社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－NL－165 (11)
2005／1／12

機械翻訳のための文簡易化

フィンチ アンドリュー　下畑 光夫　隅田 英一郎
ATR 音声言語コミュニケーション研究所
{andrew.finch, mitsuo.shimohata, eiichiro.sumita}@atr.jp

本論文では、機械翻訳のために入力文から不必要な語を取り除く方法について提案する。我々は、不必要な語を除去して簡易化した文は機械翻訳に有利であり、翻訳性能を向上させると考えている。この論文では、まず自動簡易化文と人手簡易化文の比較について述べる。次に、異なる2種類の機械翻訳システムに入力文 および自動簡易化文を与え、簡易化処理により訳質が向上したことについて述べる。自動簡易化処理では、人手による簡易化と同等の性能を得ることができた。また、入力文に自動簡易化を施した場合に、EGYPT ツールに基づく機械翻訳では大きな改善効果が、ATR の翻訳システムでもある程度の改善効果が確認できた。

Sentence Simplification for Machine Translation

Andrew Finch, Mitsuo Shimohata and Eiichiro Sumita
ATR Spoken Language Translation Research Laboratories
{andrew.finch,mitsuo.shimohata,eiichiro.sumita}@atr.jp

We propose a method for removing unnecessary words from sentences to faciliate automatic translation. The hypothesis being that the resulting simplified sentences will be easier to automatically translate, giving improved translation performance. This paper evaluates the system in isolation against a test set of human shortened sentences and also its application to assist two different machine translation systems. We show the system is able to perform at close to human performance in shortening sentences. We also show that we are able to significantly improve the performance of the publicly available EGYPT machine translation (MT) system, and make a small improvement to the ATR Translation System by pre-processing the input to these systems.

1 Introduction

Long sentences are often a problem for natural language processing tasks such as machine translation and parsing. In general, such processes require a search to find the optimal output given an input word sequence, and this search process often does not scale well with increasing input word sequence length. One strategy to mitigate this problem is to preprocess the input sequence into a form that is easier to process. Examples of this are paraphrasing strategies to shorten or normalize the input, and sentence splitting techniques that can be used to divide long sentences into shorter, more manageable sentences. In this paper we consider the strategy of word deletion as a method of simplifying sentences, with an eye to combining the technique with other simplification methods in the future. The main motivation for selecting such a simple strategy over a full paraphraser is that the task is much easier and therefore fewer errors are likely to be made by the system. We argue later in this paper that accuracy is of pivotal importance for this task.

2 Related Work

Word deletion has been used in other areas such as sentence compression for document summarization [6] and in the removal of disfluencies in real human speech to aid MT [3]. These methods both differ from the method

presented here in that they adopt a noisy channel model. In earlier experiments a maximum entropy noisy channel model was built in combination with a language model, but the results were not as good as those obtained using the direct ME model we employ here.

Other work on using paraphrasers to normalize the form of sentences to make them easier to translate has also been successful [11]. The paraphraser in this case operated by substituting word sequences in the training and test corpus for a standard normal form.

3 Methodology

3.1 The ME Model

Our system is built within a maximum entropy (ME) framework which allows the use of a combination of a diverse selection of contextual features. The model has the following form:

$$P(o,h) = \gamma \prod_{k=0}^{K} \alpha_k^{f_k(h,o)} \quad (1)$$

where:

- o is the outcome we are predicting;
- h is the history of o;
- γ is a normalization coefficient;
- K is the number of features;
- $\alpha_k (k = 1, K)$ is the weight of feature f_k;
- $f_k \in \{0,1\} (k = 1, K)$ are feature functions.

The improved iterative scaling technique [10] was used to train the parameters in the ME model.

