Example-based Bi-directional Chinese-English Machine Translation with Semi-automatically Induced Grammars

K. C. Siu, Helen M. Meng and C. C. Wong

Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Hong Kong SAR, China
{kcsiu, hmmeng}@se.cuhk.edu.hk

Abstract

We have previously developed a framework for bi-directional English-to-Chinese/Chinese-to-English machine translation using semi-automatically induced grammars from unannotated corpora. The framework adopts an example-based machine translation (EBMT) approach. This work reports on three extensions to the framework. First, we investigate the comparative merits of three distance metrics (Kullback-Leibler, Manhattan-Norm and Gini Index) for agglomerative clustering in grammar induction. Second, we seek an automatic evaluation method that can also consider multiple translation outputs generated for a single input sentence based on the BLEU metric. Third, our previous investigation shows that Chinese-to-English translation has lower performance due to incorrect use of English inflectional forms – a consequence of random selection among translation alternatives. We present an improved selection strategy that leverages information from the example parse trees in our EBMT paradigm.

1. Introduction

This work extends our previous effort in the use of semi-automatically induced grammars for bi-directional English-Chinese machine translation using an example-based approach [1,2]. Our parallel experimental corpora includes the English ATIS-3 Class A sentences (training set, test set 1993 and 1994) with their Chinese translations. Our grammar induction framework involves an *agglomerative clustering* procedure that can generate context-free grammar rules from unannotated English or Chinese sentences. The grammar rules are amenable to manual refinement, hence our approach is semi-automatic in nature. The advantages of such an approach include: significant reduction of manual effort in handcrafting grammar rules, generation of a grammar that can closely model real data and achieving enhanced portability across domains and languages. The agglomerative clustering procedure is implemented both *temporally* and *spatially*. In temporal clustering, words or multi-word entities that co-occur sequentially are clustered together based on the Mutual Information metric or the Information Gain metric [3]. In spatial clustering, words or multi-word entities with similar left and right linguistic contexts are clustered together based on the symmetrized divergence (*Div*) that is applied to the left and right linguistic contexts of the entity pair. *Div* incorporates the Kullback-Leibler *(KL)* distance metric [1,2] (See Equation 1).[1] In general, temporal clustering generates phrasal categories in the grammar and spatial clustering generates semantic categories.

$$\begin{aligned} & Dist_{KL}(e_1,e_2) = Div(p_1^{left},p_2^{left}) + Div(p_1^{right},p_2^{right}) \quad \textit{where} \\ & Div(p_1^{adj},p_2^{adj}) = \sum_{i=1}^{V} p_1^{adj}(i)\log\frac{p_1^{adj}(i)}{p_2^{adj}(i)} + \sum_{i=1}^{V} p_2^{adj}(i)\log\frac{p_2^{adj}(i)}{p_1^{adj}(i)} \end{aligned} \tag{1}$$

The clustering procedure produces parallel context-free grammars from parallel training corpora. The grammars are refined by hand-editing and then used in conjunction with example-based bi-directional machine translation (EBMT). This is illustrated in Figure 1. The EBMT module accepts as input the parse structure based on the source language, finds the closest-matching parse structure (and its corresponding sentence) from the training examples in the source language, identifies the parallel sentence in the target language and outputs the parse structure of this parallel sentence. The output parse structure is then used together with the grammar in the target language to generate one or more translations [1]. EBMT has the advantage of being rapidly retargetable to other language pairs, and the use of semi-automatically induced grammars reinforces this advantage.

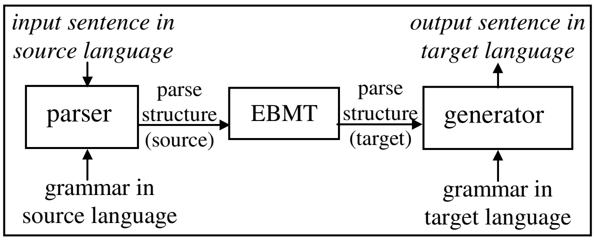
input sentence in source language
output sentence in target language
parser
parse structure (source)
EBMT
parse structure (target)
generator
grammar in source language
grammar in target language

Figure 1: Overview of our bi-directional machine translation framework that uses semi-automatically induced parallel grammars.

