Experiments of Chinese-to-English Machine Translation

1 Introduction

Chinese and English are different in many ways and here are several key differences that affect our translation system.

1. *Segmentation*: Chinese uses a logographic system for its written language, where symbols represent words. Unlike English, there's no space between words. Therefore, in order to translate Chinese into English, the first thing we need to do is to determine the boundaries of words, i.e. segmentation.

2. *Tenses*: In Chinese, tenses are usually inferred instead of marked. Sometimes there are clauses that describe time. Sometimes character "了" indicates the completion of an action (e.g. "我去看电影了" means "I went to watch a movie"), but not always (e.g. "价格太高了" means "The price is too high"). Each verb in Chinese has only one possible form, whereas in English verbs have many forms. Thus one of the challenges is to determine the right forms of verbs during the translation.

3. *Nouns and articles*: In Chinese, there's no strict distinction among plurals, singular forms, and uncountable nouns etc. Articles are usually optional. For example, it's ok to just say "苹果" (apple), or if one wants to precise: "一个苹果"(an apple). There's no character that functions as "the" in Chinese. In English, whether the subject is in plural forms or not affects the following verbs. Thus we need to add appropriate articles and try to keep the verb forms consistent.

4. *Phrases ordering*: In Chinese, postpositional phrases are used together with noun phrases, while in English, prepositions are used. For example, "在窗外" means "outside the window" but the direct translation is "window outside". So we should detect such patterns and reorder the noun and adpositional phrases. Meanwhile, relative clauses in Chinese are often before nouns, usually connected with "的", while in English they come after. For example, "在社会中的地位", which literally means "in society status", should be translated as "status in society".

5. *Stylistic and cultural differences*: In Chinese, four-character phrases are very compact and powerful expressions. If translated word by word, the meanings of those phrases tend to be strange and obscure. For example, "有所取舍", whose direct translation would be "somewhat make a choice", actually means "to make trade-offs". Besides, "when" in Chinese is often expressed in two-phrases such as "当...时" or "当...的时候". During translation, the second part could be omitted and we can just translate "当" as "when". There are many such differences, which are relatively hard to generalize and thus need special attention.

2 Working Corpus

Dev set:
1) 人们多通过社会比较，来 确定自己的优势，和在社会中的地位，他认为幸灾乐祸就来源于社会性比较。 (from http://www.guokr.com/article/437875/)
2) 有所取舍，不仅是差异化的标志，更是优秀的标志。 (from http://read.douban.com/ebook/2800884/?dcs=book-hot&dcm=douban&dct=read-subject)
3) 所以她就决定，早晚要打我一个耳光。 (from Collections of Xiaobo Wang)
4) 诗，就是以最少的语言，去表达最多的文字。 (from http://book.douban.com/review/3345946/)
5) 在漫画中，习近平穿着灰色夹克和蓝色裤子。 (fromhttp://news.ifeng.com/mainland/detail_2014_02/20/33983439_0.shtml)
6) 这是中国社会更自信和更开放的一种姿态。 (from http://news.ifeng.com/mainland/detail_2014_02/20/33983439_0.shtml)
7) 放弃美国国籍是个人的选择，但富人要放弃美国国籍则是需要付出代价的。 (from http://blog.sina.com.cn/s/blog_5d8d68c10102ef1r.html?tj=1)
8) 许多中小学、大学都只强调了知识的传授，而忽视了对于学生人格的培养。 (from http://blog.sina.com.cn/s/blog_7159859d0102f6f8.html?tj=1)
9) 所以我认为现行的音乐定价模式是有问题的，单首歌曲的价格太高了。 (from http://erdong.baijia.baidu.com/article/4613)
10) 当我们面临许多方法时，这些方法不是简单的对或者错，捷径或者弯路。 (from http://www.zhihu.com/question/22790970/answer/22665239)

