A Corpus-Centered Approach to Spoken Language Translation

Eiichiro SUMITA, Yasuhiro AKIBA, Takao DOI, Andrew FINCH, Kenji IMAMURA, Michael PAUL, Mitsuo SHIMOHATA, and Taro WATANABE

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, JAPAN
eiichiro.sumita@atr.co.jp

Abstract

This paper reports the latest performance of components and features of a project named *Corpus-Centered Computation* (C^3), which targets a translation technology suitable for spoken language translation. C^3 places *corpora* at the center of the technology. Translation knowledge is *extracted from corpora* by both EBMT and SMT methods, translation quality is *gauged by referring to corpora*, the best translation among multiple-engine outputs is *selected based on corpora* and the *corpora themselves are paraphrased or filtered* by automated processes.

1 Introduction

Our project, named *Corpus-Centered Computation* (C^3), proposes solutions for efficiently constructing a high-quality translation subsystem for a speech-to-speech translation system.

This paper introduces recent progress in C^3. Sections 2 and 3 demonstrate a competition between multiple machine translation systems developed in our project, and Sections 4 and 5 explain the features that differentiate our project from other corpus-based projects.

2 Three Corpus-based MT Systems

There are two main strategies in corpus-based machine translation: (i) Example-Based Machine Translation (EBMT; Nagao, 1984; Somers, 1999) and (ii) Statistical Machine Translation (SMT; Brown et al., 1993; Knight, 1997; Ney, 2001; Al-shawi et al., 2000). C^3 is developing both technologies in parallel and blending them. In this paper, we introduce *three* different machine translation systems: D^3, *HPAT*, and *SAT*.

The three MT systems are characterized by different translation units. D^3, *HPAT*, and *SAT* use sentences, phrases, and words, respectively.

***D^3* (Sentence-based EBMT):** It retrieves the most similar example by DP-matching of the input and example *sentences* and adjusts the gap between the input and the retrieved example by using dictionaries. (Sumita 2001)

***HPAT* (Phrase-based EBMT):** Based on *phrase*-aligned bilingual trees, transfer patterns are generated. According to the patterns, the source phrase structure is obtained and converted to generate target sentences. (Imamura 2002)

***SAT* (Word-based SMT):** Watanabe et al. (2002b) implemented *SAT* dealing with Japanese and English on top of a *word*-based SMT framework (Brown et al. 1993).

3 Competition on the Same Corpus

3.1 Resources

In our competitive evaluation of the MT systems, we used the BTEC corpus [1], which is a collection of Japanese sentences and their English translations typically found in phrasebooks for tourists. The size is about 150 thousand sentence pairs. A quality evaluation was done using a test set consisting of 345 sentences selected randomly from the above corpus, and the remaining sentences were used for learning and verification. For each source sentence in the test set, 16 reference translations were prepared by 5 bilingual translators.

[1] BTEC was called BE in the paper (Takezawa et al., 2002).

We used bilingual dictionaries and thesauri of about fifty thousand words for the travel domain.

3.2 Evaluation Measures

We used the *measures* below. The BLEU score and the RED rank are measured by referring to the *test corpus*, i.e., a set of input sentences and their multiple reference translations; the HUMAN rank and the estimated TOEIC score are judged by bilingual translators.

(1) Average of Ranks[2]:

HUMAN rank: In our evaluation, 9 translators who are native speakers of the target language ranked the MT translations into 4 ranks: A, B, C, and D, from *good* to *bad* (Sumita et al., 1999). [3]

RED rank: An automatic ranker is learned as a decision tree from HUMAN-ranked examples. It exploits edit-distances between MT and multiple reference translations (Akiba et al., 2001).

(2) BLEU score: The MT translations are scored based on the precision of N-grams in an entire set of multiple reference translations (Papineni et al., 2002). It ranges from 1.0 (best) down to 0.0 (worst).

(3) Estimated TOEIC score: It is important to interpret MT performance from the viewpoint of a language proficiency test such as TOEIC[4]. A translator compared MT translations with human ones, then, MT's proficiency is estimated by regression analysis (Sugaya et al., 2000). It ranges from 10 (*lowest*) to 990 points (*perfect*).

3.3 Results

Table 1 wraps up the results. So far, SMT has been applied mainly to language pairs of similar European languages. Skeptical opinions dominate about the effectiveness or applicability of SMT to dissimilar language pairs. However, we implemented SMT for translation between Japanese and English. They are dissimilar in many points, such as word order and lexical systems. We found that ***SAT*, which is an SMT, worked in both J-to-E and E-to-J directions.**

The EBMT systems, *HPAT* and *D*[3], surpassed *SAT* in the HUMAN rank. This is the reverse result obtained in a Verbmobil experiment (Ney, 2001) where an SMT system scored highest. We are studying these interesting contradictory observations.

