Chinese To English Translation

Clay Schubiner
Evan Shieh
Yitao Zhang

Introduction [4 properties discussed in detail, 4 points]

We chose to translate Chinese into English. Chinese to English translation has been ranked as one of the 5 hardest languages to translate to English by top translation software maker Rosetta Stone. That said, we feel like our translations approach the quality of Google Translate's translations. However, because Chinese offers some difficult translation challenges, both our system and Google's struggled to make adequate translations.

Key challenges include:

- Word/sentence order. The chinese language tends to put attributives and adverbials before their modified objects if they are short modifiers.
    - Example:
        - Chinese: 我们在教室里学习翻译
        - Direct translation: We in the classroom are studying translation
        - Correct translation: We are studying translation in the classroom

    We see in the example that that "studying translation" describes the action that is going on in the classroom. There is an inversion in word order in chinese. Naturally, this doesn't always happen and simple sentences tend to be the exceptions where translation without reordering could work. For example "我在做饭" translates to "I am make food" using direct translation.

- Ellipsis: Ellipsis refers to the intentional omission of parts of a sentence. In Chinese this happens frequently with pronouns as it is assumed that unless a different pronoun is introduced, the subject does not change. This is especially apparent in longer sentences because when sentences get long in english, we usually reaffirm the relevant pronoun. In Chinese this is assumed.
    - Example: 我买了六只钢笔，一共三十元，拿回家一看，都是用过了的。
    - Direct Translation: I bought six pens, altogether thirty yuan, took home and had a look, all used
    - Acceptable English translation: I bought six pens which cost me thirty yuan. When I took them home, I found that they were all used.

    We see in this example that in Chinese the relevant pronoun referring to the pens becomes implicit and omitted after the first reference. However, in the English translation the pronoun had to be repeatedly used to clarify that we are referring to the pens. Again, the exception here is short sentences, since the pronouns are usually only omitted in long sentences with multiple clauses.

- Lack of sentence separation: Chinese sentences do not have separations between

characters. The lack of character separation adds an additional layer of difficulty because to interpret a sentence, we not only have to translate individual words but also group characters the right way for the translations to make sense:

- Example: 里根总统是美国的总统
- Translation character by character: Inside root total leader is pretty country of total leader
- Character grouping: 里根 总统 是 美国 的 总统
- Translation by word group: Reagan President is America of President
- Correct Translation: President Reagan is the President of America

As we can see, the ungrouped characters make a lot less sense compared to the grouped characters. This is especially true for names. The Chinese language handles English names by choosing character combinations that are “homophones”, i.e. Raegan is 里根, which said in Chinese sounds like “Leegan”, a poor but nonetheless acceptable approximation. Since the words are chosen by how they sound rather than what they mean, translating by character meaning would give disastrous results, and character grouping is necessary so that we can recognize that we are referring to a name. This is a language-wide issue and has no exceptions.

- Heavy context needed: Chinese is full of phrases that are deeply rooted in historical context, even in daily speech. In Chinese these are called 成语, and their closest approximation in English would be phrases such as metaphors. However, while English metaphors such as “as red as a tomato” can be understood easily, Chinese 成语 tend to require a lot more context. For example, 悬梁刺股 translates to working hard, but a direct translation would be ‘hanging from a beam and jabbing one’s buttocks’. The story here is that back in the olden days scholars used to study really hard, and they stay awake by tying their hair with rope to the roof beam so they can’t nod off, and jabbed themselves repeatedly to stay awake. Thus the story over time started to get encapsulated in the 4 characters of 悬梁刺股. Note that this is different from something like ‘Achilles Heel’ because while Achilles Heel translated to Chinese means the same thing, most Chinese 成语 would be completely nonsensical when translated to English. Nonetheless, the exception here is that a small fraction of these 成语 (such as 胆小如鼠 which literally translates to timid as a mouse) actually sound meaningful when translated, the majority is still extremely obscure after translation.

Corpus

Development Set:

1. 从即日起，烟草税将提高10%，每年估计带来额外7000万元的税收；酒精税也将提高25%，每年带来1亿2000万元的额外税收。（http://www.zaobao.com.sg/realtime/singapore/story20140221-312774）
2. 政府正在开始一场‘透明革命‘（http://news.xinhuanet.com/politics/2014-02/22/c_119454251.htm）