Test set:
1) 他们发现，随着时间推移，听起来明显是黑人的名字越来越能可靠地表明孩子的社会经济背景。 (from http://www.guokr.com/article/437940/)
2) 每个人都可能是创业者，而这些专业级的分享，会让创业者不再孤单。 (from http://read.douban.com/ebook/2595517/?icn=index-rec)
3) 设计不仅意味着风格和品味，还融入了文化与个性。 (from http://read.douban.com/ebook/2593494/?icn=profile-guess)
4) 他向当地媒体指出，法律没有明文禁止成立这类组织。 (from http://www.bbc.co.uk/zhongwen/simp/china/2014/02/140220_hunan_gay_lawsuit.shtml)
5) 那么，现金怎么花出去，让它变得更有价值，就成为最关键的问题。 (from http://fusheng.baijia.baidu.com/article/4680)

3 Translation System

Pre-processing strategies:

1) *POS tagger*: we use the library “jieba” in python which focuses on Chinese segmentation to divide our Chinese sentences into reasonable segments and it also helps us tag each word in the segmentation. Using POS tagger can help us determine which entry in the dictionary we will use in our translation. This strategy applies in all the sentences in both sets.
2) *Dictionary building*: we built our dictionary not only from the segmentation of our corpus, but also connected consecutive words to see whether the combined one is meaningful since the

direct translation of two words sometimes is meaningless. For example, “自己“ means “oneself” and “的” means “of”, but “自己的” means “own”, so “own” will be more sensible here. So we will add “自己的“ into our dictionary. And we also add the adjective form of a noun word into the dictionary. This strategy extracts the combined word from sentence (1)(2)(4)(6)(7)(10) in the dev set, sentence (2)(5) in the test set.

3) *Remove unnecessary words*: In Chinese, some words function as auxiliary words and have no meanings. For those words, we can remove them from our sentence before translation. For example, “了” has no meanings in most cases, so we can remove it. This strategy applies in sentence (8)(9) in the dev set and sentence (3) in the test set.

4) *Remove redundant words*: In Chinese, there may exist more than one word expressing the same meaning in a sentence. In this case, we should only retain only one word in our sentence. For example, “当...时” means “when”, but “当” has already express the meaning, so we no longer need the word “时” in our sentence. This strategy applies in sentence (10) in the dev set and no sentences in test set.

5) *Special words correction*: Because some words in Chinese can have multiple meanings, so we should take care of them. “来” and “去” are translated in to “come” and “go” in general, but if they follow a verb or a preposition and precede some verb, then their meanings should be changed to “to” in most cases. This strategy applies in sentence (1)(4) in the dev set and no sentences in the test set.

6) *Merge words*: as mentioned in “Dictionary building” strategy, we will have a more reasonable word sometimes if we merge two consecutive words after segmentation. In this strategy, we traverse a sentence and check each pair of two consecutive words can be combined into a word that exists in the dictionary, and if there is any, we replace the original two words with the combined one. This strategy applied in sentence (1)(2)(4)(6)(7)(10) in the dev set and sentence (2)(5) in the test set.

7) *Preposition reordering*: Chinese has locative post-positions while English has prepositions. So we detect the occurrence of a word representing location (which is marked ‘f’ in our POS tagger system) following a noun but no noun following the preposition, and we put the word before the noun, change the mark of it to ‘p’ which means preposition and also remove the redundant words. For example, ”在社会中“ will be translated into “in society middle” in the directed translation. But after reordering, we have “in society”. This strategy applies in sentence (1)(5) in the dev set and no sentences in the test set.

8) *Subject reordering*: In Chinese, “A的B” means B of A, but the direct translation will get A of B. So we detect the pattern of “noun of noun” and exchange the two noun found. This strategy applies in sentence (1)(2)(8)(9) and sentence (1) in the test set.

9) *Adjective reordering*: sometimes the direct translation will have an adjective after a subject, so we can detect such pattern and correct it. For example, the direct translation will be “people many…” so we should change the order to make it sensible. This strategy applies in sentence (1) in the dev set and no sentences in the test set.