The features we use are functions of the history of words, part-of-speech (POS) tags, outcomes, and also of the words and POS tags in the future. The POS tags are the UPENN [7] tags provided by a maximum entropy tagger. The features are composed of unigrams, bigrams and trigrams of objects of the same type and include the word whose outcome is being predicted. The context used in prediction is illustrated in Figure 1. In this figure for example, the active word-based features would be:

$$f_1 = \begin{cases} 1 & \text{if } o_0\text{=DEL \& } w_0\text{='really'} \\ 0 & \text{otherwise} \end{cases}$$

$$f_2 = \begin{cases} 1 & \text{if } o_0\text{=DEL \& } w_0\text{='really' \&} \\ & w_{-1}\text{='really'} \\ 0 & \text{otherwise} \end{cases}$$

$$f_3 = \begin{cases} 1 & \text{if } o_0\text{=DEL \& } w_0\text{='really' \&} \\ & w_{-1}\text{='really' \& } w_{-2}\text{='i'} \\ 0 & \text{otherwise} \end{cases}$$

Where o_i is the outcome at offset i from the word whose outcome is being predicted, w_i is the word at offset i, and the outcome "DEL" means the word has been marked as deleted (or to be deleted). "KEEP" means that the word was not deleted (or is not to be deleted).

3.2 Search

The ME model provides us with the joint outcome/history probability per word, but we require the conditional probability of the outcome sequence given the history. We calculate this by combining the per word joint probabilities according to:

$$P(o_1 \ldots o_n | w_1 \ldots w_n, t_1 \ldots t_n) = \prod_{i=1}^{n} p(o_i|h_i)$$

Where $w_1 \ldots w_n$ is the word sequence for the sentence, $o_1 \ldots o_n$ the sequence of outcomes and $t_1 \ldots t_n$ is the sequence of POS tags. The conditional probability is derived from the joint probability using:

$$P(o|h) = \frac{p(o,h)}{p(DEL,h) + p(KEEP,h)}$$

It is necessary to search for the optimal sequence of outcomes. Since there are only two outcomes, for most of our sentences it is feasible to perform a full search, and this is done for sentences of less than 20 words in length. For the few long sentences (less than 1% of the test corpus) we employ a simple beam search algorithm.

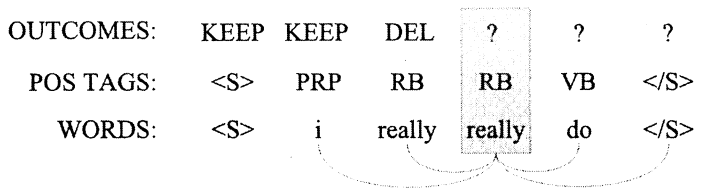

Figure 1: The Context of an Outcome

4 Corpora

The word deleter was trained on a corpus of approximately 20,000 English sentence pairs (approximately 165,000 words). One sentence in the pair being a simplified version of the other, such that it is possible to derive the simplied form from the original sentence only by deleting words. This corpus is a subset of a larger corpus of sentence pairs where the sentences are simplified in an arbitrary fashion. The guidelines for this simplifiction process are described in [12]. Example sentences from the training corpus are shown in Figure 2.

1. **a compact car** *would be okay*
2. *actually there 's going to be* **only me** *who 's playing*
3. *and* **for what date** *would it be*

Figure 2: Training data (italic words deleted)

In the MT evaluation experiments, the translation systems were trained on the ATR Basic Travel Expression Corpus (BTEC) [13], a corpus of approximately 400,000 sentences (2.5 million words).

5 Experiments

5.1 Stand-alone Evaluation

In this experiment we measure the performance of the word deleter with respect to a test set consisting of 1018 test sentences each with 9 reference sentences. The references were provided by human annotators. We compare the performance of the system against two performance benchmarks. We used the multi-reference word error rate (mWER) [8] to assess our performance, since this method has a very intuitive interpretation. Past experiments have shown that this measure has strong correlation with human performance relative to several other schemes we have tested, and importantly for these experiments, includes no brevity penalty.

The results are shown in Table 1. The ME word deleter had an mWER score of 3.25%. This seems to be a very accurate score, however many of the sentences in the test set are annotated by at least one annotator as not being compressible and it is possible to achieve a good evaluation score by simply keeping all the words. We therefore use the original sentence itself in unmodified form as our baseline. The human score in the table was calculated by jack-knifing over the human references. Each annotator was evaluated against a reference set constructed from the other annotators, and the results combined. The human scores are worse than the machine's score. This anomaly is caused by inconsistencies among the human annotators' output. Many of human's mWER scores were below 3.0% but one annotator had an mWER score of over 9%. The ME word deleter score in Table 1 is for 9 references rather than 8, the ME word deleter's score drops to 3.31% when the evaluation uses the same reference sets.

