Heuristic-Based Machine Translation from Chinese to English

Introduction

We choose Chinese as the foreign language to translate. All team members are native Chinese so we know Chinese very well.

Difficulty of Translation

Chinese is known as a language difficult to translate, even in its appearance (you have to segment the text properly before translation). To quote a piece from the web:

> The Chinese language is still a very different animal from English or other Indo-European languages, which often share linguistic cognates, grammatical similarities, and a common set of imagery. Chinese has completely different roots, so in the process of translation you often have to go “farther” away from the original, and then come back again. It’s impossible to translate word-by-word, and sometimes even sentences have to be reworked and reordered. It takes a certain confidence with the original, and also a surer hand with literary writing in your own target language.

The most conspicuous difference may be that, Chinese do not have a strict grammar like Western languages. For example, the sentence “I took classes yesterday.”, you can use any of the following:

```
昨天 我 上了课。
昨天 我 上课了。
我 昨天 上了课。
我 昨天 上课了。
我 上课了 昨天。
```

Note: 昨天: yesterday, 我: I, 上了课/上课了: took classes

The placement of the adverbial of time, `昨天` (yesterday), is very flexible in Chinese. In English, it’s uncommon to say `I yesterday took classes.`, while `我 昨天 上了课。` is actually the most idiomatic Chinese.

Some important notions of English, such as tense, genitive and single/plural forms, are weak concepts in Chinese and constructed in a totally different way. In terms of tense, instead of having a past tense for each verb, Chinese append auxiliary terms like `了` or `过` to verbs to indicate the past tense. As a result, to express `learned`, you say `学了` where `学` means to learn. And that’s not the only rule. This is also true for genitives and plurals: you append auxiliary `的` and `们` after pronouns to form them respectively:

singular	plural	genitive
我 (I)	我们 (we)	我的 (my)
你 (you)	你们 (you)	你的 (your)
她 (she)	她们 (they)	她的 (her)

Moreover, context is very important in Chinese. You will not be able to understand a sentence without a clear context. Again, take tense as example, it must be inferred from the context, mostly with adverbial of time. Consider the translation:

```
那天，我终于成功了。
```

> I finally succeeded that day.

If `那天`, the adverbial of time, is omitted, the sentence can either be understood as `I succeeded.` or `I succeed.` But in English, even if you omit `that day`, with `succeeded` it's still unambiguous that the event happened in the past.

Another example of the importance of context is determining POS. A Chinese word can have multiple part of speeches (POS) under a same meaning, and the POS must be inferred from the context.

> 世界/真/*美丽*。(美丽: adj. beautiful)
> 世界/充满了/*美丽*。(美丽: n. beauty, beautiful things)

Although there is also polysemy in English, such words are much more prevalent in Chinese.

Finally, just for fun, Chinese can even be confusing to Chinese speakers. As an extreme example, consider the sentence which has two surprisingly similar parts with distinct meanings:

> 冬夏穿衣方法：冬天：能穿多少穿多少；夏天：能穿多少穿多少。

To understand it meaningfully, you have to do the following segmentation:

> 冬夏穿衣方法：冬天：能穿/多少/穿/多少；夏天：能穿/多/少/穿/多/少。
> How much to wear in winter and summer: in winter, wear as much as you can; in summer, wear as little as you can.

Insights

Although it's impossible to get a perfect translation from Chinese to English (yes, even with human translator), there are many heuristics we can apply to get a good enough translation with the basic meaning preserved. Some examples are:

- Select word-to-word translation based on POS of Chinese word
- Select word-to-word translation based on statistical information, e.g. how likely the specific Chinese word is translated to the candidate English word, and how likely the English word appears in a sentence
- Infer the correct form of a verb based on auxiliary words as well as the general context in the clause, sentence or paragraph
- Infer the correct form of a noun (singular or plural) based on context (e.g. whether there is a number word before it, or words indicating plural like `these`)
- Infer the sentence structure and translate into corresponding English sentence by reordering words and adding/removing structural words
- Infer the logical relations between Chinese clauses, which are usually connected by implicit logics

Corpus

Our corpus of sentences is taken from various sources and covers a variety of topics. It's mostly contemporary Chinese. The sentences are neither too long nor too short and have some common patterns. The following is the sentences in development set and test set, along with source.

