Chih-Chiang Wei (ccwei)
Andy Mai (andymai)
Kevin Miller (kjmiller)
PA 7 Write-Up
03/08/2013

Translation Chinese to English

**General Observations/Introduction**

We chose to translate Chinese as our foreign language for the assignment. Chinese is non-alphabetical. Words and phrases are constructed using characters, the smallest units of the language that are analogous to the alphabet in English. The issue, however, is that there are thousands of characters rather than just 26 letters of the alphabet. As a result, we had to construct a lexicon where all characters in our paragraph, as well as any phrases that can be formed using any combination of these characters, are defined. One problem we faced was that many of the words in our paragraph exist colloquially and are not formally defined in the dictionary. Therefore, we had to resort to the definitions of the individual characters that make up the word and string the definitions together. This can be problematic because when the definitions of individual characters tend to change when they are part of a word. For example, the first sentence of our translation has the phrase “all beautiful.” In reality, the Chinese word intends to say “all America.” Since the two characters that comprise the word “all America” is not defined in the dictionary, our algorithm translated the word to “all beautiful,” since “America” in Chinese means “beautiful country” literally.

Another challenge that we face is segmentation of words. There are no whitespaces in Chinese, so we cannot use whitespace tokenization to extract words from sentences. We decided to use a greedy approach to segment the words, because the algorithm works fairly well in Chinese.

Chinese also does not have tenses. Rather than conjugating verbs to past tense, Chinese uses context words to convey past actions. For example, “it attract already approximately two-thirds adult people..” is perfectly acceptable, because the audience can infer from the sentence that the action of attracting has happened. The lack to tenses is a challenging problem to translate Chinese into English, because it is difficult to extract context clues that indicate past actions and give the correct conjugation of the verb.

Idioms also appear frequently in many Chinese writings. They are usually four characters long but have deep meanings behind them. Translating them character by character literally will lead to phrases that make little sense. For example, the idiom “流连忘返” appears in our paragraph, and it means lingering. However, our translation cannot recognize the idiom and translates it literally to “flow even forget return.”

After translation, we also have to do some post processing like capitalizing the first word of the sentence and remove space between punctuations to make it look more like English sentences.

**Original Test Document (Chinese)**

尽管Facebook是目前全美最受欢迎的社群网站，它吸引了大约三分之二的成年人流连忘返。但是最近却有研究报告指出，人们似乎有点退烧了。但是专家说，他们不是厌倦或不再使用，只是暂时给个喘息的空间。这其中又有4%的使用者表示，他们减少使用是因为他们觉得隐私没有被保障，甚至因此而砍掉自己的帐号不再使用。不过有趣的是，很多人在短暂"休息"后，又会回到Facebook的行列。在研究中，18到29岁的年轻人中，有42%的人表示他们花在Facebook的时间越来越少；同时，也有23%，年纪在50岁以上的大人也减低了使用量。专家说，这个现象称为"social reckoning"（社交微调），毕竟社群网站对人们来说还算是相对新鲜的文化冲击，大家会有不同阶段的反应与调适。渐渐地，他们可能会开始问自己，那些网路上的事情跟我有什么直接关联？我的朋友们现在到底在做什么？我多久没见到他们了？

Human Translation:
Although Facebook is currently the most popular social networking website in America, which draws in about two-thirds of adults. While recently, there is a research report pointed out, it seems that people lose some interest in it. But experts say people are not tired of it or no longer use it, just temporarily need a breathing space. Among them, there were also 4% of the users who said that they reduced the use because they felt that privacy was not guaranteed, even shutting down their accounts because of it. In this study, 42 percent of people, ages from 18 to 29, indicated that they spend less and less time on Facebook. Experts say this phenomenon is called "social reckoning" (social fine-tuning), because, after all, social websites are still a relatively new culture shock to people and there are different phases of reaction and adaptation for everyone. Gradually, people might start to ask themselves: what do those things on the internet directly relate to me? What are my friends doing now? How long has it been since I last saw them?

