Statistical Machine Translation on Paraphrased Corpora

Taro Watanabe, Mitsuo Shimohata and Eiichiro Sumita

ATR Spoken Language Transaltion Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun,
Kyoto 619-0288 JAPAN
{taro.watanabe, mitsuo.shimohata, eiichiro.sumita}@atr.co.jp

Abstract

This paper presents a statistical machine translation trained on normalized corpora. The automatic paraphrasing is carried out by inducing paraphrasing expressions from a bilingual corpus. Then, the normalization is treated as a specific paraphrase of a given input determined by the frequency in a corpus. The experimental results on Japanese-to-English translation with normalized English corpus exhibited the reduction of word-error-rate by 8% and the improvement of subjective evaluation from 70% into 72.5%.

1. Introduction

Recent success on statistical approach to machine translation demands huge bilingual corpora in good quality and broad coverage. However, such an ideal corpora is not usually available: one may contain sufficiently large number of samples, for instance, derived from web pages with translations, but not well-aligned or translation quality is low. Others may consists of translations in good quality, though the number of examples might be limited. In addition, the variety of possible translations further makes it harder to estimate parameters for statistical-based machine translation.

This paper describes a way to overcome the problems by creating a corpus that consists of normalized expressions, expressions with less variety, through automated paraphrasing. The paraphraser induces synonymous expressions from bilingual corpora, by observing the difference of a set of utterances that holds the same meaning in another language. By transforming translation target sentences of a given bilingual corpus into normalized form, it is expected the improvement of parameter estimation for statistical machine translation model.

The experimental results on Japanese-to-English translation indicated that the statistical translation model created on the target normalized sentences yield word-error-rate of 58%, while that of the non-normalized one was 66%. In addition, the subjective evaluation score was improved from 70% to 72.5%.

2. Statistical Machine Translation

Statistical machine translation regards machine translation as a process of translating a source language text ($\mathbf{f}$) into a target language text ($\mathbf{e}$) with the following formula:

$$\mathbf{e} = \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f})$$

The Bayes Rule is applied to the above to derive:

$$\mathbf{e} = \arg\max_{\mathbf{e}} P(\mathbf{f}|\mathbf{e})P(\mathbf{e})$$

The translation process is treated as a noisy channel model, like those used in speech recognition in which there exists $\mathbf{e}$ transcribed as $\mathbf{f}$, and a translation is to infer the best $\mathbf{e}$ from $\mathbf{f}$ in terms of $P(\mathbf{f}|\mathbf{e})P(\mathbf{e})$. The former term, $P(\mathbf{f}|\mathbf{e})$, is a translation model representing some correspondence between bilingual text. The latter, $P(\mathbf{e})$, is the language model denoting the likelihood of the channel source text. In addition, a word correspondence model, called alignment $\mathbf{a}$, is introduced to the translation model to represent a positional correspondence of the channel target and source words:

$$\mathbf{e} = \arg\max_{\mathbf{e}} \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e})P(\mathbf{e})$$

An example of an alignment is shown in Figure 1, where the English sentence "could you recommend another hotel" is mapped onto the Japanese "hoka no hoteru o shokaishi teitadake masu ka", and both "hoka" and "no" are aligned to "another", etc. The NULL symbol at index 0 is also a lexical entry in which no morpheme is aligned from the channel target morpheme, such as "o" and "ka" in this Japanese example.

2.1. IBM Model 4

The Translation consists of 4 models according to the IBM Model 4 (see Figure 2):

- Lexical Model — $t(f|e)$: Word-for-word translation model, representing the probability of a source word f being translated into a target word e.
- Fertility Model — $n(\phi|e)$: Representing the probability of a source word e generating ϕ words.
- Distortion Model — d : The probability of distortion. In Model 4, the model is decomposed into two sets of parameters:
  - $d_1(j - c_{\rho i}|\mathcal{A}(e_i), \mathcal{B}(f_j))$: Distortion probability for head words. The head word is the first of the target words generated from a source word a cept, that is the channel source word with fertility more than and equal to one. The head word position j is determined by the word classes of the previous source word, $\mathcal{A}(e_i)$, and target word, $\mathcal{B}(f_j)$, relative to the centroid of the previous source word, c_{ρ_i}.
  - $d_{>1}(j - j'|\mathcal{B}(f_j))$: Distortion probability for non-head words. The position of a non-head word j is determined by the word class and relative to the previous target word generated from the cept (j').