Approach

We first implemented a baseline MT system with direct translation and then apply various post-processing strategies upon it, under development set only. After we've finishing polishing the MT system, we run it under the test set.

Preprocessing

The preprocessing involves two phases: 1. segmenting and POS-tagging sentences, and 2. building dictionary.

sentences. Despite jieba's satisfactory segmentation accurary, its POS-tagging is not as reliable, so we also manually correct mis-tagged word.

Once we have all the Chinese words, we use WordReference API to build the Chinese-to-English dictionary. Again, as the API misses translations for a few words, we manually add translations to those words.

See `preprocess.py` for details.

Direct Translation

The direct translation algorithm is deadly simple: for each Chinese word in sentence, it retrieves translations from the dictionary and choose any English word as the translation of this word.

Note although the output contains POS tags, direction translation simply discards this information.

Strategy 1: Matching POS

The first strategy is simple yet powerful: pick English word as candidate only if its POS matches the POS of the Chinese word and choose any candidate English word as translation. For example, for Chinese word `成长`, possible translations are `grow/vi, grow up/vi, growth/n, growing up/gerund ...`. If the context is `成长/n ,/x 最/d 残酷/a 的/uj 部分/n 就是/d`, it's clear that `growth/n` and `growing up/gerund` are better translations than `grow/vi` and `grow up/vi` as the POS matches.

A small implementation caveat here is that jieba and WordReference use different POS naming systems. We use the following table to do POS matching:

Meaning	**ICTCLAS POS**	**WR POS**
noun	n*	n, gerund
adjective	a	adj
adverb	ad?,d	adv
verb	v	v, vtr, vi, v pres
auxiliary	u	v aux
quantity	m	n
pronoun	r	pron
preposition	p	prep
conjunction	c	conj
idiom	i, l	any
time	t	n

ICTCLAS POS system, which jieba adopts, also designates several special class for the following auxiliary words: 了/ul, 的/uj, 过/ug, 地/uv, 着/uz. We don't use these information in this strategy but will use them later.

Applicability:

data set:	**dev. set**	**test set**
sentences:	10	5

Strategy 2: Unigram model

For each Chinese word, strategy 1 simply picks the first POS-matching translation. This randomness caused quite a few unreasonable translations, for example translating `女人` (woman) to `board`, and `不` (no) to `nay`. Our second strategy makes use of unigram model in the process of choosing from candidate translations. Among all the candidate translations, we look at the unigram score of the English

words, and pick the one with the highest unigram score.

This is a very general strategy, and affected all the sentences in the dev set and the test set. For example, it correctly chooses `woman` over `board`, and `no` over `nay`.

Applicability:

data set:	dev. set	test set
sentences:	10	5

Strategy 3: Subject pronoun

In Chinese, there aren't designated objective case for pronouns. `我`, which means I in Chinese, can be used both as subject and object. This should be properly handled when translating to English. For example, in the first sentence of development set, `她` in `她表示` is clearly a subject, but our system translates it to `her`.

As a result, we should make sure pronouns which are subjects are in subjective case in translated English. To achieve this, we use the heuristic that a pronoun is subject if it's the first word of a clause - it's the first word of the whole sentence or follows a punctuation or conjunction.

Applicability:

data set:	dev. set	test set
sentences:	4	0

Strategy 4: Number date

This is a very specific strategy. We notice that our system mistakenly translates `22/x 日/x` (which means the 22nd day of a month) to `twenty-two sun` (`日` also means `the sun` in Chinese). The pattern here is that in Chinese `日` or `号` is put after a number to indicate a date, which has the same function as using ordinal form (e.g. 22nd) in English. Although we believe this is a fairly good strategy, since there's only 1 sentence in the corpus that has such pattern, it applies to that only one.

Applicability:

data set:	dev. set	test set
sentences:	1	0

Strategy 5: Remove '到' after transitive verbs

In Chinese, `到` is often used as an function word after verbs and doesn't have a concrete meaning. We see multiple examples of this in the dev set, including `打/v 到` (hit) and `联系/v 到` (contact), where the function word `到` is translated into `to`. This is not desired for transitive verbs like `hit` and `contact`.

