Similarity-based Multi-Lingual Multi-Document Summarization

Anon1
WWW, ZZZ
XX, XX YYYYY
anon1@example.com

Anon2
WWW, ZZZ
XX, XX YYYYY
anon2@example.com

Abstract

We present a new multi-document system for summarizing automatic translations of foreign-language documents supplemented with related English documents and an evaluation of the system over data from the 2004 DUC conference. The system uses text similarity to choose sentences from the relevant English documents. Applying syntactic sentence simplification is shown to improve performance, and the approach outperforms first-sentence extraction baselines.

1 Introduction

With the large amount of text available on the web, summarization has become an important tool for monitoring and keeping up with information. While multi-document summarization of English text has become more common, less attention has been paid to producing English summaries of foreign language text, a situation becoming more important with the rapid globalization of our society.

We have recently been experimenting with summarization of clusters of multi-lingual multi-document news. Our input consists of multiple documents in a foreign language on a particular topic, and English documents that are relevant, but not translations of the foreign language documents. Our task is to produce an English summary of the foreign language documents. This task was added this year to the Document Understanding Conference (Over and Yen, 2003) on summarization this year. In this paper, we introduce a new method to summarize non-English documents in English using text similarity to the relevant English documents. We use machine translation to translate the non-English text and compute similarity between the translated text and the relevant English text. The summary is built by identifying the sentences to extract from the translated text, and replacing it with a sentence that is very similar from the relevant English text. The idea is to match content in the non-English documents that is realized in the English documents, thus ensuring that the summary contains text that is grammatically correct. We present results from our similarity-based system compared to two baselines (using first sentence extraction, and an existing summarization system) as well as results on performing two types of sentence simplification on the relevant English text.

1.1 Related Research

Previous work in multilingual document summarization, such as the SUMMARIST system (Hovy and Lin, 1999) extracts sentences from documents in a variety of languages, and translates the resulting summary. (Chen and Lin, 2000) describe a system that combines multiple monolingual news clustering components, a multilingual news clustering component, and a news summarization component. Their system clusters news in each language into topics, then the multilingual clustering component relates the clusters that are similar across languages. A summary is generated

for each language based on scores from counts of terms from both languages. The system has been implemented for Chinese and English, and an evaluation over six topics is presented. Our system differs by explicitly generating a summary in English using selection criteria from the non-English text.

Other work that use similarity-based approaches to summarization, such as the MEAD multi-document summarization system (Radev et al., 2000) are related in their use of similarity to guide selection, but our work is original in using text originally from one language to guide selection exclusively on English text. The MultiGen summarization system (Barzilay et al., 1999) uses text similarity to identify "themes" in a document, and then builds a summary sentence from a theme by combining information from the similar sentences. Our application of text similarity is to improve grammaticality by selecting similar content from English text, not to use similarity to identify important content, or merge information from similar sentences.

2 Summarization Approach

The goal of our similarity-based summarization approach is to take advantage of the large amount of English text available to us to help improve the quality of summaries produced in a multilingual environment. In applications such as an online multilingual news tracking system, multiple documents in a variety of languages, some of which are in English, must be summarized. One approach is to use machine translation systems to translate the documents, and then use the English documents to guide reformulation of the translated text in the summary to improve summary grammaticality and readability.

We use machine translation to obtain English translations of the documents, and use similarity at the sentence level to identify similar sentences from the English text. Using machine translation introduces problems with later text processing as the output from such systems contain more grammatical errors and poor lexical choice than natively-produced English text. To examine the effect that the machine translation output has on sentence selection (via an existing summarization system) and the text similarity component (trained and tested on English text) we include manual translations of the non-English documents in our evaluation. The system architecture is:

- Select sentences for the summary from the machine translated documents using an existing sentence extraction summarization system
- Compute similarity of selected sentence to related English sentences.
- Replace selected summary sentences with English sentences that are very similar.

In addition, to evaluate the idea of using similarity at sub-sentential levels, we apply sentence simplification of two types to the English documents. The sentence simplification system we use splits sentences at the clause level when possible and generates two simpler sentences.

