Engine-specific Chinese-English User Parallel Corpora

Weimin Jiang
US Air Force

Abstract

This paper proposes some strategies and techniques for creating phrase-level user parallel corpora for Systran translation engine. Though not all strategies and techniques discussed here will apply to other translation engines, the concept will.

1. Introduction

Parallel corpora with text pairs that are translations of each other are playing an ever-increasing role in statistical translation (Koehn, 2005), they are invaluable resources for many NLP applications, such as machine translation, multilingual lexicography, and cross-lingual information retrieval, (Lu et al, 2009) and they are critical resources for extracting translation knowledge in machine translation (Zhang et al). And therefore it is only natural that Chinese-English parallel corpora find wide applications in Chinese-English information processing, bilingual lexicography, language research and teaching (Chang, 2002).

To meet the needs and challenges of increasing globalization and intercultural communication, great efforts have been dedicated to developing bilingual or multi-lingual phrase-level parallel corpora. However, because building parallel corpora is time-consuming and work-intensive, most research on developing parallel corpora focuses on computer-based automation, such as using a computer system for mining parallel text from the Web on a large scale, (e.g. Resnik and Smith, 2003), determining the optimal phrase length with high levels of accuracy (Koehn et al, 2003), etc. There is very little discussion and research on techniques or strategies for manual creation of phrase-level parallel corpora, to say nothing of manual development of engine-specific phrase-level Chinese-English user parallel corpora.

There is no doubt at all that "automatic and semi-automatic techniques for lexical acquisition are more critical now than ever before as it becomes infeasible to produce adequate semantic representations on a large scale by human labor alone." (Bonnie Jean Dorr et al, 2001) Automatic or computer-aided methods do have their advantages, making it possible to create many large domain-specific parallel corpora with relatively less human intervention and considerably less time. But the dark side of the picture is that quite some of the corpora are not very accurate, thus yielding unsatisfactory machine translation result. That is why many machine translations lack both clarity and fluency at phrase or sentence level, which is more obvious in Chinese-English MT due to some unique Chinese language characteristics discussed below.

After going through more than half a century, Chinese-English machine translation has made some significant progress. Available translation engines are backed by extensive word-level dictionaries as well as phrase-level and sentence-level parallel corpora. However, as a whole, Chinese-English machine translation is still low in quality and accuracy due to difference in Chinese and English sentence structures, implicit meaning in both languages, lack of inflections in Chinese language, very different patterns and positions of prepositional and adjective phrases, etc. These linguistic differences pose challenges to Chinese-English MT, but at the same time, offer a chance for manually created parallel corpora to come in and patch the holes left by computer created parallel corpora. Fully aware there is no universal recipe for designing parallel corpora (Rura et al), this paper tries to present some techniques or strategies that might be used in manual development of engine-specific phrase level Chinese-English parallel corpora, hoping to trigger further research and discussion.

2. Engine-specific phrase level parallel corpora

Because of its good compromise between word and sentence level, less ambiguity, and flexibility (Zhao et al, 2005), domain-specific and phrase-level parallel corpora have now been widely accepted as the most efficient in machine translation. But why is it better for phrase-level parallel corpora to be engine-specific? To answer this question, let's first take a look at the translations of a Chinese sentence by 3 different MT engines: Systran, Google, and Microsoft.

Example 1

Source:
他引用新华社报道称,中国此次"陆基中段反导拦截技术"试验是在"中国境内"进行的。试验达到了预期目标,

HT:
He ***quoted*** a Xinhua news ***report saying*** that this test of Chinese "ground-based midcourse missile interception technology" was conducted "within Chinese territory". The test has achieved its anticipated goal .

Systran:
He ***quotes*** Xinhua News Agency ***to report said*** that China this time "ground-based midcourse missile interception technology" experiment is conducted in "China". The experiment has achieved the anticipated target .

Google:
He ***quoted*** the Xinhua News Agency ***reported*** that the "ground-based midcourse missile interception technology" test in China conducted. Test has achieved the expected goals.