**Direct Translation Output**

Without rules:

*despite Facebook is currently all beautiful most bear welcome NULL social network website , it attract already approximately two-thirds NULL adult people flow even forget return . but recent nevertheless have study report indicate , people seem a little retreat burn already . but expert say , they not be bored or no more use , merely temporary give individual gasp for breath NULL space . this among again have 4 % NULL user express , they lessen use is because they think privacy do not have by ensure , even that is why and chop off self NULL account number no more use . however interesting NULL is , many people in short rest after , again be able to return to Facebook NULL possession . in study within , 1 8 to 2 9 year NULL young people within , have 4 2 % NULL people express they spend in Facebook NULL time less and less ; at the same time , too have 2 3 % , age in 5 0 year above NULL adult too lower already volume of use . expert say , this individual phenomenon be called social interaction tweak , after all social*

*network website pair people with regard to also regard as is relatively fresh NULL culture impact , everyone be able to have different stage NULL reaction and harmonize fit . gradually NULL , they might be able to begin ask self , those Internet on NULL thing follow I have what direct be related ? I NULL friend present finally in do what ? I how long have not see to they already*

(The NULL corresponds to the spurious character "的" that acts a glue word in Chinese but has no direct translation in English. )

With rules:
*Despite Facebook is currently all beautiful the most welcome bear social network website, it attracts already approximately two-thirds adult people flow even forget return. But recent nevertheless have study report indicates, people seem a little retreat burns already. But expert says, they are not bored or no more use, merely temporarily give individual gasps for breath space. This among again have 4% user expresses, they lessen use is because they think privacy does not have by ensure, why even that is and chop off self account number no more use. However interesting is, many people in short rest after, again is able to return to Facebook possession. In study within, 18 to 29 years young people within, there is 42% people express they spend in Facebook less time and less; at the same time, too have 23%, age in 50 years above adult too lower already volume of use. Expert says, this individual phenomenon is called social interaction tweak, after all social network website pair people with regard to also regard as is relatively fresh culture impact, everyone is able to have different stages reaction and harmonize fit. Gradually, they might be able to begin to ask themselves, what direct do I have those Internets on thing follows to is related? what do my do present friend finally in? how long I have not see to they already?*

**Rules**

Rule #1: Add 'the' in front of any RBS (ex: the most)
In English, "the" always appears before superlative adverbs like "biggest," "craziest," and "most." Chinese does not have this rule, as "the" does not translate to anything meaningful in Chinese. Therefore, our first rule is to add the article "the" before superlative adverbs in order to make the English translation more readable. One instance that the rule applied is changing ***most*** to ***the most*** in the first sentence

Rule #2: Not be -> be(are) not
In Chinese, negation is expressed as 'S 不是 O'(S not be O), but in English the order is reversed, we say 'S is not O'. An example of what this rule fixed is changing *but expert say they **not be** dreary* to *but expert say they **are not** dreary*.

Rule #3: Get rid of space between digits(_CD)
Example:  "1 8 to 2 9" becomes "18 to 29".  Our Chinese-English dictionary only contains numbers from 0 to 9, so we treat each digit as a separate word in translation.  This rule puts the

digits back together.

Rule #4: Swap adjective location with noun( NN JJ[R|S] -> JJ[R|S] NN)
Adjectives in Chinese can be placed both before and after the noun. In English, however, adjectives almost exclusively appear before nouns. This rule swaps the location of the adjective and noun, if the adjective appears after a noun. For example, this rule changes ***bear welcome*** to ***welcome bear*** in the first sentence.

Rule #5: Pluralize NN if it follows a number that is at least 2 or the words “those”, “these”, “different”, “many”, or “multiple”
Examples: “age in 50 year” becomes “age in 50 years”; also, “everyone is able to have different stage” becomes “everyone is able to have different stages”; in addition, “those Internet” becomes “those Internets” (the fact that our translation had “those Internet” in the first place is another problem in and of itself - the translation was supposed to be “those things on the Internet”). In Chinese, the reader uses such preceding keywords along with other context cues to infer that certain words should be plural - in order to translate into English, we need a rule like this to do that inference.

