - Country" (75,509). The frequency of the Company – Company domain ranks 11th with 35,567 examples.

As lower frequency examples include noise, we set a threshold that an NE category pair should appear at least 5 times to be considered and an NE instance pair should appear at least twice to be considered. This limits the number of NE category pairs to 2,000 and the number of NE pair instances to 0.63 million.

Step 2. Find keywords for each NE pair

The keywords are found for each NE category pair. For example, in the CC-domain, 96 keywords are found which have TF/ITF scores above a threshold; some of them are shown in Figure 3. It is natural that the larger the data in the domain, the more keywords are found. In the "Person – Person" domain, 618 keywords are found, and in the "Country – Country" domain, 303 keywords are found. In total, for the 2,000 NE category pairs, 5,184 keywords are found.

Step 3. Gather phrases using keywords

Now, the keyword with the top TF/ITF score is selected for each phrase. If a phrase does not contain any keywords, the phrase is discarded. For example, out of 905 phrases in the CC-domain, 211 phrases contain keywords found in step 2. In total, across all domains, we kept 13,976 phrases with keywords.

Step 4. Link phrases based on instance pairs

Using NE instance pairs as a clue, we find links between sets of phrases. In the CC-domain, there are 32 sets of phrases which contain more than 2 phrases. We concentrate on those sets. Among these 32 sets, we found the following pairs of sets which have two or more links. Here a set is represented by the keyword and the number in parentheses indicates the number of shared NE pair instances.

buy - acquire (5) buy - agree (2)
buy - purchase (5) buy - acquisition (7)
buy - pay (2)* buy - buyout (3)
buy - bid (2) acquire - purchase (2)
acquire - acquisition (2)
acquire - pay (2)* purchase - acquisition (4)
purchase - stake (2)* acquisition - stake (2)*

unit - subsidiary (2) unit - parent (5)

It is clear that these links form two clusters which are mostly correct. We will describe the evaluation of such clusters in the next subsection.

3.3 Evaluation Results

We evaluated the results based on two metrics. One is the accuracy within a set of phrases which share the same keyword; the other is the accuracy of links. We picked two domains, the CC-domain and the "Person – Company" domain (PC-domain), for the evaluation, as the entire system output was too large to evaluate. It is not easy to make a clear definition of "paraphrase". Sometimes extracted phrases by themselves are not meaningful to consider without context, but we set the following criteria. If two phrases can be used to express the same relationship within an information extraction application ("scenario"), these two phrases are paraphrases. Although this is not a precise criterion, most cases we evaluated were relatively clear-cut. In general, different modalities ("planned to buy", "agreed to buy", "bought") were considered to express the same relationship within an extraction setting. We did have a problem classifying some modified noun phrases where the modified phrase does not represent a qualified or restricted form of the head, like "chairman" and "vice chairman", as these are both represented by the keyword "chairman". In this specific case, as these two titles could fill the same column of an IE table, we regarded them as paraphrases for the evaluation.

Evaluation within a set

The evaluation of paraphrases within a set of phrases which share a keyword is illustrated in Figure 4. For each set, the phrases with bracketed frequencies are considered not paraphrases in the set. For example, the phrase "'s New York-based trust unit," is not a paraphrase of the other phrases in the "unit" set. As you can see in the figure, the accuracy for the domain is quite high except for the "agree" set, which contains various expressions representing different relationships for an IE application. The accuracy is calculated as the ratio of the number of paraphrases to the total number of phrases in the set. The results, along with the total number of phrases, are shown in Table 1.

Domain	# of phrases	total phrases	accuracy
CC	7 or more	105	87.6%
	6 or less	106	67.0%
PC	7 or more	359	99.2%
	6 or less	255	65.1%

Table 1. Evaluation results within sets

Table 1 shows the evaluation result based on the number of phrases in a set. The larger sets are more accurate than the small sets. We can make several observations on the cause of errors. One is that smaller sets sometime have meaningless keywords, like "strength" or "add" in the CC-domain, or "compare" in the PC-domain. Eight out of the thirteen errors in the high frequency phrases in the CC-domain are the phrases in "agree". As can be seen in Figure 3, the phrases in the "agree" set include completely different relationships, which are not paraphrases. Other errors include NE tagging errors and errors due to a phrase which includes other NEs. For example, in the phrase "Company-A last week purchased rival Marshalls from Company-B", the purchased company is Marshalls, not Company-B. Also there are cases where one of the two NEs belong to a phrase outside of the relation. For example, from the sentence "Mr. Smith estimates Lotus will make a profit this quarter…", our system extracts "Smith estimates Lotus" as an instance. Obviously "Lotus" is part of the following clause rather than being the object of "estimates" and the extracted instance makes no sense. We will return to these issues in the discussion section.

