Bilingual Knowledge Acquisition from the Parallel Corpus

Lina Zhou[12] James Liu[2] Shiwen Yu[1]

([1]Institute of Computational Linguistics, Peking University)
([2]Department of Computing, Hong Kong Polytechnic University)

Abstract: This paper proposes a model for acquiring bilingual knowledge from a parallel corpus. Unlike that of the statistically aligning method, the model focuses on the contrastive linguistic aspects by accommodating the idiosyncracies of one language in its rule-based analyzing process, thus reducing the effect of corpus size. Then, it applies the statistical procedure to acquire different levels of knowledge from the corpus. It shows promising results and the acquired information can be put into a variety of applications including machine translation.

并行语料库中的双语知识获取

周莉娜[12] 廖雅国[2] 俞士汶[1]

([1]北京大学计算语言学研究所, 100871)
([2]香港理工大学电子计算学系, 香港九龙红勘)

摘要: 本文提出了一种从并行语料库中获取双语知识的模型. 与常见的基于统计的对应方法不同, 该模型把重点放在了比较语言学方面. 首先, 通过基于规则的分析过程解决反映语言个性的问题, 然后, 再用一般的统计方法从语料库中获取双语知识; 因而大大减少了语料规模对获取结果的影响. 实验结果令人满意. 获取的双语知识可以应用于机器翻译等领域.

1. Introduction

Parallel corpus has contributed to many applications, particularly machine translation(MT). It gives rise to example-based (EB) method, which performs translation by referring to similar source sentences in the example base, then generating target translation for the new sentence on the basis of the target translations of the similar sentences. EB method can be employed at different stages of MT, such as source language analysis, source to target transformation and target language generation, etc. They take parallel corpus as knowledge base for translation and acquire knowledge from the example base accordingly. Therefore, bilingual knowledge is the direct and significant contribution of a parallel corpus. In this paper, we will propose a schema for knowledge acquisition from a parallel English-Chinese corpus.

To overcome the labor-intensive and 'bottleneck' problems in the area of computational linguistics, machine learning techniques are being adopted to acquire knowledge automatically. Nevertheless, considering the regularly available parallel corpus which are at most aligned on the sentence level rather than syntactically analyzed, pre-processing becomes necessary for the later knowledge acquisition process.

In an early attempt to align Asian and Indo-European language pairs, [1] used bilingual dictionary to segment the sentence. Moreover, correspondence between phrase structures couldn't be created directly but implied in the provided bilingual grammar, which is still open to generalization. [2] has suggested a statistical model which integrates the problem of deciding alignment units and aligning the candidates of different word orders. But it depends on large amount of training corpus for estimating the parameter of the model to get better

performance. However, Chinese words can' t be determined before the alignment process, for they are not uniformly defined yet. The pioneering work by [3] in aligning English-French corpus employed statistical method which can' t be transported easily to languages other than the structurally similar pairs of English and French.

The proposed model focuses on the contrastive linguistic aspects by accommodating the idiosyncracy of one language in its rule-based analyzing process while reducing the effect of the corpus size. It integrates knowledge inference with the statistical approach to improve the MT performance. Details are given below.

2. Knowledge Acquisition from the Parallel Corpus

2.1 Word translation acquisition

The most direct application of bilingual corpus is to generate bilingual word translation dictionary. However, Chinese words are not given explicitly in the corpus. We need to segment the Chinese sentence first to pave the way for knowledge acquisition.

1. Strategies

From comparison on the acquisition result of bilingual word correspondence with and without separate segmentation procedure in [6], it can be found that pre-segmentation is a better solution. With regard to the fundamental problem of segmentation in Chinese computing, there have been a lot of reports on how to address the issues of ambiguity and unknown word recognition, but little on the problem of segmentation criteria. So, MT oriented segmentation criteria is proposed in [5]. The published segmentation specification [4] is adjusted and detailed for the correspondence between the words in the bilingual corpus in the following aspects:

- Numerical structures (including time expression)
- Name (including professional title addressing)
- Reduplicated structures
- 2-syllable construction containing one structural auxiliary word
- Suffixes (including plural structure)
- Compound word structures (verb-object type, verb-complement type, head-adjunct type)

The modified segmentation criteria facilitates bilingual word translation accurately.