In this strategy, we look for `到` as a function word (auxiliary word, preposition, etc) that follow a transitive verb, and remove the translation of `到`.

Although this strategy applies to multiple sentences in the dev set, there is unfortunately no such instance in the test set.

Applicability:

data set:	dev. set	test set
sentences:	2	0

Strategy 6: Past tense of verbs

Unlike in English, Chinese verbs do not come in various forms to indicate the time that the action takes place, or the subject of the action. However, there are a few function words in Chinese, including `了` and `过` which indicate that an action is

completed or took place in the past. In this strategy, we make use of these function words to infer which verbs should be in the form of past tense.

This is a general strategy and applies to two test sentences: it translates `听/v 了` as `heard`, and `离开/v 了` as `left`.

Applicability:

data set:	dev. set	test set
sentences:	3	2

Strategy 7: 着 as function words after verbs

Another verb form related characteristic in Chinese is that people use `着` (a function word) after verbs to indicate a fact or a state. In English such meaning is usually conveyed through the present tense of verbs. Therefore our strategy is to remove `着` after verbs.

This is a general strategy that deals with a commonly used Chinese language feature, although we only see one sentence in the test set using this feature (`爱/v 着 我的 祖国` as `love my country`).

Applicability:

data set:	dev. set	test set
sentences:	1	0

Strategy 8: Third-person singular forms of verbs

This is yet another verb form related strategy that deals with third-person singular cases. We try to identify the cases where the subject of a sentence or clause is in third-person singular form. However, this is a very hard problem in Chinese, because there is no explicit plural form of nouns in Chinese, therefore it is hard to tell whether the third-person subject is singular or not. In this strategy, we only deal with the cases where the subject is a pronoun, including `他` (he), `她` (she), and `它` (it). When we see a third person pronoun as the subject of a sentence or clause, we transform the predicate (identified as the first verb) into third-person singular form. For example, `她 表示/v` is translated into `she expresses` instead of `she express`.

Applicability:

data set:	dev. set	test set
sentences:	2	0

Strategy 9: Passive voice

In Chinese, passive voice is indicated with the preposition `被`. For example, `被 <v>` means `be <v>.pp`, and `被 <n> <v>` means `be <v>.pp by <n>`. In this strategy, we detect `被` followed by `<v>` or `<n> <v>`. In the first case, we translate `被` as the appropriate form of `be` verb, and change the verb to past participle form. In the second case, in addition to the previous operations, we also switch the positions of the verb and the noun, and add a `by` between them. For example, `被 对方/n 碰/v` is translated into `is touched be the other` (where the subject is `woman`).

Applicability:

data set:	dev. set	test set
sentences:	2	0

Strategy 10: remove 的 in “adj 的 n”

In Chinese, when we use a adjective word to describe a noun, it use a `的` after the adjective while `的` itself does not stands for any meaning. However, in English, the adjective itself would include the meaning to express and there is just not additional information needed to be translated by `的` in English. Thus, we should directly remove the `的` in the pattern `<adj> 的 <n>`.

Applicability:

data set:	dev. set	test set
sentences:	2	2

Strategy 11: noun1 的 noun2 -> noun1's noun2

In Chinese, there's a pattern that 'noun1 的 noun2' which means a belongs relationship between noun1 and noun2 and this should be conveyed by the phrase 'noun1's noun2'. Thus we just fetch this pattern and substitute the '的' in the middle as "'s".

Applicability:

data set:	dev. set	test set
sentences:	6	4

Strategy 12: 'noun 的' -> corresponding adjective

This is special case of strategy 11. As if noun1 is one of [I, you, she, he], we need to translate them into [my, your, her, his] correspondingly.

[我的，你的，她的，他的] in Chinese and [my, your, her, his] in English are extremly high frequent words. This is a quite general strategy that deals with a commonly used Chinese/English language feature, although we only see one sentence in the dev set using this feature. This could concludes the fact that our dev set and test set are quite small to convey certain language features very well.