Microsoft:
He ***cited*** the Xinhua News Agency ***reported*** that China's "ground-based Midcourse missile interception technology" test was "Republic of China". Test and achieved the desired objective.

The three different machine translations above demonstrate that words used by the three MT engines are literally correct in meaning and almost identical. But at the phrase and sentence levels, there are some shades of difference in meaning, indicating that each MT engine does have its own mechanism in parsing, programming, phrase and sentence organization, etc.

For example, all the words in the phrase "他引用新华社报道称" (he quoted a Xinhua news report saying) are translated, but none of the three engines got the correct parts of speech for the three likely verbs ("引用" /quote, "报道" /report, and "称" /say), nor did they organize these words correctly according to English grammatical rules. This is because it is too difficult, at least at present, for MT engines to correctly translate all the three likely verbs in such a short phrase. Systran translated all the three words as verbs while Google and Microsoft translated the first two and dropped the third.

The above example also suggests that, due to the difference in MT engine mechanism, we might need first to identify what is incorrectly translated by an MT engine and then create engine-specific phrase-level parallel corpora of frequently repeated phrases for each domain. Based on this concept, this paper is dedicated to discussing techniques or strategies for creating phrase-level user parallel corpora or user dictionaries for Systran MT engine.

3. Associating words of the same meaning

Parallel corpora, as we all know, contain aligned source and target text pairs that are exact translations of each other. However, Texts that convey the same information will exhibit great differences at the sentence level (Munteanu and Marcu, 2005) and even at the phrase level. The latter is especially true of Chinese phrases because: 1) many Chinese words have synonyms or variants; 2) form words or particles are often used in Chinese phrases and sentences; 3) different writing styles are used.

Example 2
2a.
Source:

从信息流动的角度***讲***, 现有的许多搜索引擎技术都是不可靠的。

HT: From the angle of information flow, many existing search engine technologies are not reliable.

MT without UD:
From the angle of information flow ***said that*** existing many search engine technologies were unreliable.

MT with UD:
From information flow's angle, many existing search engine technology is unreliable.

UD entries: 角度讲=angle, 现有的许多=many existing

Since this corpus works and 角度(angle)+ 说(say), or 谈(talk) are also phrases in good Chinese that express the same meaning, we might as well create similar corpora. In addition, the Chinese character 来(lai/come) can be inserted as a particle between 角度 and 讲, 说, or 谈 to form phrases of similar meanings. And therefore more similar corpora can be created.

2b.
Source:
从信息流动的角度***谈***, 现有的许多搜索引擎技术都是不可靠的。

Without UD:
From the angle of information flow ***discussed*** that existing many search engine technologies are unreliable.

With UD:
From information flow's ***angle***, many existing search engine technology is unreliable.

2c.
Source:
从信息流动的角度***说***, 现有的许多搜索引擎技术都是不可靠的。

Without UD:
From the angle of information flow ***said that*** existing many search engine technologies were unreliable.

With UD:
From information flow's ***angle***, many existing search engine technology is unreliable.

Association like this saves time in corpora development. Generally, association may include using synonyms of the same part of speech (e.g. 角度***讲*** and 角度***说***), changing word order, (e.g. many existing:***现有的***许多 and 许多***现有的***), eliminating or adding particles (e.g. operation procedure: 操作程序要求 and 操作程序***的***要求), changing meaning to the opposite (e.g. rockets with carrying capacity ***over*** 20 tons/20吨***以上***运载能力的火箭 and rockets with carrying capacity ***under*** 20 tons/20吨***以下***运载能力的火箭), adding or removing inessential words (e.g. the strictest ***最***最严格的/最严格的), eliminating redundancy or awkwardness (e.g. change with time/随着时间***的推移而产生***变化 and 随着时间而变化), etc.

In short, association in corpora creation can help generate more text pairs with similar or different meanings, thus increasing correction probability. This technique, in a sense, can be described as creating text pairs with synonymous or antonymous meanings.

