CS124 Machine Translation Final Project

Date: February 28, 2014

INTRODUCTION

We chose Chinese as our language F. The difference in representation proved to be a challenge early on, with a significant struggle working out the encoding of Chinese characters so that we could proceed with further processing. Following that, we turned to the Stanford Word Segmenter for segmentation so that we could generate the bilingual dictionary without bias, since different grouping of characters yield different meanings. However, by virtue of characteristics like Chinese not using spaces, let alone using blocks of four nouns for poetic effect, the segmenter inevitably got tripped up on our development set. As such, the bilingual dictionary ended up with strange entries not typical of a fluent speaker, so even a flawless syntactic analysis would still be saddled with the same incorrect lexical translations.
Part-of-speech tagging was particularly useful for isolating forms specific to Chinese. For example, Chinese has measure words for counting (similar to loaves being specific to bread or reams being specific to paper in English), which led to a post-processing strategy. Tense proved a major challenge, as our sentences generally did not include time words like “yesterday” and “tomorrow” for reference. (Our one sentence with “tomorrow” actually translated into the correct tense.) Reduplication in Chinese also raises some problems; if words are not tokenized to include the doubled characters, the English translation is also doubled. It was also challenging that Chinese is a pro-drop language, since pronouns toward the end of a sentence have unclear antecedents.

CORPUS - DEVELOPMENT SET

1. 她沒有資格未經家屬同意就做這些事情。[1]
2. 新年後的第四天，一切生活復常，此時 就能外出訪友了。[2]
3. 窗臺上有一株玫瑰花，不久前它還十分嬌艷、充滿青春活力。[3]
4. 而我的選擇是，就算犯錯挨罵，也要昂首闊步。[4]
5. 他就用它僅會的中文單字或是比手畫腳和我爸媽溝通[5]

[1] http://tw.news.yahoo.com/%E5%B0%8F%E9%87%91%E8%BD%9F%E9%AB%98%E5%87%8C%E9%A2%A8%E5%8A%A9%E7%90%86%E7%84%A1%E6%AC%8A%E5%AE%A3%E5%B8%83%E9%81%BA%E7%94%A2-215051145.html

[2] http://www.mdnkids.com/calander/0115.htm

[3] http://www.epochtimes.com/b5/3/3/24/c11683.htm

[4] Friend’s facebook status update

[5] http://plum0925.pixnet.net/blog/post/235720181-%E3%80%90%E7%AB%A5%E8%A9%B1%E6%95%85%E4%BA%8B%E3%80%91%E8%80%81%E5%A4%96%E5%AD%B8%E4%B8%AD%E6%96%87%E7%9A%84%E8%83%8C%E5%BE%8C%E5%BF%83%E9%85%B8%E6%95%85%E4%BA%8B%E3%80%8E%E5%BD%B1

6. 不出一年，皇后果然生了一位公主。[6]
7. 很久很久以前，有一個可愛的小女孩，跟爸爸媽媽住在一個小村莊裏。[7]
8. 燕子在崎嶇的崖道旁築巢，在土坡上啄出了洞。[8]
9. 你要鼓起勇氣去做，否則明天一早，你將變成泡泡死去。[9]
10. 我們必須按照正確的原則辦事！[10]

CORPUS - TEST SET

11. 我正在努力找親戚朋友願意以投資目的購買，則你仍然可以繼續住下去。[11]
12. 其實最好的實驗對象應該是黑猩猩。[12]
13. 東方人的眼睛可以很分明地看出有黑色眼珠和白色眼珠。[13]
14. 安葬的日子到來了。[14]
15. 有一次坐捷運邊坐邊玩，還忘我到坐過頭很多站；[15]

OUTPUT - DEV SET

1. she not be qualifications not stand home kind agree to be these thing .
2. New Year after fourth day , everything live normalization , this time to can out friends .
3. window station on there is a rose , not long ago it also very jiao yan- , fall upon youth vitality .
4. and my choice be , even if guilty wrong scolded , and must head rich step .
5. he then take it only chinese separate character or contrast hands and feet and I mom and dad communicate .
6. no a year , empress really grow a princess .
7. very long a long time ago , there is a lovely little girl , with father mom living in a little village woman in this town .
8. swallow in rugged cliff road other nest , in slopes the pecking happen hole .
9. you must plump courage to be , else tomorrow morning , you will become finish bubble dead .
10. we must on the basis of correct principles act !

