Semi-Automatic Grammar Induction for Bi-directional English-Chinese Machine Translation

K.C. Siu and Helen M. Meng

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

Abstract

We have previously designed a methodology for semi-automatic grammar induction from un-annotated corpora belonging to a restricted domain. The induced grammar contains both semantic and syntactic structures, and experiments with the Air Travel Information Service (ATIS-3) corpus demonstrated the viability of our approach [1] for natural language understanding. This work explores the portability of our grammar induction approach to Chinese, based on a corpus of translated ATIS-3 queries. To assess grammar quality, we developed a framework bi-directional English-Chinese example-based machine translation using the induced grammars. Our translation framework can handle word order differences between the language pair during translation. Translations based on the ATIS-3 test sets showed a high percentage (76% to 91%) of user-accepted translations.

1. Introduction

We have previously designed a methodology for semi-automatic grammar induction from un-annotated corpora belonging to a restricted domain. The grammar contains both semantic and syntactic structures, and is conducive towards natural language understanding. Previous experiments compared the semi-automatically-induced grammar (G_{SA}) with a handcrafted grammar (G_H) based on the English ATIS-3 corpus [1]. It took a month to develop G_H; but only a week to develop G_{SA}, with slight degradations in language understanding. Our approach is semi-automatic because grammar rules are first inferred automatically from corpora, and then hand-edited for refinement only. Hence we can reduce manual handcrafting in grammar development, obtain a closer model of real data, and potentially achieve enhanced portability across domains and languages. Our current work explores the portability of our grammar induction approach to Chinese. We have translated the ATIS-3 corpora to Chinese in order to support this investigation. English and Chinese are of regional importance in Hong Kong, and are projected to become the two predominant languages used by the Internet user population by 2005.[1] They are also very different linguistically (e.g. in word order; and in the presence / absence of inflectional forms[2]), and therefore presents interesting challenges for natural language research.

As described above, we obtained an English grammar (G_{SA}) and a Chinese grammar (G_{CSA}) by running our semi-automatic grammar induction procedure on the English and Chinese corpora separately. Aside from evaluating the grammar quality in terms of language understanding performance; we also used the grammars to develop a bi-directional machine translation (MT) system. The use of grammars in MT is more desirable than dictionary-based, word-by-word translation using a bilingual translation dictionary. The grammars preserve word order and hence may produce higher quality translations. We present a unified, bi-directional translation framework, in which G_{SA} is used to parse an input English query, and the parsed concepts can then be used with G_{CSA} to generate a Chinese translation. Similarly, we can parse with G_{CSA} and generate with G_{SA} for Chinese-to-English translation. The framework is illustrated in Figure 1.

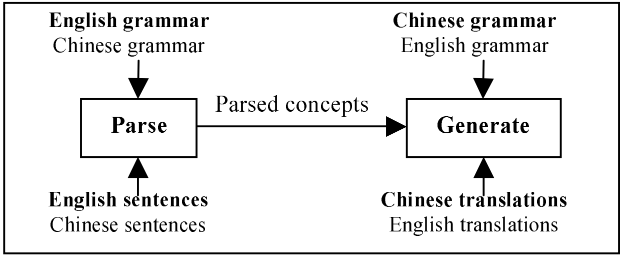

Figure 1. A unified framework for bi-directional English-Chinese machine translation, using semi-automatically induced grammars. The bold-faced words indicate English-to-Chinese translation. The remaining words indicate Chinese-to-English translation.

As will be explained later, ours is a translation-by-analogy (also known as example-based machine translation, or EBMT) approach. EBMT has the advantage of being rapidly retargetable to other language pairs, and the use of semi-automatically induced grammars (instead of handcrafted grammars) reinforces this advantage. Possible applications of this work include translation for on-line information systems, as well as speech-to-speech translation.

Much previous work exists in the area of MT. The Pangloss system [2] applies the EBMT technique to Spanish and English, and recently attempted to adapt the system for Chinese-to-English translation [3]. The Candide system uses a statistical (information-theoretic) approach for French-to-English translation [4]. The Kant system uses a knowledge-based approach (KBMT) which involves an intermediate interlingua for English-to-Spanish and English-to-French translation [5]. Efforts in speech-to-speech translation include: the ATR-MATRIX system which uses the EBMT approach that can translate recognized conversational Japanese speech to English [6]; the JANUS system which can translate conversational speech among several languages (English, German, Spanish, Japanese and Korean) with interlingua for limited domains such as travel planning and appointment scheduling [7]; and the Verbmobil system which can also translate spontaneous dialogs among German, English and Japanese, for limited domains using several different MT approaches [8].