3. 一天下午，妈妈带我去公园散步了。
(http://www.zuowenren.com/youxiu/20100806/0PE0Q42010.html)
4. 奥林匹克运动会每四年举办一次，每届会期不超过16天。（
http://www.baike.com/wiki/%E5%A5%A5%E6%9E%97%E5%8C%B9%E5%85%8B）
5. 你说我就像油条，很简单却很美好（http://mojim.com/twy102520x2x5.htm）
6. 2014年02月22日昆明：动物园饲养员疑遭大象“踩踏”死亡
(http://news.ifeng.com/society/2/detail_2014_02/22/34067707_0.shtml)
7. 该机配件为单电、充电器、耳机、数据线等标配 (http://mobile.zol.com.cn/432/4321716.html)
8. 中国若不实行“节能”政策，其巨大的能源需求将使水资源面临更大压力。
(https://www.chinadialogue.net/article）
9. 华人农历新年是新加坡一年中最主要的精彩节日之一（
http://www.yoursingapore.com/content/traveller/zh/browse/whats-on/festivals-and-events/chinese-new-year.html）
10. 未来几天，美国总统奥巴马需要送给加拿大总理哈珀一箱啤酒。
http://www.chinanews.com/gj/2014/02-22/5869820.shtml

Test Set:

1. 中国只有高举毛泽东思想才能成为强盛和公正的国家（
http://junshi.xilu.com/zhuanti/my_2221/）
2. 学习不声不响地关心她，用你的诚实和善意对待她（
http://www.360doc.com/content/11/0123/21/2219312_88575531.shtml）
3. 其实，只要动起来都比不动好。（
http://fashion.sina.com.cn/d/ft/2014-02-22/090235481.shtml）
4. 晚餐吃什么？（http://www.haodou.com/recipe/all/842）
5. 古往今来，有多少的成功者被人们赞赏。（
http://www.edudown.net/student/zuowen/fanwen/201002/28148.html）

Baseline Translations

We have two different baseline translations for our project. After we made our Chinese-to-English dictionary, we use a simple character-by-character translation to make our first baseline translation. The results are abysmal, but somewhat comedic.

Dev

1. from immediately date start ，smoke grass tax will mention high 10% ，each in estimate meter band come amount outside 7000ten thousand yuan of tax income ；wine fine tax also will mention high 25% ，each in band come 1hundred million 2000ten thousand yuan of amount outside tax income 。
2. politics prefecture positive in open beginning a field ‘through next leather life ‘
3. a day under noon ，mom mom band I go public garden scatter step 。
4. abstruse forest match gram transport move can each four in move do a secondary ，each session can period not exceed cross 16day 。
5. you say I on image oil article ， very simple single but very US good
6. 2014in 02month 22date elder brother next ：move matter garden feed support member suspect suffer large elephant “tread tread ”dead die

7. that machine match item for single electricity 、charge electricity implement 、ear machine 、number according to line wait mark match
8. in country as not real row “festival can ”politics plan ，its huge large of can source need request will make water capital source surface face more large press force 。
9. China people agriculture calendar new in be new add slope a in in most main want of fine color festival date subordinate particle a
10. not come a few day ，US country total collectively abstruse hope horse need want send to add take large total reason laughing amber a box beer wine

Test

1. in country only have high move hair benefit east think want only can become for strength hold and public positive of country home
2. learn study not sound not ring ground shut heart she ，use you of sincere real and good meaning to stay she
3. its real ，only want move start come all ratio not move good 。
4. night meal eat assorted one on dice ？
5. ancient to this come ，have many less of become merit person quilt people they praise reward 。

We then implemented a segmenter to segment the Chinese sentences into multiple words. Often, Chinese is written without spaces, which makes good translation near impossible without a segmenter. By using the segmenter and adding the resulting segments into our dictionary, we were able to significantly improve the translation:

Dev

1. from now get up ，tobacco geld check raise 10％，per year gauge bring supernumerary 7000万first aim tax ；alcohol dues too would increase 25％，every year bring 1亿2000万 Yuan Dynasty possessive particle superfluity taxing 。
2. government in the process of initial single open space ‘open revolution ‘
3. a nature afternoon ，mom striped I go park go for a walk know。
4. Olympic sports competition each four year run a secondary ，per fall due meeting day non- exceed 16nature 。
5. you talking myself at once look Fritters ，very common retreat strongly glorious
6. 2014年02月22日Kunming ：zoo cattle farm worker uncertain catastrophe elephants “stampede ”mortality
7. the aircraft replacement serve as alone give an electric shock 、replenisher 、 headphone 、data line grade standard
8. China like not practiced “saving ”policy ，he very large possessive particle power source require prod supposing water be faced change huge pressure 。
9. ethnic Chinese lunar calendar New Year exist Singapore one age central -est principal aim glittering holiday one of
10. future nearly sky ，US presidential Obama requirement sent Canada premier Harper single container brewage

Test

1. China exclusive be hold aloft Mao Zedong cogitation expert able become powerful and prosperous write a poem in reply just aim country
2. be taught quietly land regard she ，need you of honourable harmonious goodwill serve she
3. as a matter of fact ，as long as use stand up big city ratio non- arouse okey dokey 。
4. evening meal receive what ？
5. through the ages ，be amount possessive particle winners cover folk incense 。