Investigation based on this EBMT framework led to the following research issues that are addressed in this paper: First, the KL distance is very sensitive to sparse data problems. The aggregate distance is often dominated by infrequent words (or clusters) with very low probabilities [4]. In order to alleviate the influence of infrequent data on the aggregate distance, we propose two alternative distance metrics – the Manhattan-Norm and the Gini Index. Second, our previous work has relied on manual evaluation of the top-scoring translation output from our EBMT approach. However, since manual evaluation is laborious and our approach is capable of generating multiple translation candidates, we seek an automatic evaluation criterion that can account for multiple translation outputs. Third, Chinese-to-English translation has a lower performance than English-to-Chinese due to the use of errorful inflectional forms. This is because inflections are absent in Chinese (as the source language), but the appropriate

[1] In Equation 1, e_1 and e_2 are the entities under consideration, V is the vocabulary size for the left / right context, $p_1(i)$ is the probability of the entity i adjacent (adj) to entity e_1.

form needs to be selected for English (as the target language). When multiple inflectional forms are offered by a grammar rule during generation/translation, our original approach simply selects one at random [1]. Our current work attempts to incorporate an enhancement that leverages off parse structures in the example template to make a better selection. Details related to these three research issues are presented in the following.

2. Distance Metrics for Grammar Induction

The Manhattan-Norm (*MN*) distance introduced in [4,5] takes the absolute value of the difference of the two distributions p_1 and p_2 (see Equation 2).

$$MN(p_1, p_2) = \sum_{i=1}^{V} | \, p_1(i) - p_2(i) \, | \qquad (2)$$

The Gini Index (*GI*) [6] takes the square of the difference of the two distributions p_1 and p_2 (see Equation 3). Equation 4 shows how a given metric (*MN* or *GI*) can be applied to both the left and right contexts of the entities e_1 and e_2 when calculating the distance (*Dist*) for agglomerative clustering.

$$GI(p_1, p_2) = \sum_{i=1}^{V} [\, p_1(i) - p_2(i) \,]^2 \qquad (3)$$

$$Dist_{Metric}(e_1, e_2) = Metric(p_1^{left}, p_2^{left}) + Metric(p_1^{right}, p_2^{right}) \quad (4)$$

In order to compare the various metrics in grammar induction, we utilize the ATIS-3 SQL queries for comparative evaluation. The SQL expression corresponding to each ATIS utterance specifies the necessary database access action and contains the meaningful natural language structures (e.g. attribute-value pairs in the semantic frame) that should be captured in the grammar. As an illustration, examples of spatial clusters derived by our grammar are presented in Table 1 and examples of the attribute labels and values extracted from an SQL expression in Table 2.

Spatial Cluster → *Terminals*	**(Cluster Label)**
SC₀ → *milwaukee | nashville | detroit | tampa*	(CITY_NAME)
SC₇ → *baltimore | chicago | charlotte*	(CITY_NAME)
SC₁₀ → *monday | wednesday | saturday*	(DAY_NAME)
SC₁₂ → *evening | morning | afternoon*	(PERIOD)
SC₁₈ → *american | united*	(AIRLINE_NAME)

Table 1: Examples of spatial clusters produced by the grammar induction algorithm. Cluster labels are assigned by hand.

ATTRIBUTE LABEL:	***Value***
ORIGIN:	*charlotte*
DESTINATION:	*minneapolis*
DEPARTURE_TIME:	*14:30 – 15:30*

Table 2: Examples of attribute-value pairs extracted from an SQL expression in the ATIS domain (c.f. values with grammar terminals in Table 1).

At a given iteration in the grammar induction process, we can evaluate the interim grammar by using it to parse all training queries. For each training query, we examine the (meaningful) structures extracted from its SQL expression and compare them with the structures in the query's parse. Hence we can measure the overall *precision* (*P*) and *recall* (*R*) rates of meaningful structures in the SQL. This is essentially the PARSEVAL framework. Similarly, we can evaluate the final grammar based on the test queries. Details are provided in [3].