Evaluation of links

A link between two sets is considered correct if the majority of phrases in both sets have the same meaning, i.e. if the link indicates paraphrase. All the links in the "CC-domain are shown in Step 4 in subsection 3.2. Out of those 15 links, 4 are errors, namely "buy - pay", "acquire - pay", "purchase - stake" "acquisition - stake". When a company buys another company, a paying event can occur, but these two phrases do not indicate the same event. The similar explanation applies to the link to the "stake" set.

We checked whether the discovered links are listed in WordNet. Only 2 link in the CC-domain (buy-purchase, acquire-acquisition) and 2 links (trader-dealer and head-chief) in the PC-domain are found in the same synset of Word-Net 2.1 (http://wordnet.princeton.edu/). This result suggests the benefit of using the automatic discovery method.

Domain	Link accuracy	WN coverage
CC	73.3 %	2/11
PC	88.9%	2/8

Table 2. Evaluation results for links

4 Related Work

The work reported here is closely related to [Hasegawa et al. 04]. First, we will describe their method and compare it with our method. They first collect the NE instance pairs and contexts, just like our method. However, the next step is clearly different. They cluster NE instance pairs based on the words in the contexts using a bag-of-words method. In order to create good-sized vectors for similarity calculation, they had to set a high frequency threshold, 30. Because of this threshold, very few NE instance pairs could be used and hence the variety of phrases was also limited. Instead, we focused on phrases and set the frequency threshold to 2, and so were able to utilize a lot of phrases while minimizing noise. [Hasegawa et al. 04] reported only on relation discovery, but one could easily acquire paraphrases from the results. The number of NE instance pairs used in their experiment is less than half of our method.

There have been other kinds of efforts to discover paraphrase automatically from corpora. One of such approaches uses comparable documents, which are sets of documents whose content are found/known to be almost the same, such as different newspaper stories about the same event [Shinyama and Sekine 03] or different translations of the same story [Barzilay 01]. The availability of comparable corpora is limited, which is a significant limitation on the approach.

Another approach to finding paraphrases is to find phrases which take similar subjects and objects in large corpora by using mutual information of word distribution [Lin and Pantel 01]. This approach needs a phrase as an initial seed and thus the possible relationships to be extracted are naturally limited.

There has also been work using a bootstrapping approach [Brin 98; Agichtein and Gravano 00; Ravichandran and Hovy 02]. The basic strategy is, for a given pair of entity types, to start with some examples, like several famous book title and author pairs; and find expressions which contains those names; then using the found expressions, find more author and book title pairs. This can be repeated several times to collect a list of author / book title pairs and expressions. However, those methods need initial seeds, so the relation between entities has to be known in advance. This limitation is the obstacle to making the technology "open domain".

5 Discussion

Keywords with more than one word

In the evaluation, we explained that "chairman" and "vice chairman" are considered paraphrases. However, it is desirable if we can separate them. This problem arises because our keywords consist of only one word. Sometime, multiple words are needed, like "vice chairman", "prime minister" or "pay for" ("pay" and "pay for" are different senses in the CC-domain). One possibility is to use n-grams based on mutual information. If there is a frequent multi-word sequence in a domain, we could use it as a keyword candidate.

Keyword detection error

Even if a keyword consists of a single word, there are words which are not desirable as keywords for a domain. As was explained in the results section, "strength" or "add" are not desirable keywords in the CC-domain. In our experiment, we set the threshold of the TF/ITF score empirically using a small development corpus; a finer adjustment of the threshold could reduce the number of such keywords.

Also, "agree" in the CC-domain is not a desirable keyword. It is a relatively frequent word in the domain, but it can be used in different extraction scenarios. In this domain the major scenarios involve the things they agreed on, rather than the mere fact that they agreed. "Agree" is a subject control verb, which dominates another verb whose subject is the same as that of "agree"; the latter verb is generally the one of interest for extraction. We have checked if there are similar verbs in other major domains, but this was the only one.

Using structural information

As was explained in the results section, we extracted examples like "Smith estimates Lotus", from a sentence like "Mr. Smith estimates Lotus will make profit this quarter…". In order to solve this problem, a parse tree is needed to understand that "Lotus" is not the object of "estimates". Chunking is not enough to find such relationships. This remains as future work.

