Direct MT of Chinese

Introduction

Translating Chinese to English presents significant challenges for any MT system. Below we discuss some of the linguistic properties of Chinese.

1. Chinese words are not delimited by spaces to indicate clear word boundaries.

2. Chinese has very little inflectional morphology.

- nouns are not in inflected with person or number
- verbs are not in inflected with tense
- there is no subject-verb agreement
- pronouns do not conjugate, e.g.," 他" can be translated as either *his*, *him*, or *he*

3. Chinese makes heavy use of particles to mark grammatical relations

- Aspect particles, e.g., perfective particle"我玩[了]" → I play 了 → *I played*
- Manner particle 地 to mark adverbial phrases: "高兴[地]" → happy 地→ *happily*
- 的 as a genitive marker/associative marker: "她是他[的]" → she is he 的 → *she is his*

4. Chinese has SVO word order and modifiers generally precede noun phrases, like in English. However, there are also notable differences:

- Whereas an English relative clause follows the noun phrase it modifies and starts with a complementizer (e.g., *that*, *which*), a Chinese relative clause precedes the noun that it modifies, and ends with the relative particle 的. For instance, "使用 筷子 [的] 人"→ use chopsticks [的] person → *the person who uses chopsticks*
- Locatives such as 上 appear after noun phrases in Chinese, whereas English usually uses prepositions which precede noun phrases: 太阳[上] → sun [on]→ *on the sun*
- In Chinese, prepositional phrases about time and location usually occur between subject NPs and VPs, while in English they tend to occur after the VP: 我 [昨天] [在公园里] 跑步→I [yesterday] [in the park] run→ *I ran in the park yesterday*

5. Chinese has no articles such as *a(n)* and *the*. Instead, it uses measure words as determiners, such as 个 after a number in a noun phrase: 一[个]人 → one 个 person → one person.

6. Chinese differentiates between attributive (e.g., *the happy boy*) and predicative (e.g., *the boy is happy*). In Chinese, predicative adjectives do not require the copula 是 ("be") to precede them: e.g., 他 浪 高兴→ he very happy → *he is very happy*. In contrast, predicative adjectives in English must be preceded by a copula.

7. Chinese words can generally be associated with multiple syntactic categories. Many words are both adjectives and adverbs; prepositions can be used as verbs and vice versa; et cetera. In particular, Chinese noun-noun compounds are highly productive, whereas in English, the same complex NPs are likely to be modified by adjectives, possessives, and prepositional phrases.

Corpus and Dictionary

We chose 15 sentences from the "2014 Level 4 Translation Expected Topic Drills", a series of sentences used to help Chinese students prepare for the College English Test Level 4 (CET-4). There are 3 passages, containing 6, 4, and, 5 sentences respectively. The first two passages constitute our development set, and the last passage forms our test set. Please see our appendix for the corpus.

Prior to constructing our dictionary, we used the Stanford segmenter to tokenize the words in our corpus. We then hand-built our dictionary, pulling definitions from wordreference.com and linedict.com. For words that possessed multiple translations, we listed the various options in the order that they appeared on the websites. We also stored the part-of-speech (POS) tags of the English translations provided by the online dictionaries.

This led us to our baseline system, which translated each Chinese word into its first entry in the dictionary. Henceforth, we will refer to pulling words from the dictionary as "lookup".

Pre-processing Translation Strategies

Please note: Many of our strategies utilize a language model to score the likelihood of a phrase's translation. We coded a stupid backoff (trigram) model that we trained on the NLTK Brown corpus.

1. Using a Chinese parser to reorder constituents

As Chinese and English differ in word order, prior to lookup, we re-ordered the source sentences to make them more English-like. We used the Stanford parser to parse the source sentences and used the following rules to reorder the constituents.

1.1 Reordering relative clauses

The Stanford parser outputs the following parse tree for a noun phrase (NP) modified by a relative clause:

(NP ... (CP (IP 使用 筷子) (DEC 的)) (NP 国家))

(literally: *that/who use chopsticks country*)

Our reordering occurs in two steps. First, our system moves a complementizer phrase (CP) headed by the complementizer "的" (tagged and referred to as DEC) *behind* the NP it modifies. Then, it moves DEC to the front of the CP. We thereby derive (NP 国家) (CP (DEC 的) (IP 使用 筷子)). This directly translates to *country that/who use chopsticks*. (We later use a language model to determine which complementizer to use.)