2.1 Sentence Selection

The focus of this work is not extraction-based summarization, so instead of re-implementing some form of sentence selection, we used the DDDD summarization system (Anon, XXXXa) to select the sentences to use for the similarity computation process.

2.2 Sentence Simplification

The eventual aims of a similarity-based summarization system is to combine the machine translated text with the existing English text in a way that improves the readability of the summary. It is extremely difficult to find sentences in the related English documents containing exactly the same information as the translated sentences, and in fact we would prefer to perform similarity computation at a clause or phrase level. Since it is difficult to parse the output of the machine translation systems, we opted to use full sentences from the translated text, but wanted to perform some more sophisticated processing on the English text. We ran the English text through sentence simplification software (Siddharthan, 2002) to reduce them in the hope that a single concept would be expressed by each resulting sentence, allowing us to mix and match simplified sentences that might

have originally been from a single, more complicated sentence. The sentence simplification software breaks a long sentence into two separate sentences by removing embedded relative clauses from a sentence, and making a new sentence of the removed embedded relative clause.

We present results showing the effects of using full sentences, simplified sentences, and simplified sentences with pronoun resolution in Section 4.

2.3 Similarity Computation

Text similarity between the translated and relevant text is calculated using XXX (Anon, XXXXb). XXX is a tool for clustering text based on similarity computed over a variety of lexical and syntactic features. The main features used in XXX are the overlap of word stems, nouns, adjectives, verbs, WordNet (Miller et al., 1990) classes, noun phrase heads, proper nouns, and compound features that combine two features, possibly requiring the two instances to appear in the same order for a match to occur. The feature values for each pair of sentences is computed, and a final similarity value is assigned via a log-linear regression model with feature exponents learned from a corpus of news text. The sentence-clustering phase of XXX was skipped as our system only requires the final similarity values for the sentence selection stage. No modifications were made to XXX to try to compensate for using machine translated text as input, although we suspect that the machine translated text used is quite different from the news text used to train XXX.

2.4 System Integration

We participated in the DUC conference this year, which set a limit of 665 bytes for each summary. In order to ensure that we are able to produce a summary of that length, we have two levels of "fall back" for the system to use. The default uses DDDD for sentence selection, collects similar (possibly simplified) English sentences to each translated sentence selected by DDDD, and replaces the sentence with the most similar sentence from the English text. Only sentences with a similarity of over 0.30 are considered. The final summary is truncated to 665 bytes, as required by the DUC submission criteria.

If the full version of the system fails to generate a large enough summary (arbitrarily set as more than 600 bytes) the first fall-back is to re-run without the sentence selection phase. All of the input machine translated sentences are used in the similarity computation phase. This removes any intelligence in the content selection phase, instead selecting the most similar sentences between the machine translated text and the relevant English text. The machine translated sentences are sorted based on the highest similarity value to an English sentence, and the summary is built by taking the top-ranking English sentence for each machine translated sentence. If the summary is still less than 600 bytes, the second fall-back is to use DDDD to perform summarization over just the machine translated input. All summaries were limited to 665 bytes since that was the size threshold that was used for the DUC evaluation.

2.5 DUC 2004 Multilingual Data Set

For the 2004 DUC a new multilingual summarization task was added: 24 sets of Arabic news text were collected and translated by machine translation systems. Participating groups were asked to generate summaries for the machine translation output, human translations of the articles, as well as a third optional run with the option to make use of additional relevant English articles for some of the sets. We applied our similarity-based summarization system to the 10 sets that contained supplementary English articles. Each of the 10 sets were summarized by 4 human assessors after reading the manual translations of the Arabic documents. These 4 summaries are used as the basis for evaluating the automatic summaries produced by the system via the Rouge metric.

3 Experiments

We ran our system in a variety of configurations over the DUC 2004 multilingual data set, and compared it to multiple baseline systems. The evaluation was performed using three automatic evaluation systems, BLEU (Papineni et al., 2001), NIST (NIST, 2002) and Rouge (Lin and Hovy, 2003). We include results from two baseline systems: a first-sentence system, and runs of the DDDD system. The first-sentence summarization

baseline takes the first-sentence from each document in the set until the maximum of 665 bytes is reached. If the first-sentence has been included from each document in the set, the second sentence from each document is included in the summary, and so on. Summaries were generated over the manual translations (1st.manual), the relevant English documents (1st.relevant), the IBM translations (1st.ibm), and the ISI translations (1st.isi). The DDDD system was also run over the manual translation, relevant English documents, IBM and ISI translations.

