Paraphrase Fragment Extraction from Monolingual Comparable Corpora

Rui Wang

Language Technology Lab

DFKI GmbH

Stuhlsatzenhausweg 3 / Building D3 2

Saarbruecken, 66123 Germany

`rwang@coli.uni-sb.de`

Chris Callison-Burch

Computer Science Department

Johns Hopkins University

3400 N. Charles Street (CSEB 226-B)

Baltimore, MD 21218, USA

`ccb@cs.jhu.edu`

Abstract

We present a novel paraphrase fragment pair extraction method that uses a monolingual comparable corpus containing different articles about the same topics or events. The procedure consists of document pair extraction, sentence pair extraction, and fragment pair extraction. At each stage, we evaluate the intermediate results manually, and tune the later stages accordingly. With this minimally supervised approach, we achieve 62% of accuracy on the paraphrase fragment pairs we collected and 67% extracted from the MSR corpus. The results look promising, given the minimal supervision of the approach, which can be further scaled up.

1 Introduction

Paraphrase is an important linguistic phenomenon which occurs widely in human languages. Since paraphrases capture the variations of linguistic expressions while preserving the meaning, they are very useful in many applications, such as machine translation (Marton et al., 2009), document summarization (Barzilay et al., 1999), and recognizing textual entailment (RTE) (Dagan et al., 2005).

However, such resources are not trivial to obtain. If we make a comparison between paraphrase and MT, the latter has large parallel bilingual/multilingual corpora to acquire translation pairs in different granularity; while it is difficult to find a "naturally" occurred paraphrase "parallel" corpora. Furthermore, in MT, certain words can be translated into a (rather) small set of candidate words in the target language; while in principle, each paraphrase can have infinite number of "target" expressions, which reflects the variety of each human language.

A variety of paraphrase extraction approaches have been proposed recently, and they require different types of training data. Some require bilingual parallel corpora (Callison-Burch, 2008; Zhao et al., 2008), others require monolingual parallel corpora (Barzilay and McKeown, 2001; Ibrahim et al., 2003) or monolingual comparable corpora (Dolan et al., 2004).

In this paper, we focus on extracting paraphrase fragments from monolingual corpora, because this is the most abundant source of data. Additionally, this would potentially allow us to extract paraphrases for a variety of languages that have monolingual corpora, but which do not have easily accessible parallel corpora.

This paper makes the following contributions:

1. We adapt a translation fragment pair extraction method to paraphrase extraction, i.e., from bilingual corpora to monolingual corpora.

2. We construct a large collection of paraphrase fragments from monolingual comparable corpora and achieve similar quality from a manually-checked paraphrase corpus.

3. We evaluate both intermediate and final results of the paraphrase collection, using the crowdsourcing technique, which is effective, fast, and cheap.

52

Corpora		Sentence level	Sub-sentential level
Paraphrase acquisition			
Monolingual	Parallel	e.g., Barzilay and McKeown (2001)	This paper
	Comparable	e.g., Quirk et al. (2004)	e.g., Shinyama et al. (2002) & This paper
Bilingual	Parallel	N/A	e.g., Bannard and Callison-Burch (2005)
Statistical machine translation			
Bilingual	Parallel	Most SMT systems	SMT phrase tables
	Comparable	e.g., Fung and Lo (1998)	e.g., Munteanu and Marcu (2006)

Table 1: Previous work in paraphrase acquisition and machine translation.

2   Related Work

Roughly speaking, there are three dimensions to characterize the previous work in paraphrase acquisition and machine translation, whether the data comes from monolingual or bilingual corpora, whether the corpora are parallel or comparable, and whether the output is at the sentence level or at the sub-sentential level. Table 1 gives one example in each category.