This strategy alone resulted in 3 instances of improvement in the development set and 2 instances of improvement in the test set.

1.2 Reordering time and locational phrases

The parser outputs the following structure:

(LCP (NP 中国 和 世界 其他 国家) (LC 之间))

(literally: *China and world other country between*)

When our system encounters localizer phrases (LCPs), if the LCP is not the first phrase in the clause, it will move it to the end of clause. In addition, localizers (e.g*., between*) are moved to the front of LCPs (e.g., yielding (e.g., yielding *between China and world other country*). Time phrases, which are headed by NT, are treated similarly.

This strategy led to 2 non-trivial improvements in *each* the development and test sets. For example, in our test set, "2009 年 毕业 的 学生" *(2009 graduate DEC student*) became "学生 的 毕业 2009 年" (*student DEC graduate 2009*).

2. Omitting particles that have no English counterparts

As mentioned previously, Chinese has many particles that have no English counterparts. These particles often share the same form as content words, causing our baseline system to wrongly translate them into content words. For example, the aspectual particle 了 is the same character as the Chinese word for *finish* . As a result, our baseline system wrongly translates "达到 [了#AS] 3150 亿 美元" (*reaching 315 billion USD*) in the development set as *reach [finish] 315 billion USD*. Using the Stanford POS-tagger, we identified the follow particles and omitted them prior to lookup:

MSP (所/以/而/来), DEV(地) AS(了/过/着/的), SP(了/吧/呢/吗)

This led to 3 improvements in *each* the development and test sets, in both of which all of these particles were correctly omitted.

3. Using POS to select and modify translations

As Chinese is flexible with parts of speech (POS), our dictionary often included several POS tags for each entry. We therefore used the POS tags provided by the Stanford parser to help select the best translations among these entries, and modified them if needed:

- If the parser tagged a Chinese word as an adverb, we only selected adverbs from the word's dictionary entry. If there were none, we searched for adjectives and conjugated them into adverbs.
- For parser-tagged adjectives, we only selected adjectives. If they were followed by the adverbial (manner) particle DEV "地", we conjugated them into adverbs.
- For parser-tagged verbs and prepositions, we only selected verbs and prepositions.

This led to 5 improvements in the test set. For instance, *get a job market* became *employment market*, as the system correctly identified the first word as a noun and not a verb.

4. Using context to disambiguate DEG 的

The particle "的" is the most frequent word in Chinese and is also highly versatile. We have seen that it can mark relative clauses (DEC). Depending on what precedes it, "的" can also be a genitive or associative marker, in which it is tagged as DEG. We distinguish between the following three contexts in which DEG occurs:

- If there is an adjective before DEG 的, it is an associative marker and will be dropped.
- If there is a pronoun before DEG 的, the pronoun is turned into the possessive form and DEG 的 is dropped. E.g., "他们#PN 最#AD重要#JJ [的#DEG]" (*they most important DEG*) → *their most important*. NB: here both the 1st and 2nd cases apply.
- Otherwise, it is translated as the possessive *'s*. For example "中国#PR 的#DEG 贸易 赤字" → *China's trade deficit*.

This led to 5 improvements in *each* of the development and test sets. In the test set, "他们 的 (*they of*) → *their*, "合适 的" (*appropriate of*) → *appropriate*, and "毕业生 的#DEG" (*graduate of*) → *graduate's*). Note that the baseline translation of DEG 的 as "of" is wrong in every instance.

5. Using a language model in lookup

The lookup strategy in (3) returns multiple translations for given word, so long as their POS tags matches the Chinese's word's POS tag. To refine lookup, we used our language model to select among these possible entries. This led to 6 non-trivial improvements in the test set (e.g., *below most circumstances → under most circumstances*.)

Post-processing Translation Strategies

1. Genitive alternation

Our dictionary translates the genitive marker "的" as *of*, however, it is most comparable to the English possessive *'s*. As this construction appeared 3 times in our development set, we generated candidate sets of the form *noun2 of noun1* and *noun1's noun2*, and used the language model to determine which was the most fluent English construction on a case-by-case basis. For instance, this derived *beginning of chopsticks* from *chopstick's beginning* in the development set. It also derived *raise of the university enrollment rate* from *university enrollment rate's raise* in the test set. (This strategy applied once more in the test set, but it yielded an ill-formed phrase in this instance – see the error analysis.)