Statistical Language Model

1) *Ngram Model*: In order to distinguish between multiple candidate translations in the dictionary, we used the nltk Ngram Model trained on the Brown corpus. When we are able to find the part of

speech of a Chinese word in its dictionary entry, the candidates are limited to those in the dictionary that have the same part of speech. If we cannot find the part of speech, we let all translations in the dictionary be candidates. Before the language model, we would just choose the first translation. Thus, the language model plays a huge part in picking a decent translation to use and, as a result, gives non-trivial improvement in all the sentences that we translate. This is especially true in the case where we can't find the corresponding POS. If we didn't have the language model, we could be picking a translation that has a very obscure definition and POS that doesn't make sense in the context. With the language model, we can at least guarantee that we pick a common translation, which gives us a higher probability of having a translation that makes sense both in context and as a translation for the original word. To be more precise, the language model gives us a better chance of picking a translation that is both more fluent and more faithful.

2) *Noun/Pronoun/Verb Classification*: Due to the lack of multiple written forms for a single Chinese character, it is very hard to determine how nouns, pronouns, and verbs should be translated. Specifically, for nouns, it is difficult to decide whether the noun should be singular and plural. Sometimes, one can tell due to characters like "们" and "多", but in many cases, there is no demarcation. For pronouns, there is no indication of whether the pronoun is in its subject form, object form, or possessive form. One must discern this from the context of the sentence. Lastly, for verbs, it is very hard to determine what tense a verb should be because there are usually no marking characters. As mentioned above, the only occasional marker for time is "了", and that marker is not very consistent in meaning. Before classification, there was very little consistency among sentences with nouns very often not matching the verbs at all. For instance, a couple of examples from the dev set before classification:

> "So she then determine, sooner or later want make I a slap on the face."
> "In caricature center, Xi Jinping wearing gray jacket and blue pants."

In order to do classification, we first keep track of the plurality and person of the most recent noun/pronoun. Plurality is determined by the markers mentioned above, and person is determined by whether it contains the Chinese for "I" or "you". From there, we choose the tense of the verb based on the characteristics of the most recent noun/pronoun. To be more specific, if the verb is say "eat" and the last noun was singular and third person, we would choose "eats". For pronouns, if it is followed by the Chinese for "of", we use the possessive. Otherwise, we use the object form unless it is the first noun of the sentence. Since we may not always get the noun classification correct, we can end up using incorrect verb tenses. However, we were most likely not going to get the noun and verb forms correct before classification. Further, the classification provides a consistency among nouns, pronouns, and verbs that was definitely not there beforehand. Thus, this strategy provides us with much more fluent translations that are also often more faithful due to the fluency. This strategy was general and significant and allowed for non-trivial improvement in all sentences. Here are the same two sentences from above after:

> "So she then determines, sooner or later wants makes me a slap on the face."
> "In comics center, Xi Jinping wears gray jacket and blue pants."

Post-processing strategies:

1) *Modal Verbs Check*: In English grammar, verbs following modal verbs such as “can” and “must” should be in their original forms, regardless of the tenses. This is also true for “to” when followed by verbs in most cases, except when “to” is used as preposition, such as “apply methods to improving the situation”. In the previous strategies, such as language models and tenses determination, this rule is not guaranteed. Therefore, we’ve included this strategy after the translation, which checks the verbs following, not necessarily immediately, the modal verbs and “to”. In the dev set,

Poetry, is with minimal word, to expresses most character.

is then corrected into

Poetry, be with minimal word, to express most character.

Similarly, in the test set, we found that:

Each people entirely may be entrepreneur…

where “be” is in the correct form. This strategy applies to (1)(3)

2) *Plurality Check*: To complement the strategy that determines plurality of nouns mentioned above, we’ve added this post-processing strategy to further improve the precision. After some words are translated in English, they might indications of plurality. For example, “these” translated from “这些” should be followed by nouns in their plural forms. With this strategy, in the dev set,

When we are confronted with many way, these way no simple yes or wrong, shortcut or roundabout.

is corrected as:

When we are confronted with many ways, these ways no simple yes or wrong, shortcut or roundabout.

4 Final output

1) They discover, along with the passage of time, sounds obvious is first name of black people more can reliably show society of child economy background.
2) Design not only mean style and taste, also integration culture and the individuality.
3) Each people entirely may be entrepreneur, and these professionals of shares, can make entrepreneur no longer alone.
4) He to local medium indication, law have not expressly prohibited sets up this type of organization.
5) Like that, cash how spends goes out, makes it became more valuable, then became most problem of key.