Paraphrase acquisition is mostly done at the sentence-level, e.g., (Barzilay and McKeown, 2001; Barzilay and Lee, 2003; Dolan et al., 2004), which is not straightforward to be used as a resource for other NLP applications. Quirk et al. (2004) adopted the MT approach to "translate" one sentence into a paraphrased one. As for the corpora, Barzilay and McKeown (2001) took different English translations of the same novels (i.e., monolingual parallel corpora), while the others experimented on multiple sources of the same news/events, i.e., monolingual comparable corpora.

At the sub-sentential level, interchangeable patterns (Shinyama et al., 2002; Shinyama and Sekine, 2003) or inference rules (Lin and Pantel, 2001) are extracted, which are quite successful in named-entity-centered tasks, like information extraction, while they are not generalized enough to be applied to other tasks or they have a rather small coverage, e.g. RTE (Dinu and Wang, 2009). To our best knowledge, there is few focused study on *general* paraphrase fragments extraction at the sub-sentential level, from comparable corpora. A recent study by Belz and Kow (2010) mainly aimed at natural language generation, which they performed a small scale experiment on a specific topic, i.e., British hills.

Given the available parallel corpora from the MT community, there are studies focusing on extracting paraphrases from bilingual corpora (Bannard and Callison-Burch, 2005; Callison-Burch, 2008; Zhao et al., 2008). The way they do is to treat one language as an pivot and equate two phrases in the other languages as paraphrases if they share a common pivot phrase. Paraphrase extraction draws on phrase pair extraction from the translation literature. Since parallel corpora have many alternative ways of expressing the same foreign language concept, large quantities of paraphrase pairs can be extracted.

As for the MT research, the standard statistical MT systems require large size of parallel corpora for training and then extract sub-sentential translation phrases. Apart from the limited parallel corpora, comparable corpora are non-parallel bilingual corpora whose documents convey the similar information are also widely considered by many researchers, e.g., (Fung and Lo, 1998; Koehn and Knight, 2000; Vogel, 2003; Fung and Cheung, 2004a; Fung and Cheung, 2004b; Munteanu and Marcu, 2005; Wu and Fung, 2005). A recent study by Smith et al. (2010) extracted parallel sentences from comparable corpora to extend the existing resources.

At the sub-sentential level, Munteanu and Marcu (2006) extracted sub-sentential translation pairs from comparable corpora based on the log-likelihood-ratio of word translation probability. They exploit the possibility of making use of reports within a limited time window, which are about the same event or having overlapping contents, but in different languages. Quirk et al. (2007) extracted fragments using a generative model of noisy translations. They show that even in non-parallel corpora, useful parallel words or phrases can still be found and the size of such data is much larger than that of
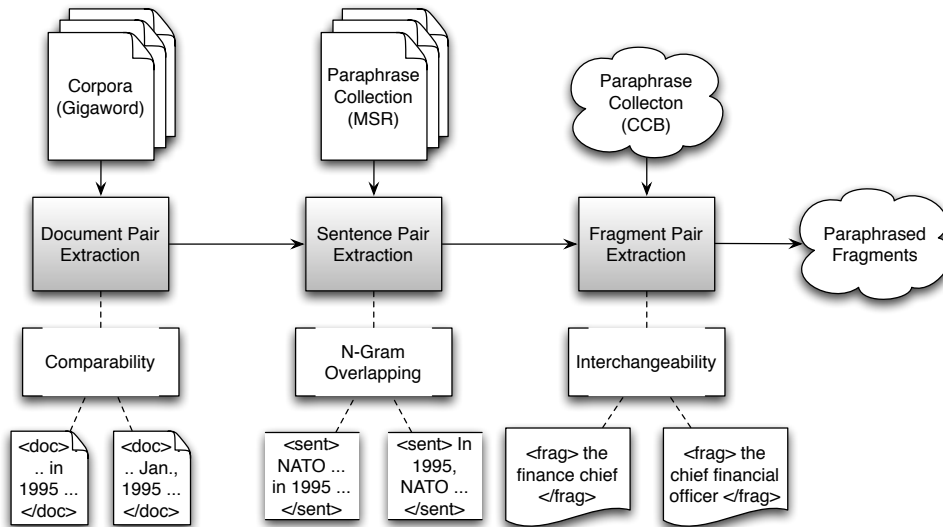
Figure 1: A three stage pipeline is used to extract paraphrases from monolingual texts

parallel corpora. In this paper, we adapt ideas from the MT research on extracting sub-sentential translation fragments from bilingual comparable corpora (Munteanu and Marcu, 2006), and use the techniques to extract paraphrases from monolingual parallel and comparable corpora.