We ran three versions of our similarity-based system: one without any sentence simplification, one with simplified sentences without pronoun resolution and replacement, and one with pronoun resolution and replacement on simplified sentences. The types of runs are listed in Table 1, with the simplification type listed in the "Simplification" column. "Both" means syntactic simplification and pronoun resolution and replacement was performed. We ran with each type of simplification over all three translations of the Arabic text (IBM machine translations, ISI machine translations, and the manual translations from the LDC) paired with either the relevant English text, or the manual translations. Running with the manual translations as the input Arabic text shows the performance given "perfect" translation systems, while running with the manual translations as the relevant English text attempts to give us perfect "clustering" for finding relevant documents. The runs are listed in Table 1 as the data set used for the Arabic translations, and the data set used for the matching English text.

4 Results

4.1 Rouge scores

Table 1 lists both ROUGE-L and ROUGE-W-1.2 metrics. The ROUGE-L score is a longest common subsequence score, while ROUGE-W-1.2 weights the longest common subsequence. While the two metrics do result in slightly different rankings, they are very strongly correlated with a Spearman's Rho of $\rho = 0.9832$ at p-value less than 0.001. When taking the 95% confidence intervals into account, it is difficult to really separate the systems according to their score; a statistically significant distinction can be made between the systems in the bottom of the group (systems 1, 5, 8, 9, 10, 16) and the ones at the top, but many of the differences between the systems are not significant at the 95% confidence level. The top systems are DDDD over the ISI translations, manual translations to relevant English with syntactic simplification, manual translations to relevant English with no simplification, and ISI translations to relevant English with syntactic simplification. The first-sentence and DDDD runs tend to have the largest confidence intervals, while the similarity-based runs have smaller confidence intervals. The chart for the ROUGE-W-1.2 metric is similar to the chart for the ROUGE-L metric, so has not been included. While the highest scoring system is the first-sentence extraction over ISI translations run, the large confidence interval makes it not significantly better than systems that the other top systems beat. Of the top four systems, two used manual translations to extract from the relevant English articles, one used ISI's translations with the relevant articles, and one was a run of DDDD over the ISI translations. Three of the top 4 runs used the relevant English articles and beat extraction systems that had access to manual or machine translations of the data that was being summarized. This indicates that the approach of summarizing a set of non-English articles using

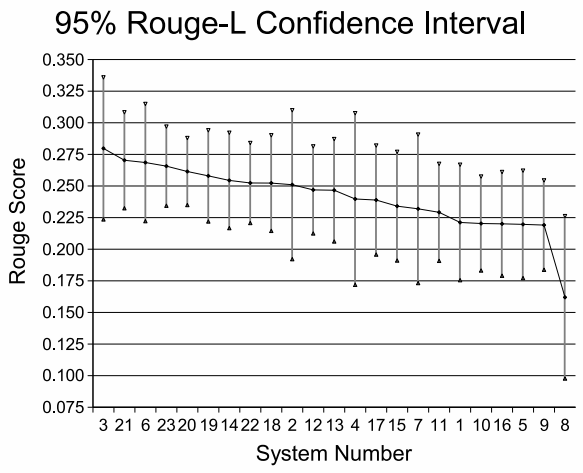

Figure 1: Graph of 95% Confidence Intervals for Rouge-L Metric.

