CS124 PA6 - Machine Translation

February 28, 2014

1 Introduction

Our project is a Chinese-English MT system. Chinese is a language that is vastly different from English. Here are some challenges that we have encountered during the development.

1. Chinese does not have spaces between words, so there is an extra step of identifying the words in the sentence. There are several automated methods for tokenization, but none of them are perfect, so we have one extra source of error.

2. Chinese uses almost no word morphology, so it is hard to tell if a noun is a subject or an object, or if a verb is in present or past tense. Therefore a direct translation almost always results in a wrong form of the verb. We can only conjugate later on by inferring the correct form using context. On the other hand, many Western languages make use of morphology, so the translation to English is a bit easier.

3. Inferring plural form is also hard. Chinese uses a special type of word ”measure word” to quantify the nouns, but this can not be directly translated to English, because English does not have such measure words. Some nouns are not countable in English, but Chinese quantify them with a measure word. Sometimes, such translation is even hard for human to do right.

4. 4-character phrases are common and popular in Chinese. Those phrases usually have a story behind them and have a complicated meaning. Translating them precisely would require the knowledge of the context, which is hard for us to implement in this project.

5. The punctuation usage is also different in Chinese. Chinese uses period loosely and usually uses that to separate sentences that are logically different. Whereas in English, two sentences with two subject must be separated by period or similar punctuation. Thus in Chinese we often see very long sentences if we define sentence boundary with period. During translation, we need to oftentimes break one Chinese sentence into several English sentence to make them grammatically correct.

6. Omitting subject is common in Chinese, because Chinese grammar is not as strict as English. As long as the meaning can be inferred, sometimes the subject can be omitted or even implicitly changed in a couple of sentences. This is very problematic, because English requires subject in every complete sentence. A correct translation would have to infer the subject from the context and insert them into the English text.

Some key insights that we have gained during the translation are as follows.

1. Chinese has a certain set of words that indicate the tense of the action. By looking for those words, we can identify the tense pretty confidently.

2. Some structures such as ”one of the most ...” is always expressed in the opposite order in Chinese. We can easily match those structures and get a decent translation by reversing the order.

3. Sometimes a Chinese word can be translated into either a verb or a noun in English. Since Chinese has the same subject-verb-object model as English, if tag the Chinese sentence first, we can have a better idea of what to choose when there is ambiguity.

2 Working Corpus

The source of the corpus is as follows.

1. http://baike.baidu.com/view/13725.htm
2. http://baike.baidu.com/view/5051.htm
3. http://vsac5.blog.163.com/blog/static/68753737200822106115571/

2.1 Dev set

1. 斯坦福大学，是一所坐落于美国加利福尼亚州斯坦福市的私立研究型大学。
2. 该校位于硅谷的西北方邻近帕罗奥图。
3. 斯坦福培养了不少著名人士，并因学术声誉而获评为世上最著名的高等学府之一。
4. 大学辖下目前共有7所学术学院。
5. 1901年12月，物理学家伦琴获得了诺贝尔奖。
6. 斯坦福校友创办了众多著名的公司机构，如：谷歌、雅虎、惠普、耐克、升阳电脑等。
7. 马丁路德手写原稿保存在该校。
8. 斯坦福长期以来也是最难入读的高等学府之一。
9. 在36631名2016年的本科课程申请者当中，只有2427人获得录取，是全国第二选择性最高的大学。
10. 凭借学校的人才优势和丰富的资金、设备条件，斯坦福大学在科学研究方面成为了最好的机构之一。

2.2 Test set

1. 位于谷歌西北方的斯坦福大学创校时并不十分有名。
2. 1891年10月1日，斯坦福大学举行了开学典礼。
3. 其校友涵盖30名富豪企业家及17名太空员，亦为培养最多美国国会成员的院校之一。
4. 硅谷有很多重要组织，比如：斯坦福大学、谷歌等。
5. 1924年3月,老斯坦福出生在最富裕的农场主家庭之一。

3 Output of the system

1. Stanford University, is a located in Stanford California United States private research University.
2. This school locating in the Northwest of Silicon Valley closes to Palo Alto.
3. Stanford trained many famous people, and because academy reputation was rated is world one of the most famous advanced university.
4. University department currently totally has 7 academy colleges.
5. December 1901, physicist Rontgen got Nobel prize.
6. Stanford alumnus established numerous famous companies organizations, such as Google, Yahoo, HP, Nike, Sun Microsystems, etc.