5 Comparison with Google Translate

1) They found that, over time, it sounds more and more obvious is the name of Negro can reliably indicate the child's socio-economic background.

Comment: Google translates “随着时间的推移” as the common phrase “over time”, which is much more succinct and fluent in English. Our translation, though grammatically correct, is obviously too verbose due to word by word translation from Chinese. “Black people” is probably more polite than “Negro”. “越来越” (more and more) should be describing “reliably”, this is not expressed in both translations. Google Translate does better in ordering the phrases, like: the child's socio-economic background.

2) Design does not only mean style and taste, but also into the culture and personality.

Comment: Clearly Google Translate has got the verb form and negation right, and ours failed. But both failed to translate “融入了” as “integrates”. Ours has found the right word with the right meaning, but not the right form. Google Translate uses “into”, which is a preposition and clearly doesn’t fit in this context. “Personality” is the right word to use here. However our language model has chosen “individuality”, which should be a less common word.

3) Everyone may be entrepreneurs, and these professional-grade sharing, entrepreneurs will no longer be alone.

Comment: Since our segmenter decides to treat “每个” and “人” separately, our system outputs “Each people” rather than the correct “Everyone”. “都” here doesn’t have much meaning here, and Google is right to neglect it. Both Google Translate and our system got the plurality half right, where the first “entrepreneur” should be in singular form because of “Everyone”, the second “entrepreneurs”. “professional-grade sharing” is the right meaning and forms, whereas our system failed to realize “专业级的” is an adjective phrase and “分享” means the action of sharing.

4) He pointed out to the local media, the law does not expressly prohibit the establishment of such organizations.

Comment: Google correctly translates the Chinese phrase “向...指出” as “point out to...”. Our system has tackled similar issues, but is not comprehensive enough to accommodate more cases like this. Google also predicted the correct tense. Our translation failed to keep the subject and verb consistent, “law have” should be “law has”. Google has a better understanding of the grammatical structures, so it knows after verb “prohibit”, there should be a noun phrase. Whereas our system is not able to detect that. “这种” is literally closer to “this type of”, but “such” is definitely more fluent in English.

5) So, how to spend the cash to go out, it becomes more valuable, it becomes the most critical issues.

Comment: “那么” should be translated as “so” in this context. Since our system is based on a unigram model, it’s less context-aware. “How” should be placed in the beginning of the clause. Our system doesn’t consider this ordering issue. Both Google Translate and our system haven’t

recognized this cultural difference, that in Chinese, things are spent “out”. In English, this directional phrase is unnecessary. “花出去” should be translated altogether as “spend”. Google Translate has arbitrarily omitted “让”, and our system has translated as “makes”. Here the correct translation should be “how to spend the cash, and how to make it become more valuable”. Our segmenter and dictionary failed to recognize “关键的” as one adjective of “critical”. Thus “critical issues” should be a better translation.

6 Error Analysis

The more reasonable translation of the first sentence in the test set should be “They discover that if a name that sounds like Black’s obviously can show the child’s social and economic background more reliably.”

1) The first error is for the “name” in the translation. The translation of our system didn’t reorder the words those function as modifier, e.g. “sounds like Black’s obviously” should be used to describe the word “name”, but the machine translation didn’t detect that. To fix the error, we should add a function to detect the subject of a clause and identify the other words describing our subject, and the translation could be “Subject that…”.

2) Another error in this sentence is that we exchange the position of “child” and “society” by the pre-processing strategy (8) because we didn’t include the definition of “孩子的”, which means “child’s”, in our dictionary according to the “Dictionary building” strategy. The way to fix the problem is to make our dictionary more complete to include different versions of a word as many as possible. If we have “孩子的” in our dictionary, the translation would not make the mistake.

3) And in this sentence, we should use “name” instead of “first name”. The problem is that we used Brown corpus to train our language model and then use it to decide which translation we should use. Maybe the frequency of “first name” is bigger than “name”, but the “名字” in Chinese means “name” in most cases. Instead we should use a corpus which is bilingual and indicates the corresponding translation of Chinese word to train our language model and find the most likely translation of a word rather than translating a Chinese word firstly and then use English corpus to find which translation is better.