4. Defining verbs by expansion

Chinese language, with the exception of very few pronouns that have declension, does not really have conjugation, declension, or other inflections, thus making it very different from English. What is more, many words can be used both as verbs and nouns, making it difficult for MT engines to determine their parts of speech, tense, and voice.

Chinese verbs usually consist of one or two characters and are always in their bare forms or root infinitives. Generally speaking, when a verb is used in the active voice and the sentence is short and simple, Systran engine together with its dictionaries can handle it fairly well, and in most cases can use the right tense. But, Systran engine fails to offer correct translation when the sentence is long and has several likely “verbs”, or when

there is a passive relation between the verb and its logic subject.

To expand the bare form or root infinitive of verbs by incorporating more words in corpora can improve translation because the expansion can serve one of the following three purposes or a combination: 1) expressing the correct tense and voice (e.g. 3c); 2) reorganizing word order (e.g. 3c); 3) defining part of speech.

But the expansion is suggested to be kept at a reasonable length. If a source phrase consists of too many words, its recurrence probability in MT translation is very low. And therefore, the length of a verb phrase, though highly flexible, is suggested to be about 8 Chinese characters. Below are some examples with the word develop/研制 which can be used as a noun or a verb, in the active voice or the passive voice, or in a phrase to serve as a predicative, adjective, or adverb.

Example 3:

3a. Develop/研制 as a noun or a verb

Source:
一项新产品的***研制***过程不可避免包含研究、试用的过程。(noun)

HT:
The ***development*** of a new product inevitably goes through research and trial.

MT:
A new product's ***development*** process contains the process of research and test inevitably.

Source:
美国军方正在***研制***一款名为“非洲猎豹”的机器人。(verb)

HT:
The US military is ***developing*** a robot named “African Cheetah.”

MT:
The US military is ***developing*** named “the African cheetah” robot.

3b. ***研制的*** as an adjective with a particle

Source:
中国***研制的***高速铁路试验车将于明年进行速度试验。

HT:
The high-speed test train ***developed*** by China will see a speed test next year.

Without UD:
China ***develops*** the high-speed railroad testing car will carry on the speed trial next year.

With UD:
The Chinese ***developed*** high-speed railroad testing car will carry on the speed trial next year. (with UD entry: ***研制的***=developed)

3c. ***研制*** as a verb in present continuous tense or as an adjective phrase

Source:
我国***正在研制***时速 400-500 公里的高速列车。(active voice)

HT:
Our country ***is developing*** a high-speed train of 400-500km/h.

MT:
Our country ***is developing*** a speed 400-500 kilometers high-speed train.

Source:
时速 400 公里的高速列车***正在研制***中。(passive voice)

HT:
A high-speed train of 400 km/h ***is being developed***.

MT:
A speed 400 kilometers high-speed train ***is being developed***.

Source:
正在研制的高速列车时速可达 400 公里。(passive relation with its logic subject)
HT:

The train ***being developed*** can reach a speed of 400 km/h.

MT without UD:
The high-speed train speed ***that developed*** may amount to 400 kilometers.

MT with UD:
The high-speed train ***being developed*** may amount to 400 kilometers. (with UD entries: 正在研制的高速列车= the high-speed train being developed)

Example 3c shows that Systran can use the correct tense and even voice of the verb 研制/develop when the sentence is short and its structure is simple. But when sentence length grows and sentence structure becomes complicated, Systran fails to fully express the conjugation and other inflection of its corresponding English verb. The above examples also show that MT quality is improved by incorporating bare verb forms with other words to express inflections.

5. Editing verb phrases to clarify meaning or to simplify sentence structure

A sentence is like a jigsaw puzzle consisting many pieces. The fewer pieces there are in a puzzle, the easier to solve the puzzle. In machine translation, long sentences are usually more difficult for translation engines as parsing gets more difficult. This is especially true of Chinese-English MT because Chinese language does not have rigid grammar like English, nor does it have inflections as mentioned above. Many ambiguities are caused by uncertain boundary of right association in parsing (Yoon-Hyung Roh et al), and verbs are among the most difficult in parsing because they require either an actual or a logic subject and sometimes an object as well.