We ran our grammar induction procedure with the distance metrics *KL, MN* and *GI* to generate three grammars respectively – G_{KL}, G_{MN} and G_{GI}.[2] We compared these three grammars at every tenth iteration until the stopping criterion is met.[3] When we evaluated with the ATIS training set and rank the grammars in decreasing order of *precision (P)*, we observed ($G_{MN} > G_{GI} > G_{KL}$) across the various iterations. When we ranked with decreasing *recall (R)*, we observed ($G_{GI} > G_{MN} > G_{KL}$). We also evaluated with the ATIS test sets and compared the resulting grammars (without hand refinement) against a handcrafted grammar G_H. Results are in Table 3.

	G_{KL}	G_{MN}	G_{GI}	G_H
Precision 93	0.308	0.341	0.323	0.634
Recall 93	0.757	0.786	0.803	0.889
Precision 94	0.313	0.346	0.325	0.626
Recall 94	0.736	0.766	0.779	0.909

Table 3: Comparison across G_{KL}, G_{MN}, G_{GI}, and G_H based on *P, R* and *F*-measure in extracting attributes and semantic values derived from SQL expressions. Results from both the ATIS test set 1993 (shaded) and 1994 (no shading) are shown.

We also compared the grammars in terms of natural language (NLU) understanding performance on the ATIS-3 test sets. Again, no manual effort is involved in developing the grammars except for G_H. In Table 4, FULL refers to the percentage of test sentences that have a full match between the attribute-value pairs derived from the SQL (illustrated in Table 2) and those derived from parsing with a grammar. PARTIAL refers to a partial match and NO refers to no match.

Based on all the above comparisons, we observe that both *MN* and *GI* outperform *KL* in efficacy for grammar induction.

NLU	G_{KL} (%)	G_{MN} (%)	G_{GI} (%)	G_H (%)
	1993 Test Set			
FULL	7.6	49.3	50.9	85.5
PARTIAL	52.0	34.4	31.9	14.5
NO	40.4	16.3	17.2	0.0
	1994 Test Set			
FULL	9.7	50.5	51.3	78.6
PARTIAL	63.7	41.7	41.7	20.2
NO	26.6	7.9	7.2	1.1

Table 4: Comparison across G_{KL}, G_{MN}, G_{GI}, and G_H based on natural language understanding (NLU) performance. Results from both the ATIS test set 1993 (shaded) and test set 1994 (no shading) are shown.

3. Automatic Machine Translation Evaluation

Machine translation systems have mostly been evaluated by human judges based on such criteria as completeness and fluency [7]. However, human evaluation is labor-intensive and tends to be a lengthy process with subjectivity. In comparison, automatic evaluation is inexpensive, fast and hence more desirable. The BLEU (Bilingual Evaluation Understudy) metric was recently proposed by IBM [8] for automatic evaluation of machine translation. BLEU compares variable length phrases of the translated result against *multiple*

[2] All other experimental parameters are controlled. No. of merges / iteration =5 for both spatial and temporal clustering.

[3] The automatic stopping criterion is defined to be the point when the relative growth in training vocabulary coverage per iteration falls below 1%.

reference translations. The use of multiple reference translations allows for the differences in word choices and word orders. BLEU involves the computation of a *modified n-gram precision score* p_n (see Equation 5) and the *sentence brevity penalty BP* (see Equation 6). p_n is computed as the candidate counts clipped by their corresponding reference maximum value, summed, and divided by the total number of candidate *n*-grams. *BP* is a multiplicative penalty computed over the entire corpus that enforces comparable lengths between candidate and reference, and penalizes short candidates whose precision values may be deceptively high.