7. Martin Luther handwriting originals save in this school.

8. Stanford long time since also is one of the most difficult enrolling advanced university.

9. In 36631 2016 undergraduate course applicant among, only 2427 people gets admits, am entire nation second selective the highest University.

10. Relying on school talented person superiority and rich funds, equipment condition, Stanford University in one of science research aspects became the best organization.

11. Locating in the Northwest of Google Stanford University found school time not at all very famous.

12. October 1 1891, Stanford University held opening ceremony.

13. Its alumnus cover 30 magnate entrepreneurs and 17 astronauts, also is train one of the most United States congress member institutions.

14. Silicon Valley has many important organizations, such as Stanford University, Google, etc.

15. March 1924, old Stanford is born in one of the richest farmer family.

4 Corpus Segmentation and POS Tagging

Chinese is written without spaces between words. Thus, tokenization of the Chinese raw text is a pre-processing step for machine translation. We used the Stanford Word Segmenter to tokenize our corpus. The segmenter is available for download at http://nlp.stanford.edu/software/segmenter.shtml.

The POS tagging is to read text in some language and assign parts of speech to each word, such as noun, verb, adjective, etc.. After tokenizing the raw Chinese sentences in our corpus, we used Stanford Log-linear Part-Of-Speech Tagger to tag each Chinese words in the corpus. The tagger is available for download at http://nlp.stanford.edu/downloads/tagger.html.

5 Dictionary

We created a bilingual Chinese-English dictionary for each word in our working corpus. The corresponding translation for English was retrieved from Google Translate. For each English word, we also labeled it with its POS tag. The dictionary was stored in dictionary.txt.

6 Translation System

The first part of the system is try to pick the most appropriate English words for the segmented Chinese sentence, there are two steps. Notice these two rules are very general rules, it actually take effects on every word of sentences. However it only take effects on certain words.

6.1 Pick Word By POS tag result

A Chinese word maybe translated to multiple English words. Baseline just randomly choose a word. In the postprocessing system, we use the POS tag result to choose the appropriate word. We annotate the word in our dictionary with the word type like the real dictionary. And we choose English words with the same tag as the Chinese word. Later, it is also being used to remove the measure word which English doesn't really have.

Where the rule applies:

- Dev Set: Sentence 4 - 大学辖下:”University in” was translated to ”University Department”; Sentence 6 - 等等: ” “Sun Microsystems, wait” ”Sun Microsystems, etc.”

- Test Set: Sentence 4 - 比如: ”For example” was translated to ”Such as”

6.2 Applying bigram language model

If there are still more than one option after filtered by the POS tag, we use a bigram model to choose the most possible word to fill in by consulting the bigram dictionary with distribution frequency. Where the rule applies:

- Test Set: Sentence 1 - 创校:”create school” was translated to ”found school”

*The bigram rule only took effect on one word in our corpus. However it is still a very general rule.

6.3 Reorder date representation

The way Chinese represents date is different from English. In Chinese, Year always comes first, then Month, and then Day; whereas English uses a Month-Day, Year pattern. This can be easily translated by matching the pattern with regular expression.

Where the rule applies:

- Dev Set: Sentence 5 - ”1901年12月”
- Test Set: Sentence 2 - ”1891年10月1日”; Sentence 5 - ”1924年3月”

6.4 Correct listing

The baseline translation for listing items is not good. Although both language uses the same pattern for listing (A, B, and C etc.), the direct translation will pick the wrong word for ”etc” because the Chinese word for ”etc” also means ”wait”. It’s also more natural to use ”such as” for short items instead of ”for example”. Therefore we use a regular expression to match ”for example A, B, and C wait.” and convert it into ”such as A, B, and C etc.”