Strategies

Part of Speech Tagging [5 points]

Part of speech tagging was a moderate improvement to our translation system that changed a few words in almost every sentence of our development set and most of the sentences in the test set. Our implementation of part of speech tagging works as follows:

1. Input a Chinese sentence to Stanford's NLTK POS tagger and get back parts of speech for each word
2. For each part of speech received, choose only the English dictionary translations for each word that contain those parts of speech. E.g., If the word in Chinese is tagged as a noun, and the English translation can be nouns or verbs, we will only consider the ones that are nouns which corresponds to the Chinese POS.
3. Append each word candidate to a list of possible sentence variations for the English language model to choose among

After incorporating part of speech tagging, we saw quite a few differences in our results, including:

Test corpus:

> **Without POS tagging:** *Learning* quietly think she, take you *honest* and goodwill serve she.
> **With POS tagging:** *Learn* quietly think she, take you *honesty* and goodwill serve she.
>
> Notice how the POS tagging system corrected "learning" to "learn" and correctly took the noun form of the word "honest."
>
> **Without POS:** Actually, as long as move stand up all ratio a- move like.
> **POS:** In fact, as long as move stand up all compare not move well.
>
> This sentence is supposed to translate to: "It's better to move, than to not move at all." We can see that the POS tagging system is better, as it adds the words "compare" and "not," which are both relevant to the ideal translation.
>
> **Without POS:** Through the ages, have number winners people praise.
> **POS:** Through the ages, have how many winners people praise.

This sentence is supposed to translate to: “Throughout the ages, how many times have the winners been praised [for their efforts].” The POS tagger helps by identifying that we are looking for a “phrase” after the word “have”, not a noun. Without POS, the translation system chose a noun (the word “number”) to follow the word “have,” but the POS system correctly chooses “how many.”

Dev corpus:

Without POS: Government while first a 'open revolution '.
POS: Government while start a 'open revolution '.

The POS tagger correctly identifies that we need a verb after the word “while,” so we correctly choose “start” instead of “first.”

Without POS: 2/22/2014 Kunming: zoo breeder suspect lot elephants "stampede" death.
POS: 2/22/2014 Kunming: zoo breeder suspect receive elephants "stampede" death.

The POS tagger correctly swaps “lot” for “receive,” which is better as it indicates that the zoo keepers actually died as a result of the stampeding, instead of including the nonsensical word “lot.”

Additionally, without the part of speech tagger, some of the longer Chinese sentences took an unfeasibly long time to translate, because, without the tagger, all definitions of a word are evaluated in the language model, instead of just the ones of the correct part of speech. There were also many other dev/test sentences affected by the POS tagger, but we omitted those for sake of space. Overall, the POS tagger changed about 6-7 sentences in the dev set and all sentences in the test set.

Utilizing an English Language Model [5 points]

We found that when taken in isolation, many of the translations returned by direct dictionary lookup of word segments did not make sense as a coherent sentence. This phenomenon becomes especially prevalent for Chinese words that have a large number of possible meanings, in which it becomes important to disambiguate. For instance, the Chinese word ‘带’ could take on multiple English translations even given its part of speech as a verb (ranging from ‘lead’ to ‘wear’). Thus, utilizing the context of the sentence is necessary to determine the most likely English translation. Simply finding the most likely translation for a word/phrase atom might not correspond to the most accurate translation for a sentence. As a result, we employ an English language model to choose a sentence that is most likely to occur among multiple candidates in English. We utilized a Stupid Backoff Bigram model as coded in class, where the model was trained on English words from the Holbrook corpus. The results we discovered even for this simple model were encouraging - when disambiguating the word ‘带’,

our model chose “lead” as the most likely verb (corresponding to the action of a mother taking her child to the park - more details below). To integrate our English language model into our workflow, we modified our program to generate a set of possible candidate sentences based on word variations (including part of speech, tenses, etc.) instead of choosing one output sentence.

Results of our work are shown below. We note that choosing this strategy has multiple tradeoffs - while the advantages are quite pertinent, there are some notable disadvantages in the modeling assumptions we make. For instance, we notice that the Stupid Backoff Bigram Model isn’t necessarily the strongest English language model especially when it comes to the most correct grammatical ordering, a challenge that is particularly pertinent for the Chinese language where the POS arrangements differ greatly from those in English. For instance, our model returns “Dinner eat what” as a most likely English sentence when “What would you like to eat for dinner” might be even better (the latter wasn’t a generated candidate sentence, but nonetheless receives a lower score via our model). Switching from a bigram model to a tree-based one that is more grammatically accurate would be a natural extension here (in addition to choosing a model that might not as heavily penalize uncommon words).