In the following, we present a review of our semi-automatic grammar induction algorithm, describe its portability from

[1] Source: Global Reach.

[2] English is an inflected language but Chinese is not.

English to Chinese ATIS, and present our work in bi-directional English-Chinese machine translation.

2. The Parallel ATIS Corpora

Our task corpus was based on the Air Travel Information Service (ATIS-3) domain [9]. We prepared a parallel ATIS corpus to support our investigation. A large number of subjects were recruited to translate ATIS-3 queries. Translators were asked to read the English query, and then formulate a (Cantonese) Chinese translation freely as long as the meaning is preserved. Cantonese is the key dialect of Chinese used in Hong Kong, Macau, South China and many overseas Chinese communities; and it is very conversational in style. For example, the English query: "*show me one way flights from detroit to westchester county*" is translated as: "話俾我知由底特律飛去西赤斯特城既單程航班". Our training set has 1564 queries, test set 1993 and test set 1994 have 448 and 444 queries respectively.

3. Semi-Automatic Grammar Induction

A detailed description of this procedure is presented in [1]. We provide a brief review in the following for the sake of continuity. Grammar rules are induced by an iterative agglomerative clustering procedure. Each iteration involves spatial clustering to form semantic clusters, and temporal clustering to form phrasal structures. Spatial clustering adopts a distance measure computed from a symmetrized divergence which incorporates the Kullback Liebler distance (see Equation 1).[3]

$$Dist(e_1,e_2) = \sum_{i=1}^{V} p_1^{left}(i)\log\frac{p_1^{left}(i)}{p_2^{left}(i)} + \sum_{i=1}^{V} p_2^{left}(i)\log\frac{p_2^{left}(i)}{p_1^{left}(i)} + \sum_{i=1}^{V} p_1^{right}(i)\log\frac{p_1^{right}(i)}{p_2^{right}(i)} + \sum_{i=1}^{V} p_2^{right}(i)\log\frac{p_2^{right}(i)}{p_1^{right}(i)} \qquad (1)$$

Temporal clustering adopts mutual information as the distance measure (Equation 2), to indicate the degree of co-occurrence of two consecutive entities (e_1 and e_2).

$$MI(e_1,e_2) = P(e_1,e_2)\log\frac{P(e_2 \mid e_1)}{P(e_1)} \qquad (2)$$

Two free parameters are involved in the clustering process: M is the pre-set minimum count threshold in the corpus, below which the entity will not be considered for clustering. N is the number of merges allowed for each iteration, i.e. the N entity pairs with lowest values for $Dist(e_1,e_2)$, and the N pairs with highest values for $MI(e_1,e_2)$ will be merged. Both are empirically set to 5.

Clustering is allowed to run for 100 iterations. From the output grammar, we selected the 20 categories that we regard as basic semantic classes for the ATIS domain. These correspond to classes such as AIRLINE_NAME, DIGIT, FARE_CLASS, etc. We manually complete the terminals for these semantic classes, and use them as seed categories to catalyze the re-run of the agglomerative clustering procedure. The output grammar of this run is then hand-edited for refinement. Hand-editing involves (i) giving meaningful labels to the grammar rules, e.g. CITY_NAME, MONTH_NAME, etc.; (ii) completing their set of terminals; (iii) consolidating similar rules and (iv) pruning irrelevant rules.

This semi-automatic grammar induction procedure was applied to the ATIS-3 corpus. The grammar has 36 non-terminals and 446 terminals. The semi-automatic approach sped up grammar development dramatically, suffering only slight degradations in language understanding performance compared to a handcrafted grammar.

3.1. Portability to Chinese

We applied the semi-automatic grammar induction approach to the Chinese ATIS queries. Since the Chinese language lacks explicit word delimiters, and our clustering algorithm operates on the word unit, we pre-processed all Chinese queries by word tokenization. This is a greedy string matching procedure that references a Chinese word lexicon, CULEX.[4] We have also augmented CULEX with translated airport and city names found in the ATIS-3 training set. Referring to our previous example, the Chinese query is tokenized (with a space delimiter) as:

話俾我知 由 底特律 飛去 西赤斯特城 既 單程 航班

(Approximate translation: <show me> <from> <detroit> <fly to> <westchester county> <particle> <one way> <flight>)

We applied the same procedures as was used for English, and obtained a Chinese grammar (G_{CSA}) with 44 non-terminals and 292 terminals. The discrepancy between the sizes of the Chinese and English grammars are due to a different method of counting. For example, *salt lake city* is counted as three terminals, but its Chinese form, 鹽湖城 is counted as a single tokenized terminal. Tables 2a and 2b present results on Chinese language understanding using the induced grammars.