Applicability:

data set:	dev. set	test set
sentences:	1	0

Strategy 13: 还/d 会/v ->还会

This is a strategy designed specifically for the limitation of our word segmentation mechanism, which would always segment '还会' into '还' and '会'.

'还会' is a very frequent phrase in Chinese. It would be translated in English as 'would also'. Additionally, as long as '还' and '会' occurs together, there is no exception that we need to combine them together and regard as one phrase.

Thus, this would be applied well in Chinese-English translation in general though this strategy was only used once in dev set.

Applicability:

data set:	dev. set	test set
sentences:	1	0

Final Result of Test Set

```
  Original: 对/p 这个/r 问题/n 因为/c 好/a 的/uj 回答/n 而/c 留下/v 好/a 印象/n 很/
d 难/a ，/x 关键/n 是/v 避免/v 留下/v 坏/a 印象/n
Translated: to this case for fine response whilst stay fine show only hard , k
ey is help stay bad show

  Original: 人性/n 的/uj 阴暗/n 、/x 政治/n 的/uj 残酷/n 、/x 美式/a 政治/n ....../
x 美国/ns 政治/n 题材/n 的/uj 影视作品/n 往往/d 把/p 这些/r 抽象概念/l 变得/v 生动/a
、/x 具体/a 。/x
Translated: humanity 's obscurity 、 action 's cruelty 、 American-style action
...... republicanism action subject 's produce often handle these abstract run
vivid 、 specific .
```

```
  Original: 宝玉/nr 听/v 了/ul 这些/r 话/n ，/x 气/v 的/uj 浑身/n 乱战/v 。/x
Translated: Baoyu hearded these word , catch great all over the body quiver .

  Original: 滑冰场/n 关了门/v ，/x 政府/n 几乎/ad 没有/v 任何/a 资金投入/n ，/x 许多
/m 教练/vn 都/d 离开/v 了/ul 俄罗斯/ns 。/x
Translated: skating rink was closed , state about lack any capital investment
, much coach already parted Russia .

  Original: 有人/r 担心/v ，/x Facebook/eng 这种/r 更/d 重视/v 用户数量/n 而/c 不是
/c 营收/n 的/uj 做法/n ，/x 是/v 在/p 给/p 新/a 的/uj 互联网/n 泡沫/n 推波助澜/i 。/
x
Translated: somebody concern , Facebook of this kind more prize number of user
s whilst but revenue 's course , is in for new web head fuel .
```

Error Analysis

After applying our heuristic strategies, there still exist significant errors in the translations.

One of the most common errors is that we choose the wrong word-to-word translation from the dictionary. In our system, for each Chinese word, we are choosing the POS-matching translation with highest unigram score. While this helps in many cases, it introduces new errors to the translation. For example, `嘴` is translated into `bill`, while it really means `mouth` in the sentence. Apparently the dictionary has both `mouth` and `bill` for the same Chinese word, but `bill` wins out because it has higher unigram score in our unigram training corpus. One way to solve this problem is, instead of just choosing the word with highest unigram score, also take into consideration the probability that the Chinese is translated into each of the candidate translations. In the same example, we would look at P(`嘴` is translated to `bill`) and P(`嘴` is translated to `mouth`), and score word `x` with a formula like `unigramScore(x) * P(Chinese word is translated to x)`. Since P(`嘴` is translated to `mouth`) is potentially much higher than P(`嘴` is translated to `bill`), this strategy will be more likely to pick the right translation. However the dictionary that we use do not provide the probability of each translation.

More generally speaking, it is common that one Chinese word has multiple meanings, which in some cases are very different from each other. This makes it a challenge to pick the right word that fits the context, and requires more knowledge than just the probability that the Chinese word is translated to the English word. For example, it may help to use a bigram or trigram model to help pick the correct translation of each (assumword ing the order of words are adjusted properly), or look at other words in the sentence and consider how likely each candidate translation co-occurs with the other words in the sentence.