Test corpus:

Without English Model: In fact, as long as *act* stand up all *contrast* not *arouse* fine.
English Model: In fact, as long as *move* stand up all *compare* not *move* well.

The Google translation for this sentence is “In fact, as long as the move does not move than good.” In Chinese, the sentence most accurately means “In fact, just a little bit of movement is better than no movement”. We notice that our language model chooses the candidate that maps the verbs “act” and “arouse” to “move”, which makes more sense contextually and is a more consistent and accurate translation.

Without English Model: *Evening meal bear* what?
English Model: *Dinner eat* what?

The correct translation for this sentence should be “What’s for dinner?”. Here, we see our language model handling ambiguity in verbs well, choosing the verb “eat” as a more likely verb than “bear”.

Without English Model: Through the ages, *be* how many of winners quilt folk *rhapsodize*.
English Model: Through the ages, *have* how many of winners quilt people *praise*.

Our language model again performs best in choosing the most accurate verbs to fit the context - the Google translation is “Throughout the ages, many of the winners are the people appreciated”. Praise is a significantly closer match.

Without English Model: China only *exist* hold high Mao Zedong *ideology* only can *turn into* powerful and prosperous and fair of country.
English Model: China only *be* hold high Mao Zedong *thought* only can *become* powerful and prosperous and fair of country.

This Chinese sentence was difficult for both our model and Google translator, and it appears that the language model helped incrementally. The correct translation is “Only by embracing Mao Zedong’s philosophy can China become a powerful and prosperous country”. Our language model finds the nuance of the verb “be” as greater in likelihood than “exist” (“turn into” and “become” seem comparable).

Two of the sentences that originally inspired this change in our development set are shown below:

Dev corpus:

Without English Model: A *nature* afternoon, mom *bring up myself leave public* park go *for a* walk know.
English Model: A *day* afternoon, mom *lead me go* park *take a* walk know.

The correct translation is, “One afternoon, my mom took me to the park to take a walk”. Our model disambiguates between both nouns (nature vs. day, myself vs. me) and verbs (bring up vs. lead, leave vs. take).

Without English Model: Olympic *sports competition* each four *harvest conduct* a *sequence*, each session *meeting day* not *surpass* 16 *season*.
English Model: Olympic *games* each four year *run* a *order*, each session *session* not *exceed* 16 *day*.

Our model here disambiguates between the verb “harvest” vs. “run” (the latter being more likely within context of running the Olympic games).

We consider this strategy a sweeping, global strategy that provided non-trivial, significant improvements to every sentence in the test and development sets.

Punctuation and Basic Fixes [0 points]

There were punctuation problems with most of the sentences in the corpus. This strategy did a simple find and replace on some Chinese punctuation, replacing the following characters:

- ‘ ： ’ with ‘:’
- ‘。’ with ‘.’
- ‘、’ with ‘,’
- ‘%’ with ‘%’

We also ensured in this strategy that:

- the first letter of each sentence was capitalized
- there was never more than one space between each word
- there was always a space after a comma
- there was no space before any comma
- that quotes and apostrophes always were next to a non-whitespace character

Overall, this strategy affected almost every sentence in the development set and 4 out of 5 sentences of the test set. Nonetheless, it is pretty cosmetic and we are just including this to standardize the format of our output.

Chinese to English Date Formatting [1 point]

Chinese date formats come in the following format: 2014年 2月 3日. We see this in line 6 of the dev set, which is an excerpt for a dated news article. This translates directly to “2014 year 2 month 3 day”. The new translation doesn’t make much sense, and also differs from the order in which dates are expressed in English (MM/DD/YYY in the U.S and DD/MM/YYYY in the UK). We aim to implement a correction that will put date expressions in the right order and in a form that makes sense.

The correction is implemented through grabbing the specific date regex of r'.{,3}?(\d{2,4})年.{,3}(\d+)月.+?(\d+?)日' in Chinese, and replacing it with the standard MM/DD/YYYY expression using the numbers.

> **Sentence before**: *2014 year 02 month 22 day* Kunming ：zoo breeder suspect receive elephants "stampede" death.
>
> **Sentence after**: *02/22/14* Kunming: zoo breeder suspect receive elephants "stampede" death.

We consider date constructions common in Chinese. Even though this did not appear in the test set, we believe this to be general, clearly explained strategy deserving of a point.

Chinese to English Number Formatting [1 point]

While English numbers go by “ones, tens, hundreds, thousands, millions, billions”, Chinese numbers go by “个，十，百，千，万，亿”, which is equivalent to “ones, tens, hundreds, thousands, ten thousands, hundred millions”. While direct translations work under the “thousands” denomination, a direct translation of 7000万 as it appears in our test set would translate to “7000 ten thousand” and “1亿2000万’ would translate to “1 hundred million 2000 ten thousand”. We aim to implement a correction that would convert the chinese numbers into their numerical values, and then reconvert those numerical values into reasonable english

translations.