System	Simplification	Number	NIST	BLEU	ROUGE-L	ROUGE-W-1.2
1st.ibm		1	3.9725	0.0778	0.22118	0.09481
1st.isi		2	4.0354	0.1058	0.25103	0.10686
1st.manual		3	4.5706	0.168	0.27975	0.12177
1st.relevant		4	4.4016	0.1104	0.23973	0.10091
DDDD.ibm		5	4.1112	0.0658	0.21966	0.09408
DDDD.isi		6	4.8671	0.137	0.26856	0.11325
DDDD.manual		7	4.6803	0.1057	0.23196	0.09927
DDDD.relevant		8	3.2906	0.0762	0.16197	0.07016
IBM Manual	none	9	4.6952	0.1278	0.2191	0.09531
IBM Manual	both	10	4.5836	0.1075	0.22033	0.09527
IBM Manual	syntactic	11	4.8294	0.1221	0.22918	0.09988
IBM Relevant	none	12	4.6717	0.0847	0.24691	0.1047
IBM Relevant	both	13	4.5262	0.0692	0.24664	0.10419
IBM Relevant	syntactic	14	4.6695	0.0797	0.25441	0.10772
ISI Manual	none	15	4.7148	0.1222	0.23405	0.10187
ISI Manual	both	16	4.5144	0.1111	0.22007	0.09434
ISI Manual	syntactic	17	4.7737	0.1283	0.23889	0.10211
ISI Relevant	none	18	4.7642	0.109	0.2523	0.10694
ISI Relevant	both	19	4.7359	0.1078	0.25801	0.1089
ISI Relevant	syntactic	20	4.847	0.1091	0.26145	0.1108
Manual Relevant	none	21	5.0696	0.1307	0.27035	0.11539
Manual Relevant	both	22	4.8784	0.1195	0.25239	0.10791
Manual Relevant	syntactic	23	4.9899	0.1226	0.26571	0.11371

Table 1: Summary evaluation results. 1st is a set of runs using the first-sentence extraction baseline, DDDD is an extractive-based summarizer, the remaining runs are the similarity summarizer runs.

related relevant articles in English is a promising approach.

4.2 BLEU and NIST scores

The reasons for also evaluating against the BLEU and NIST scores are twofold: first, we wanted to see if these metrics popular with the machine translation community correlated to the ROUGE scores, which are used for evaluation in the DUC. Second, the BLEU and NIST metrics are said to score more favorably for grammaticality (Lin and Hovy, 2003), and since this particular summarization task is using output from machine translation systems, we would like to see if there is any discernible difference in the scores based on the extraction source for the summary. Using the similarity-based summarization approach to choose an English sentence to represent content from the machine translated source should result in a more grammatical summary than one constructed from extracts from the machine translation system output.

The BLUE and NIST metrics are correlated with each other (Spearman's Rho is $\rho = 0.6482213$ at p-value 0.001.) Comparing BLEU to Rouge-L there is a weaker positive correlation of $\rho = 0.4407115$ at p-value 0.036. NIST and Rouge-L correlate more strongly ($\rho = 0.5701581$ at p-value 0.005) but the BLEU and NIST metrics are moderately correlated with Rouge, and might be evaluating something different.

The evaluation metrics we have used are totally automatic, and as such we don't have any explicit judgments on grammaticality or understandability of the summaries that were produced. We examined the BLEU and NIST scores compared to the Rouge scores of the systems that use machine translation output to see if there is any difference

Num	ROUGE-L	95% CI Lower - Upper
3	0.27975	0.22357 - 0.33593
21	0.27035	0.23237 - 0.30833
6	0.26856	0.22219 - 0.31493
23	0.26571	0.23440 - 0.29702
20	0.26145	0.23493 - 0.28797
19	0.25801	0.22193 - 0.29409
14	0.25441	0.21668 - 0.29214
22	0.25239	0.22074 - 0.28404
18	0.25230	0.21448 - 0.29012
2	0.25103	0.19213 - 0.30993
12	0.24691	0.21247 - 0.28135
13	0.24664	0.20612 - 0.28716
4	0.23973	0.17186 - 0.30760
17	0.23889	0.19576 - 0.28202
15	0.23405	0.19105 - 0.27705
7	0.23196	0.17323 - 0.29069
11	0.22918	0.19080 - 0.26756
1	0.22118	0.17551 - 0.26685
10	0.22033	0.18315 - 0.25751
16	0.22007	0.17903 - 0.26111
5	0.21966	0.17715 - 0.26217
9	0.21910	0.18376 - 0.25444
8	0.16197	0.09774 - 0.22620

Table 2: 95% Confidence Intervals for Rouge-L metric

between how the metrics rank them compared to systems that use English text. In our data, four of the systems used extraction from the machine translation output (1st.ibm, 1st.isi, DDDD.ibm, and DDDD.isi) while the other systems used the relevant English text, or the manual translations.

