A Parallel Corpus-Based Study of Chinese Arabic Verb Phrase Alignment

Alaa Mamdouh Akef, Yingying Wang and Erhong Yang (Beijing Language and Culture University, China)

1. Introduction

The Bilingual Alignment Corpus is composed of five categories that vary according to their alignment units: paragraph, sentence, lexical, chunk, and phrase alignments. Over the years, researchers have been committed to the phrase alignment of bilingual corpora based solely on the syntax and lexical alignments.

In this paper, we used an *"Analysis - Analysis - Matching"* alignment strategy to extract the Arabic-Chinese alignment pairs. First, we used a lexical alignment for a parallel Arabic-Chinese corpus, and in accordance with this result, we extracted the phrase alignment pairs based on syntactic analysis and we built a parallel Arabic-Chinese phrase corpus. Then, we selected Arabic verb phrases as the object of our study, afterwards we examined the Arabic verb phrases and their corresponding Chinese phrases' internal structures and after discovery of the Arabic-Chinese verb phrase alignment rules, we summarized them; which built translation rules for instance-based and rule-based machine translation systems.

2. Experiment

First, 7,125 Chinese-Arabic aligned sentences with 827,500 words were drawn from UN official transcripts; the Arabic data included 7,125 sentences and 827,500 words, and the Chinese data included 7,523 sentences and 265,427 characters. The total of sentence pairs after sentence alignment was 7,125. Second, the errors of the parallel corpus were simply modified manually, including typos, Arabic collocation, punctuation and titles.

The experiment was carried out as follows:

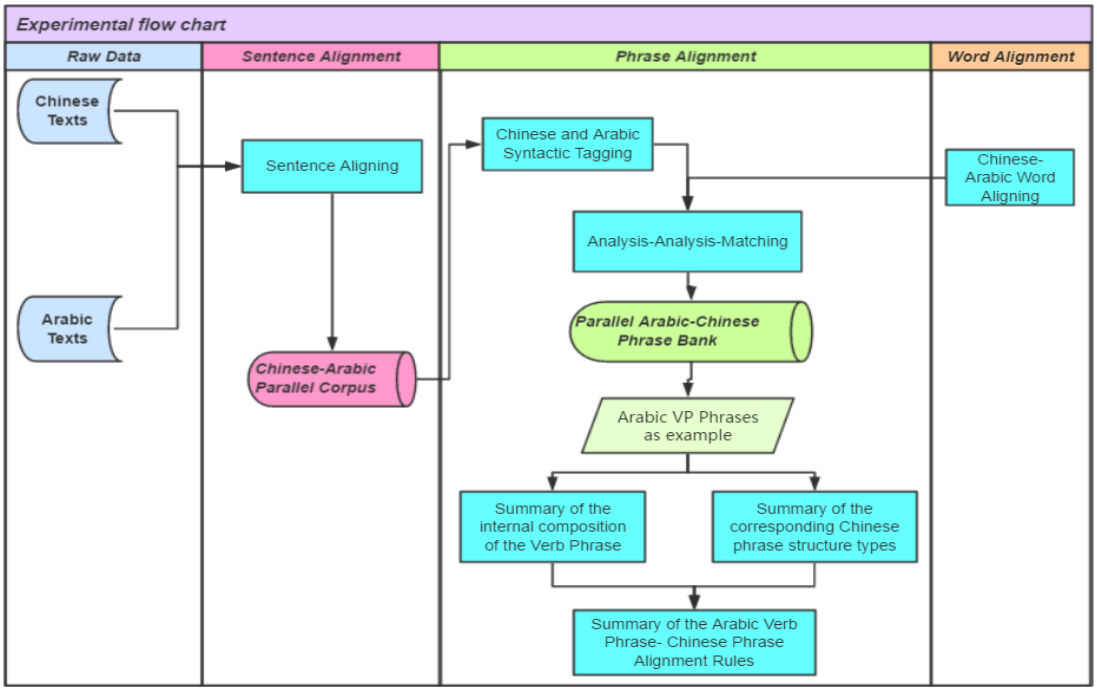

Figure 1 Experimental flow chart

This paper used the word alignment tool GIZA ++ (Och & Ney, 2003) to complete the two-direction (Arabic-Chinese and Chinese-Arabic) lexical alignment. The GIZA ++ tool process specifies that each source language word corresponds to only a single target language word. When the source language word does not correspond to the target language word, it is assumed to be aligned to a null word. We observed the existence of null words, which are caused by the huge lexical, word order, and syntactical differences between the Arabic and Chinese linguistic systems.