Evaluation is another challenge for resource collection, which usually requires tremendous labor resources. Both Munteanu and Marcu (2006) and Quirk et al. (2007) evaluated their resources indirectly in MT systems, while in this paper, we make use of the crowd-sourcing technique to manually evaluate the quality of the paraphrase collection. In parcitular, Amazon's Mechanical Turk[1] (MTurk) provides a way to pay people small amounts of money to perform tasks that are simple for humans but difficult for computers. Examples of these Human Intelligence Tasks (or HITs) range from labeling images to moderating blog comments to providing feedback on relevance of results for a search query. Using MTurk for NLP task evaluation has been shown to be significantly cheaper and faster, and there is a high agreement between aggregate non-expert annotations and gold-standard annotations provided by the experts (Snow et al., 2008).

3 Fragment Pair Acquisition

Figure 1 shows the pipeline of our paraphrase acquisition method. We evaluate quality at each stage using Amazon's Mechanical Turk. In order to ensure that the non-expert annotators complete the task accurately, we used both positive and negative controls. If annotators answered either control incorrectly, we excluded their answers. For all the experiments we describe in this paper, we obtain the answers within a couple of hours or an overnight. Our focus in this paper is on fragment extraction, but we briefly describe document and sentence pair extraction first.

3.1 Document Pair Extraction

Monolingual comparable corpora contain texts about the same events or subjects, written in one language by different authors (Barzilay and Elhadad, 2003). We extract pairs of newswire articles written by different news agencies from the GIGAWORD corpus, which contains articles from six different agencies. Although the comparable documents are not in parallel, at the sentential or sub-sentential level, the paraphrased fragments may still exist.

To quantify the comparability between two documents, we calculate the number of overlapping words and give them different weights based on TF-IDF (Salton and McGill, 1983) using the *More-*

LikeThis[2] function provided by *Lucene*.

After collecting the document pairs, we asked annotators, "Are these two documents about the same topic?", and allowing them to answer "Yes", "No", and "Not sure". Each set of six document pairs contained, four to be evaluated, one positive control (a pair of identical documents) and one negative control (a pair of random documents). We sampled 400 document pairs with the comparability score between 0.8 and 0.9, and another 400 pairs greater than 0.9. We presented them in a random order and each was labeled by three annotations. After excluding the annotations containing incorrect answers for either control, we took a majority vote for every document pair, and if three annotations are different from each other.

We found document pairs with >0.9 were classified by annotators to be related more than half the time, and a higher threshold would greatly decrease the number of document pairs extracted. We performed subsequent steps on the 3896 document pairs that belonged to this category.

3.2 Sentence Pair Extraction

After extracting pairs of related documents, we next selected pairs of related sentences from within paired documents. The motivation behind is that the standard word alignment algorithms can be easily applied to the paired sentences instead of documents. To do so we selected sentences with overlapping n-grams up to length n=4. Obviously for paraphrasing, we want some of the n-grams to differ, so we varied the amount of overlap and evaluated sentence pairs with a variety of threshold bands[3].