Table 3 shows the ranks of the 4 systems that used machine translation output, and their average ranking. Both BLEU and NIST on average ranked those systems lower than either Rouge-L or Rouge-W, and except for DDDD.isi, cluster the machine translation based systems near the bottom of the rankings. Predictably, the Rouge-L and Rouge-W rankings are statistically significantly positively correlated using Pearson's product moment correlation coefficient (cor=0.99, 95% confidence interval 0.90 - 1.0, p-value = 0.001,) as are BLEU and NIST (cor=0.98, 95% CI 0.51 - 1.0, p-value = 0.01.) BLEU and Rouge-W are correlated (cor=0.93, 95% CI: 0.018 - 1.0, p-value = 0.03) but the confidence interval almost contains 0. NIST and Rouge-W are not positively correlated at the 0.05 level: (cor=0.84, p-value = 0.08, 95% CI: -0.41 - 1.0) so we are not confident that the NIST ranking is the same as the Rouge rankings. It is ranking the translation-based summaries worse than the other metrics.

NIST	BLEU	Rouge-L	Rouge-W
4	2	3	4
20	16	10	10
21	20	18	20
22	23	21	22
16.75	15.25	13	14

Table 3: Rankings and average rank of Machine Translation based summaries

4.3 Comparison to baselines

We have two baseline systems to compare against: a first-sentence extraction system, and an extraction-based multi-document summarization system. Are the similarity-based methods that match to the relevant English documents better than using first-sentence extraction over the relevant documents or the machine translation systems?

When looking only at the IBM and ISI to relevant English runs, comparing to the 1st.isi, 1st.ibm, or 1st.relevant systems, the similarity based systems perform better than the first-sentence extraction baseline under most metrics. For the NIST metric, all similarity-based systems perform better than any of the first-sentence systems. ISI and IBM translations with relevant English text and syntactic simplification outperform all first-sentence systems under Rouge metrics. The first-sentence runs over the relevant English text and IBM translations ranked at the bottom of the list, while the ISI first-sentence run performed better than the IBM translations to relevant English text with no sentence simplification or both syntactic simplification and pronoun resolution and replacement. Under BLEU, the first-sentence extraction system run over the relevant English text performed best, with the ISI similarity

runs next, followed by 1st.ISI, and the IBM similarity runs. The interesting point to note is that under Rouge, the metrics tailored to summarization, the similarity-based runs outperform the first-sentence baseline, and also does so under NIST which was shown to not treat the machine translation based runs the same as Rouge.

If you have manual translations of the Arabic documents, the first-sentence extraction technique performs better than any of the similarity-based techniques, and even better than extraction using DDDD under all but the NIST metric. Under BLEU and both ROUGE metrics, the next best performing system is the similarity-based approach using manual translations to compare to the relevant English without sentence simplification. Under NIST, the manual translations to relevant English perform the best, followed by ISI translation to relevant English and IBM translations to the manual translations. While the manual translations in either the source language or target language improve performance of the similarity-based summarization approach, under most metrics the first-sentence baseline scores the best overall. The only metric where this isn't the case is the NIST metric, where the manual translations coupled with the relevant English documents score the best.

The manual translations coupled with the relevant English outperform the DDDD baselines in all but the BLEU metric. This shows that the approach is viable given good translations. DDDD.isi outperforms DDDD.manual, which seems unusual since the human reference summaries were made by people who had read the manual translations of the articles. The IBM translations running over the relevant English or manual translations always outperform the DDDD run over the IBM translations. An informal examination of the output of the ISI and IBM translation systems indicate that the ISI system does a better job translating and generating coherent noun phrases, while IBM does a better job at lexical choice for verb selection. ISI's edge in generating coherent noun phrases might reflect in improved scores, especially if the noun phrases are repeated in part or whole in the reference summaries. Looking at just the machine translations to relevant text, while DDDD.isi still performs the best, the similarity-based system outperforms DDDD over all the other sets. Under all metrics, the ISI translations to relevant text perform better than DDDD over manual translations, and only DDDD over the manual translations beats the IBM translations to relevant text. The other DDDD runs are beaten by the similarity-based systems.