Where the rule applies:

- Dev Set: Sentence 6 - ”如：谷歌、雅虎、惠普、耐克、升阳电脑等”
- Test Set: Sentence 4 - ”比如：斯坦福大学、谷歌等。”

6.5 Superlative

In Chinese, superlative is always expressed by appending ”最”(most) before an adjective or adverb. Therefore the direct translation method will not produce the correct superlative form, e.g. ”most rich” instead of ”richest”. This rule fixes this by finding adjective or adverb with a preceding ”most” then convert it into the correct superlative form with NLTK package.

Where the rule applies:

- Dev Set: Sentence 3 - ”最著名”; Sentence 8 -””最难”; Sentence 9 - ”最高”
- Test Set: Sentence 2 - ”最多”; Sentence 5 - ”最富裕”

6.6 Reorder ’one of’

It is very unlike English, Chinese append ”之一” in the end of a plural noun phrase to express the concept of ”one of” we set some simple rule to move one of the front of the noun phrase. The simple rule try to dictate the nearest boundary of the plural noun phrase like ”the” or superlative words. Where the rule applies:

- Dev Set: Sentence 3 - ”最著名的高等学府之一”; Sentence 8 - ”最难入读的高等学府之一”; Sentence 10 -”最好的机构之一”
- Test Set: Sentence 3 - ”最多美国国会成员的院校之一”; Sentence 5 - ”最富裕的农场主家庭之一”

6.7 Inferring past tense

In Chinese, the concept of past tense is not really existing. However, usually Chinese use "了" as an indication fo the action happened already, therefore, we put every verb preceding a "了" in past tense by using the pattern.en package. Where the rule applies:

- Dev Set: Sentence 3 - "培养了"; Sentence 6 - "创办了"
- Test Set: Sentence 1 - "举行了"

6.8 Conjugating verb in subclauses

In the direct translation, it couldn't tell the subclauses apart from the main sentence. We observe some sentence have two verbal phrases. Since it is hard for us to made them to be actual cause, we conjugate the verb in the subclause to made the clause to be a noun phrases instead.

The method used the POS tag result. If a word in wordnet doesn't have any lexical name other than verb and word is not being tagged as a verb, then we use the present form of the verb instead of infinitive form by using the pattern.en package. Where the rule applies:

- Dev Set: Sentence 2 - "位于" was translated to "locating in"; Sentence 10 - "凭借" was translated to "relying on"
- Test Set: Sentence 1 - "位于" was translated to "locating in"

6.9 Pluralization

Since the Chinese language does not mark plurals, one of our post-processing strategy is to detect where we need to pluralize a corresponding English word and then change this word to its plural form. We used the CLiPS package to pluralize English words. The package is available for download at http://www.clips.ua.ac.be/pattern.

With pyStatParser, we generated a parse tree for each English sentence. We analyzed the parse tree, and found that if there are a Cardinal number before a noun, and the number is larger than one, then we need to pluralize the noun. Another similar case is that a noun's adjective word is 'many', 'numerous' and so on.

Where the rule applies:

- Dev Set: Sentence 3 - "斯坦福培养了不少著名人士"
  In this sentence, the rule changes "Stanford trained many famous person" to "Stanford trained many famous people".
- Dev Set: Sentence 4 - "7所学术学院。"
  In this sentence, the rule changes "7 academy college" to "7 academy colleges".
- Dev Set: Sentence 6 - "众多著名的公司机构"
  In this sentence, the rule changes "numerous famous company organization" to "numerous famous companies organizations".
- Test Set: Sentence 3 - "30名富豪企业家及17名太空员"
  In this sentence, the rule changes "30 magnate entrepreneur and 17 astronaut" to "30 magnate entrepreneurs and 17 astronauts".
- Test Set: Sentence 4 - "很多重要组织"
  In this sentence, the rule changes "many important organization" to "many important organizations".

6.10 Removing Measure Words

Another difference between English and Chinese is that Chinese requires a measure word for every noun. For example, in English we can say, “three cars”, but in Chinese, we need to say “three (measure word) cars”. There are more than a hundred Chinese measure words. Moreover, some Chinese measure words have other meanings. For instance, “one university” in Chinese is “一所大学”, where “一” is “one”, “大学” is “university” and “所” is a measure word. But “所” also have other meanings, such as “place”, “office” and “spot”. Thereby, one of our pre-processing strategy is to detect which Chinese words are measure words, and then remove them in the translation. For all the measure words in our corpus, besides listing their other meanings, we also marked them with a special label “M” to indicate that they could serve as measure words. During translation, we assume a word is a measure word if it is marked ”M” in the dictionary, and it follows a cardinal number in the sentence. In this case, we remove this measure word from the sentence.