This is implemented through a 2 step process. We first use regular expressions to pick out the 亿and 万 in the sentences because these are the ones that mess up in english translations. We then append the appropriate number of zeros behind. For example, 7000万 would be converted to 7000000. Next, we use a linguistic package to convert the 7000000 to its English equivalent of seventy million. This process was slightly more challenging when both 亿 and 万 were used together, as in the case of 1 亿 2000 万. In this case, we had to add another layer of post processing to add 100000000 to 20000000 before converting that into its English equivalent of a hundred and twenty million.

> **Sentence before**: From now start, tobacco tax will improve 10%, each year estimate bring additional *7000 ten thousand yuan* of tax; alcohol tax also will improve 25%, each year bring *1 hundred million 2000 ten thousand* yuan of additional tax.
>
> **Sentence after**: From now start, tobacco tax will improve 10%, each year estimate bring additional *seventy million* yuan of tax; alcohol tax also will improve 25%, each year bring *one hundred and twenty million* yuan of additional tax.

We consider number constructions common in Chinese. Even though this did not appear in the test set, we believe this to be general, clearly explained strategy deserving of a point.

Handling the Word "一" [1 point]

In Chinese, when we refer to countable objects, we have to include both the number and type of the object. For example, when referring to an apple, we have "一粒苹果", which translates to "A piece of apple"; when referring to a table, we have "一张桌子" which translates to "A sheet of table; when referring to a war, we have "一场战争" which translates to "a field of war". In the case of the dev set sentence 2, "一场透明革命" got translated to "a field "transparent revolution"".

The Chinese use of the word after 一 throws off the English translations, and it would be a lot more effective if we just ignored the second word which is unique to Chinese. This is implemented through regex matching of 一 followed by an independent word, and the a substitution that removed the independent word

> **Sentence before:** Government while start *a field* 'open revolution '.
> **Sentence after:** Government while start *a* 'open revolution '.

As described above, we consider '一' constructions common in Chinese (it appears pretty much just as frequently as the "a + noun" conjugation appears in English) . Even though this did not appear in the test set, we believe this to be general, clearly explained strategy deserving of a point.

English “a/an” Disambiguation [1 point]

In chinese, there is no concept of vowels since all words are based on characters. Thus “一天下午” would translate to “A afternoon” instead of “An afternoon” as in the case of sentence 3. We handle for this through post processing.

Post processing is handled through regexes that look for individual “a”s in front of vowel words and modify the a to an accordingly. This change affects sentences 2 and 3 in the dev corpus.

> **Sentence before:** *A afternoon*, mom lead me go park take a walk know.
> **Sentence after:** *An afternoon*, mom lead me go park take a walk know.

Since this strategy affects all words with the word ‘a’, which is a frequent occurrence in the English language, we believe this to be general, clearly explained strategy deserving of a point.

Removing Chinese Filler Words [1 point]

Chinese includes a number of filler words that sometimes mean something useful, but usually are just there as a connective word that serves no use when translating to English. The Chinese word “被” is a prime example of a filler word. 被 can be used as a noun to mean “quilt” or “blanket,” or as a verb to mean “cover” or “wear.” However, most often, it means none of these things and is simply a word that should be deleted.

To remove these filler words, we looked at the part of speech of common filler words. For example, if “被” was marked as a noun in our part of speech tagger, we would say that the word was not a filler word, and would translate it as “quilt” or “blanket,” depending on what the language model chooses. However, the POS tagger did not mark it as a verb or noun, we’ll say that the word is a filler word, and remove it from the translation.

Test corpus:

> **Without Filler Removal**: Learn quietly *ground* think she, take you of honesty and goodwill serve she.
> **Filler Removal**: Learn quietly think she, take you honesty and goodwill serve she.
>
> **Without Filler Removal**: Through the ages, have how many of winners quilt people praise.
> **Filler Removal**: Through the ages, have how many winners people praise.

Though removing filler words in the development set barely changed any of the sentences, we knew when choosing to implement the filler word removal system that filler words are common in Chinese and would affect many translations of other sentences. Additionally, our filler word removal system is general enough to be applied, in its current state, to any Chinese sentence and achieve reasonable results.

We were pleased to see that the test set had two sentences (the ones above) that changed significantly as a result. Notice how we remove “ground” from the first sentence and “quilt” from the second sentence.