Rule #6: Move WP to the beginning of the sentence if it’s a question sentence.
The way question sentence is expressed is different from English, English almost always starts with WH, but Chinese doesn’t put them in the beginning of the sentence or sometimes even ignore those WH words. For example the sentence ‘*I* ***how long*** *have not see to they already?*’. The correct translation should be ‘How long has it been since I last saw them?, but it’s a fairly complex sentence because it involves in past perfect tense as well. But at least we should move ‘how long’ or other _WP to the beginning for the sentence and move the subject next the WP as well.

Rule #7: Personal pronoun + “NULL” ⇒ Possessive pronoun:
Example: “I NULL friend” becomes “my friend”. “NULL” is a sentinel for “的”, which is a “glue” word that turns personal pronouns into possessive pronouns. In this context, it is analogous to the English “ ‘s “, although that’s not the only role this “glue” word can play.

Rule #8: JJ + VB ⇒ RB + VB
Example: “temporary give” becomes “temporarily give”. In Chinese, one can infer whether a word is an adjective or an adverb based on whether it is modifying a noun or a verb; our dictionary always translates to the adjectival form, so this rule simply carries out the adjective-versus-adverb inference and expresses the result in English using a somewhat rudimentary spelling change rule, in which words ending in “y” get an “ily” ending and otherwise get a “ly” ending.

Rule #9 - "have" (VB) in the beginning of the sentence or after comma or semicolon or conjunction ⇒ “there is/are”
Whether we choose “is” or “are” depends on whether the next noun we run into in the sentence

is an NN/NNP or NNS/NNPS. If we don't run into any noun, then by default we choose "is". We think this rule makes sense because "there is/are" most likely is referring to the subsequent noun that is closest to it.
Examples: "have study report indicate" becomes "there is study report indicate", and "have 42 NULL people" becomes "there are 42 NULL people".

Rule #10 Change be-VB to the correct form (is/are)
Chinese uses the infinitive "be" exclusively, but English uses the conjugations of the word such as "is" and "are." This rule changes the "be" to the correct conjugation by considering a window of words around "be." If plural nouns and plural pronouns are present in this window, then "be" is changed to "are." Else, it is changed to "is." However, if the previous word is "to," "can," and similar words, then it will not be changed. One example that is affected by this rule is changing *they not **be** dreary* to *they **are** not dreary*.

Rule #11: Change 3rd Person PRP | NN,NNP + VB/VBP to the correct form.
There's no 3rd person singular verb in Chinese, so we need to add the conjugation to those verbs. The way we add it is when we see a VB/VBP, we check if there's a 3rd person pronoun, including 'he', 'she', 'it' in front of it within two words. If yes, we add 's' or 'es' to it depends on what verb is it. For example, it changed *it **attract*** to *it **attracts*** and *report **indicate*** to *report **indicates.***
Some of the sentences are past tense, and it would be wrong to add conjugation to the verb. But because we didn't work on tense so we think make it present tense would at least make it more fluent.

Rule #12: VB VB -> VB to VB
This rule adds the word "to" between two consecutive verbs. An example that this rule applied is changing *be able to **begin ask** self* to *be able to **begin to ask** self*. "begin" and "ask" are both verbs in this sentence.

Rule #13: Adj./ Adv + NULL => Adj./ Adv.
The NULL corresponds to the spurious character "的" that acts a glue word in Chinese but has no direct translation in English. It has multiple functions and it's very hard to write out all the rules for it. One major functionality of the '的' word in Chinese is to make the phrase become adjective, for example, in our paragraph '新鲜的' means 'fresh', and '新鲜' also means 'fresh', so when translate it into English it becomes 'fresh NULL' . Most of time Adj./ Adv + '的' has the same meaning as the Adj./ Adv, so we remove the NULL for this case.

Rule #14: If we see the word "self", then we backtrack until we run into a personal pronoun. If we run into one, then we change "self" into a reflexive pronoun appropriate for that pronoun.
Example: "they might be able to begin to ask self" become "they might be able to begin to ask themselves".
Chinese reflexive pronouns do not explicitly have number or gender, so they all translate to "self". We think that a good heuristic for figuring out what the reflexive pronoun refers to is to find the

closest preceding pronoun.  From that we can deduce the number and gender of the reflexive pronoun in English.

