E-53

# Paraphrasing by Statistical Machine Translation

フィンチ　アンドリュー† 渡辺　太郎† 隅田　英一郎†
Andrew Finch Tarō Watanabe Eiichirō Sumita

## 1. Introduction

The ability to paraphrase text has many practical applications, for example, in the fields of text summarization and machine translation. This is particularly so in the case where the paraphrased sentence is shorter than the original sentence, since this may aid a human reader, or simply the task of a machine in processing the data. Long sentences cause problems with many language processing tasks; for example, machine translation and parsing, and there are large advantages in being able to simplify sentences, whilst preserving their meaning.

This paper presents work investigating the direct application of statistical machine translation (SMT) techniques to automatically paraphrase Japanese sentences, with the objective of shortening the sentences. There has been much recent interest in the area of text summarization using machine translation techniques (E.g. Knight and Marcu, 2000).

## 2. Paraphrase Corpus

The data we used for these experiments is a subset of the ATR's Paraphrase Corpus (Sugaya et al., 2002). The corpus used for these experiments consists of about 650,000 sentences (7.4 million words) of paraphrased sentences drawn from the kind of phrasebooks produced to aid travelers (Takezawa et al., 2002). There are approximately 2700 'seed' sentences that have been paraphrased to produce this data.

## 3. Methodology

For these experiments we treat the task of paraphrasing as a task of translation. The system is required to translate from one 'language' (long sentences) into another language (shorter sentences that convey the meaning of their longer counterparts). We chose to use the EGYPT machine translation system (El Onaizan et al., 2000) together with an in-house developed decoder to perform the translation task. The system is able to train using only a corpus consisting of pairs of sentences (one sentence from each 'language'). The data was divided into training and test sets. The output of the system was evaluated for adequacy by human evaluators.

## 4. Data Generation

In order to generate the training data for the machine translation system we first clustered the paraphrased sentences. This was to ensure that the sentence pairs used for training were as similar to each other as possible in terms of edit distance: the number of insertion, deletion or word-for-word substitution operations required to transform one sentence into another. We used the following agglomerative clustering algorithm to cluster sentences according to edit distance.

† ATR Spoken Language Translation Research Labs.

1) Assign each sentence in the set of paraphrased sentences to its own cluster.
2) For each possible pair of clusters $C_1$ and $C_2$, calculate the distance between them (the average edit distance between members of the clusters).

$$\text{distance}(C_1, C_2) = \frac{\sum_{c_1 \in C_1} \sum_{c_2 \in C_2} \text{editdist}(c_1, c_2)}{|C_1||C_2|}$$

3) Merge the two closest clusters
4) Repeat from 2) until there is only one cluster

The result of this clustering is a tree, or dendrogram. The leaf nodes of this tree are sentences. Leaves that are close to each other in to tree, are also similar in terms of edit distance. The motivation behind this approach is that sentences that are similar in terms of edit distance will also make a good sentence pair in the training of a machine translation device. Non-leaf nodes, define sets of similar sentences.

Clusters of similar sentences of arbitrary granularity can be extracted from the dendrogram as follows:

1) For each node in depth first a traversal of the tree
2) Calculate the average intra-cluster edit distance (I.e. The average edit distance between pairs of sentences within the cluster).
3) If this distance is less than some threshold value, extract the cluster of sentence defined by the leaf nodes under the node being considers, and exclude all nodes in this sub-tree from the traversal.

Once these clusters of similar sentences have been defined, we generate training data by pairing all sentences in the cluster (except the shortest) with the shortest sentence in the cluster. This procedure generated approximately 650,000 sentence pairs. The long form of the sentences averaged 10.8 words, the short form 7.3 words, an average sentence shortening of approximately 32%.

## 5. Machine Translation

The machine translation system, which uses EGYPT together with GIZA++ was trained directly on the sentence pairs produced by the generation procedure described above.

## 6. Results

The sentences in figure 1 show examples (both good and bad) of the input and output of the system. The system has learned to transform common long-winded turns of phrase into a more compact form. However, mistakes made by the system are

omitted words that render the sentence ungrammatical, or cause some of the meaning to be lost, and rarely additional words which lengthen the sentence. Sometimes the sentence is left unchanged by the system; however in 99% of all cases for these experiments, sentences are changed by the SMT paraphraser.