Performing Tense Switching Based on Trigger Words [1 point]

Through comparison of Chinese and English linguistics, we observed that verb tense was one of the areas that differed most greatly. Indicating tense in English can be done through (familiar) changes in conjugation that are specific to word or phrase atoms (such as the difference between adding an “ed” or changing a word stem such as from “go” to “went”). In Chinese, tense might not always be determinable through such explicit, local cues. For instance, just the additional presence of the character ‘了’ could change an otherwise present tense sentence to one in the past tense, involving no changes in actual character spellings or intonations. As a result, tense in Chinese relies on cues that are more global (and in many cases, implicit). We therefore employ the following rule-based strategy to approximate Chinese tense based on the presence of “trigger” words.

We keep a list of phrases that, when taken individually, are generally sufficient to indicate past tense. Therefore, if one of the atoms ["了", "昨天", "前"] (corresponding to <filler>, <yesterday>, and <before>) are observed in a sentence, we mark the entire sentence as past tense. Of course, there are exceptions and limitations to this assumption (discussed below). We furthermore perform contextual matching for words that, in certain scenarios, could also indicate past tense. For instance, the word "过", when appended to a verb, could indicate that the said verb or action was performed in the past (when taken alone, it means to <cross> or <pass by>). For future tense, something very similar is performed, where the presence of any element in a list of indicators (the phrases ["未来", "将来", "后天", "明天"], corresponding to <in the future>, <future>, <the day before yesterday>, and <yesterday>) is sufficient to mark a sentence as future tense. We note here that we considered performing contextual matching in the future tense for the word “会”, but chose not to since it is grammatically impossible to disambiguate its structure from indicating “will do” or “can do” in English (and its meaning relies entirely on semantic context, which is a harder problem). Once a tense is determined, we utilize the NodeBox::Linguistics library in Python’s NLTK to change the tense of all verbs in the sentence.

Dev corpus:

> **Without Tense Switching**: An afternoon, mom *lead* me *go* park *take* a walk know.
> **Tense Switching**: An afternoon, mom *led* me *went* park *took* a *walked* know.
>
> The correct translation for this sentence is “One afternoon, my mom took me to the park to take a walk”. We notice that the main tense was caught (as indicated by the presence of the ‘了’ trailing the sentence) - however, we were unable to disambiguate parts of speech between “walk” as a noun versus as a verb.

Without Tense Switching: Future a few day, America president Obama *want sent* Canada premier Harper a box beer.
Tense Switching: Future a few day, America president Obama *will want will send* Canada premier Harper a box beer.

The correct translation involves the verb “will want to send”, which our tense switching logic brought us one step closer to by identifying that the sentence refers to an event in the future.

Unfortunately, none of the tense changes appeared in our test corpus (most likely as result of a small sentence selection size). However, we claim that tense is a highly important and generalizable change, especially given the vast differences between tense specification in Chinese and English.

Removing Duplicate Word Translations in English [1 point]

We observed in testing our translation pipeline that even after segmenting and removing Chinese filler words, duplicate words would appear in our translated sentences. The root cause of this phenomenon is the role that reiteration plays in Chinese. Chinese sentences often repeat word synonyms in adjacent usages for purposes of fluency (preferring phrases that contain multiple syllables). One example of this is the structure of “研究”, which is the most common term for “to research”. “研” individually means “to research”, and “究” could mean “to research” or “to investigate”. While the segmenter may catch this particular phrase by lumping the two together, it cannot comprehensively catch all possible permutations of scenarios where two synonyms are juxtaposed. As a result, we employ the strategy of simply eliminating one of the elements of a duplicate pair (with the exception of cases where punctuation surrounds one or both of the elements in order to allow for cases such as proper nouns).

Dev corpus:

Without Duplicate Removal: Olympic games each four year ran a, each session *session* not exceeded 16 day.
Duplicate Removal: Olympic games each four year ran a, each session not exceeded 16 day.

Our fix involves the Chinese phrase “会期”, where “会” alone can indicate “session”, but the two when taken together also mean “session”.

Without Duplicate Removal: Chinese lunar calendar New Year be Singapore a centre most main of bright one *one* of.
Duplicate Removal: Chinese lunar calendar New Year be Singapore a centre most main of bright one of.

Our fix here involves the term “之一”, which means “one of” when taken together.

Error Analysis

Context-Specific Tenses + Ambiguous Tense Trigger Words [5 points]

Our assumption that the tense of an entire could be perfectly determined by the presence of trigger words introduced several limitations in determining the correct tense of individual verbs. Firstly, we assumed that the tense of an entire sentence is homogeneous, which may not be true when referring to one tense within context of another. For instance, “Yesterday I ate apples, but I will normally eat bananas” is the correct translation for “昨天我吃了苹果，但我通常会吃香蕉”, but our system would treat the whole sentence as past tense and translate “eat bananas” to be “ate bananas”. We also found that edge cases exist for several indicator words in Chinese. For instance, a less common usage of the word "过" is as the second half of a compound verb (such as in the English word “surpass”). While "前" most commonly means “before”, it also could be taken to mean “in front of”. Examples of these errors in action are shown below:

Dev corpus:

> **Without Tense Switching:** Olympic games each four year run a, each session session not exceed 16 day.
> **With Tense Switching:** Olympic games each four year ran a, each session session not *exceeded* 16 day.
>
> We observe that tense switching here failed on the term “超过”, which when taken together means “to exceed” as a compound verb. As a result, our tense switching rules need to encode special cases for when the trigger word “过” is itself a verb.