Within a clause centered by a verb, Chinese mostly uses a SVO or SV structure like English (Yamada and Knight 2001). However, if there are multiple likely verbs or verb-centered clauses in a sentence, Systran engine often fails to determine which is the main verb and what should be the appropriate verb forms for the rest.

As the main verb usually plays a crucial role in conveying the meaning of a sentence (Ma et al), it is only natural that ambiguity and poor translation quality will be generated if the main verb of a sentence is not correctly recognized in parsing. On the contrary, if we give each verb its correct verb form in corpora, ambiguity can be greatly reduced, thus yielding better translation quality. Strategies here include eliminating a structural word that might serve as a verb in a Chinese phase but not necessarily in its English translation (4a), expressing passive verb forms (4b), merging verbs to eliminate a passive verb form (4c), minimizing verb forms in each verb-centered phrase (4d), etc.

Example 4:

4a.
Source:
时速***达***(reach)400公里的高速列车正在研制中。

HT:
A high-speed train of 400 km/h is being developed.

MT without UD:
The speed ***reaches*** 400 kilometers high-speed train to develop.

MT with UD:
The high-speed train of 400km/h is being developed.
UD entry: 时速***达***400公里=400km/h

4b.
Source:
F-22从1986年***开始研制***，直到2005年才装备部队。

HT:
The project of F-22started in 1986, but the fighter did not go into service until 2005.

MT without UD:
F-22 project ***started to develop*** from 1986, equipped the army until 2005.

MT with UD:
F-22 ***started to be developed in 1986,*** not until 2005 equips the army.
(UD entries: 从1986年开始研制, =started to be

developed in+number operator, 直到2005才= not until + number operator)

4c.
Source:
该机***开始研制的时间***大约是在上世纪九十年代中叶。

HT:
The development of the aircraft started approximately in the mid 1990s.

MT without UD:
This machine ***the time that started to develop*** approximately in the mid-1990s.

MT with UD:
This aircraft ***development started*** approximately in the mid-1990s.
(UD entries: 该机=this aircraft, ***开始研制的时间*** =development started)

4d.
Source:
他在***接受***(accept)本刊记者***采访***(interview)时***介绍***(introduce)***说***(say), 中国***发射***(launch)导弹***是***(is)防御性的。

HT:
In an interview with our reporter, he said China's missile launch is defensive in nature.

MT without UD:
He when ***accepting*** this publication reporter ***interviewed says***, China ***fired*** the missile ***is*** defensive.

MT with UD:
He said during an interview with our reporter, China's missile launch is defensive.
With UD entries: 在接受本刊记者采访时介绍说=said during an interview with our reporter, 中国发射导弹=China's missile launch

5. Compromise

The concept of machine translation came into being in 1940s. In the past 70 years, machine translation has seen many significant changes and improvements, but compared with human translation and from the perspective of translation quality, the former is still a crying infant in its cradle. Computer chess started in 1950s and now it can beat world champions. The difference is obvious and the reason is simple: though there are many possible options for each chess move, the options are limited after all. But machine translation is challenged by languages that have life and change all the time, dealing with multiple meanings of the source and the target languages, word order, phrase order, writing styles, tense, voice, OCR quality, etc.

Systran has a hybrid translation engine which combines the strength of rule-based and statistics-based engines. But it is a machine after all and can only translate in line with programmed rules and collected statistics. That is, they can only work mechanically and are not capable of changing the same word or phrase into different meanings according to context, especially when the same word or phrase plays different grammatical functions.

Systran-specific phrase-level user parallel corpora or user dictionaries can address considerable amount of translation errors, but in many cases they can only offer compromises until we work more closely with Systran developers.