We evaluated 10 pairs of sentences at a time, including one positive control and two negative controls. A random pair of sentential paraphrases from the RTE task acted as the positive control. The negative controls included one random pair of non-paraphrased, but highly relevant sentences, and a random pair of sentences. Annotators classified the sentence pairs as: paraphrases, related sentences,

---

[2]http://lucene.apache.org/java/2_9_1/
api/contrib-queries/org/apache/lucene/
search/similar/MoreLikeThis.html

[3]In the experiment setting, the thresholds (maximum comparability and minimum comparability) for the 4 groups are, {0.78,0.206}, {0.206,0.138}, {0.138,0.115}, {0.115,0.1}.
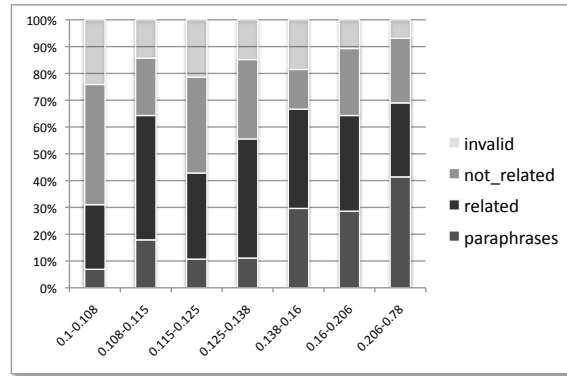

Figure 2: Results of the sentence pair extraction. The x-axis is the threshold for the comparability scores; and the y-axis is the distribution of the annotations.

and non-related sentences.

We uniformly sampled 200 sentence pairs from each band. They are randomly shuffled into more than 100 HITs and each HIT got three annotations. Figure 2 shows the distribution of annotations across different groups, after excluding answers that failed the controls.

Our best scoring threshold band was 0.2-0.8. Sentence pairs with this overlap were judged to be paraphrases 45% of the time, to be related 30% of the time, and to be unrelated 25% of the time. Although the F2 heuristic proposed by Dolan et al. (2004), which takes the first two sentences of each document pair, obtains higher relatedness score (we evaluated F2 sentences as 50% paraphrases, 37% related, and 13% unrelated), our n-gram overlap method extracted much more sentence pairs per document pair.

One interesting observation other than the general increasing tendency is that the portion of the related sentence pairs is not monotonic, which exactly reflects our intuition about a good comparability value (neither too high nor too low). However, some errors are difficult to exclude. For instance, one sentence says "The airstrikes were halted for 72 hours last Thursday..." and the other says "NATO and UN officials extended the suspension of airstrikes for a further 72 hours from late Sunday...". Without fine-grained analysis of the temporal expressions, it is difficult to know whether they are talking about the same event. The F2 method does provide us a fairly good way to exclude some unrelated sentence pairs, but note that the pairs collected by this method are
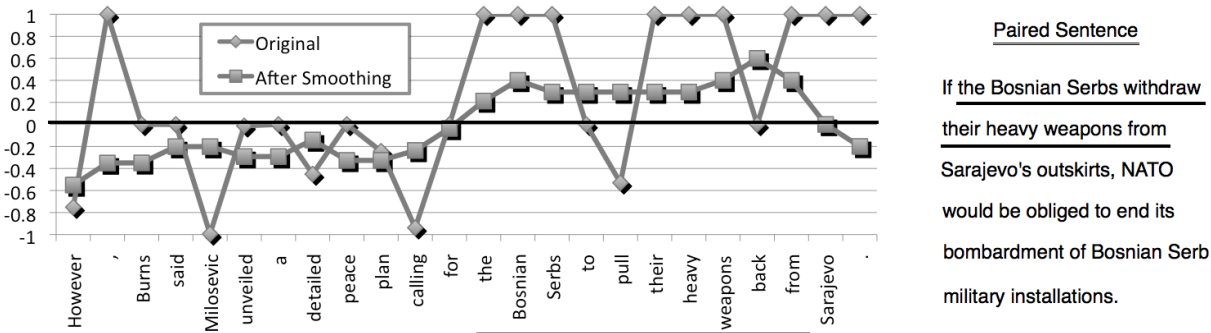
Figure 3: An example of fragment pair extraction. Stop words are all set to 1 initially. Zero is the threshold, and the underscored phrases are the outputs.

only about 0.5% of using the comparability scores.