2. Repositioning copulas

In Chinese, the copula ("是") can come after adverbs and occur before prepositional phrases. However, in English, adverbs cannot precede the copula (e.g., the ungrammaticality of **the boy very is quick*) and the copula cannot directly precede a prepositional phrase. We thus moved copulas in front of adverbs that preceded them, and deleted them when they arose in front of prepositions. This resulted in 5 successful uses of the copula in the development set that would have otherwise been ill-formed (e.g., *also increasingly be → be also increasingly*). This strategy also applied accurately to the only copula that appeared in the test set: *opportunity **be** because → opportunity because*.

3. Conjugating superlatives

The superlative morpheme "最", which modifies adjectives in Chinese, appeared twice in our development set. Using the *pattern.en* Python package, we correctly conjugated both of these adjectives into their superlative form (e.g., *MOST big → biggest*). However, this strategy did not apply in the test set, as it did not contain any superlatives.

Please note: Although "repositioning copulas" and "conjugating superlatives" were less utilized in the test translation, they are still two crucial strategies for any translation into English, as both constructions are rampant in English. Furthermore, parallels of these constructions exist in every language.

4. Pluralizing nouns

Unlike in English, Chinese seldom marks plurality on nouns, and did not do so in our development set. Therefore, whenever we spotted a cardinal number modifying a singular noun, as indicated by the NLTK POS-tagger, we pluralized the noun, successfully pluralizing 3 nouns in the development set (e.g., *3000 year → 3000 year**s***). It also applied accurately in one case in the test set: *3 million student → 3 million student**s***.

5. Inflecting verbs

While Chinese never inflects verbs, verb inflection is very common in English. To recover these inflections, we used *pattern.en*'s conjugator to generate candidate sets for each verb we encountered, then used the language model to pick the best inflection. This strategy correctly inflected 12 verbs in the development set (e.g., *use chopsticks have a meal → us**ing** chopsticks **to** have a meal*). It also inflected 4 verbs in the test set, but only 2 of these were inflected with the correct tense: *strikingly descend. → strikingly descend**ed**.*, and *that trade need. → that trade need**ed**.*

6. Inserting determiners

In contrast to English, in which the and a are two of the most frequent words, Chinese does not contain definite or indefinite articles. Accordingly, we attempted to insert a determiner in front of each noun phrase, generating candidate sets of the form *NP*, *a NP*, and *the NP*. We then used the language model to decide which determiner to insert (if any). This inserted 6 necessary determiners in the development set (e.g., *beginning of chopsticks → the beginning of chopsticks*), and successfully inserted 3 determiners in the test set: *raise of university enrollment rate → **the** raise of **the** university enrollment rate*, *student number → **the** student number.*

As inflecting verbs and inserting determiners both impact the POS-tagger's ability to detect nouns and verbs, strategies 6 and 7 executed successively and iteratively until the output no longer changed.

Error Analysis

Sentence 1: The first sentence revealed that our translation system struggles with translating Chinese compounds. Specifically, the NP "2008全球 经济 衰退" was translated as *2008 the whole worlds economy decline*. While the last three words are all nouns in Chinese, a fluent English translation would have turned the two penultimate words into adjectives to form *the global economic decline.* This error arose because our system operates on a word-by-word lookup, which will rarely capture a compound relationship between nouns. In the future, we should consider building a phrase-level dictionary and greedy lookup to capture compounds.

In addition, the translation *2008 the whole worlds economy* showed us that we neglected to consider that the POS-tagger would identify years as cardinal numbers. Accordingly, our "pluralizing nouns" strategy pluralized *world*, the first noun to follow *2008*. In hindsight, we should have constrained the context in which this strategy applied by using regular expressions to capture (and ignore) dates, and using dependency parsing to detect the boundaries of NPs.

Sentence 2: Our system translated the phrase "2008年 毕业 仍 在 找 工作 的 300万 学生 之中" as *graduate still in from 3 million students who to find work on 2008* (cf. among three million students who graduated in 2008 and are still seeking job opportunities). This mistranslation illustrates how our "reordering" strategy crucially relies on the parser to accurately recognize constituents. In this example, the parser misinterprets "在...之中" as a single constituent, and treats "2008年" as modifying the entire phrase. As a result, "reordering" failed to move "的 300万 学生" (*from 3 million students*) to the front of the clause, and failed to move "2008年" immediately after the verb毕业 (*graduate*).