[6] http://www.epochtimes.com/b5/3/3/9/c11211.htm
[7] http://www.epochtimes.com/b5/3/3/2/c10978.htm
[8] http://www.epochtimes.com/b5/3/2/24/c10800.htm
[9] http://www.epochtimes.com/b5/3/3/18/c11478.htm
[10] http://www.epochtimes.com/b5/3/5/29/c13130.htm
[11] Friend's facebook status update
[12] http://www.mdnkids.com/101science/1.shtml
[13] http://www.mdnkids.com/101wonder/1.shtml
[14] http://www.epochtimes.com/b5/3/4/16/c12264.htm
[15] http://gamedb.yahoo.greatone.com.tw/2012_gamedb/gamenews.php?id=17713

OUTPUT - TEST SET

11. tomb when reach coming .
12. eastern eye can very clear see there is black eye and white eye .
13. I while great effort call relative friend want with money invested object buy , follow you yet can continue stop continue .
14. really most like experiment appearance should be chimpanzee .
15. there is a put MRT edge put while playing , return ecstasy reach sat head very multistation ;

TRANSLATION STRATEGIES

Note: References to Chinese words will be written in character form, with romanization in parentheses.

Post-Processing Strategy 1 (Unigram Frequency):

Our first strategy was to look at unigram frequency; we used the 10 billion-word WebText corpus and a 20K+ children's story from Project Gutenberg. This strategy helped eliminate the rare lexical translations; the character 未 (WEI) that can be used as an abbreviation of "1-3pm" (according to our Bilingual Dictionary via Google Translate), for example, translated at this stage as the much more common "not". We saw major improvements in the test set, with "as a matter of fact most good ablative cause suffix experiment take after out to be jocko ." now translated as "really most like of experiment seek should be chimpanzee ." The "ablative cause suffix" tag was very common, appearing in the majority of the corpus. Since the trigram "ablative cause suffix" is itself incredibly rare and in fact would not appear unless within a linguistic context, we were able to substitute in the generally more appropriate translation "of" for significant improvement in fluency.

Post-Processing Strategy 2 (Part-of-Speech Tagging):

The second strategy was to tag parts of speech in the original Chinese. Since the bilingual dictionary is fairly extensive, covering several different parts of speech for each word, the tagging isolated possible lexical translations to a particular part of speech. It was helpful to bring the focus away from just word frequency. For Example, words like "and" had been the dominant translation simply because of their frequency in English. Rather, in the particular case of "and", the word translated as the more appropriate "with" because it matched the part of speech designated by the tagger. This strategy yielded noticeable improvement in 4 of the 5 test set sentences, most commonly helping to differentiate between verbs and nouns. For example, "fund" became "money invested", "seek" became "appearance", and "up" became "reach."

Post-Processing Strategy 3 (Processing Measure Words and "的" (DE)):

The third strategy was to remove parts of speech that handled aspects particular to Chinese. We mentioned measure words earlier; at this stage, we drop measure words from the translation. For example, "three (measure word) apples" makes more sense in English as "three apples." However, we also found that time-related words like "month" and "year" were tagged as measure words, likely because they followed a cardinal number. This is a special case in Chinese, though; time-related words do not take measure words and are not measure words themselves. Similarly, the character 的 (DE) is frequently used as a possessive particle. We initially thought that a reasonable strategy would be to change instances of 的 to the genitive '*s*, but judging by our dev set, it was significantly more effective to drop 的 entirely. "Lovely (possessive particle) little girl", for example, translated at this stage to "lovely little girl." Furthermore, in three test set sentences, 的 translated to "ablative cause suffix", and fluency improved dramatically with the removal of this character.

Post-Processing Strategy 4 (Processing Chinese Adverbs "就" (JIU) and "要" (YAO) with respect to sentence grammatical structure):

The fourth strategy was to locate specific verb patterns: [$VERB^+$ 就 $VERB^+$] and [要 VERB]. Working from prior knowledge of Chinese, we determined that in the first pattern, 就 (JIU) should translate to "to", as in an infinitive, and in the second pattern, 要 (YAO) should translate to "must". We tried to work with more general rules, but words that we considered verbs for the purposes of our rules were not always tagged as such by the Stanford POS Tagger. 用 (YONG), for example, which can mean "to use", was tagged as a preposition, so our rule could not apply. Nevertheless, we successfully modified an occurrence of "then be" in the first dev set sentence to become "to be". Instances of 要 were already translated as "must" in the dev set, but we felt that it was a helpful rule that might pay off in the test set.

Post-Processing Strategy 5 (Processing "有" and "也"):

The fifth strategy was to set specific translations for common words in more general contexts. We specified that 有 (YOU) should always be translated as "there is", unless followed by an adverb. This was motivated by previous translation of 有 as "be", which yielded phrases in the dev set like "be a rose" and "be a little girl" which were more appropriate with "there is". The dev set saw the changes "there is a put MRT edge" and "there is black eye" from the corresponding "be" forms. Furthermore, we translate 也 (YE) as "and" whenever it follows punctuation, which we use as indication for 也 being the beginning of a clause. This rule was inspired by a sentence in the dev set where 也 begins a clause; while "still" would have been a more appropriate translation in this particular case, we felt that "and" was more generally applicable.