Another important issue is that the sentence structure in Chinese is usually very different from that in English. For example, the word by word translation of `Facebook这种更重视用户数量而不是营收的做法` would be `Facebook this more emphasize user number but not revenue way`, but it actually means `The approach that Facebook is adopting which emphasizes more on user number than revenue`. There is a clause in the sentence. In the Chinese sentence, the clause comes before the subject, whereas in English the subject goes first. Our system does not deal with the sentence structure, therefore is not capable of adjusting the order or words on a sentence level. One way of addressing this error is to analyze the common patterns of Chinese sentence structure (e.g. clauses), and construct a mapping to the corresponding English sentence structures.

Finally, we think that the Chinese language is too complicated to be accurately modeled by tens, even hundreds, of handwritten strategies, and we need a much more sophisticated model (e.g. statistical translation model) to produce reasonable translations.

Google Translate

In this section we compare the translations for the five test sentences from our system and Google Translate. We also include a manual translation for each sentence so that you the reader can have a better sense of the performance of each translation.

Comparative Analysis with Google Translate

```
1. 对/p 这个/r 问题/n 因为/c 好/a 的/uj 回答/n 而/c 留下/v 好/a 印象/n 很/d 难/a ，/
x 关键/n 是/v 避免/v 留下/v 坏/a 印象/n
Ours: to this case for fine response whilst stay fine show only hard , key is
avoid stay bad show
Google: Good answer to this question because while a good impression is diffic
ult, the key is to avoid a bad impression
Manual: It's really difficult to leave a good impression with an excellent ans
wer to this problem. The key is to avoid a bad impression.
```

Analysis: the two systems both do a good job translating the adjectives. The translation google and our MT agreed with makes sense. And since the context is not quite specific, both `question` and `case` are acceptable translations to `问题`.

However, the structure of this sentence is very complex, which Google Translate handles better than ours. Also, Google Translate has a good mechanism for adding articles (notice `the` before `key`). Our system has the problem of not choosing the optimal verb as well, e.g. it selects `stay/vi` instead of `leave/vt` with `impression`.

For Google Translate's result, it would make more sense to use 'because of' rather than 'because', which is a conjunction, where 'for' makes more sense here.

Finally, both systems don't handle `很`, which means `very` or `really`, well.

```
2. 人性/n 的/uj 阴暗/n 、/x 政治/n 的/uj 残酷/n 、/x 美式/a 政治/n ....../x 美国/ns
政治/n 题材/n 的/uj 影视作品/n 往往/d 把/p 这些/r 抽象概念/l 变得/v 生动/a 、/x 具体/
a 。/x
Ours: humanity 's obscurity 、 action 's cruelty 、 American-style action .....
. republicanism action subject 's produce often handle these abstract run vivi
d 、 specific .
Google: Humanity's dark, brutal politics, American politics ..... American po
litical movies and television work is often put these abstractions become vivi
d and specific.
Manual: The darkness of humanity, the cruelty of politics, politics of the Uni
ted States... movies and television works about American politics usually make
these abstract concepts vivid and concrete.
```

Both systems do generally well on translating the nouns and adjectives, although Google Translate did slightly better.

Google Translate translates `阴暗/n` to `dark/adj`, which is clearly an error, while our system gets the POS right but selects an improper word.

Google Translate does better on the rest of words. For example, `politics` is clearly more suitable than `action` for `政治` in this context. Our MT system pick `action` from all the translation of `政治` because it has the highest unigram score. However, it does not make sense since we did not count in the probability of Chinese-to-English translation. Although P(action) > P(politics), P(action|政治) < P(politics|政治).

Google's translation 'movies and television work' must be a work of statistical model since this is quite hard to get from a single word translation of `影视作品`.

There's a pattern of `把 <noun> <vt>` in this sentence where `把` is an auxiliary and the structure should be reversed to `<vt> <nou>`. Google Translate captures this structure correctly and does not translate the auxiliary, while our system mistakenly translates `把` to `handle/n`.

```
3. 宝玉/nr 听/v 了/ul 这些/r 话/n ，/x 气/v 的/uj 浑身/n 乱战/v 。/x
Ours: Baoyu hearded these word , catch great all over the body quiver .
Google: Baoyu heard these words, the gas did shake.
Manual: Baoyu was so angry that his body quivered after hearing these words.
```

Again, both systems succeed getting the basic meaning correct. For example, `shake` and `quiver` are both valid translations for `乱战`. Since we do not have a greater context, choosing either is fine.