Sentence 3: "毕业生过剩也可以 ... 教育机构的增加" was translated as *Graduates excessive increase who also can… education body* (cf. The graduate glut can also … education body's increase). This mistranslation is also due to a parsing error. The Chinese parser mistakenly tagged the last 的#DEG as a complementizer and mistook the noun 过剩 (*excess*) as a verb. Had this sentence been parsed accurately, reordering would not have applied, and the verb *increase* would not have moved directly behind the word *excessive*.

Sentence 4: "大学的学生人数" was translated as *the student number of the university of people.* Here, our post-processing strategy "genitive alternation" went awry. By design, the strategy does not attempt to swap a possessive construction for a genitive construction when multiple possessive markers appear within the same clause. (This is because it is much more complicated to change *Sally's dog's tail* into *the tail of the dog of Sally*.) However, we failed to consider that the English preposition *of* would be fed into our input from the dictionary ("人数": "number of people"), causing this strategy to apply where it shouldn't have: *the university's student number of people* became *the student number of the university of people*.

In the future, this strategy should either not apply when it encounters any combination of genitive and possessive markers within the same clause, or handle these constructions more delicately.

Sentence 5: Our system translated the phrase "在大多数情况下" as *under the most circumstances.* In this one instance, our "inserting determiners" strategy over-applied, changing *under most circumstances* into *under the most circumstances*, a less fluent translation. Here, our language model simply failed to pick the best candidate. We might have pre-empted this had we trained our model on longer n-grams.

Lastly, our "inflecting verbs" strategy produced a happy accident in sentence 5. The POS-tagger misidentified the noun *graduate* as a verb and our system then generated and scored a candidate set (i.e., graduate unable…, *graduates unable*…, *graduated unable*…, etc.) using our language model. However, because our language model does not distinguish between parts of speech, it scored *graduates* as the best candidate, as a pluralized form of the noun was most likely in that context.

In sum: Although both our pre- and post-processing strategies are linguistically well motivated and highly generalizable, the pre-processing strategies heavily rely on the parser to *accurately* recognize constituents and POS tags, which is not always the case. This, in turn, impacts post-processing, which relies on the pre-processing strategies to *accurately* re-order the source sentences.

Additionally, the test set translation revealed just how much our language model struggles with capturing long-distance dependences.

Future improvements to our system should include improving the parser and training the language model on n-grams longer than trigrams. Finally, instead of making clear-cut decisions at each processing stage, we should keep track of all of the candidates generated by each strategy, and choose the best candidate only at the very end. This might have circumvented earlier strategies from thwarting later ones.

Translate (GT) v. Our Translator (OT)

Unsurprisingly, Google Translate (G) does a better job than our system (O). Google produces almost perfect translations of the 1st, 4th, and 5th sentences (except missing a *the* in sentence 1 and failing to delete a *but* in sentence 4). However, Google's translation suffers in sentences 2 and 3 due to the highly ambiguous nature of the particle 的 and the parsing errors that likely stem from it. Overall, Google systematically outperforms our system in the following respects:

- verb inflection…导致 → G: *led to*, O: *lead to*
- passivization… 归因于 → G: *be attributed to*, O*: put down to*
- noun plurality… 教育机构 → G: *educational institutions*, O: *education body*
- compound nouns… 2008全球经济衰退→ G: *2008 global recession*, O: *2008 the whole worlds economy decline*
- determiners… 技能→ G: *the skills*, O: *hand*
- noun selection (see above example)

To their advantage, Google likely has much larger "dictionary," more powerful and robust languages models, and sophisticated methods for capturing long-distance dependencies (e.g., an English PCFG).

Please see appendices II and III for more visually-pleasing comparisons.