Where the rule applies:

- Dev Set: Sentence 1 - ”一所坐落于美国加利福尼亚州斯坦福市的私立研究型大学”
- Dev Set: Sentence 4 - ”7所学术学院”
- Dev Set: Sentence 9 - ”在 36631名2016年的本科课程申请者当中”
- Test Set: Sentence 3 - ”30名富豪企业家及17名太空员”

6.11 Arranging Place Names

In Chinese, big place name usually comes first. Thus, we often see such order, Country, State/Province, City, Street name, Apartment number. So one of our post-process is to rearrange these place names according to English grammar. Our translation system detects these place names, and analyze which place categories they belong to. Then, based on this information, the system rearranges the order of these place names.

In Chinese, positional nouns can also be used after some nouns to indicate position. For example, ”在硅谷的西北” which can be directly translated into ”in Silicon Valley Northwest”. But this is not correct in terms of English grammar. Therefore, we forward these positional nouns before the place names. So the above example phrase is translated to ”in the northwest of Silicon Valley”.

Where the rule applies:

- Dev Set: Sentence 1 - ”美国加利福尼亚州斯坦福市”
  In this sentence, the rule changes ”United States California State Stanford City” to ”Stanford, California, United States”.
- Dev set: Sentence 2 - ”位于硅谷的西北方邻近帕罗奥图”
  In this sentence, the rule changes ”located in Silicon Valley Northwest close to Palo Alto” to ”located in the Northwest of Silicon Valley closes to Palo Alto”.
- Test Set: Sentence 1 - ”位于谷歌西北方” In this sentence, the rule changes ”locate in Google Northwest” to ”Locate in the Northwest of Google”.

6.12 Changing One to a/an

In English, there are some situations where we need to use ”a/an” instead of ”one”. Thereby, we have a post-process to deal with it. The system detects such situations where a cardinal number ”one” is followed by a noun, and then changes ”one” to ”a/an” based on the noun.

Where the rule applies:

- Dev Set: Sentence 1 - ”一所坐落于美国加利福尼亚州斯坦福市的私立研究型 大学。”
  In this sentence, the rule changes ”is one located in United States California State Stanford City private research University.” to ”is a located in United States California State Stanford City private research University.”

7 Comparison with Google Translate

- Ours: Locating in the Northwest of Google Stanford University found school time not at all very famous.
- Google: Located north-west of Google when founding of Stanford University is not very famous.
- Compare: Google use ”located” instead of ”locating” and it add a verb to the main sentence. It also picks no word for ”十分” since not very famous is enough for expressing this meaning. And it add ”when” to construct a subclause for ”found school”.

- Ours: October 1 1891, Stanford University held opening ceremony.
- Google: October 1, 1891, the opening ceremony was held at Stanford University.
- Compare: Google use a more common expression for the sentence by changing it to passive.

- Ours: Its alumnus cover 30 magnate entrepreneurs and 17 astronauts, also is train one of the most United States congress member institutions.
- Google: Its alumni covers 30 richest entrepreneurs and 17 astronaut, is also one of the institutions up to train members of the U.S. Congress.
- Compare: It seems Google use incorrect word ”up to” for Chinese word ”最” and didn't figure out it should use the superlative form.

- Ours: Silicon Valley has many important organizations, such as Stanford University, Google, etc.
- Google: Silicon Valley has many important organizations, such as: Stanford University, Google and so on.
- Compare: two sentences are actually very similar except 'etc' v.s. 'so on'

- Ours: March 1924, old Stanford is born in one of the richest farmer family.
- Google: March 1924, the old Stanford was born in one of the richest farmer families.
- Compare: Google figure out the past tense correctly by looking at the time information. It also pluralize ”family” while ours are not.