2. Classification of word translations and the solutions

There are many exceptional cases other than the ideal one-to-one correspondence in the statistical acquisition result. They can be classified into:

- One-to-many: e.g. six (六 , 6 点)[1], morning (早晨, 早上), coat (外衣, 大衣, 上衣).
- Many-to-one: e.g. fault, wrong (错), consented, agree, agreed (同意).
- Many-to-many: e.g. choose, elect, select (选, 选择); capital (首都, 资金); suggest, recommend, suggestion, proposal (建议, 提议).
- Whole-to-part: e.g. cyclist (骑/ 自行车/的/人), in (在/里), unthinkable (没/ 想到).
- Part-to-whole: e.g. but (也, 而且), order (为了), take (起飞), disease (病因), Yellow (黄河).
- Part-to-part: e.g. clear (清理/ 干净), hand (交/ 上来).
- Null-to-existence: e.g. ε[2] (只), ε (了).
- Existence-to-null: e.g. be (ε), to-INF (ε)

In view of the above corresponding types, three kinds of strategies are adopted. First, null correspondence can be identified directly from the statistical result. If some words only appear together with the high frequency words in the other language, it is very likely that they have no translation correspondences. Secondly, the problem of partial or incomplete matching cases

[1] The bilingual word pair are represented as English word (Chinese word).

[2] ε means null correspondence in one language version.

could be resolved by double checking co-occuring frequencies of the partial words and their distributions in the corpus. Lastly, more translation candidates are extracted in case of multiple translation appears, accordingly a filtering process should follow the acquisition process. Generally, the threshold on word frequency as well as word-level heuristics information may then be utilized. The former is easier to be implemented and modified, but it is better to capture such information and formalize them into the filtering rules whenever there exist bilingual word translation regularities, which might be relevant to the words themselves or their grammatical attributes, such as part-of-speech information incorporated in our model.

2.2 Phrase structure translation knowledge

From the viewpoint of alignment, the smaller the bilingual correspondence unit, the more flexible the unit could be used in later procedure; while the more difficult to create the correspondence and more ambiguities involved. Besides the bilingual dictionaries, another essential component for an MT system is the structural transformation relationship between the source and target language. Therefore, we need to produce correspondence for phrase structures.

1. Strategies

Most of the present parallel corpus is quite primitive instead of having been analyzed into phrase structures or parsed into dependency structures. Even with the parsed corpus, it is highly probable that the languages involved may not have been formalized in the same type of grammar due to their pecularities, and earlier parsing hasn't taken the component of improving structural correspondence between the languages into consideration. Furthermore, parsing is a basic step in natural language understanding. Though there are several alternative methods for parsing, it still seems necessary to adopt a rule-based element. The statistical method performs well on the large scale corpus, but such parallel corpus is still quite few. In the example-based one, time efficiency is an important issue to be dealt with. With the tagged bilingual corpus, we have to parse them syntactically by ourselves. We consider parsing to be suitable for both the corpus and the input in this study.

The suggested parsing is performed on each language separately. It can be classified as a shallow one, but devised from the bilingual point of view.

2. Parsing Principles and the Types of Constructions Concerned

The overall parsing principles are stated as follows:

- Analyze into flat structure which is appropriate for both the structurally similar and dissimilar languages and easier to be adjusted later.
- The correspondence between translation units takes higher precedence than that of word order, i.e. try to accommodate mismatch of the translation unit in the parsing result.
- Bi-directional parsing to take advantage of both the left-boundary and right-boundary determined properties of some phrase structures.
- Avoid analysis of structural ambiguities by free-riding or grouping.
- Collect the constituents of noun phrases together except the clausal modifier, while single out the verb groups.
- Attach the coordination structures with the modifying (or modified) part while detaching the conjunction from the embedded clause which will be parsed further.

According to the above guidelines, the following constructions are covered in a particular way herein.

Structure Type	Description		Strategies	
	English	Chinese	English	Chinese
Noun Phrase	Articles	Classifiers	NP recognition	
Verb Groups	Inflection of verbs for voice, tense and aspect, etc.	Particles		Group verb with particles with foatures as tense,modality, etc.