**Error Analysis**

- The glue word: '的'
As we mentioned above, the "的" word that acts a glue word in Chinese has no direct translation in English, and it has as multiple functions. It can either express possession similar to " 's " in English, or it can also be appended to many words to turn them into adjectives. The multi-purpose of this word '的' makes it difficult to figure out which function it is serving based on a given sentence. In this project, we decided to map '的' to NULL and discard it for our final translation. We came up with rules to patch up the broken phrases that should have been linked by the glue word '的'.

- Tense
As previously mentioned, Chinese does not have different tense verbs. Tenses are inferred by the context of the sentence or other words. It is difficult to extract context clues that indicate different tenses and give the correct conjugation of the verb. In our paragraph, the sentence '*it **attracts** already approximately two-thirds adult people*'. It should be '*it **attracted** already approximately two-thirds adult people*'.

- Multiple VB or be-VB in a sentence
Because there's no strict grammatical limitation like English that sort of limit where the verb could appear in a sentence, there could be multiple verbs in a Chinese sentence. When translate that into English, it doesn't sounds fluent because there are many verbs in a sentence. For example, '*they **lessen use is** because'* contains multiple verbs.

- Question sentences
The way questions is expressed is different from English, English almost always starts with WH, but Chinese doesn't put them in the beginning of the sentence or sometimes even ignore those WH words. And Chinese doesn't add 'do/does' into the question sentences. Verbs and tenses are important factors that make translate questions harder. Even we tried to reorder it, the sentence doesn't read very fluent, for example: *what do my do present friend finally in?* While some reads better after reordering: *how long I have not see to they already?*

- Pluralization
Chinese doesn't have pluralization, and many times readers have to infer the word is singular or plural by the context. It's hard to get the context because sometimes it's in a sentence that is many sentences away, or it's not even mentioned. For example, ***expert** says* should be ***experts** say.* But our simple system is not able to catch this.

- Idioms

As previously mentioned, idioms are difficult to translate unless we have large bitext of them. For example, the idiom “流连忘返” appears in our paragraph, and it means lingering. However, our translation cannot recognize the idiom and translates it literally to “flow even forget return.”

-Colloquialism/cultural differences
Similar to idioms but somewhat on the opposite end of the linguistics spectrum is colloquialism. These words are not officially in any dictionary, but they are commonly accepted vernacular words. Many of these words have references to aspects of the Chinese culture. For example, our translation has the phrase “**retreat burn**,” when in fact the Chinese paragraph meant to say “**cool down**.” The two phrases are similar, but if the system does not understand the colloquialism, it can produce a very perplexing translation.

**Google Translate Output**

Although Facebook is currently the nation's most popular community website, which attracted about two-thirds of adults linger. But recently, there is a research report pointed out, it seems a little fever. But experts say they are not tired or no longer used just temporarily give a breathing space.Which has 4% of the users said that they reduce the use is because they feel that privacy is not guaranteed, or even cut their own account is no longer in use. However, it is interesting to note that a lot of people will return to the ranks of Facebook after a brief "rest".18-29 year-olds in the study, 42% of people said they were less and less time spent on Facebook,; same time, 23% of older adults over the age of 50 to reduce the usage. Experts say, this phenomenon is called "social reckoning" (social fine-tuning), after all, the community website for people to still relatively fresh culture shock, that there will be at different stages of the reaction and adaptation. Gradually, they may begin to ask yourself those things on the Internet directly associated with me? My friends in the end doing? How often do I did not see them?

**Comparative Analysis**

Both our system and Google Translate seem to agree on the overall breakdown of the words in the paragraph, which means that our segmentation algorithms are very similar. While there are apparent differences on the ordering of words in the two translations, the many words themselves are very similar and have many overlaps in the two systems.