Our system has better translation for `浑身`, which means `all over the body` or `body`. Google Translate, with no reason, just omits the word.

On the other side, Google Translate does better in:

1. it uses correct past tense of `hear` where our system mistakenly added `ed` to it, and
2. it transforms `word` to `words` because `这些`, which means `these`, indicates a plural noun

Finally, this sentence is from a novel in the 19th century. `的` is used as the adverbial modifier (contemporary Chinese uses `得` to indicate a certain degree) instead of adjective modifier, both systems failed to handle it. Our system successfully translates `气` to a verb but failed to omit the auxiliary, while Google

Translate mistakenly translates `气` to the noun `gas` but successfully omits the auxiliary.

```
4. 滑冰场/n 关了门/v ，/x 政府/n 几乎/ad 没有/v 任何/a 资金投入/n ，/x 许多/m 教练/vn
都/d 离开/v 了/ul 俄罗斯/ns 。/x
Ours: skating rink was closed , state about lack any capital investment , much
coach already parted Russia .
Google: Skating rink shut the door, almost no government funding, many coaches
have left Russia.
Manual: With skating rinks closed and little funding from the government, many
coaches left Russia.
```

Both systems preserve the original meaning of the original sentence and have accurate translations for nouns.

Our system did better in translating `关了门`, which means `be closed` or `stop doing business` in this context. Google Translate, however, does a direct translation.

Google Translate does better in treating `coach` as a countable noun and translating it to `many coaches`. Plus, it translates `离开`, which means `to leave (a place)` in this context, to more idiomatic `leave`. Our system doesn't do as good.

Finally, both systems lack an understanding of the structure of the whole sentence. There's a casual-effect relationship - the coaches left Russia because of depression of skating in Russia. So either translate the first two clauses to a `because` clause or using `with` to make them a cause is better.

```
5. 有人/r 担心/v ，/x Facebook/eng 这种/r 更/d 重视/v 用户数量/n 而/c 不是/c 营收/n
的/uj 做法/n ，/x 是/v 在/p 给/p 新/a 的/uj 互联网/n 泡沫/n 推波助澜/i 。/x
Ours:
somebody concern , Facebook of this kind more prize number of users whilst but
revenue 's course , is in for new web head fuel .
Google:
Some worry, Facebook is more emphasis on the number of users of this revenue r
ather than practice, is to the new Internet bubble fueled.
Manual: Some worry that, Facebook's practice of overemphasizing number of user
s instead of revenue is helping intensify a new wave of dot-com bubble.
```

Both systems failed to translate this sentence well. Considering the complicatedness of this sentence, this result is reasonable. The first part is the subject and predicate of the main sentence and the following two parts are the objective clause. In the clause, the first sub-clause is the subject and the second is the predicate.

To be fair, Google Translate does slightly less bad by getting the subject of the main sentence right and having better translations for certain words like `重视 -> emphasis`, `做法 -> practice`.

In particular, `推波助澜`, which means to help intensify or aggravate, `is translated to` fuel `in both systems`. Fuel `does have a meaning of to intensify but just doesn't fit with bubbles. Also,` 互联网泡沫 `is a term which is` dot-com bubble` in English.

Sub-Conclusion

It is clear that Google Translate is a better Chinese-to-English translation system than our system which is built upon heuristics from 10 sentences. Meanwhile, it is surprising to see our simple translation system has a satisfactory performance on several sentences.

We can see that statistical translation is a powerful way to do machine translation. The state-of-the-art NLP model adopted by Google Translate often can produce translation from Chinese to English of human-level quality.

Meanwhile, it is again worthwhile to mention that Chinese is indeed a difficult language to translate. Correct POS tagging, a decent vocabulary and basic grammar checking is just a start. Context dependency and long-range dependency have to be considered and handled as well. With advancing ML research, we hope to see machine translation of Chinese become better and better.

Conclusion

Our machine translation system adopts various heuristic strategies, and performs significantly better than the baseline translation. However, we only tackled a tiny fraction of the Chinese-to-English translation problem. There are many ways that our system may be furthered improved, including introducint more statistical information, analyzing sentence structures, adopting statistical translation models.