Proper Noun	Capital-initialised		Morphological cues and NP recognition.	Word Segmentation
Main predicate verb	Explicit finite verb	Serial verbs without explicit indications for main verb, e.g. 开车/v 出去/v 玩/v, 倒/v 下来/v 横/v 在/p 路上/s, etc.		Refer to subcategory, e.g. directional, auxiliary or common verb, etc. to be divided into smaller units.
Coordination structures			Combine	
Ending particles		Function as mood at the end of the imperative or interrogative sentences		Group with the end mark.
Movement Phenomena	*e.g. pay close attention to, Have read ...?*	Discontinuous characters of a verb, e.g. 鞠/了/一/躬, 吃/过/了/饭	Combine except that preceded auxiliaries in interrogative sentences are singled and marked with movement tags	Refer to the Chinese grammatical dictionaries.
Special Sentence Type1	Imperative sentence Passive voice	BA-type sentence BEI-type sentence		Taken as the starting of PP or classify BEI into the verb group when the agent of the action is omitted.
Special Sentence Type2	Causative usage in English, e.g. *have my hair cut.*	Double constituent sentence	Correspond to fixed word group in Chinese, so combine them, e.g. 令人激动, 让人理发	Map onto the object complement structure in English directly
Clause	Initialized with explicit conjunctions	Conjunctions are immersed but some word cues such as "的", conjunctive or adverbs are identification.	Picked out from the sentence and marked with special tags. Further analyze remaining parts beyond relative clause.	Make use of word heuristics and part-of-speech information, while checking the corresponding English sentence.

Table 1. Structures of Particular Concern in Bilingual Parsing

3. Types of structural correspondences and Solutions

After analyzing the syntactic structure of the bilingual corpus, sentences are indexed with their flat phrase structures to reduce the matching time later, and structural correspondence is expected to be obtained from the corpus with the aid of the bilingual dictionary. More information can be induced from the structural correspondence as shown below:

- Automatic recognition of English phrasal verbs and quantifiers: e.g. throw away (扔掉), a glass of (一杯), wait for (等待), a great deal (许多), no less than (不少于).
- Automatic recognition of English proper nouns: e.g. the Atlantic Ocean (大西洋), New York (纽约).
- Word gaps are narrowed: e.g. a small spoon (一把小勺), happy and healthy (快活而健康), it (把它).
- Time expression alignment: e.g. on October 1, 1949 (1949年/10月/1日), 8 a.m. (早上8 点).
- Capture special usage of adjectives in English phrases: e.g. in blue (穿/蓝/衣服/的), the poor (穷人).
- Determine the prepositional phrases in English by referring the aligned Chinese sentence: e.g. , pay special attention to][the world (密切/][注意), sit][beside me (坐/][在/我/旁边).
- Separate the Chinese verb series with reference to the English sentence: e.g. [开车/v][出去/v 玩/v (to come][for a drive), 倒/v 下来/v][横/v 在路上 (had fallen][across the road).
- Structural Correspondence between bilingual sentences

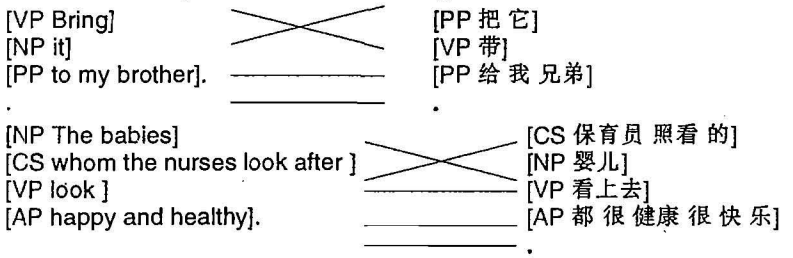

Figure 1. Structural Correspondence Samples

Apart from complementing the word correspondence which is beyond the scope of word acquisition process, we emphasize on the structural correspondence. In the above examples, though the phrases in each pair of sentences are not aligned in the same order, they are really matched one by one in different orders. These information is very helpful for the later translation process.

- Inconsistency Process

There are some movement phenomena in contemporary English, such as interrogative sentence, it-type formal subject sentence, the conjunctions in the relative clause, etc. Moreover, there are some particles in Chinese which play auxiliary function on mood, which are incorporated into morphology or word order in English, e.g.