The Stanford parser *arabicFactored* was used for the Arabic data syntax parsing (Green & Manning, 2010), while *xinhuaFactoredSegmenting* was used for the syntactic parsing of the Chinese data (Levy & Manning, 2003). During the analysis of the parsing results, it was found that both Arabic and Chinese had parsing errors, such as the tagging of a blank element with some part of speech or syntactic element, e.g. *(NP (NN))*. After removal of the blank elements within the sentence pairs, 6,387 sentence pairs were left.

Based on the syntactic analysis and word alignment results, this paper used the *"Analysis - Analysis - Matching"* strategy to write a program to automatically extract Arabic-Chinese aligned phrase pairs. In 1992, Kaji first used this method for Japanese-English bilingual phrase alignment, and afterwards it was applied to other language pairs. In this paper, a total of 19,586 alignment phrase pairs were extracted from 6,387 Chinese sentences, and were used as data to build the Arabic-Chinese parallel phrase bank.

Arabic is a VSO language, and this experiment's results also show that VSO word order accounts for 82.7% of the whole Arabic corpus, leading us to choose Arabic verbal phrase as our research object. After deleting the 1,088 phrases which had syntactic parsed errors, we had 4,833 verbal phrase pairs,

on which the basis of the internal composition of the Arabic verbal phrase and its corresponding Chinese phrase structures were summarized. Arabic verbal phrase and Chinese phrase alignment rules were also summarized.

3. Results and analysis

The syntax parsing errors for the above-mentioned 1,088 Arabic verb phrases mainly included tagging prepositions, nouns, Arabic ann particles, function words, demonstratives, pronouns, and other words as verbs. The author describes this question in more detail in another paper about Arabic-Chinese phrases (Akef, Yang, & Wang, 2016).

3.1 The internal composition of Arabic verb phrases

The three main types of verb phrases are verb + noun phrase (V + NP), verb + prepositional phrase (V + PP), and function word + verb (RP + V). The tense of the verbs in each type can vary.

Table 1 The Internal Composition of the Arabic Verb Phrase

Internal composition	Examples	Number	Percent
VP->VBD+NP	*(VP (VBD صارت) (NP (DTNN التجارة) (DTJJ الخارجية))*	3,032	69.1%
VP->VBP+NP	*VP (VBP يمثل) (NP (DTNN السلام))*	869	19.8%
VP->RP+VBP	*VP (PRT (RP لا)) (VBP يستغني) (NP (NN ازدهار) (NP (NN واستقرار) (NP (DTNN العالم))))*	136	3.1%
VP->VBP+PP	*VP (VBP تحافظ) (PP (IN علي) (NP (CD 100) (PUNC %)))*	61	1.4%
VP->VBD+PP	*VP (VBD ورفعت) (PP (IN من) (NP (NN مستوي) (NP (NN ادارة) (NP (DTNNS المؤسسات)))))*	61	1.4%
VP->VBN+PP	*VP (VBN وشاركت) (PP (IN في) (NP (DTNN الثمار) (DTJJ الناتجة)))*	61	1.4%
VP->VBN+NP	*(VP (VBN وتوسعت) (NP (NN حصة) (NP (NN وارداتها) (JJ وصادراتها)))*	61	1.4%
VP->VN+PP	*(VP (VN ويحتاج) (PP (IN الي) (NP (NN بذل) (NP (NN جهود) (JJ شاقة)))))*	75	1.7%
VP->VBD+ADJP	*(ROOT (S (VP (VBD وقع) (ADJP (JJR اكثر) (PP (IN من) (NP (NP (CD 150) (NP (NN بلدا)))*	32	0.7%
Total	**4,388**		

3.2 Chinese Phrase Structure with Verb Phrase Alignment

Through the analysis of the structure of the corresponding Chinese phrases for the three types of internal composition of the Arabic verb phrases, we may find that the structure of the Chinese phrase is different for each one, as follows:

1) Verb + noun phrase (V + NP)

When the internal structure of the verb phrase is verb + noun phrase, there are five corresponding Chinese phrase structures, as shown in the following table:

Table 2 Arabic V+NP corresponding Chinese phrase structure

Structure of Chinese Phrase	**Arabic verbal phrase**	**Chinese phrase**
VV+NP+VP	*(VP (VBP يسيطر) (PP (IN على) (NP (DTNN المرض) (DTJJ ويمكن))*	*(VP (VP (VV 使) (NP (NN 疾病)) (VP (VV 得到) (NP (NN 控制)))))*
VV+VP	*(VP (VBP تساعد) (PP (IN على) (NP (NN توسيع) (NP (DTNN المسالك) (DTJJ الهوائية) (DTJJ الرئيسية))*	*(VP (VV 帮助) (VP (VV 扩张) (NP (NP (NN 肺部))*
VV+NN	*(VP (VBP يؤدي) (PP (IN الى) (NP (NP (NN تضييق) (NP (DTNN المسالك) (DTJJ التنفسية))*	*(VP (VV 导致) (NP (NN 气道狭窄))*
V+PP	*(VP (VBP يظهر) (PP (IN في) (NP (NOUN_QUANT جميع) (NP (DTNN البلدان))*	*(NP (PN 它)) (VP (VV 发生) (PP (P 在) (NP (DP (DT 所有)) (NP (NN 国家))))*
NP+ADJP	*(VP (VBP تختلف) (PP (IN في) (NP (NP (NN شدتها) (JJ وتواترها))*	*(NP (PN 其)) (ADJP (JJ 严重)) (NP (NN 程度)))*