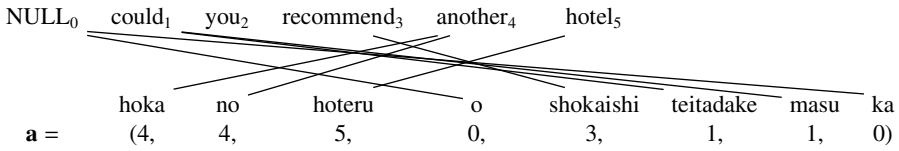

Figure 1: An example of alignment for Japanese and English sentences

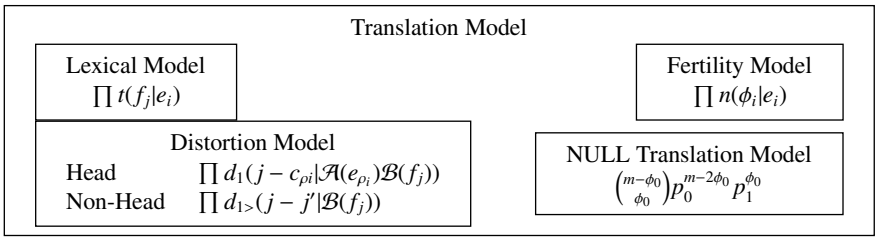

Figure 2: Conponents of translation model (IBM Model 4)

- NULL Translation Model — p_1 : A fixed probability of inserting a NULL word after determining each target word f.

For details, refer to Brown et al. (1993).

2.2. Problems of Statistical Machine Translation

The parameter estimation of the translation model relies on the EM-algorithm by accumulating counts on all the possible alignments given a pair of two language sentences **e** and **f**. Therefore, if there exists sufficiently larger number of samples with variety of translation, it is expected that the estimation algorithm would converge to reasonable set of parameters. However, an actual corpus usually consists of variety of translations without counts not sufficient for the above algorithm. This paper suggests normalization of corpora for translation target sentences through automatic paraphrasing, which can suppress the diversity of translations.

3. Normalization by Paraphrasing

The normalization of corpora is treated as a specific paraphrasing of a given input by transforming into the most frequently occurring expressions. This paper propose an automatic paraphrasing/normalization by exploiting knowledge from bilingual corpora (Shimohata and Sumita, 2002).

3.1. Extraction of Synonymous Expressions

Synonymous expressions are defined to be consisting of a sequence of variant words with surrounding common words. The expressions are extracted from bilingual corpora by the following procedures (refer to Figure 4):

1. Collect sentences that share the same translation in another language. The accumulated sentences are defined as *synonymous sentences*.

2. For all pairs of *synonymous sentences*, apply DP-matching and collect sequences of words,*synonymous expressions*, that consists of variant words preceded/followed by common words.

3. Remove pairs of *synonymous expressions* with the frequency lower than a given threshold.

E1	$< s >$	**Could**	**you**
	$< s >$	Would	you
	$< s >$	Can	you
	$< s >$	Will	you
E2	$< s >$	**Nice**	**to**
	$< s >$	Glad	to
	$< s >$	Pleased	to
	$< s >$	Happy	to
E3	**a**	**guarantee**	$< /s >$
	a	warranty	$< /s >$
E4	**the**	**toilet**	$< /s >$
	the	bathroom	$< /s >$
	the	lavatory	$< /s >$
	the	restrooms	$< /s >$
E5	**what**	**'s**	**wrong**
	what	is	wrong
E6	**I**	**'m**	**a**
	I	am	a
E7	**a**	**bad**	**cough**
	a	terrible	cough

Figure 3: Examples of cluster of expressions extracted though automated paraphrase. The expressions in bold faces are those with the highest frequency among each cluster.

4. Cluster the pairs of *synonymous expressions* by transitive relation.

Examples of synonymous expressions extracted by the above procedure is presented in Figure 3.