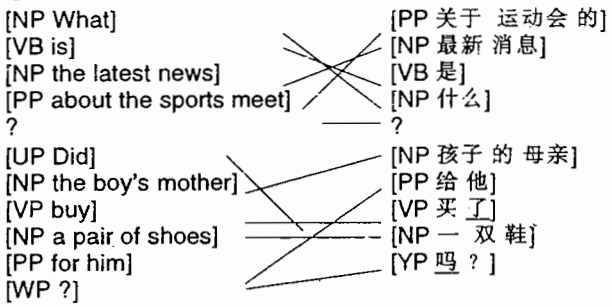

Figure 2. Structural Inconsistency Samples

The first example case can be processed as the preceding type. It has shown in the second example, the separated UP in English will be connected with the next VP with finite verb, while the particles in Chinese are grouped with the word they assist. However, we can't avoid mismatch when aligning the bilingual sentences, such as *it* in the it-type English sentences. In terms of MT, there are still inserting and deleting operations in the transformation of the sentences. It should be noticed that these 'unpleasant' cases don't matter much as long as there exist correspondences on the whole pair of sentences and they are identified as unmatched part.

3. Algorithms Involved

3.1 Statistical Algorithm for Word Translation Acquisition

It is assumed that the translation probability between a pair of words is proportional to their co-occurrence frequency. So, the expected number of times t that any particular word e in an English sentence $\mathbf{e} = e_1 e_2 .. e_l$ generates any particular word c in the corresponding Chinese sentence $\mathbf{c} = c_1 c_2 \ldots c_m$ is given by:

$$t(c \mid e; \mathbf{c}, \mathbf{e}) = \sum_{j=1}^{m} \delta(c, c_j) \sum_{i=1}^{l} \delta(e, e_i) \qquad (1)$$

The total number of co-occurrence frequency s_e for English word e in the whole corpus is:

$$s_e = \sum_{c} \sum_{c,e \in corpus} t(c \mid e; \mathbf{c}, \mathbf{e}) \qquad (2)$$

The algorithm is listed in Table 2:

1. Count $t(c \mid e; \mathbf{c}, \mathbf{e})$ for all pairs of words (e, c) in the sentence pair $(\mathbf{c}, \mathbf{e})$ using equation (1).
2. Computer s_e for every English word e with equation (2).
3. Sort Chinese word translation candidates of every English word in the descending order of s.

4. Extract the first five Chinese candidates for every English word, filtering some vocabulary gap between Chinese and English, e.g. "把", "被", "着", "得" , etc. which have no correspondence in English. Also with other special word heuristics and filtering strategies
5. Filter English words with frequency threshold of 3.

Table 2. Algorithm for extracting word translation candidates

3.2 Bilingual Parsing Model and Phrase Correspondence Extraction Algorithm

Our parsing model is based on context free grammar and driven to work separately on each language in both directions. Conventionally, the syntactic analysis for an input sentence starts from the beginning of the sentence, no matter what type of parsing method is employed. However, it was found from our observation that reverse parsing is more efficient for the flat structure. Since some of the linguistic structures are left-boundary determined; while others are right-boundary determined, we scan the new sentence in both directions at different scanning times based on the heuristics below.

1. Scan the sentences forward to produce small fragments, which should be combined in the later parsing process, while recognize the starting of prepositional phrase and clause in English.
2. Scan the sentences from the end to the beginning to extend the former smaller fragments into larger ones while assigning or adjusting phrase types, in which noun phrase, prepositional phrase, conjunctive phrase, adjective phrases, verb groups, etc. are recognized.
3. Create dictionary for the correspondences between bilingual sentence structures and every phrase structures individually. If all the phrases in a sentence pair can be matched with each other, extract the matches and put them into the corresponding Chinese phrase dictionaries; otherwise, using bilingual dictionary to extract more from the remaining fragments.
4. Run through the acquired results and move the pairs which have only one word on one side to a temporary bilingual word dictionary.
5. Check the acquisition results manually.

Table 3. Algorithm for extracting phrase correspondence

With this parsing process, it is easy to align the constituents in a pair of Chinese and English sentences as well as set up a better foundation for deeper monolingual parsing.