	Compression	mWER
Baseline	0%	8.04%
ME Deleter	12.0%	3.25%
Human	20.0%	4.66%

Table 1: Stand-alone Evaluation Results

5.2 Improving EGYPT

For this experiment a well-known and publicly available MT system was trained on the BTEC English-Japanese corpus described in Section 4. The system used was composed of GIZA++ [9] to build the models, and the ISA rewrite decoder [2]. The system was evaluated against a 510 sentence BTEC reference set with 16 references per sentence. The results are shown in Table 2.

The baseline system performed no processing on the input sentence. The 'Compressed Test' system deleted words from the test sentences before passing them to the translation system, and in 'Compressed Train and Test' the MT system was trained on sentences compressed by the ME word deleter, and also tested on the compressed test sentences. Compressing the input sentences improved translation quality, however surprisingly training the MT system on compressed data made the translations worse. The explanation for this is likely to be that the problems caused by errors introduced by processing the training corpus outweigh the gains from making the training corpus consistent with the output from the word deleter.

5.3 Improving the ATR Translation System

In this experiment we evaluated the effect on MT accuracy of our word deletion system when applied to the input to the ATR Translation System. This system has considerably higher performance on the translation task, and is based on two MT systems: HPATR [4] and SAT-greedy [14]. The final translation is selected from candidates from both systems using a language model and a translation model [1].

The experimental methodology is very similar to the previous experiment in that we pre-processed the input before handing it to the MT system, however in this case we only process segments of length greater than 6 words, since this translation system has little difficulty with short sentences. The effectiveness of this strategy for a different paraphraser is discussed in [11]. The test data in this case consisted of 502 sentences from the MAD4 corpus [5], a corpus of transcribed speech data that is less perfect than the phrase-book data from the BTEC corpus. Again 16 references per segment were used in the evaluation. The results are shown in Table 3 and show only a very small improvement in performance for this system, we will discuss the reasons for this in the following sections.

	Compression	mWER
Baseline	0%	53.64%
ME Deleter	5.84%	53.40%

Table 3: ATR Translation System Results

6 Discussion

Figure 3 shows example unseen sentences with the deleted word sequences shown in italics. The first two sentences are fairly typical of the operation of the word deleter on travel data. Quantifiers and phrases that express politeness are commonly deleted. In the third sentence, a more ambitious deletion is undertaken, the deleter is relying on the "i" at the start of the sentence to combine with the "ll" later on. This kind of deletion is successful in this example, but can lead to errors. The fourth sentence in the figure is an example where the deletions have caused an ungrammatical sentence to result. Although this is a clear error, it is not necessarily going to prevent the machine translation from generating correct translations since the grammaticality of the input sentences is not relied upon. The final sentence is a pronounced illustration of perhaps the most damaging negative effect of word deletion. In this sentence, the system has actually produced a perfectly valid short form of the original sentence, however some of the information in the original sentence has been removed, and translation of this sentence is unlikely to convey this information because the translation system does not have the chance to use it. In fact, all word deletion simplifications remove information from the sen-

	Compression (%)	mWER (%)
Baseline	0	61.00
Compressed Test	7.36	58.66
Compressed Train and Test	7.36	62.28

Table 2: EGYPT Evaluation Results

tence, and this information varies from virtually meaningless interjections, though words that convey subtle nuance, to the removal of valuable content words or phrases that convey the main meaning of the sentence.

1. **we would like** *some* **sake** *please*
2. *and ah* **how many** *people will be* **in your party**
3. **i** *think i* **'ll** *just* **take a taxi**
4. *i 'm afraid* **that my feet may be too big**
5. **the restaurant is on the left** *hand side of the road*

Figure 3: Example output (italics deleted)

This approach is also affected by the nature of MT evaluation methods. Both the NIST and BLEU scoring systems explicitly include a term that penalizes shorter sentences, and we therefore adopt the mWER score to evaluate our system's performance, although even this carries an implicit length penalty by virtue of the references in the reference set having not been constructed in a concise form.