3.2. Normalization

After the acquisition of clusters of synonymous expressions, normalization is carried out by transforming the expressions into major ones, selected according to the frequency in corpora. For instance, the cluster obtained in Figure 4 consists of expressions "$< s >$ would you", "$< s >$ how do you" and "$< s >$ do you." Suppose an expression "$< s >$

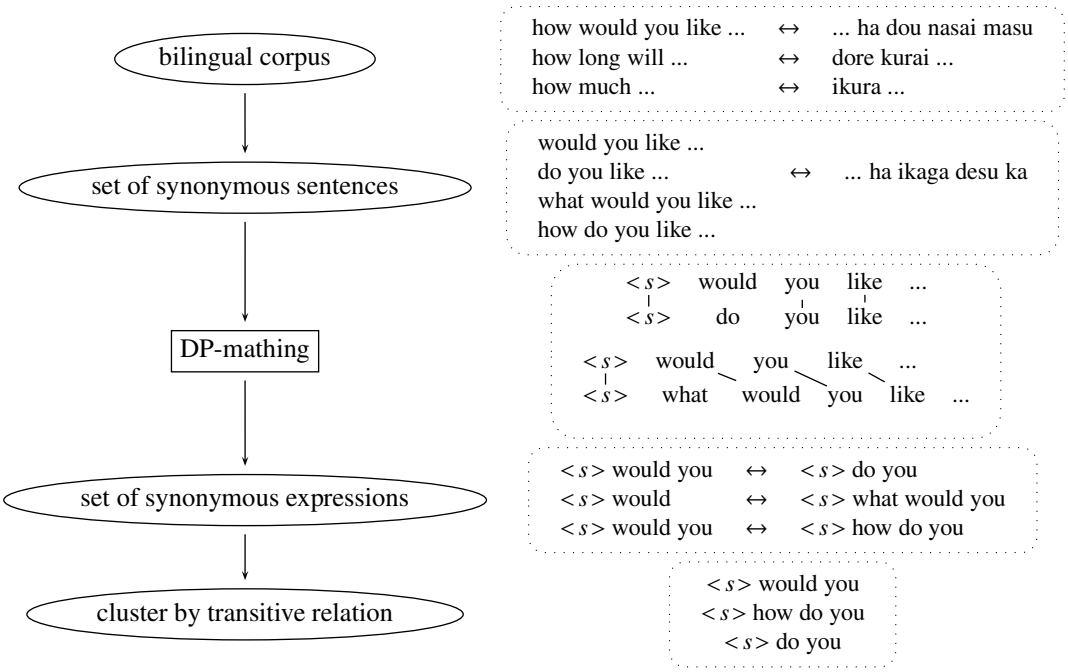

Figure 4: Extraction of synonymous expressions given a set of synonymous sentences

do you" occurred most frequently in a given corpus, an input "how do you like your coffee" could be normalized into "do you like a coffee" (refer to Figure 3 for sample major expressions).

The normalization/paraphrasing method proposed here locally replace word sequences, hence will not affect the syntactical coherence. Therefore, the normalization will not affect the distortion model, which accounts for reordering of bilingual texts. In addition, reduction of the vocabulary size will greatly help improve the parameter estimation for lexical model.

3.3. Related Works

Barzilay and McKeown (2001) and Jacquemin et al. (1997) proposed automatic paraphrasing method, though the types of acquired paraphrases are limited to technical terms and adjective-noun phrases, respectively. Lapata (2001) suggested various types of paraphrases based on variant context containing common words. Our method enforce different criteria for expressions, that consists of variant words surrounded by common words. In addition,it does not rely on additional morphological information, such as POS etc., therefore, is applicable to any raw corpora for many languages.

4. Experiments

The experiments was carried out for Japanese-to-English translation on a large-scale travel conversation corpus that consists of nearly 200,000 Japanese and English sentence pairs. The 152,181 utterances were extracted for the training set and 10,129 were used for cross-validation both for translation models and language models. Table 1 summarize the statistics on the corpus.

The normalization procedure was carried out for English sentences utilizing Japanese as a pivot to collect a set of synonymous sentences[1]. The result of normalization is summarized in Table 2. The trigram perplexity were evaluated for the language model of original corpus, tested both on original test set and normalized one. The language model of the normalized corpus was tested on normalized test set.

Table 1: Statistics of a travel conversation corpus

	English	Japanese
# of sentences	152,181	
# of words	835,048	896,302
vocabulary size	13,162	20,348
average sentence length	5.74	6.16

Table 2: Statistics of normalized corpus for English

# of expressions	988
# of clusters	439
perplexity (original)	35.37
perplexity (original on normalized)	36.52
perplexity (normalized)	32.00

The statistical translation model was created both for the original corpus and English-normalized corpus, bootstrapping from IBM Model 1 to 4 with intermediate HMM Model. The translation experiments were carried out on 240 Japanese sentences with a decoder based on Tillmann and Ney (2000).