We show in Figure 1 that we also use an additional sentence-level paraphrase corpus as the input of this module. We take all the positive instances (i.e. the two sentences in a pair are paraphrase to each other) and pass them to the later stage as well, as for comparison with our paraphrase collection extracted from the comparable sentence pairs. In all, we used 276,120 sentence pairs to feed our fragment extraction method.

3.3 Fragment Pair Extraction

The basic procedure is to 1) establish alignments between words or n-grams and 2) extract target paraphrase fragments. For the first step, we use two approaches. One is to change the common substring alignment problem from string to word sequence and we extend the longest common substring (LCS) extraction algorithm to multiple common n-grams. An alternative way is to use a normal word aligner (widely used as the first step in MT systems) to accomplish the job. For our experiments, we use the BerkeleyAligner[4] (Liang et al., 2006) by feeding it a dictionary of pairs of identical words along with the paired sentences. We can also combine these two methods by performing the LCS alignment first and adding additional word alignments from the aligner. These form the three configurations of our system (Table 2).

Following Munteanu and Marcu (2006), we use both positive and negative lexical associations for the alignment. The positive association measures

[4] http://code.google.com/p/berkeleyaligner/

how likely one word will be aligned to another (value from 0 to 1); and the negative associations indicates how *unlikely* an alignment exists between a word pair (from -1 to 0). The basic idea to have both is that when a word cannot be aligned with any other words, it will choose the *least unlikely* one. If the positive association of w_1 being aligned with w_2 is defined as the conditional probability $p(w_1|w_2)$, the negative associations will simply be $p(w_1|\neg w_2)$. Since we obtain a distribution of all the possible words aligned with w_1 from the word aligner, both $p(w_1|w_2)$ and $p(w_1|\neg w_2)$ can be calculated; for the LCS alignment, we simply set $p(w_1|w_2)$ as 1 and $p(w_1|\neg w_2)$ as -1, if w_1 and w_2 are aligned; and vice versa, if not.

After the initialization of all the word alignments using the two associations, each word takes the average of the neighboring four words and itself. The intuition of this smoothing is to tolerate a few unaligned parts (if they are surrounded by aligned parts). Finally, all the word alignments having a positive score will be selected as the candidate fragment elements. Figure 3 shows an example of this process.

The second step, fragment extraction, is a bit tricky, since a *fragment* is not clearly defined like a *document* or a *sentence*. One option is to follow the MT definition of a *phrase*, which means a sub-sentential n-gram string (usually n is less than 10). Munteanu and Marcu (2006) adopted this, and considered all the possible sub-sentential translation fragments as their targets, i.e. the adjacent n-grams. For instance, in Figure 3, all the adjacent words above the threshold (i.e. zero) will form the target

56

Configurations			
Aligner+ Phrase Extraction	LCS+ Chunk	Word+ N-Gram	LCS+Word+ Chunk
Our Corpus			
PARAPHRASE	15%	36%	32%
RELATED	21%	26%	21%
SUM	36%	**62%**	53%
The MSR Corpus			
PARAPHRASE	38%	44%	49%
RELATED	20%	19%	18%
SUM	58%	63%	**67%**

Table 2: Distribution of the Extracted Fragment Pairs of our Corpus and MSR Corpus. We manually evaluated 1051 sentence pairs in all. We use LCS or word aligner as the initialization and apply n-gram-based or chunk-based phrase extraction. The first column serves as the baseline.

paraphrase, "the Bosnian Serbs to pull their heavy weapons back from" and those aligned words in the other sentence "the Bosnian Serbs withdraw their heavy weapons from" will be the source paraphrase. The disadvantage of this definition is that the extracted fragment pairs might not be easy for human beings to interpret or they are even ungrammatical (cf. the fourth example in Table 5). An alternative way is to follow the linguistic definition of a *phrase*, e.g. noun phrase (NP), verb phrase (VP), etc. In this case, we need to use (at least) a chunker to preprocess the text and obtain the proper boundary of each fragment and we used the OpenNLP chunker.