7 Conclusion

In this paper we have presented an effective method for removing words from English sentences. The system performed well when its output was compared to a reference set of concise sentences. The system was also able to improve the performance of MT systems by pre-processing their input.

The success or failure of this approach hinges on whether the gains in translation performance made by simplifying the input, are able to offset the losses incurred by damage caused by erroneous word deletions and missing information. For this reason, it is necessary for our word deletion system to be very accurate, making few errors in its output, and this motivated our choice of model. Our system has achieved a high standard of accuracy, as shown by the stand-alone evaluation results. The differences between the small improvements to the ATR Translation System and the larger improvements to the EGYPT system reflect the differences in translation quality between the two systems. The ATR Translation System is more capable and therefore receives less benefit from the word deleter.

In theory it is possible to apply this word deleter to any language for which we have training data, but in practice there are difficulties with other languages since not all languages have a clear definition of "word". We trained and tested our word deleter on a sister corpus of Japanese data paraphrased in the same way as the English data, but our word deleter in its present form was not able to paraphrase Japanese as well as it can paraphrase English. The unit we used to represent the "word" in these experiments was the morpheme. The problem was due to the fact that in Japanese there are often long sequences of morphemes, all of which need to be deleted in order to preserve the grammaticality of the sentence. Work is currently underway to address this issue by making a word deleter which deletes at the level of bunsetsu. The idea being that busetsu can usually be removed without damaging the grammaticality of the sentence. This word deleter will be combined with a second word deleter trained to delete morphemes from within bunsetsu.

Future versions of this word deleter will incorporate other models within a log-linear framework. In particular, a language model of the target sentence and an target length model

will be included. The latter can be used to control the length of sentences produced by the system, should a greater or lesser shortening be required for a particular application. We intend to use this system in combination a sentence splitter and other paraphrasing devices to further improve its effectiveness.

Acknowledgments

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled "A study of speech dialogue translation technology based on a large corpus".

References

[1] Yasuhiro Akiba, Taro Watanabe, and Eiichiro Sumita. Using language and translation models to select the best among outputs from multiple mt systems. In *In Proc. COLING-2002*, pages 8–14, 2002.

[2] Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. Fast decoding and optimal decoding for machine translation. In *Meeting of the Association for Computational Linguistics*, pages 228–235, 2001.

[3] M. Honal and T. Schultz. Correction of disfluencies in spontaneous speech using a noisy-channel approach. In *Proceedings of Eurospeech-2003*, Genf, Schweiz, 20031.

[4] Kenji Imamura, Hideo Okuma, Taro Watanabe, and Eiichiro Sumita. Example-based machine translation based on syntactic transfer with statistical models. In *In Proc. COLING-2002*, pages 99–105, 2004.

[5] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. Creating corpora for speech-to-speech translation. In *Proceedings of EUROSPEECH-03*, pages 381–384, 2003.

[6] K. Knight and D. Marcu. Statistics-Based Summarization - Step One: Sentence Compression. In *National Conference on Artificial Intelligence (AAAI)*, pages 703–710, Texas, USA, 2000.

[7] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[8] S. Niessen, F. Och, and H. Ney. An evaluation tool for machine translation: Fast evaluation for machine translation research. In *Proceedings of the LREC Conference*, Athens, Greece, 2000.

[9] F. J. Och and H. Ney. Improved statistical alignment models. In *ACL00*, pages 440–447, Hong Kong, China, 2000.

[10] Della Pietra, Stephen, Della Pietra, Vincent J., and John D. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

[11] M. Shimohata and E. Sumita. Automatic paraphrasing based on parallel corpus for normalization. In *Proceedings of the LREC Conference*, Las Palmas, Gran Canaria, 2002.

[12] M. Shimohata, E. Sumita, and Yuji Matsumoto. Building a paraphrase corpus for speech translation. In *Proceedings of LREC-04*, pages 1407–1410, 2004.

[13] F. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World. In *Proceedings of the LREC Conference*, Las Palmas, Gran Canaria, 2002.

[14] Taro Watanabe and Eiichiro Sumita. Example-based decoding for statistical machine translation. In *Proceedings of Machine Translation Summit IX*, pages 410–417, 2003.