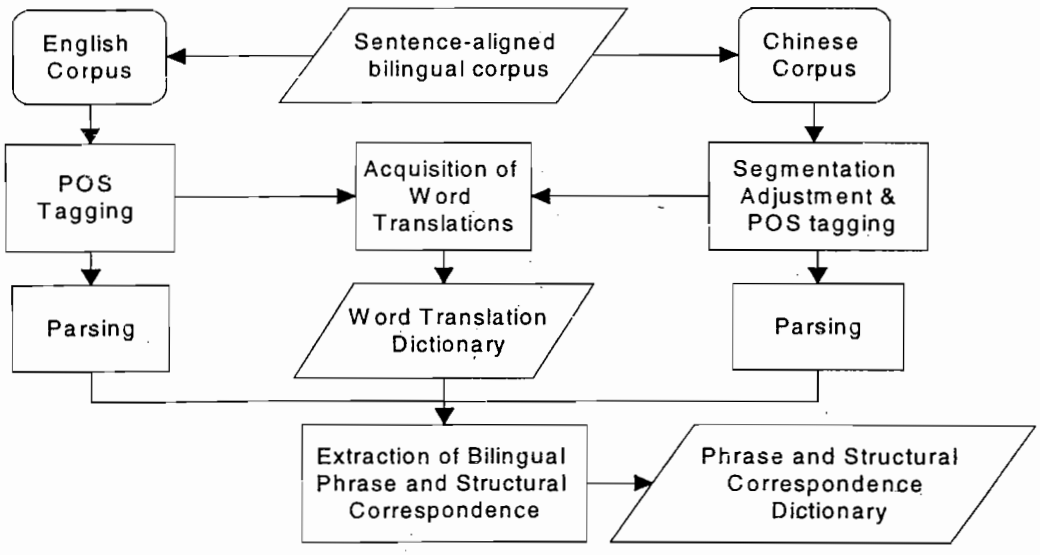

Figure 3. Overview of the Acquisition Process

4. Experiments

Experiments was implemented on a small but comprehensive corpus, covering all the linguistic phenomena that are expected to meet in English-Chinese translation. Segmentation was done on the 3,000 Chinese sentences. Then the acquisition procedure begins as shown in Figure 3.

In our experimental results, the accuracy for word translation amounts to a little more than 91%. While the accuracy for phrase structure correspondence hasn' t been evaluated yet for the undecided computerized criteria, we believe that the parsing result is good enough for future bilingual structural alignment from our observations.

5. Conclusion

In this paper, we propose a model for acquiring bilingual knowledge from the original bilingual corpus step by step. It can be regarded as the second development upon the bilingual corpus. The knowledge base obtained is valuable resource for a variety of applications:

1. The obtained knowledge base can be put into the development of MT system directly.
2. Both the bilingual word dictionary and phrase dictionary can be a good reference and supplementation for compilation of bilingual dictionaries.
3. The dictionary with correspondence between bilingual sentence structures will improve the research on contrastive linguistics.

One of the factors affecting the usability of the corpus is its processing depth, so many efforts have been made upon analyzing the corpus in our work. However, based on the result of coarse processing on the bilingual corpus, other meaningful new knowledge is obtained through statistical or rule-based procedure, which justified our past efforts. The suggested model may be well applied to other language pairs whether with similar structures or not. The former case mainly depends on the statistical method; while the latter strengthens the rule-based module with the difference between the language pairs being enlarged.

We identify some areas for future enhancement or improvements. These include 1. Efficient algorithm for filtering wrong word translations from the statistical result. 2. Exploration for mutual disambiguation, i.e. resolve the ambiguities in one language by resorting to its correspondence in the other. 3. Further align the bilingual sentences and create the bi-directional structural transformation relationship between them.

References

[1] D.K. Wu and P. Fung, 《 Improving Chinese Tokenization with Linguistic Filters on Statistical Lexical Acquisition 》 . Proc. of Fourth Conf. on Applied Natural Language, Stuttgart, October 1994, pp. 180-181.

[2] Jung H. Shin, Young S. Han & Key-Sun Choi, 《 Bilingual Knowledge Acquisition from Korean-English Parallel Corpus Using Alignment Method 》 . Proceedings of COLING' 96, Denmark.

[3] Peter F. Brown, Stephen A. Deela Pietra, Vincent J. Della Pietra, Robert L. Mercer. 《 The Mathematics of Statistical Machine Translation: Parameter Estimation》 . Computational Linguistics, 19(2):263-311, 1993.

[4] 《 Contemporary Chinese Language Word Segmentation Specification for Information Processing》 . Beijing: National Bureau of Technology Control (in Chinese), 1993.

[5] Lina Zhou, James Liu, 《 Extracting more word translation pairs from samll-sized bilingual parallel corpus: integrating rule and statistics-based method》' . ICCPOL' 97.

[6] P. Fung, and D.K Wu, 《 Statistical Augmentation of a Chinese Machine-Readable Dictionary, Proc. Workshop on Very Large Corpora》 . Kyoto, August 1994, pp. 69-85.