$$p_n = \frac{\sum_{C\in\{candidates\}} \sum_{n-gram\in C} Count_{clip}(n-gram)}{\sum_{C\in\{candidates\}} \sum_{n-gram\in C} Count(n-gram)} \tag{5}$$

$$BP = \begin{cases} 1 & if \quad c > r \\ e^{(1-r/c)} & if \quad c \le r \end{cases} \tag{6}$$

In Equation 6, *c* is the length of the candidate translation corpus and *r* is the sum of the best-matching reference lengths for each candidate sentence in the corpus. The BLEU score for an entire test corpus is computed with Equation 7 [8]:

$$BLEU = BP \times \exp(\sum_{n=1}^{N} w_n \log p_n) \tag{7}$$

where *N* is the length of the *n*-grams and w_n is uniform weight 1/*N*. Hence BLEU computes the geometric mean across several modified *n*-gram precisions. p_n decays roughly exponentially with *n* [8]. A geometric mean (c.f. arithmetic mean) maintains sensitivity to longer *n*-grams.

3.1. The Adapted BLEU Score

BLEU is mainly designed for evaluating a single candidate translation against multiple references. Our current work involves a single reference and multiple candidate translations. This is because our experimental corpus contains only a single reference translation for each sentence in the source language. However, our EBMT framework can generate multiple translation outputs depending on the number of examples available. Hence the number of translation outputs also varies from one sentence to another. We propose an *adapted* BLEU score that involves the computation of a *modified n-gram recall score* r_n (see Equation 8) and the *sentence length penalty LP* (to be described in Equation 9).

$$r_n = \frac{\sum_{C\in\{references\}} \sum_{n-gram\in C} Count_{clip}(n-gram)}{\sum_{C\in\{references\}} \sum_{n-gram\in C} Count(n-gram)} \tag{8}$$

In Equation 8, all candidate *n*-gram counts and their corresponding maximum reference counts are collected. The candidate counts are clipped by their corresponding reference maximum value, summed, and divided by the total number of reference *n*-grams. Table 4 shows an example for three candidate translations with one reference translation. If we consider unigram recall, there are 9 words in the reference sentence and all of them are matched in the three candidate sentences. Hence, the unigram recall is 9/9 (=1). If we consider bigram recall, there are 8 bigrams in the reference sentence of which only "*find flight*" is not matched in any of the three candidate sentences. Hence, the bigram recall is 7/8.

It is conceivable that longer candidate translations and a larger number of candidate translations for a given sentence will lead to higher values of r_n. Hence we introduce the sentence length penalty (*LP*) to balance this effect (see Equation 9).

C1: "*what flight on wednesday from saint louis to houston*"
C2: "*i'd like to have the flight from saint louis to houston*"
C3: "*find the flight from saint louis to houston on wednesday*"
Ref: "*find flight from saint louis to houston on wednesday*"

Table 4: An example for the three-best translation outputs with the reference sentence.

$$LP = \begin{cases} 1 & if \quad c \le r' \\ e^{(1-c/r')} & if \quad c > r' \end{cases}, \quad r' = \sum_{i=1}^{C} n_i r_i \tag{9}$$

where *C* is the number of test sentences, *i* is the index of the test sentence that has a reference translation of length r_i and n_i (n_i-best) available candidate translations, *c* is the sum of the lengths of these candidate translations, and *r'* is an adjusted length of the reference translation based on the number of available candidates.

LP penalizes cases where the total length of candidate translations exceeds the adjusted total length of the reference translations. We see that *LP* is analogous to *BP* except that the former adjusts for length based on recall and the latter adjusts for brevity based on precision. *LP* can also be extended with a multiplicative factor $e^{(1/n'-1)}$ (that ranges between e^{-1} and e^{0}) to adjust for the use of more translation alternatives (see Equation 10):

$$LP_{extended} = \begin{cases} e^{1/n'-1} & if \quad c \le r' \\ e^{(1-c/r')+(1/n'-1)} & if \quad c > r' \end{cases}, \quad n' = \sum_{i=1}^{C} n_i / C \tag{10}$$

The adapted BLEU score (*ABLEU*) for an entire test corpus is formulated in Equation 11:

$$ABLEU = LP \times \exp(\sum_{n=1}^{N} w_n \log r_n) \tag{11}$$

where *N* is the length of the *n*-grams and w_n is uniform weight 1/*N*. If we consider up to the 4-grams, N = 4 and w_n = 1/4. ABLEU is used to evaluate our Chinese-to-English machine translation outputs, as will be described in the next section.