We finalize our paraphrase collection by filtering out identical fragment pairs, subsumed fragment pairs (one fragment is fully contained in the other), and fragment having only one word. Apart from sentence pairs collected from the comparable corpora, we also did experiments on the existing MSR paraphrase corpus (Dolan and Brockett, 2005), which is a collection of manually annotated sentential paraphrases.

The evaluation on both collections is done by the MTurk. Each task contains 8 pairs of fragments to be evaluated, plus one positive control using identical fragment pairs, and one negative control using a pair of random fragments. All the fragments are shown with the corresponding sentences from where they are extracted[5]. The question being asked is

"How are the two highlighted phrases related?", and the possible answers are, "These phrases refer to the same thing as each other" (PARAPHRASE), "These phrases are overlap but contain different information" (RELATED), and "The phrases are unrelated or invalid" (INVALID). Table 2 shows the results (excluding invalid sentence pairs) and Table 5 shows some examples.

In general, the results on MSR is better than those on our corpus[6]. Comparing the different settings, for our corpus, word alignment with n-gram fragment extraction works better; and for corpora with higher comparability (e.g. the MSR corpus), the configuration of using both LCS and word alignments and the chunk-based fragment extraction outperforms the others. In fact, PARAPHRASE and RELATED are not quite comparable[7], since the boundary mismatch of the fragments may not be obvious to the Turkers. Nevertheless, we would assume a cleaner output from the chunk-based method, and both approaches achieve similar levels of quality.

Zhao et al. (2008) extracted paraphrase fragment pairs from bilingual parallel corpora, and their log-liner model outperforms Bannard and Callison-Burch (2005)'s maximum likelihood estimation method with 67% to 60%. Notice that, our starting corpora are (noisy) comparable corpora instead of parallel ones (for our corpus), and the approach is almost unsupervised[8], so that it can be easily scaled up to other larger corpora, e.g. the news websites. Furthermore, we compared our fragment pair collection with Callison-Burch (2008)'s approach on the same MSR corpus, only about 21% of the extracted paraphrases appear on both sides, which shows the potential to combine different resources.

4 Analysis of the Collections

In this section, we present some analysis on the fragment pair collection. We show the basic statistics of the corpora and then some examples of the output.

---

[5]We thought about evaluating pairs of isolated fragments, but later found out it was difficult to make the judgement.

[6]A sample of the corpus can be downloaded here: `http://www.coli.uni-saarland.de/~rwang/resources/paraphrases`.

[7]Thanks to the anonymous reviewer who pointed this out.

[8]The MTurk annotations can be roughly viewed as a guide for parameter tuning instead of *training* the system

As for comparison, we choose two other paraphrase collections, one is acquired from parallel bilingual corpora (Callison-Burch, 2008) and the other is using the same fragment extraction algorithm on the MSR corpus.

4.1 Statistics of the Corpora

Stage	Collection Size	%
GIGAWORD (1995)	600,000	10%
Documents Retrieved	150,000	2.5%
Document Pairs Selected	10,000	0.25%
Sentence Pairs Extracted	270,000	0.1%
Fragment Pairs Extracted	90,000	0.01%

Table 3: The size of our corpus. We only used ca. 10% of the GIGAWORD corpus in the experiments and the size of the collection at each stage are shown in the table.

Table 3 roughly shows the percentage of the extracted data compared with the original GIGAWORD corpus at each stage[9]. In the experiments reported here, we only use a subset of the news articles in 1995. If we scale to the full GIGAWORD corpus (19 Gigabytes, news from 1994 to 2006), we expect an order of magnitude more fragment pairs to be collected.