8 Error Analysis

1. The Chinese particle de(的) is most commonly used as a possessive modifier. It can be used between two nouns to indicate a relationship of possessor/possession. However, as this character is used in many places, we found it is difficult to identify what kind of relationship the character is indicating. For example, ”我的” is ”mine”, ”你的” is ”your”, ”马丁路德的” is ”Martin Luther's”. But in other situation, the relationship is different. For instance, ”硅谷的西北” is ”the northwest of Silicon Valley”. One of the errors our system made is failed to identify the ownership relation. The system translates ”马丁路德的手写原稿” to ”Martin Luther handwriting originals”, whose correct translation is ”Martin Luther's handwriting originals”. One way to improve this would be identifying the owner object, which usually is people's name, or a pers pron. Then, recognize if the relationship is ownership.

2. The English language has several subordinate clauses, which the Chinese language does not have. Thus, it is difficult for us to build subordinate clauses based on Chinese sentences. For example, ”斯坦福大学，是一所坐落于 美国加利福尼亚州斯坦福市的私立研究型大学。”, is ”Stanford University is a private research university, which is located in Stanford, California, United States.” But our system translates

it to ”Stanford University, is a located in Stanford California United States private research University.” Simple rules cannot analyze semantic meanings of a sentence, but we have to extract some parts of a sentence to build a subordinate clause based on the semantic meaning. This error also occurs in Google translation. We believe that in order to solve this error, the machine translator must be intelligent enough to understand the semantic meanings.

3. English has several ways to express tense. The most common are verb conjunctions which change the form of the verb depending on the time frame. However, Chinese does not have any verb conjugations. All verbs have a single form. For example the verb ”吃” (eat), can be used for the past, present, and future. Our system can only change the form of the English verbs to past tense in some simple cases, which was discussed in the translation system section. In most cases, we have to analyze the semantic meaning of a sentence to recognize the time frame. For example, ”老斯坦福出生在最富裕的农场主家庭之一” which is ”Old Stanford was born in one of the richest farmer family”, but our system translates it to ”Old Stanford is born in one of the richest farmer family”. Our system cannot understand that ”is born” is used for the past. Similar to the previous error, solving this error also require the machine translator to be intelligent enough to recognize the time frames.

4. Passive voice is not correctly translated. For example, ”原稿保存在” becomes ”handwriting saves in” instead of ”handwriting is saved in”. Since the Chinese sentence is valid without passive voice, the direct translation wouldn’t produce a passive voice. One way to detect this is to check if the a transitive verb is followed by a preposition. If so, it’s likely that there is a grammar error and passive voice is needed. However this simple rule has many complications. For example a verb can be either transitive and intransitive, or the preposition that follows may be unrelated. One way to improve this would be parsing the sentence and identify the object of the verb. If the object is missing, we probably need passive voice.

5. Our translation system does not handle missing or implicitly changing subject well. It’s not uncommon to omit the subject in Chinese sentences if the subject can be inferred. For example, this sentence ”是全国第二选择性最高的大学。” should have a subject of Stanford University, but the subject is omitted because previous sentences have probably mentioned ”Stanford University”. Our system produces this result ”is entire nation second most selective University”, which lacks the subject. Subject can also change implicitly in Chinese, for example, this sentence ”其校友涵盖30名富豪企业家及17名太空员，亦为培养最多美国国会成员的院校之一。”. The first half’s subject is ”its alumni”, but the second half is ”Stanford University” and is omitted. Our result is ”Its alumnus cover 30 magnate entrepreneurs and 17 astronauts, also is train one of the most United States congress member institutions.”, which does not produce the correct subject. Such error is hard to fix with simple rules, because the subject heavily depends on the context.

9 Conclusion

This project of machine translation focuses on the post processing rules of the system. As the reader can tell from the output, the result still contains many errors. Post processing does improve a lot in comparison with the naive baseline algorithm, but it is nowhere near good natural English. Given the translation from Google Translate, we can see that there is still a lot room for improvement. Most of the problems we think lie in the overall knowledge of the sentence. Our rules lack the ability to see the bigger picture and only work on a micro level, e.g. one word or a phrase structure. The future work of this project includes putting more effort into pre-processing, where we can parse and tag the input sentence and use certain language model to help us break down the elements of the sentence. This way we perserve as much information as possible, and we believe we can only achieve a good translation with a good understanding the context.