Appendix I: Corpus

Development set

说到筷子的起源，中国是世界上第一个使用筷子的国家，用筷子吃饭已经有至少 3000 年的历史了。
筷子看起来很简单，只有两根小细棒。
但它有很多功能，比如挑选，移动，夹，搅拌或者挖。
此外，它便于使用，价格便宜。
而且筷子也是世界上独有的餐具。
使用筷子的人，无论是中国人还是外国人，都无不钦佩筷子的发明者。
http://cet4.koolearn.com/20150203/789828.html

对于世界上很多国家来说，中国正迅速成为他们最重要的双边 贸易伙伴。
然而，中国和世界其他国家之间贸易不平衡的 问题已经引发了关注。
尤其是美国对中国的贸易赤字是最大的，达 到了 3150 亿美元，这个数字是十年前的三倍还多。
贸易纠纷也越来越多，主要是关于倾销、知识产权和人民币的估价。
http://cet4.koolearn.com/20150203/789829.html

Test set

2008 全球经济衰退导致中国的新毕业生的就业市场显著下降。
2009 年毕业的学生将加入到 2008 年毕业仍在找工作的 300 万学生之中。
毕业生过剩也可以归因于大学入学率的提高和教育机构的增加。
虽然大学的学生人数增加了，但是他们的质量并没有明显地提高。
在大多数情况下，毕业生无法在 2008 年找到合适的就业机会是因为他们没有行业所需的技能。
http://cet4.koolearn.com/20150203/789830.html

Appendix II: Comparison of translations

Test Set	Human	Our System	Google
2008 全球经济衰退导致中国的新毕业生的就业市场显著下降。	The 2008 global recession resulted in a significant drop in the job market for China's new graduates.	2008 the whole worlds economy decline lead to China's new graduate's employment market strikingly descended.	2008 global recession led to a significant decline in the job market of new graduates in China.
2009 年毕业的学生将加入到 2008 年毕业仍在找工作的 300 万学生之中。	Students graduating in 2009 will join around three million students who graduated in 2008 and are still seeking job opportunities.	Student who graduate in 2009 will add to graduate still in from 3 million students who to find work on 2008.	2009 graduates will be added to the 2008 graduates are still looking for work three million students.
毕业生过剩也可以归因于大学入学率的提高和教育机构的增加。	The graduate glut can also be attributed to a rise in the number of college enrollments and educational institutions.	Graduates excessive increase who also can put down to the raise of the university enrollment rate and education body.	Graduates can also be attributed to excess increase increase college enrollment and educational institutions.
虽然大学的学生人数增加了，但是他们的质量并沒有明显地提高。	Although the number of college students has increased, there has not been any significant improvement in their quality.	Although the student number of the university of people increase, but their quality really not have obviously to raise.	While the number of university students has increased, but their quality has not improved significantly.
在大多数情况下，毕业生无法在 2008 年找到合适的就业机会是因为他们沒有行业所需的技能。	In most cases, graduates were unable to find suitable employment in 2008 because they did not have the skills required by the industry.	Under the most circumstances, graduates unable in 2008 finds appropriate employment opportunity because they have not hand that trade needed.	In most cases, graduates can not find suitable jobs in 2008 because they did not have the skills required for the industry.

Appendix III: Google Translation versus Our Translation

S	Google Translate (G)	Our system (O)	Winner	Reason
1	2008 global recession	2008 the whole worlds economy decline	G	G: correct; O: wrong plural; awkward compound
	led to a significant decline	lead to China's … strikingly descended.	G	G: correct inflection
	the job market of new graduates in China	China's new graduate's employment market	G	G: correct; O: no plural; less fluent
2	2009 graduates	Student who graduate in 2009	?	G: ambiguous; O: plural/tense
	will be added to	will add to	G	G: correct passivization
	the 2008 graduates are still looking for work 3 million students.	graduate still in from 3 million students who to find work on 2008.	?	Both incorrect; perhaps G is less bad.
3	Graduates	Graduates excessive	O	O: faithful, though wrong POS
	can also be attributed to	also can put down to	G	G: correct passivization
	excess increase increase college enrollment	the raise of the university enrollment rate	O	O: almost correct except "rise"
	educational institutions	education body	G	O: no plural, less fluent
4	the number of university students	the student number of the university of people	G	G: correct; O: additional "of people"
	has increased	increase	G	G: correct inflection
	has not improved significantly.	really not have obviously to raise.	G	G: correct inflection
5	In most cases	Under the most circumstances	G	G: correct; O: additional "the"
	graduates were unable to find suitable employment in 2008	graduates unable in 2008 finds appropriate employment opportunity	G	G: correct inflection & time phrase movement
	they did not have the skills required by the industry.	they have not hand that trade needed	G	G: correct inflection & "skill" and "the industry" are correct