Apart from the size of the corpus, we are also interested in the composition of the corpus. Table 4 shows the proportions of some n-grams contained in the corpus. Here CCB denotes the paraphrase collection acquired from parallel bilingual corpora reported in (Callison-Burch, 2008), and MSR' denotes the collection using the same algorithm on the MSR corpus.

In Table 4, the four columns from the left are about the fragments (one part of each fragment pair), and the six columns from the right are about paraphrases. For example, 1 & 2 indicates the paraphrase contains one single word on one side and a 2-gram on the other side. Since we deliberately exclude single words, the n-gram distributions of OUR and MSR are "flatter" than the other two corpora, but still, 2-grams fragments occupy more than 40% in all cases. The n-gram distributions of the paraphrases are even more diverse for the OUR and MSR corpora. The sum

---

[9]All the numbers in the table are roughly estimated, due to the variations of different settings. This just gives us an impression of the space for improvement.

of the listed proportions are only around 45%, while for CCB and MSR', the sums are about 95%.

4.2 Examples

Table 5 shows some examples from the best two settings. From our corpus, both simple paraphrases ("Governor ... said" and "Gov. ... announced") and more varied ones ("rose to fame as" and "the highlight of his career") can be extracted. It's clear that the smoothing and extraction algorithms do help with finding non-trivial paraphrases (shown in Figure 3). The extracted phrase "campaign was" shows the disadvantage of n-gram-based phrase extraction method, since the boundary of the fragment could be improper. Using a chunker can effectively exclude such problems, as shown in the lower part of the table, where all the extracted paraphrases are grammatical phrases. Even from a parallel paraphrase corpus at the sentence level, the acquired fragment pairs (w/o context) could be non-paraphrases. For instance, the second pair from the MSR corpus shows that one news agency gives more detailed information about the launching site than the other, and the last example is also debatable, whether it's "under $200" or "around $200" depending on the reliability of the information source.

5 Summary and Future Work

In this paper, we present our work on paraphrase fragment pair extraction from monolingual comparable corpora, inspired by Munteanu and Marcu (2006)'s bilingual method. We evaluate our intermediate results at each of the stages using MTurk. Both the quality and the quantity of the collected paraphrase fragment pairs are promising given the minimal supervision. As for the ongoing work, we are currently expanding our extraction process to the whole GIGAWORD corpus, and we plan to apply it to other comparable corpora as well. For the future work, we consider incorporating more linguistic constraints, e.g. using a syntactic parser (Callison-Burch, 2008), to further improve the quality of the collection. More importantly, applying the collected paraphrase fragment pairs to other NLP applications (e.g. MT, RTE, etc.) will give us a better view of the utility of this resource.

N-grams	Phrases				Para-phrases					
	1	2	3	4	1 & 1	1 & 2	2 & 2	1 & 3	2 & 3	3 & 3
OUR	N/A	**43.4%**	30.5%	16.4%	N/A	N/A	**20.0%**	N/A	16.7%	8.8%
MSR	N/A	**41.7%**	30.5%	16.0%	N/A	N/A	**20.1%**	N/A	16.6%	9.4%
CCB	10.7%	**42.7%**	32.0%	10.9%	**34.7%**	16.3%	24.0%	2.5%	9.4%	6.9%
MSR'	8.1%	**41.4%**	37.2%	10.0%	**29.0%**	16.6%	26.8%	2.8%	10.7%	9.6%

Table 4: The (partial) distribution of N-grams (N=1-4) in different paraphrase collections.