4. Inflectional Forms in Chinese-to-English Machine Translation

As we mentioned in [1], Chinese-to-English translation has lower performance than English-to-Chinese even though our bi-directional translation system is trained on parallel corpora. This is because in Chinese-to-English translation, the source language has no inflectional forms but the target does, hence a source word may be mapped to multiple inflectional forms of the same target word (e.g. 航機 → *flight* | *flights*) and further specification will be necessary to select the appropriate inflection. However, our initial system prototype was implemented with semi-automatically induced context-free grammar rules and the generator does not propagate features related to inflectional forms. When the generator encounters the case of a one-to-many cross-lingual mapping, it simply picks one of the translations at random [1]. Hence Chinese-to-English translation outputs often have errors in inflectional forms, e.g. "*what's the cheapest flights from cleveland to miami on american airlines depart on may seventh*".

In order to address this problem, we leverage off the best-matching example parse structure in our EBMT approach. Figure 2 shows a pair of parallel English and Chinese queries from the training set, aligned with parsed English-Chinese concept-value pairs. When we translate the test sentence, "邊班係美國航空公司五月七號由克里夫蘭去邁阿密最平個班

機”, the Chinese sentence in Figure 2 scored highest [1] as an example parse. Grammar rules that are relevant to this translation present multiple alternatives that are shown in Table 5. For example, the Chinese word sequence “邊班係” has four possible English translations (*what's* | *what're* | *what is* | *what are*). Under this condition, our translation procedure follows the example English template from Figure 2 and selects the option *what's*. Similarly, “班機” is translated as *flight* between the options (*flight* | *flights*). When our procedure translates “五月七號”, there are two possible options (*on the seventh of may* | *on may seventh*). However, the reference value (*on the eighteenth of may*) does not match any of the options at the word level, hence our procedure performs a random selection. The final translation output is “*what's the cheapest flight from cleveland to miami on american airlines on may seventh*”. We compared the performance of our system with and without this enhancement in terms of the ABLEU score, which is set up to geometrically average among *n*-grams (where *n* varies from 1 to 4), and using up to *N*-best translation outputs (where *N* varies from 1 to 5).[4] Results are shown in Table 6. As a point of reference from [1], an ABLEU score of 0.2432 (Table 6, Baseline 93) and 0.2960 (Baseline 94) corresponds to over 70% user-accepted translations in the test corpus.

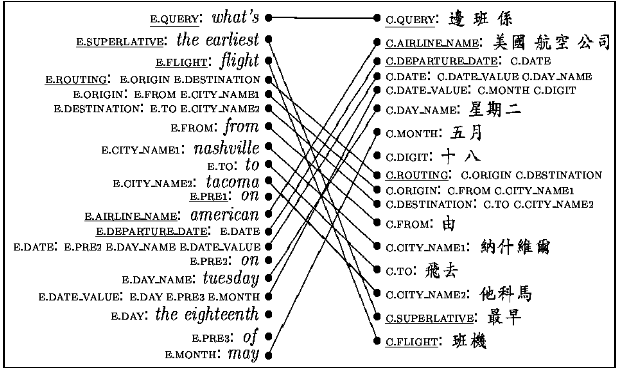

Figure 2: Example template with aligned English-Chinese concept-value pairs. Top level concepts are underlined.

5. Conclusions

In this paper, we have presented two distance metrics, Manhattan-Norm and Gini Index as alternatives over Kullback-Leibler to improve the quality of grammars produced by our semi-automatic grammar induction approach. We observe improvements in precision and recall of meaningful linguistic structures extracted by the grammars. There is also improvement in natural language understanding performance. We also propose an automatic machine translation evaluation metric (ABLEU) adapted from BLEU to cater for the multiple-candidates-single-reference setup in our problem. Finally, we introduced an enhancement to Chinese-to-English translation in an attempt to select appropriate inflectional forms for translation. This technique leverages off lexical choices of the best-matching example translation from the training corpora. Evaluation results based on ABLEU reflect the benefits of the enhancement.