As a result, two immediate tense extensions we would consider if given more time would be to consider global context to tolerate heterogeneity in addition to utilizing semantic context (or even more specific context rules) to determine tense to further disambiguate the meaning of trigger words. The data sources we could pull from this include a semantic analyzer, which would help to determine the semantic context of the usage of implicitly ambiguous words such as “过” and “会”.

Lack of prepositional phrases in Chinese [5 points]

Chinese translations often lack proposition. This is amply demonstrated by the translation of the dev sentence #10 containing “一箱啤酒“. The correct English translation for this would be “A box of beer”, but since the prepositional equivalent is not existent in Chinese, we get the translation “A box beer”. The same problem of a missing preposition also exists in sentence 3. Instead of translating “去公园散步” as “go to the park to take a walk”, the translation model cuts out all of the prepositions and we are left with “go park take a walk”. We also see the same problem in the test set sentence “In fact, as long as move stand up all compare not move well.” where the right use of proposition will have output “compare to” instead of just compare. The lack of prepositions can be attributed to the lack of such prepositions in the Chinese language

when constructing the sentences. Since our system works on word by word translations using a language model, it is unable to intelligently fill in words that do not have a equivalent in Chinese since there is nothing to translate from. The assumption that every word will have a equivalent is a problematic simplifying assumption.

To counteract this problem, we can adopt the strategy of complementization. This means that for words that could come with prepositions (e.g. cup *of beer, to take* a walk), we will add the word and preposition pair to the dictionary. This means that 啤酒 will have translations "beer" and "of beer". This will greatly improve accuracy when we cycle this through our language model which tries every translation in the dictionary and chooses the one that gives the highest score to the sentence. For example, the translator would choose "of beer" in situations like "I want a cup of beer" since that sentence would get a higher score than "I want a cup beer". Meanwhile, in situations that translate to "get me beer", the translation would not choose "of beer" because "get me of beer" will have a lower score. We can get the information on potential complements through crawling large tracts of documents to figure out the most common complementation pairs, and invoke them when we hit the words that are commonly used with complements.

Lack of present continuous tense in Chinese [5 points]

In English, we indicate that something is going on by appending an -ing, e.g. "I am eating" vs. "I eat". However, in Chinese this usually expressed implicitly as in "我 走路 时 摔倒 了". This translated directly to "I walk time fell down". The correct translation would be "I fell down when I was walking", but we have no way of telling that we should be using 'walking' instead of 'walk'. Chinese assumes that you know the only reasonable way to interpret this is for 'walk' to be present tense. This problem pops up in our dev set as well. In sentence 4, "政府正在开始一场'透明革命" is correctly translated to "The government is starting a "transparent revolution"" in Google translate but our translation gives us "Government while start an 'open revolution '." The simplifying assumption here is that we assume POS are consistent across both language. E.g. if there is a continuous tense in English there is also a continuous tense in Chinese, which turns out to be not the case here.

To counteract this problem, we can leverage information about trigger words. While Chinese doesn't have suffixes to denote continuous tenses, words that are supposed to be translated to continuous tense often are near 'trigger words' that indicate that something is happening. In the example of "I walk time fell down", the trigger word "时" (translated to time) hints to us that the sentence is trying to say "that time when I was walking". In fact, the same sentence without the character 时 would mean "I fell when I walked (sometime in the past)." The same trigger word strategy also applies to the sentence in our dev set. 正在 is the trigger word in this case, and it means "is doing". So 正在 开始 translates to "is doing start", and means starting when correctly translated. The trigger word 正在 tells us that the subject of the word "is doing" something, and the verb that follows should usually be in the continuous tense. The trigger word strategy can be implemented by identifying the most common set of trigger words through crawling many documents and using that to guide our translation. However, while identifying trigger words helps us with present tense translation, many of these trigger words have multiple

other meanings, and they don't always behave as trigger words. Purely classifying them as such would result in a lot of false positives and decrease translation quality. As a result, we would need to document edge cases for these trigger words thoroughly.

Our Implementation v. Google Translate [4 points]

Development Set

Google Translate: From now on, the tobacco tax will increase by 10% annually, with an estimated additional 70 million yuan of taxes; alcohol taxes will also increase by 25%, bringing additional revenue 100,000,000 20,000,000 yuan annually.
Us: From now start, tobacco tax will improve 10%, each year estimate bring additional seventy million yuan of tax; alcohol tax also will improve 25%, each year bring one hundred and twenty million yuan of additional tax.