Let's consider the relationships among the HUMAN rank, the RED rank, and the BLEU score. **While RED accords with HUMAN, BLEU fails to agree with HUMAN in the EJ evaluation.** One reason for this is that the BLEU score favors *SAT* translations in that they are more similar to the reference translation from the viewpoint of N-grams.

Table 1 Quality Evaluation of Three MTs[5]

pair	MT	Average of *HUMAN*	Average of *RED*	*BLEU*
JE	D[3]	3.21	3.44	0.49
	SAT	2.66	2.61	0.43
EJ	HPAT	3.17	3.13	*0.48*
	SAT	2.89	2.91	*0.56*

Let's move on to the estimated TOEIC score of the most accurate JE system in the experiment. ***D*[3] achieved a high score of 870.** This is more than one hundred points higher than the average score of a Japanese businessperson in an overseas department of a company.

4 Corpus-based Selector

This section introduces a feature of *C*[3]: *selection* of the best from outputs produced by multiple translation engines.

No single system can achieve complete translation of every input. The quality rank of a given input sentence changes system by system. We show a sample of different English translations

[2] Average is calculated: A, B, C, and D are assigned values of 4, 3, 2, and 1, respectively, and their sum is divided by the sentence count (345 in the experiment).

[3] The final rank for each translation is the *median* of the nine ranks given by independent evaluators.

[4] TOEIC is an acronym for 'Test of English for International Communication', which is an *English* language proficiency test for people whose native language is not English (http://www.chauncey.com/).

[5] *HPAT* for JE and *D*[3] for EJ also work well, but we omitted them from the table because we could not afford the time and cost of the human evaluation for them.

obtained by the three systems for the Japanese sentence, 'o-shiharai wa genkin desu ka kurejitto kaado desu ka' (**Table 2**). The brackets show the HUMAN rank, as described above.

Table 2. Sample of Translation Variety

[B] Is the payment cash? Or is it the credit card?
[A] Would you like to pay in cash or with a credit card?
[C] Could you cash or credit card?

In our experiment, while D^3, *HPAT*, and *SAT* for the E-to-J direction have A-ratios of 0.62, 0.55, and 0.53, respectively, the ideal selection would have an interestingly high A-ratio of 0.79. Thus, we could obtain a large increase in accuracy if it were possible to select the best one of the three different translations for each input sentence.

Unlike other approaches such as (Brown and Frederking, 1995), we do not *merge* multiple results into a single one but we *select* the best one because the large difference between multiple translations for distant language pairs such as Japanese and English makes merging infeasible.

Methods using N-gram statistics of a target language corpus have been proposed before (Brown and Frederking, 1995; Callison-Burch et al., 2001). They are based on the assumptions that (1) the naturalness of the translations is effective for selecting good translations because they are sensitive to the broken target sentences due to errors in translation processes, and (2) the source and target correspondences from the semantic point of view are maintained in a state-of-the-art translation system. However, the second assumption does not necessarily hold. To solve this problem, Akiba et al. (2002) used not only a language model but also a translation model of SMT derived from a corpus, and Sumita et al. (2002) exploited a corpus whose sentences are converted into semantic class sequences. These two selectors outperformed *conventional selectors using the target N-gram* in our experiments.

5 Paraphrasing and Filtering

This section introduces another feature of C^3: *paraphrasing* and *filtering corpora*.

The large variety of possible translations in a corpus causes difficulty in building machine translation on the corpus. For example, the variety makes it harder to estimate the parameters for *SAT*, to find appropriate translation examples for D^3, to extract good transfer patterns for *HPAT*. We propose ways to overcome these problems by *paraphrasing corpora* through automated processes or *filtering corpora* by abandoning inappropriate expressions.

Two methods have been investigated for automatic paraphrasing. (1) Shimohata et al. (2002a) group sentences by the equivalence of the translation and extract rules of paraphrasing by DP-matching. (2) Finch et al. (2002) cluster sentences in a handcrafted paraphrase corpus (Sugaya et al., 2002) to obtain pairs that are similar to each other for training SMT models, then by using the models the decoder generates a paraphrase. The experimental results indicate that (i) the EBMT based on normalization had increased coverage (Shimohata et al., 2002b) and (ii) the SMT created on the normalized sentences had a reduced word-error-rate (Watanabe et al., 2002a).