The translation results were evaluated by word-error-rate (WER), that penalize insertion/deletion/replacement by one. The position independent word-error-rate (PER) was also introduced for the evaluation, ignoring the positional disfluency. In addition, the translations were scored by subjective evaluation (SE) with 4-point ranking ranging

[1]frequency threshold was set to 2.

Table 3: Experimental results on Japanese-to-English translation

Model	WER	PER	SE			
			A	B	C	D
original	65.9	58.3	29.2	23.8	17.1	30.0
normalized	58.0	52.6	27.9	28.8	15.8	27.5

WER: word-error-rate
PER: position independent word-error-rate
SE: subjective evaluation (A: perfect, B: fair, C: acceptable, D: nonsense)
original: translation model on original bilingual corpus
normalized: translation model on translation target normalized bilingual corpora

Table 4: Experimental results on Japanese-to-English translation with various input lengths

Model	WER			PER			SE(A+B+C)		
length	6	8	10	6	8	10	6	8	10
original	65.0	64.6	68.1	58.0	58.3	58.7	71.3	68.8	70.0
normalized	56.9	60.1	57.0	52.2	54.8	50.8	77.5	71.3	68.8

from A to D (Sumita et al., 1999) [2].

Table 3 summarize the results and Table 4 detailed by input length, assuming that the SE scores from A to C are good translation.

5. Discussion

From Table 2, the perplexity of language model for the original corpus was slightly increased when tested on the normalized corpus. This indicates that the slight disfluency was inserted in the normalization process, although the language model perplexity for the normalized corpus was decreased. This is due to the lack of syntactic knowledge during the acquisition of paraphrasing expressions.

The translation results from Table 3 exhibited the reduction of WER from 65.9% to 58.0% together with the drop of PER from 58.3% to 52.6%. Although the percentage of A-ranked sentences was slightly degraded (29.2% to 27.9%), the boost of B-ranked sentences raise the ratio of good-quality translation (A+B+C) from 70% into 72.5%. The quality reduction of A-ranked ratio is probably due to the syntactical disfluency as explained above, though still normalization help improve the quality of translation.

The detailed analysis by differentiating input length (refer to Table 4) showed that the improved WER/PER on all the input lengths, but the increase of SE(A+B+C) was observed only for shorter length (length of 6 and 8), but not for longer sentences (length of 10). The proposed method only accounts for local expression paraphrase without observing longer contextual information, therefore the translation quality was degraded for longer sentences. This is the nature of the statistical machine translation model with bilingual correspondence represented by alignment, in which longer positional distortion is harder to be captured.

6. Acknowledgment

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".

[2] the meanings of the symbol are follows: A — perfect: no problem in either information or grammar; B — fair: easy to understand but some important information is missing or it is grammatically flawed; C — acceptable: broken but understandable with effort; D — nonsense: important information has been translated incorrectly.

7. References

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proc. of ACL/EACL*, Toulouse, France.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Christian Jacquemin, Judith L. Klavans, and Evelyne Tzoukermann. 1997. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 24–31, Somerset, New Jersey. Association for Computational Linguistics.

Maria Lapata. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proceedings of the 2nd Meeting of the NAACL*, Pittsburgh, PA.

Mitsuo Shimohata and Eiichiro Sumita. 2002. Automatic paraphrasing based on parallel corpus for normalization. In *Proceedings of the Third International Conference on Language Resources and Evaluation (to appear)*, Las Palmas, Canary Islands, Spain, May.

Eiichiro Sumita, Setsuo Yamada, Kazuhide Yamamoto, Michael Paul, Hideki Kashioka, Kai Ishikawa, and Satoshi Shirai. 1999. Solutions to problems inherent in spoken-language translation: The ATR-MATRIX approach. In *Machine Translation Summit VII*, pages 229–235.

Christoph Tillmann and Hermann Ney. 2000. Word reordering and dp-based search in statistical machine translation. In *Proceedings of the COLING 2000 : The 18th International Conference on Computational Linguistics*, July-August.