Google Translate: The government is the beginning of a 'transparency revolution'
Us: Government while start an 'open revolution '.

Google Translate: One afternoon, my mother took me to the park for a walk.
Us: An afternoon, mom led me went park took a walk know.

Google Translate: Olympic Games held every four years, each period not exceeding 16 days.
Us: Olympic games each four year ran a, each session not exceeded 16 day.

Google Translate: You say I'm like fritters, very simple and very nice
Us: You work on me on look Fritters, very simple but very fine.

Google Translate: February 22, 2014 in Kunming: elephant zoo keepers suspected to have been "trampled" death
Us: 2/22/2014 Kunming: zoo breeder suspect receive elephants "stampede" death.

Google Translate: Machine parts for the single power, charger, headset, data cable, etc. Standard
Us: The aircraft fitting for single electricity, charger, headset, data line wait standard.

Google Translate: China Without the implementation of "energy" policy, and its huge demand for energy will allow water resources face greater pressure.
Us: China as not practiced "saving" policy, they huge of energy demand will make water be face more out pressure.

Google Translate: Chinese New Year is one of Singapore's most exciting holiday of the year
Us: Chinese lunar calendar New Year be Singapore a centre most main of bright one of.

Google Translate: The next few days, U.S. President Barack Obama, Canadian Prime Minister Stephen Harper needs to send a case of beer
Us: Future a few day, America president Obama will want will send Canada premier Harper a box beer.

Test Set

Google Translate: China Mao Zedong Thought to be the only strong hold high and equitable national
Us: China only be hold high Mao Zedong thought only can become powerful and prosperous and fair of country.

This sentence is supposed to translate to: “Only by upholding Mao Zedong’s way of thinking can China become a strong, prosperous, and fair country”. Neither translation does a great job of distinguishing the intended meaning of the sentence. That said, our translation does a much better job than Google’s translation. Though we do have the word “only” twice in our translation and Google correctly has it in only once, we include the words “powerful,” “prosperous” and “fair,” whereas Google includes “strong,” “high,” and “equitable,” which clearly are worse word choices for this sentence. Finally, we also include “country” instead of “national,” which is also more accurate. Overall, our translation here is much better than Google’s.

Google Translate: Learning about her quietly, with your honesty and kindness treat her
Us: Learn quietly think she, take you honesty and goodwill serve she.

This sentence is supposed to translate to: “Learn to care for her in subtle ways; use your honesty and kindness to treat her.” Both translations here are inadequate, yet Google Translate’s is better than ours in most ways. Ours correctly includes the right version of “learn” (“learn” instead of “learning”), but the second half of our sentence is much worse than Google’s. Google correctly conveys the idea that we should treat her with our honesty and kindness. Conversely, we nonsensically output “take you honesty and goodwill serve she.” Additionally, we choose “goodwill” instead of “kindness,” which is only a slightly worse choice. Overall, Google’s translation here is much better than ours.

Google Translate: In fact, as long as the move does not move than good.
Us: In fact, as long as move stand up all compare not move well.

This sentence is supposed to translate to: “It’s better to move, than to not move at all.” Both translations are pretty poor and are hardly understandable. The Google version does not include the phrase “stand up,” which is a plus for their version. On the other hand, we use the word “compare” to indicate that we’re comparing two options, and Google uses “than,” which does a worse job of indicating a comparison is taking place. Overall, we consider both translations to be of equal quality.

Google Translate: What's for dinner?
Us: Dinner eat what?

Google does a fantastic job of translating this sentence. Conversely, our translation does not make much sense. Our translation appears to translate each word from Chinese literally, whereas Google seems to understand the meaning behind the sentence. This is likely due to the fact that the pronoun was omitted in this sentence 晚餐吃什么? and Google correctly treats this case as a colloquial one.

Google Translate: Throughout the ages, many of the winners are the people appreciated.
Us: Through the ages, have how many winners people praise.

This sentence is supposed to translate to: “Throughout the ages, how many times have successful people been praised by the masses?” Our translation is actually better than Google’s in this case, as we include the word “praise” when Google instead chooses “appreciated.” However, Google chooses “throughout” instead of “through,” which makes a little more sense. Still, people reading our translation may be able to distinguish the intended meaning (that this sentence is posed as a rhetorical question), yet it is less likely than one will get the correct meaning from Google’s translation.

As you can see, the translations that Google provides are barely adequate. We include these translations to show that even Google has a hard time translating Chinese. Overall, we feel that our translations, with a few exceptions, were on par with Google’s translations.

Thank you for grading our assignment!