5 Discussion

The results from the Rouge metric show that the similarity-based summarization approach is competitive with DDDD and the first-sentence extraction baseline. Using the 95% confidence intervals, the only systems that are shown to perform worse than any other systems are 1st.IBM, DDDD.ibm, and DDDD.relevant. It is interesting that a state-of-the-art summarization system run over the relevant English articles performs statistically significantly worse than the similarity-based summarization systems run over the same data. The similarity-based selection driven by the machine translations are able to select the good sentences from the relevant text. In fact, the ISI translations with relevant text performed better than either baseline over the manual translations on the NIST metric, or all metrics for just DDDD.

The NIST metric, with the exception of ranking DDDD.isi first, scores all similarity-based systems better than the two baselines. BLEU does not so clearly prefer the similarity-based systems. Unlike Rouge, we do not have confidence intervals for scores with the BLEU and NIST metrics; while it is possible to use bootstrapping to generate confidence intervals for BLEU and NIST scoring,[1] we do not have sentence level scores, as we simply allocated one summary per segment. Still, BLEU and NIST seem to indicate that the similarity-based methods performed better than the two baselines.

5.1 Sentence Simplification

A motivating factor behind using sentence simplification is to try to find the most relevant information to the translated Arabic text as possible, without looking only at sentence-sized chunks of

[1] http://projectile.is.cmu.edu/research/public/tools/bootStrap/tutorial.htm

text. A single sentence may easily contain multiple facts realized in different clauses, but we would like to include only the information that is relevant to the text in the translation in the summary. A first attempt at breaking down information within a sentence is to just simplify the sentences. Overall, syntactic sentence simplification improved scores over no sentence simplification. Under the ROUGE-L and ROUGE-W-1.2 metrics, scores were increased for all but the Manual-Relevant case. In every case, syntactic simplification with pronoun resolution and replacement performed worse than doing no sentence simplification or syntactic simplification only. This is likely because pronoun resolution is only successful about 70% of the time, introducing errors into the summary sentences.

6 Future Work

There are many areas that we would like to pursue for this work. In terms of evaluation, we would like to perform a qualitative evaluation that examines whether or not the similarity-based summaries have improved readability and/or intelligibility when compared with extraction methods performed over machine translation output. While Rouge metrics have been shown to correlate well with human judgments of summaries, it isn't clear that Rouge is as directly applicable when applied to summaries with mixed machine translation output and native English text.

We are also working on text similarity computation across languages, and would like to apply this summarization approach without the machine translation stage. Using a system that computes similarity between texts in different languages we could directly use the non-English text to guide selection from relevant English articles.

In this paper, we used sentence simplification to attempt to break down the similarity computation beyond the sentence level. In future work we plan to improve similarity computation to enable sub-sentential similarity. Instead of choosing entire sentences (or simplified sentences) we would choose a clause or phrase from the relevant text, and use text generation techniques to merge together information from multiple sources.

References

Anon. XXXXa. Removed for anonymous review. In *Removed*.

Anon. XXXXb. Removed for anonymous review. In *Removed*.

Regina Barzilay, Kathy McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Association of Computational Linguistics*, Maryland, June.

Hsin-Hsi Chen and Chuan-Jie Lin. 2000. A multilingual news summarizer. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 159–165.

E.H. Hovy and Chin-Yew Lin. 1999. Automated text summarization in summarist. In I. Mani and M. Maybury, editors, *Advances in Automated Text Summarization*, chapter 8. MIT Press.

Chin-Yew Lin and E.H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada, May.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography*, 4(3):235–244.

NIST. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.

Paul Over and J. Yen. 2003. An introduction to duc 2003 intrinsic evaluation of generic news text summarization systems. National Institute of Standards and Technology.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In *Proceedings of ANLP/NAACL 2000 Workshop*, pages 21–29, April.

Advaith Siddharthan. 2002. Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs. In *Proceedings of the Student Workshop, 40th Meeting of the Association for Computational Linguistics (ACL'02)*, pages 60–65, Philadelphia, USA.