Post-Processing Strategy 6 (Working with Translated English - Identifying Pronoun Possessives):

The sixth strategy was to identify English POS-tagged instances of PRP followed by NN and then make the personal pronoun possessive. This rule was motivated by an instance of "me mom and dad" which should be "my mom and dad". The difficulty comes from the fact that the initial bilingual translation would not assume a possessive form.

ERROR ANALYSIS

As mentioned in the introduction, the most difficult problem that followed us throughout the process was segmentation. Since Chinese does not have spaces, the choice of how to group characters is fundamental to the translation process. We felt that manually segmenting some problematic points would be introducing our own biases to the process, so we relied solely on the output of the Stanford Word Segmenter. These results, combined with the incomplete knowledge of Google Translate, yielded an impoverished bilingual dictionary. For example, the phrase 實驗對象 appeared in our test set. Taken as a single token, it could be translated as "subjects", which would have been the most appropriate in our context. However, broken up into 實驗 對象, the translation becomes "experiment objects", which is less faithful, but still conveys some of the idea. The segmentation we were dealing with, however, was 實驗對 象. Google Translate likely did not know what to do with the third character and simply ignored it; the first half translates again as "experiment." The final character can actually mean "elephant", but our system at least chose the more likely translation: "appearance." However, the idea of "subjects" never entered our bilingual dictionary. To derive "subjects" from "experiment appearance" in post-processing would have required absurdly specific rules, tuning to that particular phrase in that particular sentence. Given more time, it would be useful in the future to train the segmenter on further data, especially since our corpus, pulled mainly from children's stories, may have been unfamiliar. This would likely require some manual segmentation since it is not common to find children's stories in segmented form. Nevertheless, it would certainly be feasible to at least train on common phrases like 很久很久以前, which is a set form for introducing a story, much like "Once upon a time" in English.

Grammatical fluency also proved to be a major issue. We generally focused on open class parts of speech, targeting semantic content before worrying about connecting the lexical items with modals, determiners, conjunctions, etc. Furthermore, cultural contexts and idioms aside, we can generally find analogous nouns, verbs, and other open class words from one language to another. The closed class parts of speech, however, are much more rooted in the structure of the language itself, making them more difficult to work with. As a result, our sentences ended up generally being strings of disconnected nouns, verbs, and adjectives. It would be helpful to work with a parser and spend more time dissecting the underlying structure of the sentence to be translated. With a clear understanding of how the clauses relate to one another, it would be easier to generate the necessary conjunctions and other functional words

based on the grammar of language E rather than relying on simple lexical translation for each word.

GOOGLE TRANSLATE - OUTPUT AND ANALYSIS

1. 我正在努力找親戚朋友願意以投資目的購買，則你仍然可以繼續住下去。
   I'm trying to find relatives and friends willing to buy for investment purposes, then you can continue to afford to live.
   Analysis: Google's translation is far superior here. Our translation is barely coherent, with most bigrams being very rare due to their disfluency. Google recognized that the purchase is for investment purposes; this statement of cause-and-effect is remarkable, especially as compared to our "want with money invested object buy". We also saw that the last clause involves continuing of some kind, but it is unclear what is continuing, while Google produces "continue to afford to live".

2. 其實最好的實驗對象應該是黑猩猩。
   In fact, the best subjects should be chimpanzees.
   Analysis: Both translations recognized "should be chimpanzees". Our translation struggled more with the beginning, though; note the "experiment appearance" struggle we mentioned in the error analysis section on segmentation. Google also added a comma after "in fact", while our system would not add punctuation where it did not exist in the original. Google successfully produced a natural English sentence rather than a stilted translation.

3. 東方人的眼睛可以很分明地看出有黑色眼珠和白色眼珠。
   Asian eyes can be seen very clearly with black eyes and white eyes.
   Analysis: Google's output is again a more fluent English sentence. In this case, however, our output is more faithful in the original meaning. Google implies that both black eyes and white eyes can see Asian eyes, while our system is more indicative of the fact that given an eastern eye, "there is black eye and white eye". These are characteristics of the Asian eye rather than something the Asian eye is seeing.

4. 安葬的日子到來了。
   Burial day arrived.
   Analysis: Google is succinct and fluent. Our translation only alludes to a burial by referencing a tomb; Google's translation provides more information for the reader while also giving the more appropriate past tense.

5. 有一次坐捷運邊坐邊玩，還忘我到坐過頭很多站；
   Once MRT side sit while playing, but also a lot of ecstasy to stop sitting on the head;
   Analysis: Both translations are comparably poor; neither are fluent, and neither convey coherent semantic content. Both recognize the MRT subway, but since we used Google

Translate for our bilingual dictionary, that only establishes that Google is consistent with itself. As with “in fact” in sentence 2, Google inserts the conjunction “but” after the comma for a more natural flow. In addition, it is curious that the token “忘我’ is translated as “ecstasy,” which bears only a slim, if any, connection to the original meaning of “doing something so intently that you “forget” yourself”; instead, “ecstasy” implies a state of euphoria.