Limitations

There are several limitations in the methods. The phrases have to be the expressions of length less than 5 chunks, appear between two NEs. Also, the method of using keywords rules out phrases which don't contain popular words in the domain. We are not claiming that this method is almighty. Rather we believe several methods have to be developed using different heuristics to discover wider variety of paraphrases.

Applications

The discovered paraphrases have multiple applications. One obvious application is information extraction. In IE, creating the patterns which express the requested scenario, e.g. "management succession" or "corporate merger and acquisition" is regarded as the hardest task. The discovered paraphrases can be a big help to reduce human labor and create a more comprehensive pattern set. Also, expanding on the techniques for the automatic generation of extraction patterns (Riloff 96; Sudo 03) using our method, the extraction patterns which have the same meaning can be automatically linked, enabling us to produce the final table fully automatically. While there are other obstacles to completing this idea, we believe automatic paraphrase discovery is an important component for building a fully automatic information extraction system.

6 Conclusion

We proposed an unsupervised method to discover paraphrases from a large untagged corpus. We are focusing on phrases which have two Named Entities (NEs), as those types of phrases are very important for IE applications. After tagging a large corpus with an automatic NE tagger, the method tries to find sets of paraphrases automatically without being given a seed phrase or any kind of cue. In total 13,976 phrases are assigned to sets of phrases, and the accuracy on our evaluation data ranges from 65 to 99%, depending on the domain and the size of the sets. The accuracies for link were 73% and 86% on two evaluated domains. These results are promising and there are several avenues for improving on these results.

7 Acknowledgements

This research was supported in part by the Defense Advanced Research Projects Agency as part of the Translingual Information Detection, Extraction and Summarization (TIDES) program, under Grant N66001-001-1-8917 from the Space and Naval Warfare Systems Center, San Diego, and by the National Science Foundation under Grant IIS-00325657. This paper does not necessarily reflect the position of the U.S. Government.

We would like to thank Prof. Ralph Grishman, Mr. Takaaki Hasegawa and Mr. Yusuke Shinyama for useful comments, discussion and evaluation.

References

Agichtein, Eugene and Gravano, Luis. 2000. Snowball: Extracting reations from large plain-text collocations. *In Proc. 5th ACM Int'l Conf. on Digital Libruaries (ACM DL00)* pp 85-94.

Barzilay, Regina and McKeown, Kathleen. 2001. Extracting paraphrases from a parallel corpus. *In Proc. 39th Annual Meeting Association for Computational Linguistics (ACL-EACL01),* pp 50-57.

Brin, Sergey. 1998. Extracting patterns and relations from world wide web. *In Proc. WebDB Workshop at 6th Int'l Conf. on Extending Database Technology (WebDB98),* pp172-183.

Hasegawa, Takaaki, Sekine, Satoshi and Grishman, Ralph. 2004. Discovering Relations among Named Entities from Large Corpora, *In Proc. 42nd Annual Meeting Association for Computational Linguistics (ACL04), pp 415-422*

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. *In Proc Fourteenth Int'l Conf. on Computational Linguistics (COLING92).*

Lin, Dekang and Pantel, Patrick. 2001. Dirt – discovery of inference rules from text. *In Proc. 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD01)*, pp323-328

Ravichandran, Deepak and Hovy, Eduard. 2002. Learning Surface Text Patterns for a Question Answering System. *In Proc. Annual Meeting Association for Computational Linguistics (ACL02)*

Riloff E. 1996. Automatically Generating Extraction Patterns from Untagged Text. *In Proc. 13th National Conf. on Artificial Intelligence (AAAI96),* 1044-1049.

Sekine, Satoshi and Nobata , Chikashi. 2004. Definition, Dictionary and Tagger for Extended Named Enties. *In Proc. of the Fourth Int'l Conf. on Language Resource and Evaluation (LREC 04)*

Shinyama, Yusuke and Sekine, Satoshi. 2003. Paraphrase acquisition for information extraction. *In Proc. 2nd Int'l Workshop on Paraphrasing (IWP03)*

Sudo, Kiyoshi, Sekine, Satoshi and Grishman, Ralph. 2003. An improved extraction pattern representation model for automatic IE pattern acquisition. *In Proc. 41st Annual Meeting Association for Computational Linguistics (ACL03)*

```
# COMPANY COMPANY : 22535
@ 314   , a unit of
10 NBC                 General Electric Co.
 9 Citibank            Citicorp
 7 Smith Barney        Travelers Group Inc.
 6 20th Century Fox    the News Corp.
 5 Salomon Brothers    Salomon Inc.
 5 Fidelity            FMR Corp.
 5 GTE Mobilnet        GTE Corp
 4 Smith Barney        Travelers Inc.
…
@ 108   , a subsidiary of
 5 U.S. Ecology Inc.   American Ecology Corp.
 3 Pulte Home Corp.    Pulte Corp.
…
```