Example 5

5a.
Source:
雷达通常***由发射机、发射天线、接收机、接收天线以及显示器组成***。

HT:
Radar is generally composed of a transmitter, transmitting antenna, receiver, receiving antenna, and display screen.

MT:
The radar usually ***is composed of the transmitter, transmitting antenna, receiver, receiving antenna as well as the monitor***.

5b.
Source:
由发射机 、发射天线 、接收机 、接收天线以及显示器组成的雷达在军事行动中起到很重要的作用.

HT:
Radar composed of a transmitter, transmitting antenna, receiver, receiving antenna, and display screen plays a critical role in military operations.

MT:
In the military action plays very vital role from the radar that ***the transmitter, the transmitting antenna, receiver, the receiving antenna as well as the monitor are composed***.

A text pair can be created for the phrase "由发射机 、发射天线 、接收机 、接收天线以及显示器组成 ". But if we consider different orders of the five components, according to the permutation formula, the available options are (5!)=120, not including its combination with other possible variations with form words such as using "和" or "及" instead of "以及". Obviously, it is not really feasible to create corpora for phrases like this. And therefore, the compromise sometimes we have to accept is an awkward sentence structure that can be corrected in post-editing.

A more effective way to solve this lexicalized pattern is an "operator" or a wildcard for the string between the Chinese characters "由(particle/you)" and " 组成(compose/zu cheng)". With "operators" or wildcards like that, Systran users will be able to create more advanced user dictionaries. What is worth pointing out is that this lexicalized pattern has many variations, which are not to be discussed in the interest of brevity.

Another compromise is accuracy. Take "found" and "establish" as an example. Though the two words have overlapping meanings, they do have shades of difference depending on their subject or logic subject. For example, the former generally refers to establishment with provision for continuing existence while the latter basically means bringing something into existence.

In Chinese, "成立/chengli" is usually translated as "found/founding" while"建立/jianli" as "establish/establishment". However, most Chinese writers prefer using "成立" (found/founding) even when "建立" (establish/establishment) would be more appropriate because the former sounds more formal such as: 成立应急小分队 (to found vs. to establish an emergency team), 成立信访处(to found vs. to establish an office of complaints and appeals), and 成立环境监察队(to found vs. to establish an environment inspection team). Systran has different translations for "成立", depending on the position of the word rather than its subject or logic subject. And due to reasons like this, sometimes the best thing that even engine-specific user parallel corpora can offer is just a compromise. But the compromise can be easily fixed or understood by non-linguist editors or analysts.

5c.
Source:
1959年, 中国第一支战略导弹部队"地地导弹营"成立。

HT:
In 1959, China's first strategic missile troops "surface-to-surface missile battalion" was established.

MT without UD:
In 1959, the Chinese first strategic missile force "surface-to-surface missile camp" establishment.

MT with UD:
In 1959, the Chinese first strategic missile force "surface-to-surface missile battalion" was established.
(UD entry: 成立。= was established.)

5d.
Source:
成立于 1959 年的"地地导弹营"是中国第一支战略导弹部队。

HT:
The "surface-to-surface missile battalion" established in 1959 is China's first strategic missile troops.

MT:
Was established in 1959 “surface-to-surface missile camp” was the Chinese first strategic missile force.

Source:
中国第一支战略导弹部队“地地导弹营”成立于1959年。

HT:
China’s first strategic missile troops “surface-to-surface missile battalion” was established in 1959.

MT:
The Chinese first strategic missile force “surface-to-surface missile camp” was founded in 1959.

The third compromise is paraphrase. Whether translation should be paraphrase or metaphrase is beyond the discussion of this paper. What we mean by paraphrase is using a phrase that conveys a similar meaning even when a better translation is available in human translation.

Dryden said in his brilliant essay On Translation: “But since every language is so full of its own proprieties, that what is beautiful in one, is often barbarous, nay sometimes nonsense, in another, it would be unreasonable to limit a translator to the narrow compass of his author’s words: ’tis enough if he choose out some expression which does not vitiate the sense.” Dryden was talking about poetry translation but what he said is also true of corpora, especially when source and target have very different sentence structures.