	Test 93		**Test 94**	
Understanding	G_{CSA}	G_{SA}	G_{CSA}	G_{SA}
Full	77.7 %	80.4 %	74.1 %	76.8 %
Partial	16.3 %	16.5 %	22.5 %	21.8 %
No	6.0 %	3.1 %	3.9 %	1.4 %

Table 2a. Test set coverage of the semi-automatically induced Chinese grammar (G_{CSA}) in language understanding. Full understanding refers to the percentage of queries with exact matches between the generated semantic frame and the reference SQL in ATIS-3. Partial understanding refers to partial matches. No match is often caused by out-of-vocabulary words in the test set.

	Test 93		**Test 94**	
	G_{CSA}	G_{SA}	G_{CSA}	G_{SA}
Concept Error Rates	13.8 %	14.0 %	13.9 %	12.2 %

Table 2b. Concept error rates for the semi-automatically-induced Chinese grammar. Reference values from the English grammar are also provided.

Since we did not handcraft a Chinese grammar, these results should be compared with the English results. We see that G_{CSA} trails G_{SA} in language understanding performance. Analysis shows that this is caused by manual translation errors in the parallel corpora. The small discrepancy between G_{SA} and G_{CSA} suggests that our semi-automatic grammar induction approach is portable to Chinese. It is also interesting to compare the English and Chinese grammar rules. Inflectional rules, e.g. (SC_i → serve | serves), are common for English but rare for Chinese. Certain rules exhibit word-for-word correspondences between English and Chinese, e.g. (CITY_NAME → atlanta | baltimore | boston | …) versus (CITY_NAME → 亞特蘭大 | 巴的摩爾 | 波士頓 | …); but others exhibit reverse word order, e.g. (FLIGHT_NUMBER →

[3] p_1^{left} is the probability distribution to the left of the first entity in the pair, and the definitions of p_1^{right}, p_2^{left}, p_2^{right} follow accordingly.

[4] CU LEX is part of the CU Corpora, a Cantonese speech resource developed by the Chinese University of Hong Kong. (http://dsp.ee.cuhk.edu.hk/speech).

FLIGHT NUMBER) in English; versus (FLIGHT_NUMBER → NUMBERS FLIGHT) in Chinese.[5]

4. Bi-directional Machine Translation

As mentioned earlier, we attempt to assess the quality of the semi-automatically induced grammars (for English and Chinese) by using them simultaneously in an example-based machine translation (EBMT) system, as depicted in Figure 1. We began by creating a bilingual term list from the grammars. For each English and Chinese training query pair, we parse for the corresponding pair of concept sequences, which are then aligned and stored. Given a test query in the source language, we also parsed for its concept sequence with the source language's grammar, and then search for the training query (in the source language) which has the same (or a similar) concept sequence. Then the concept sequence that corresponds to the training query (in the target language) is used to generate the translation of the test query in the target language. Hence our translation approach draws from training query examples using concept alignments In the following we will describe our bilingual term list, which is used in the subsequent procedures of concept alignment and translation generation.

4.1. Bilingual Term List

Our bilingual term list is extracted from the two grammars. For example, referring to the two grammar rules on CITY_NAMES presented in the previous section, we can generate the following mappings (by hand):

atlanta ↔ 亞特蘭大 ; baltimore ↔ 巴的摩爾 ; etc.

There are also cases with many-to-many mappings, such as los angeles | l a ↔ 洛杉磯 ; salt lake city | salt lake ↔ 鹽湖城 ; and washington is mapped to 華盛頓 as a city; but to 華盛頓州 as a state. Overall, our bilingual term list has 362 translation pairs extracted from the two grammars.

4.2. Concept Alignment in the Parallel Training Corpora

The essence of our EBMT approach lies in a set of aligned concept sequences in the parallel training corpora. For each pair of English and Chinese translations in the training set, we obtain a pair of concept sequences by shallow parsing with the induced grammars. Consider the following pair of translations:

English: *what's the earliest flight from nashville to tacoma on american on tuesday the eighteenth of may*
Concepts (English): <E.QUERY><E.SUPERLATIVE><E.FLIGHT><E.ROUTING><E.AIRLINE_NAME><E.DEPARTURE_DATE>
Chinese: 邊班係美國航空公司五月十八星期二由納什維爾飛去他科馬最早個班機
Concepts (Chinese): <C.QUERY><C.AIRLINE_NAME><C.DEPARTURE_DATE><C.ROUTING><C.SUPERLATIVE><C.FLIGHT>[6]

Differences in concept ordering reflect the differences in word order between English and Chinese. Concepts may be nested, e.g. ORIGIN and DESTINATION are nested in ROUTING.