From Our Corpus: using word aligner and n-gram-based phrase extraction	
... unveiled a detailed peace plan calling for the Bosnian Serbs **to pull their heavy weapons back from** Sarajevo. If the Bosnian Serbs **withdraw their heavy weapons from** Sarajevo's outskirts, ...	Paraphrase
In San Juan, **Puerto Rico, Governor Pedro Rosello said** the the storm could hit the US territory by Friday, ... In **Puerto Rico, Gov. Pedro Rossello announced** that banks will be open only until 11 a.m. Friday and ...	Paraphrase
Kunstler **rose to fame as** the lead attorney for the "Chicago Seven," ... **The highlight of his career** came when he defended the Chicago Seven ...	Paraphrase
... initiated **the air attacks** in response to Serb shelling of Sarajevo that killed 38 people Monday. The **campaign was** to respond to a shelling of Sarajevo Monday that killed 38 people.	Invalid
From MSR Corpus: using both LCS and word aligner and chunk-based phrase extraction	
O'Brien's attorney, Jordan Green, **declined to comment**. Jordan Green, the prelate's private lawyer, said he **had no comment**.	Paraphrase
Iraq's nuclear program had been dismantled, and there "was no convincing evidence **of its reconstitution**." Iraq's nuclear program had been dismantled and there was no convincing evidence **it was being revived**, ...	Paraphrase
... to blast off between next Wednesday and Friday **from a launching site** in the Gobi Desert. ... to blast off as early as tomorrow or as late as Friday **from the Jiuquan launching site** in the Gobi Desert.	Related
... Super Wireless Media Router, which will be available in the first quarter of 2004, **at under $200**. The router will be available in the first quarter of 2004 and will cost **around $200**, the company said.	Related

Table 5: Some examples of the extracted paraphrase fragment pairs.

Acknowledgments

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*.

R. Barzilay and N. Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of EMNLP*.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL*.

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

Regina Barzilay, Kathleen McKeown, and Michael El-hadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of ACL*, College Park, MD.

Anja Belz and Eric Kow. 2010. Extracting parallel fragments from comparable corpora for data-to-text generation. In *Proceedings of the 6th International Natural Language Generation Conference*, Stroudsburg, PA, USA.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of EMNLP*.

I. Dagan, O. Glickman, and B. Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the RTE Workshop*.

Georgiana Dinu and Rui Wang. 2009. Inference rules and their application to recognizing textual entailment.

In *Proceedings of EACL*, Athens, Greece. Association of Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the IWP2005*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of COLING*.

Pascale Fung and Percy Cheung. 2004a. Mining very non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of EMNLP*.

Pascale Fung and Percy Cheung. 2004b. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of COLING*.

P. Fung and Y. Y. Lo. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of ACL*.

Ali Ibrahim, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of ACL*.

Philipp Koehn and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In *Proceedings of the National Conference on Artificial Intelligence*.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of NAACL*.

Dekang Lin and Patrick Pantel. 2001. Dirt - discovery of inference rules from text. In *Proceedings of the ACM SIGKDD*.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of EMNLP*, Singapore.

Dragos Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting comparable corpora. *Computational Linguistics*, 31(4), December.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of ACL*.

Chris Quirk, Chris Brockett, and William B. Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Chris Quirk, Raghavendra Udupa, and Arul Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *Proceedings of MT Summit XI*, Copenhagen, Denmark.

G. Salton and M. J. McGill. 1983. *Introduction to modern information retrieval*. ISBN 0-07-054484-0. McGraw-Hill.

Y. Shinyama and S. Sekine. 2003. Paraphrase acquisition for information extraction. In *Proceedings of International Workshop on Paraphrasing*.

Yusuke Shinyama, Satoshi Sekine, and Kiyoshi Sudo. 2002. Automatic paraphrase acquisition from news articles yusuke shinyama satoshi sekine automatic paraphrase acquisition from news articles. In *Proceedings of Human Language Technology Conference*, San Diego, USA.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association of Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? Evaluating non- expert annotations for natural language tasks. In *Proceedings of EMNLP*.

Stephan Vogel. 2003. Using noisy bilingual data for statistical machine translation. In *Proceedings of EACL*.

Dekai Wu and Pascale Fung. 2005. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Proceedings of IJCNLP*, Jeju Island, Korea.

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2008. Pivot approach for extracting paraphrase patterns from bilingual corpora. In *Proceedings of ACL*.