5e.
Source:
太空力量通过太空防御***确保不受***敌方进攻性武器的打击。

HT:
Through space defense, space force ensures that attacks of enemy’s offensive weapons will be prevented.

MT without UD:
The outer space strength ***does not guarantee*** through the outer space defense by the attack of enemy side offensive weapon.

MT with UD:
The space force eliminates the risk of the enemy side offensive weapon through the outer space defense the attack. (UD entries: 太空力量= space force, 确保不受= eliminates the risk of)

6. Conclusion

This paper proposes some strategies or techniques of creating phrase-level user parallel corpora for Systran translation engine. Because text-pairs are based on identified incorrect translations by Systran, the created user dictionaries significantly improve MT accuracy and fluency. In a test, sample translations were sent to non-Chinese speaking people for post editing, and all of them said the translations were easy to understand and came up with satisfactory versions with post-editing, showing engine-specific user corpora are an effective means to improve MT quality.

Though not all strategies and techniques discussed here will apply to other translation engines, the general concept might do. But that is subject to further study.

In addition to joint efforts from all MT users, better communication and coordination with translation engine developers will definitely lead to better engine-specific phrase level user parallel corpora.

Acknowledge

This article is based on Mr. David Barber’s concept of phrase-level user parallel corpora for Systran translation engine. His constructive advice, suggestion, and other input are greatly appreciated. Ms. Jin Yang, Systran computational linguist, has offered a lot of insight into Systran MT mechanism, which has greatly contributed to this paper.

Reference

1. Koehn, Philipp (2005). Europarl: A Parallel Corpus for Statistical Machine Translation, MT Summit 2005.

2. Bin Lu, Benjamin K. Tsou, Jingbo Zhu, Tao Jiang, and Oi Yee Kwong (2009), The Construction of a Chinese-English Patent Parallel Corpus. MT Summit XII, 3rd Workshop on Patent Translation August 25-30, 2009. Ottawa, Canada.
3. Yujie Zhang, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. Building an Annotated Japanese-Chinese Parallel Corpus–A Part of NICT Multilingual Corpora, pp.85-90
4. Chang Baobao (2002), Chinese-English Parallel Corpus Construction and Its Application. PACLIC 18, December 8th-10th, 2004, Waseda University, Tokyo, pp.283-290.
5. Resnik, Philip and Smith, Noah A, (2003) The Web as a Parallel Corpus, Computational Linguistics, No. 3, Vol. 29, pp349-380
6. Koehn, P., Josef Och, F., and Marcu, D. (2003) Statistical Phrase-Based Translation, proceedings of HLT-MAACL 2003 Main Papers, pp. 48-54.
7. Bonnie Jean Dorr and Gina-Anne Levow (2001) Construction of a Chinese-English Verb Lexicon for Embedded Machine Translation in Cross-Language Information Retrieval
8. Rura, L., Vandeweghe, W., Perez, Maribel Montero, (2007) Designing a Parallel Corpus as a Multifunctional Translator's Aid.
9. Zhao, J., Liu, F., and Liu, D. Two-Phase Base Noun Phrase Alignment in Chinese-English Parallel Corpora. Proceedings of 2005 IEEE International conference, pp.360-365.
10. Dragos Stefan Munteanu and Daniel Marcu (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora, Computational Linguistics, Volume 31, Number 4
11. Long Sentence Partitioning using Structure Analysis for Machine Translation (Yoon-Hyung Roh et al)
12. Kenji Yamada , Kevin Knight. 2001. A syntax-based statistical translation model, ACL
13. Wei-Yun Ma, Kathleen McKeown, Where's the Verb? Correcting Machine Translation During Question Answering (2009:333) (Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 333–336, Suntec, Singapore, 4 August 2009)
14. John Dryden "On Translation" (Theories of Translation 1992:20-21)