We applied two rules in aligning concept sequences from the source and target languages:

(i) If the concept XXX appears once in the English sequence and once in the Chinese sequence, E.XXX and C.XXX are mapped to each other directly (e.g. refer to the concept AIRLINE_NAME in Figure 2.)

(ii) If the concept XXX appears multiple times in the English / Chinese concept sequences, we map them to each other only if their terminals form translation pairs according to our bilingual term list. As an example, refer to Figure 2, where E.CITY_NAME1 is mapped with C.CITY_NAME1 because *nashville* and 納什維爾 form a translation pair; and E.CITY_NAME2 is mapped with C.CITY_NAME2 because *tacoma* and 他科馬 form a translation pair. The resulting alignment is shown in Figure 2.

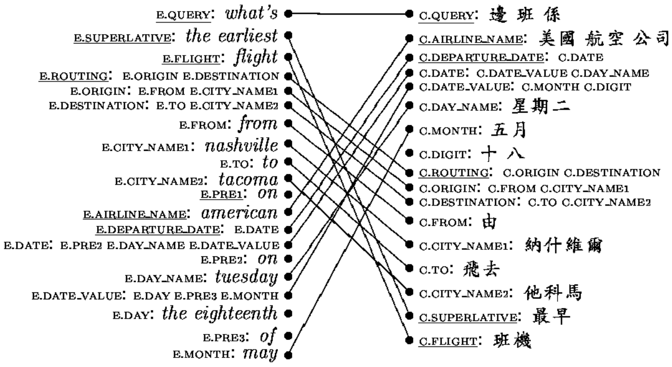

Figure 2. An example of concept alignments from the English and Chinese versions of a training query.

We also store the score corresponding to every alignment. The alignment score (S_A) is defined as:

$$S_A = 1 - \frac{C_N}{C_M + C_N} \quad (3)$$

where C_M is the number of matched concepts, and C_N is the number of concepts without a match. Hence the alignment in Figure 2 scores 0.97, since the only mismatched concept is the preposition E.PRE1. The other orphan concepts (e.g. E.DAY, E.PRE2, etc.) are nested in higher-level mapped concepts.

4.3. In Search of an Example Translation

Given a test query in the source language, we obtain its concept sequence by shallow parsing. Then we refer to our concept alignments in the training set to search for an example translation. We compare the test query's concept sequence with each training query's concept sequence, and may encounter the following cases:

(i) *Exact match* – this is most desirable as we can find a training query with the same concept sequence as our test query. Hence we take the corresponding training concept sequence (in the target language) and proceed to generate a translation.

(ii) *Robust match* – if we fail to find an exact match, we remove non-content-carrying concepts from the test query's concept sequence, and search for a match again. Examples of removable concepts include PRE (for prepositions), and FILLER (e.g. please, okay, etc.).

(iii) *No match* – if we fail to find a robust match, the test query's concept sequence (in the source language) is used directly in the next step to generate a translation.

Each matched example produces a concept sequence in the target language, and the sequence is scored in terms of the original alignment score. It may also occur that multiple

[5] For example, in English we say "flight four seventeen," but in Chinese we say "四一七航班".

[6] The prefixes E.XXX and C.XXX denotes that the concept XXX is obtained from the English or Chinese query respectively.

matched examples (*n>1*) produce the same (target language) concept sequence. Under this situation, we score the concept sequence by average across all the matching alignment scores.

4.4. Generating a Translation

The matching (full / robust) training examples provide a concept sequence in the target language for generating a translation output. For each concept in the sequence, we refer to the target language grammar, the bilingual term list, its aligned source language concept and its terminal to generate a portion of the translation output. If we obtain multiple possible grammar terminals, e.g. (E.FLIGHT → flight | flights), a selection is made at random. If we encounter nested concepts, this generation procedure is performed recursively. The outputs from all concepts are then appended (in the order of the target language concept sequence) to produce the overall output. Hence we preserve the word order in the target language as we generate our translation output. However, the word order is not preserved if we generate directly from the source language's concept sequence, i.e. in the case of no match. Examples of translation outputs are shown in Table 3. As can be seen, outputs in Chinese-to-English translation may suffer from the use of errorful inflectional forms. This is due to the random selection of viable terminals during the generation process.