Imamura et al. (2003) proposed a calculation that measures the literalness of a translation pair and called it TCR. After the word alignment of a translation pair, TCR is calculated as the rate of the aligned word count over the count of words in the translation pair. After abandoning the non-literal parts of the corpus, the acquisition of *HPAT* transfer patterns is done. The effect has been confirmed by an improvement in translation quality.

6 Conclusion

Our project, called C^3, places *corpora* at the *center* of speech-to-speech technology. Good performance in translation components is demonstrated in the experiment. In addition, the corpus-based processes of translation, evaluation, and paraphrasing have synergistic effects. Therefore, we are optimistic about the further progress of components and their integration.

Acknowledgements

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus."

References

Alshawi, H., Bangalore, S. and Douglas, S. 2000. Learning Dependency Translation Models as Collections of Finite-State Head Transducers, Computational Linguistics, 26 (1), pp. 45--60.

Akiba, Y, Imamura, K., and Sumita, E., 2001. Using multiple edit distances to automatically rank machine translation output. In *Proc. of MTSummit VIII*, pages 15–20.

Akiba, Y., Watanabe, T., and Sumita, E. 2002. "Using Language and Translation Models to Select the Best among Outputs from Multiple MT Systems, Proc. of Coling, pp. 8—14.

Brown, P., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Laffetry, J. D., Mercer, R. L., and Roossin, P. S., 1993. A Statistical Approach to Machine Translation, Computational Linguistics 16, pp. 79—85.

Brown, R. and Frederking, R., 1995. Applying Statistical English Language Modeling to Symbolic Machine Translation. In Proc. of the 6th TMI, pp. 221-239.

Callison-Burch, C. and Flournoy, S., 2001. A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines, Proc. of MT-SUMMIT.

Finch, A, Watanabe, T., and Sumita, E., Paraphrasing by Statistical Machine Translation, 2002, Proc. of FIT, E-53, pp.187—188.

Imamura, K., 2002. Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment, Proc. of TMI.

Imamura, K., Sumita, E. and Matsumoto, Y., 2003. Automatic Construction of Machine Translation Knowledge Using Translation Literality, Proc. of EACL.

Knight, K., 1997. Automating Knowledge Acquisition for Machine Translation, AI Magazine, 18 (4), pp. 81--96

Nagao, M., 1984. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, in A. Elithorn and R. Banerji (eds), Artificial and Human Intelligence, Amsterdam: North-Holland, pp. 173--180.

Ney, H., 2001. Stochastic Modeling: From pattern classification to language translation, in Proc. of the ACL 2001 Workshop on DDMT, pp. 33--37.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J., 2002. Bleu: A Method for Automatic Evaluation of Machine Translation, Proc. of the 40th ACL, pp. 311—318.

Shimohata, M. and Sumita, E., 2002a. Automatic paraphrasing based on parallel corpus for normalization, Proc. of LREC.

Shimohata, M. and Sumita, E., 2002b. "Identifying Synonymous Expressions from a Bilingual Corpus for Example-Based Machine Translation," Proc. of the Workshop on Machine Translation in Asia, Coling, pp. 20--25.

Somers, H., 1999. Review Article: Example-based Machine Translation, Journal of Machine Translation, pp. 113—157.

Sugaya, F., Takezawa, T., Yokoo, A., Sagisaka, Y., and Yamamoto, S., 2000. Evaluation of the ATR-MATRIX Speech Translation System with a Pair Comparison Method Between the System and Humans, Proc. of ICSLP, pp. 1105—1108.

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.

Sumita, E., 2001. Example-based machine translation using DP-matching between word sequences, Proc. of ACL 2001 Workshop on DDMT, pp. 1—8.

Sumita, E., Akiba, Y., and Imamura, K., 2002. "Reliability Measures For Translation Quality," ICSLP, pp. 1893-1896.

Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishikawa, K., and Shirai, S., 1999. Solutions to Problems Inherent in Spoken-language Translation: The ATR-MATRIX Approach, Proc. of MT Summit, pp. 229—235.

Takezawa, T. et al., 2002. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World, Proc. of LREC.

Watanabe, T. et al., 2002a. Statistical Machine Translation Based on Paraphrased Corpora, Proc. of LREC.

Watanabe, T. et al., 2002b. "Bidirectional Decoding for Statistical Machine Translation," Proc. of Coling, pp. 1079—1085.