Input Sequence	Translation Alternatives	Reference from EBMT	Final Output
邊班係	*what's* | *what're* | *what is* | *what are*	*what's*	*what's*
最平	*the cheapest*	*the earliest*	*the cheapest*
班機	*flight* | *flights*	*flight*	*flight*
由克里夫蘭去邁阿密	*from cleveland to miami*	*from nashville to tacoma*	*from cleveland to miami*
美國航空公司	*american airlines*	*american airlines*	*american airlines*
五月七號	*on the seventh of may* | *on may seventh*	*on the eighteenth of may*	*on may seventh* (random selection)

Table 5: Chinese-to-English translation of word sequences in a sentence. Column 2 shows translation alternatives derived from the grammar, column 3 shows the reference translation alternative extracted from the best-matching parse structure and column 4 shows the final output. Random selection is used if no match is found between columns 2 and 3.

ABLEU scores using up to N-best translation outputs (with LP)					
	N=1	***N=2***	***N=3***	***N=4***	***N=5***
Baseline 93	0.2432	0.2925	0.3001	0.3143	0.3327
Baseline 94	0.2960	0.3022	0.3185	0.3276	0.3359
Enhanced 93	0.2818	0.3353	0.3455	0.3512	0.3547
Enhanced 94	0.3056	0.3374	0.3476	0.3552	0.3587
ABLEU scores (with $LP_{extended}$)					
Baseline 93	0.2432	0.1475	0.1249	0.1149	0.1093
Baseline 94	0.2960	0.1795	0.1520	0.1398	0.1330
Enhanced 93	0.2818	0.2033	0.1774	0.1659	0.1594
Enhanced 94	0.3056	0.2046	0.1785	0.1678	0.1612

Table 6: ABLEU scores on the Chinese-to-English translation for the ATIS-3 1993 and 1994 test sets. “Baseline” selects randomly among translation alternatives (including inflectional forms) and “Enhanced” selects with reference to the closest-matching example parse tree.

6. Acknowledgments

This research is supported by the Hong Kong SAR Government Research Grant Council Earmarked Grant No. CUHK/4177/00E. We thank Professor Fred Jelinek for suggesting our investigation into the use of the Gini index.

7. References

[1] Siu, K. C. and Meng, H. M., "Semi-Automatic Grammar Induction for Bi-Directional English-Chinese Machine Translation", *Proc. of Eurospeech 2001.*

[2] Siu, K. C. and Meng, H. M., "Semi-Automatic Acquistion of Domain-Specific Semantic Structures", *Proc. of Eurospeech 1999.*

[3] Wong, C. C. and Meng, H. M., “Improvements on a Semi-Automatic Grammar Induction Framework,” *Proc. of ASRU-2001.*

[4] Pargellis, A., Fosler-Lussier, E., Potamianos, A., and Lee, C., "A Comparison of Four Metrics for Auto-Inducing Semantic Classes", *Proc. of ASRU, 2001.*

[5] Dagan, I., Lee, L., and Pereira, F., “Similarity-Based Methods for Word-Sense Disambiguation”, *Proc. ACL, 1997.*

[6] Jelinek, F., “Statistical methods for speech recognition”, *MIT Press, Cambridge, Massachusetts, 1997.*

[7] Gates, D. and et al, “End-to-End Evaluation in JANUS: A Speech-to-Speech Translation System”, *Proc. of ECAI, 1996.*

[8] Papineni, K., Roukos, S., Ward, T., Zhu, W. J., “BLEU: A Method for Automatic Evaluation of Machine Translation,” *IBM Research Report RC22176 (W0109-022), 2001.*

[4] Please note that the maximum number of translation outputs that can be generated varies from one test sentence to another.