English input	:	*yes i'd like to find a flight from memphis to tacoma stopping in los angeles*
Chinese output (Exact Match)	:	我想由孟斐斯飛去他科馬停洛杉磯既航機 (<i'd like> <from> <memphis> <to> <tacoma> <stop in> <los angeles> <flight>)
Chinese input	:	我要一班由邁阿密起飛大約下晝五 點到芝加哥既美國航空班機 (<i want> <from> <miami> <to> <around> <five p m> <to> <chicago> <american airlines> <flight>)
English output (Robust Match)	:	*i want from miami to chicago on american flights depart about five p m*
Chinese input	:	星期日紐約去拉斯維加斯同孟斐斯去拉斯維加斯 (<sunday> <new york> <to> <las vegas> <and> <memphis> <to> <las vegas>)
English output (No Match)	:	*depart on sunday from new york to las vegas and from memphis to las vegas*

Table 3. Three examples of translation outputs: (1) A case of *Exact match* for English-to-Chinese translation; (2) A case of *Robust match* for Chinese-to-English translation; (3) A case of *No match* for Chinese-to-English translation.

4.5. Evaluating the Translations

We generated translations for both ATIS-3 test sets (1993 and 1994), and also translated from English-to-Chinese as well as from Chinese-to-English. We also recruited an impartial subject to evaluate the output of the translations. The evaluator is asked to grade each translation with four levels: FULL 1 indicates a fully acceptable translation; FULL 2 indicates a translation that preserves meaning, but lack in fluency; PARTIAL indicates some concepts are missing; BAD indicates a translation is nonsensical. Evaluation results are shown in Table 4.

	English-to-Chinese		**Chinese-to-English**	
Grade	**1993 Test**	**1994 Test**	**1993 Test**	**1994 Test**
Full1	85.5%	84.2%	78.3%	73.2%
Full2	6.3%	3.6%	10.9%	3.2%
Partial	7.1%	11.0%	10.0%	20.9%
Bad	1.1%	1.1%	0.7%	2.7%

Table 4. Evaluation of translation outputs by a human subject.

5. Summary and Conclusions

In this paper, we have presented an approach for semi-automatic grammar induction from un-annotated corpora. The approach has been applied to English ATIS queries, and in this work we demonstrate its portability to Chinese queries. The induced Chinese grammar achieved the same level of language understanding performance as the English grammar. The semi-automatic nature of our approach greatly reduces the amount of handcrafting in grammar development, and the use of un-annotated corpora implies minimal resources are needed for corpora preparation. The induced grammars were also incorporated in a bi-directional EBMT framework. Our grammars preserve the word order differences between English and Chinese. Most of the translation outputs (73% to 85%) and were fluent and acceptable by an impartial human evaluator. An advantage of the EBMT approach is that the translation quality can be incrementally improved as more training data is available. In the future, we plan to extend the Chinese-to-English translation mechanism with the ability of grammar checking to generate appropriate inflectional forms.

6. Acknowlegdements

This research is supported by the Hong Kong Government Research Grant Council Earmarked Grant No. CUHK/4177/00E.

7. References

[1] Siu, K. C., Meng, H. M., "Semi-Automatic Acquistion of Domain-Specific Semantic Structures", *Proc. of Eurospeech'99, Vol. 5, pp. 2039-2042, 1999*.

[2] Brown, R., "Example-Based Machine Translation in the Pangloss System", *Proc. of COLING 96, pp. 169-174*.

[3] Zhang, Y., Brown, R., and Frederking, R., "Adapting an Example-based Translation System to Chinese", *HLT2001, 2001*.

[4] Berger, A., Brown, P., Della Pietra, S., Della Pietra, V., Gillett, J., Lafferty, J., Mercer, R., Printz, H. and Ures L., "The Candide System for Machine Translation", *Proc. of the ARPA HLT Workshop, 1994*.

[5] Nyberg, E., Mitermua, T., and Carbonell, J., "The Kant Machine Translation System: from R&D to Initial Deployment", *Proc. of the LISA Workshop on Integrating Advanced Translation Technology, 1997*.

[6] Takezawa, T., Morimoto, T., Sagisaka, Y., Campbell, N., Iida, H., Sugaya, F., Yokoo, A. and Yamamoto, S., "A Japanese-to-English Speech Translation System: ATR-MATRIX", *Proc. of ICSLP '98, 1998*.

[7] Waibel, A., Lavie, A. and Levin, L., "Janus: A System for Translation of Conversational Speech", *Kuenstliche Intelligenz (KI), Vol. 4, 1997.*

[8] Bub, T., Wahlster, W., and Waibel, A., "Werbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation", *Proc. Of ICASSP '97, pp. 71-74, 1997.*

[9] Price, P., "Evaluation of Spoken Language Systems: The ATIS Domain", *Proc. of the ARPA HLT Workshop, pp. 91-95, 1990.*