Figure 2. Extracted NE pair instances and context

4846.2	519	44778	buy
3682.8	205	261	share
3609.1	354	18186	unit
2949.2	289	18021	parent
2850.6	258	8523	acquire
2709.9	275	25541	agree
1964.1	163	4020	subsidiary
1237.9	119	14959	purchase
1036.9	94	8649	acquisition
593.7	40	843	sell
585.6	55	12000	stake
581.3	63	50868	pay

Figure 3. High TF/ITF words in "Com-Com"
(Numbers are TF/ITF score, frequency in the collection (TF), frequency in the corpus (TF) and word)

```
=== buy ===
97      agreed to buy
84      bought
50      said it will buy
45      said it agreed to buy
25      will buy
23      to buy
20      plans to buy
16      , which bought
14      is buying
11      said it would buy
11      offered to buy
10      's agreement to buy
9       , which is buying
9*      agreed to be bought by
8       is offering to buy
7       said it wants to buy
7       was buying
6       tried to buy
6       said it plans to buy
6       said it intends to buy
```

6*	was bought by
5	is offering to buy the portion of
5	is expected to announce plans to buy
5	is in talks to buy
5	would buy
5	succeeds in buying
5	, said it 's buying

=== unit ===

314	, a unit of
24	is a unit of
6*	's New York-based trust unit ,
5	a unit of

=== parent ===

108	, the parent of
81	, parent of
56	, the parent company of
14	, parent company of
10*	's parent ,
9*	's parent company ,
6*	, whose parent company is

=== acquire ===

70	acquired
38	said it will acquire
23	agreed to acquire
16	will acquire
16*	agreed to be acquired by
14*	, has agreed to be acquired by
13	to acquire
9	said it agreed to acquire
8*	was acquired by
8*	, which agreed to be acquired by
7	would acquire
7	said it would acquire
7*	is being acquired by
6*	, which was acquired by
6*	, which is being acquired by
5	, which acquired
5	succeeds in acquiring

=== agree ===

(8)	agreed to merge with
(8)	said it agreed to purchase
(8)	, agreed to accept any offer by
(6)	agreed to pay $ 19 billion for
(6)	has already agreed to make
(5)	agreed to pay
(5)	agreed to sell

=== subsidiary ===

108	, a subsidiary of
10	is a subsidiary of
(8)	's Brown & Williamson subsidiary ,
7	, a wholly owned subsidiary of
5	a subsidiary of
5*	's U.S. subsidiary ,
5	, both subsidiaries of
5	will become a subsidiary of

=== purchase ===

51	's purchase of
7	purchased
7	an option to purchase
6	for a six-year option to purchase
(6)	purchased Sterling Winthrop from
6	recently completed its purchase of
6	completes its purchase of
6	's purchase of the 37 percent of
(6)	last week purchased rival Marshalls from
(5)	's purchase of S.G. Warburg Group Plc ,
5	's $ 5.4 billion purchase of

=== acquisition ===

41	's acquisition of
21	's proposed acquisition of
11	's planned acquisition of
6	's $ 3.7 billion acquisition of
(5)	, Dresdner Bank AG 's planned acquisition of
5	's pending acquisition of
5	completed the $1 billion stock acquisition of

Figure 4. Gathered phrases using keywords
(* indicates reverse relation, () indicates it is not paraphrase of the other phrases in the set)

=== Union Pacific Corp. Southern Pacific Rail Corp.

8	-	in its takeover by
2	+	agreed to buy
2	+	said it will buy

=== United Airlines UAL

26	-	, the parent of
5	-	, parent of
4	-	, the holding company for

=== Eastern Group Plc Hanson Plc

13	+	, has agreed to be acquired by
8	+	, now owned by
2	-	to acquire
2	-	's agreement to buy

=== American Airlines AMR

18	-	, the parent of
4	-	, the holding company for
2	-	, the parent company of

=== International Business Machines Corp. Lotus Development Corp.

3	+	said it would buy
2	+	's bid for
2	-	agreed to be bought by

Figure 5. Examples of NE instance pairs for links
"+" indicates the same order of NEs,
"-" indicates the reverse order