Google Translate does a better job at picking up idioms and translating their meanings. It recognized the one idiom in our paragraph and translated it correctly to “linger,” while our system translates it to “flow even forget return.” The lack of corpus data contributes to why our system could not translate the idiom. Google Translate, on the other hand, probably has access to many parallel corpora that allows them to learn and memorize many idioms.

Google Translate does a better job at reordering and correcting part of speech. For example, it correctly translate ***directly*** *associated* and *But* ***recently***, while our system has *But* ***recent*** and ***direct*** *be related*. In order to get those adverb right, we must have the correct word order first. But because some of our word order are not correct so it's hard to get those part of speech right.

Google Translate also does better that figuring out the correct tense of the verb. As mentioned, in Chinese there are many context characters that give clues to whether the action is past or present. Our system focuses heavily on the structure and word orderings of the translation, rather than detecting clues to figure out the time of the action. The difficulty of figuring the context is a severe limitation for direct translation models.

One instance in which our system outperforms Google Translate is the phrase "他们可能会开始问自己" which Google Translate translates to "they may begin to ask yourself", whereas our translator translates it to "they might be able to begin to ask themselves". '自己' is a reflexive pronoun, and with different subjects, the meaning changes according to the subject. Here since the subject is 'they', it should be 'themselves', instead of 'yourself'. Rule 11 is able to find the "they" preceding what is originally a "self" and change the "self" to "themselves". Google Translate, on the other hand, doesn't seem to do such coreference resolution, or if it does, it doesn't do it correctly, as it gets both the number and person of the reflexive pronoun wrong.

**Errors in Google Translation**

*Although Facebook is currently the nation's most popular community website, which attracted about two-thirds of adults linger.*
- It translates '全美' to 'nation' instead of 'the United States' or 'America', it might cause some confusion to people that are not in the US.
- It translates social networking website to community website. Maybe because social networking website is still a relatively new term and there're not many bitext about 社群 being mapped to social networking.

*But recently, there is a research report pointed out, it seems a little fever.*
- '退烧' is a Chinese idiom, the literal meaning is 'the fever is gone', and is used here to describe that some people lose interest in the website. It translated it as 'a little fever', doesn't convey the idea at all.

*But experts say they are not tired or no longer used just temporarily give a breathing space.*
- Sometimes in Chinese, it omits the object in a sentence, like 'no longer use (it)' 不再使用(它). Google translation didn't catch that so it also omit the object in the sentence.
- The last sentence should be 'need' a breathing space, but since it uses word by word translation here, in Chinese it use 'give', but it means 'need' in the context.

*Which has 4% of the users said that they reduce the use is because they feel that privacy is not guaranteed, or even cut their own account is no longer in use.*
- The first sentence should be ‘Among them’, ‘them’ refers to those people who temporarily stop using Facebook in previous paragraph. Because they’re separate paragraph so Google translation didn’t catch that and it use ‘Which has’.
- ‘砍掉’ is a Chinese slang, the literal meaning is cut off, but when it’s used with ‘account’, it means shutting down the accounts. Google translation didn’t catch this.

*However, it is interesting to note that a lot of people will return to the ranks of Facebook after a brief "rest".*
- ‘行列’ is a Chinese phrase, the literal meaning of it is ‘parade’, but it’s often used to describe ‘join something’, like here it means join Facebook. But Google translated it wrong as ‘ranks of Facebook’.

*Gradually, they may begin to ask yourself those things on the Internet directly associated with me? My friends in the end doing? How often do I did not see them?*
- ‘自己’ is a reflexive pronoun, and with different subjects, the meaning changes according to the subject. Here since the subject is ‘they’, it should be ‘themselves’, instead of ‘yourself’.
- The way questions is expressed is different from English, English almost always starts with WH, but Chinese doesn’t put them in the beginning of the sentence or sometimes even ignore those WH words. Google translation didn’t get correct translation for those questions.

**Responsibilities**

Chih-Chiang Wei: Built dictionary, Wrote starter code/ rules, translation, Came up with rules, write-up
Andy Mai: Built dictionary, Came up with rules, Wrote rules, translation, write-up
Kevin Miller: Came up with rules, Wrote rules, write-up