The amount of shortening of the sentences depended on their length. As would be expected, longer source sentences were shortened more than short source sentences. For example, on average, source sentences of length 6 words were shorted by 22% to a length of 4.68 words, and sentences of length 9 words were shortened by 30.5% to a length of 6.25 words.

[Source] ラケットを貸してくれないでしょうか。
[Human] ラケットを貸し出してください。
[Baseline] ラケットをないでしょうか。
[SMT] ラケットを貸してください。

[Source] どのような種類の料理がご希望でしょうか。
[Human] どんな料理が好きでしょうか。
[Baseline] 料理でしょうか。
[SMT] どんな料理がいいでしょうか。

[Source] 全部少しずつもらいたいのですけど。
[Human] 全部少しずつもらいたいのです。
[Baseline] たいのです。
[SMT] 少しレモネードがほしいのです。

Figure 1: Examples of paraphrasing

## 7. Evaluation

We evaluated the paraphrases produced by the system as machine translation output using the following *adequacy* test (Doyon at al., 1998). Sentences were graded (from 1 to 5) by a native Japanese speaking evaluator using the following adequacy scale:

[Grade 5] All meaning expressed in the source sentence is present in the paraphrased sentence.
[Grade 4] Most of the meaning expressed in the source sentence is present in the paraphrased sentence.
[Grade 3] Much of the meaning expressed in the source sentence is present in the paraphrased sentence.
[Grade 2] Little of the meaning expressed in the source sentence is present in the paraphrased sentence.
[Grade 1] None of the meaning expressed in the source sentence is present in the paraphrased sentence.

In addition, a sample of shortened correctly paraphrased sentences from the corpus was also mixed with the evaluation data to provide a human-labeled reference. Finally, following (Knight and Marcu, 2000) we use a baseline model based on maximum word-bigram probability of the target sentence. 100 sentences from each of the sources were mixed randomly and graded at the same time by a single judge. The results are shown in table 1. The results were subjected to a T-test to determine whether the differences between the techniques were significant. The tests show that at $p < 0.01$ the difference between the performance of the SMT technique and the performance of the baseline model is statistically significant. According to the same criteria, the human's score is also significantly different from that of the SMT system.

| | Compression | Adequacy |
|---|---|---|
| **Human** | 32% | 4.35±1.34 |
| **SMT** | 27% | 3.37±1.29 |
| **Baseline** | 36% | 1.96±1.25 |

Table 1: Experimental results

## 8. Conclusion and Future Directions

The results presented here are very encouraging. The system performed well; its score being much closer to the human's score than that of the baseline model. The paraphrasing task defined in this way carries a significant drawback however, in that the amount of training data generated can be large in size, and based on only a small number of seed sentences, even when using the clustering technique described above. This can limit the number of seed sentences used and cause problems with data sparseness, since the number of different tokens seen by the system is restricted. Work is already underway to address this problem by introducing the POS tag into the translation process. We believe that this will enable the system to generalize to words it has not seen in the training data for translation. We expect to be able to accurately assign a POS tag to these unseen words, and as a consequence construct a system that will be able to deal better with a wider range of source sentences.

## 9. Acknowledgements

This research was supported in part by the Telecommunications Advancement Organization of Japan.

## 10. References

Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R., 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, 19(2):263-311.

Doyon, J., Taylor, K., and White, J.S., 1998. The DARPA MT Evaluation Methodology: Past and Present, *Proceedings of the ATMA Conference*, Philadelphia, PA.

Al-Onaizan, Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, I.D., Och F.J., Purdy, D., Smith, N.A., and Yarowsky, D., 1999. Statistical machine translation, final report, *JHU workshop*.

Knight, K. and Marcu, D., 2000. Statistics-Based Summarization - Step One: Sentence Compression, *AAAI*, 703-710.

Sugaya, F., Takezawa, T., Kikui, G. and Yamamoto, S., 2002. Proposal of a very-large-corpus acquisition method by cell-formed registration, *Proceedings of the LREC Conference*, Las Palmas, Gran Canaria.

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