2) Verb + prepositional phrase（V+PP）

When the internal structure of the verb phrase is verb + prepositional phrase, there are five corresponding Chinese phrase structures, as shown in the following table:

Table 3 Arabic V+PP corresponding Chinese phrase structure

Structure of Chinese Phrase	**Arabic Verbal Phrase**	**Chinese Phrase**
NP+VV+NP	*(VP (VBP يضم) (NP (NP (NNS منظمات) (JJ وطنية) (JJ ووكالات) (JJ ودولية))*	*(NP (PN 它)) (VP (VC 是) (NP (DNP (NP (NP (QP (CD 一) (CLP (M 个))) (NP (NN 国家) (CC 和) (NN 国际) (NN 组织)))*
PP+NP	*(VP (VBP يصيب) (NP (DTNN الرجال) (DTJJ والنساء))*	*(PP (P 对) (NP (DNP (NP (NN 男性) (CC 和) (NN 女性)) (DEG 的)) (NP (NN 影响))))*
VV+NP	*(VP (VBP يعرقل) (NP (NN عملية) (NP (DTNN التنفس) (DTJJ العادية))*	*(VP (VV 减缓) (NP (DNP (NP (NN 疾病)) (DEG 的)) (NP (NN 发展))))) (PU 。)))*
NP+VP	*(VP (VBP تتولى) (NP (NN منظمة) (NP (DTNN الصحة) (DTJJ العالمية))*	*(NP (NN 世卫) (NN 组织)) (VP (ADVP (AD 还)) (VP (VV 领导)*
VP	*(VP (VBP يركز) (PUNC ,) (NP (NN تحديدا))*	*(VP (ADVP (AD 具体)) (VP (MSP 来) (VP (VV 说)))))*

3) Function word + verb（RP+V）

Function words are a very unique language phenomenon within the Arabic language, not having a specific meaning by themselves, but playing a very important part of grammatical function.

When the internal structure of the verb phrase is function word + verb, there are five corresponding Chinese phrase structures, as shown in the following table:

Table 4 Arabic RP+V corresponding Chinese phrase structure

Structure of Chinese Phrase	**Arabic verbal phrase**	**Chinese phrase**
VP	*(VP (RP لا) (VBP يرقيان))*	*(VP (VA 不足))*
NP	*(VP (PRT (RP قد)) (VBP تظهر) (NP (NN فورا))*	*(NP (ADJP (JJ 直接)) (NP (NN 并发症)))*
IP	*(VP (VBD اصبحت) (NP (NN اعادة) (NP (NN استخدام) (NP (DTNN المياه) (DTJJ العادمة))*	*(IP (NP (NN 废水)) (VP (ADVP (AD 再)) (VP (VV 利用))))*

4. Conclusion

The internal composition of verb phrases in Arabic and the structure of their corresponding Chinese phrases were measured with a parallel Arabic-Chinese phrase bank, taking 4,388 Arabic verb phrases as objects of the study composition with regards to the verb phrases. We found that apart from verb phrases, when translating Arabic verb phrases into Chinese, noun phrases and prepositional phrases can be used as well, and every type of Arabic verb phrase can correspond to various Chinese phrase structures. The summarized correspondences obtained from an authentic corpus can be used as phrase-alignment rules in an Arabic-Chinese machine translation, and can be applied to instance-based or rule-based machine translation systems.

References

Akef A., Yang E., & Wang, Y. (2016). An analyzing of Arabic phrases for Chinese Arabic syntax phrase database study. *The 12th China Workshop on Machine Translation*. (in Chinese)

Green, S., & Manning, C. D. (2010). Better Arabic Parsing: Baselines, Evaluations, and Analysis. *International Conference on Computational Linguistics* (pp.394-402). Association for Computational Linguistics.

Kaji H., Kida Y., &Morimoto Y. (1992), Learning translation templates from bilingual tuts, *In Proceedings of the 14th International Conference on Computational Linguistics.*

Levy, R., & Manning, C. (2003). Is it harder to parse Chinese, or the Chinese Treebank? . In Proceedings Of The 41st Annual Meeting Of The Associational For Computational Linguistics (pp.439--446).

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29*(1), 19-51.