CS124 Programming Assignment 6 Report

Dated: February 23rd, 2014

Part One: A comment on the language F that you chose. You should make a brief statement of particular challenges in translating your choice of F language to English (relative to other possible choices for F), and key insights about the language that you made use of in your post-processing strategies.

We use mandarin Chinese as the source language. We faced the following challenges in translating it into English:

1. Mandarin Chinese is a logographic language consisting of individual characters [1]. As a consequence, in a Chinese sentence, white space is not used as a delimiter between words. For example, "我下个礼拜三去学校" forms a contiguous chunk even though it has five words: [我 I] [下 next] [礼拜三 Wednesday] [去 will go to] [学校 school]. To add to the challenge of finding the correct separation between words, we note that each word in Chinese can have a variable number of characters from one (e.g. 灯) to six (e.g. 拉赫玛尼诺夫). In our model, we tackled this challenge by performing the longest prefix match with pre-compiled dictionary words to determine the next Chinese word. This strategy works in the vast majority of time and it works very well in this case as the dictionary size is small, hence the probability of retrieving a longer but wrong Chinese words is small. By "long", we meant the word contains more characters.

2. Unlike English, Chinese verbs do not carry tense morphologies. For example, in the following three sentences, "我刚吃完晚饭" (I just had dinner), "我正在吃晚饭" (I am having dinner), "我等会儿吃晚饭" (I will have dinner later), the same verb "吃" is used in Chinese, even though it is used in the context of different tenses. The tense of a Chinese sentence is almost entirely determined by time signals, aspect particles, and the sentence context. Time signals can be an explicit time, e.g., 明年 (next year), or an implicit temporal adverbs, e.g., 正在 (right now). Aspect particles include "了", "的", "过" and "着" [2]. For example, if "着" follows a verb, such as "穿着" (be wearing), it signals a progressive tense. In many other cases, both time and aspect particle signals are absent, and we have to figure out the tense from the context. We tackle it using heuristics such as following the same tense as other clause in a sentence.

3. Similar to the problem of lacking tenses, the Chinese languages also do not distinguish verb morphologies for different grammatical persons, i.e., first, second and third persons. For example, "我说" (I speak), "你说" (you speak) and "他说" (he speaks) all have the same verb form for "说"; in contrast, we have a different morphology "speaks" in the case of the third person in English. To complicate the story, this grammatical person-dependent verb morphology is also tense-dependent. For example, in the past tense in English, there is no distinction in verb form for different grammatical persons (except for the special case of "be"). In translation, we attempted to choose the correct English verb by first determining the grammatical person of the sentence, and then do a dictionary look-up for various verb morphologies.

4. As mentioned, there is a single form for verbs in Chinese. Besides posting difficulties in translation with regard to tenses and grammatical persons, this single form often breaks the rule that a sentence (or a subclause connected by coordinating conjunctions or subordinating conjunctions [2]) can only accommodate one verb. For example, "He started cry" is incorrect; instead, it should be "He started

crying" or "He started to cry". In contrast, the Chinese version "他开始哭泣" is perfectly valid. To tackle this difference, we look at verb-verb combinations and signalling prepositions, such as "to", "by", "with", to return the correct verb forms in English.

5. The overall sentence structure of Chinese is also very different from that of English. To be specific, for example, the location of spatiotemporal modifiers are different. For example, "I next week will go to school" sounds weird; "Next week I will go to school" or "I will go to school next week" are better alternatives. In contrast, these adverbial modifiers are almost never put at the end of the subject-verb-object phrase, and they are commonly put in between subject and verb. For example, "我下周去学校" has a temporal modifier "下周" sandwiched between subject "我" and verb "去". To tackle these structural differences, we employ an English language model to reorder phrases in the direct translation and put the elements in a sentence in their natural locations.

6. The cultural difference inherited in Chinese and English also poses a challenge to the translation task. In the case of Chinese-to-English translation, many Chinese vocabularies have no counterparts in English, leading to an inevitable distortion in meaning, or unnatural translation. For example, in English, "cousin" is a generic term to denote the child of one's uncle or aunt; note that gender, age, and paternal/maternal origin of the cousin are not specified. Hence, to translate, "我的堂姐", we would have to employ a rather clunky phrase, "an elder female cousin of mine who is a child of my father's sibling and who shares the same surname as mine". As another example, the Chinese language has many commonly used idioms, whose meanings are often traced back to the ancient Chinese history. For example, "鸿门宴" can be literally translated to "a feast at Hong Men", though it really means (depending on the context as well), some lure that covers a greater danger beneath. The best way to accommodate such non-overlapping vocabulary and idioms between Chinese and English is to polish the dictionary, though we admit awkwardness in translation is sometimes inevitable.

7. Measure words (or classifiers) are unique to Chinese. They almost always precedes noun or noun phrases (except for proper noun or temporal noun), and is determined by the characteristics of the noun, such as its shape, usage, etc. Exceptions happen when the noun follows a preposition (e.g. "在") or the noun precedes a locomotive particle (e.g. "上") [1]. Here are some examples of measure words (underlined): "一条船" (a boat), "三把椅子" (three chairs), "很多只鸟飞过" (many birds flew by). Measure words carry no meanings, and therefore can be dealt with by simply omitting them in translation. To improve the accuracy of identifying measure words, we also employ the Stanford Log-Linear Part-Of-Speech (POS) tagger to identify the measure words tag in the "Determiner + Measure words" structure [2].

8. Like many other languages, Chinese is polysemous, i.e. a single word carries multiple meanings. In particular, a word can serve different part-of-speech functions and therefore the corresponding meaning and grammatical structure also varies tremendously. Two common scenarios are: (i) the same word serves as both noun and verb, e.g. in "我感谢您的帮助" (I thank you for the help) and "我收到您的感谢" (I received your appreciation), the word "感谢" serves as a verb in the former case but a noun in the latter. (ii) An adjective serves as a verb. This scenario is complicated by the fact that there is no *be*-verb preceding an adjective, for example, in "他学识渊博" (He is knowledgeable), the adjective "学识渊博" effectively plays the role of a verb. To tackle this problem, we again utilize the Stanford POS tagger to select the correct form of English to pick from the dictionary. For example, we can use "VA"

tag (predicative adjective) to identify adjectives in Chinese corresponds to a verbial phrase in English.

Part Two: Your corpus of 15 sentences, with clear indication of the dev-test split.

All the sentences in training and test set is obtained from Ref. [3], a mainstream Chinese news article site.

Dev Set:

1. 2月15日，在美国弗吉尼亚州中部夏洛茨维尔市派拉蒙剧院，中国青年钢琴演奏家郎朗演奏钢琴。
2. 中国青年钢琴演奏家郎朗当晚在美国弗吉尼亚州中部夏洛茨维尔市的派拉蒙剧院礼堂举行的新春独奏音乐会，受到热烈欢迎。
3. 尽管中国青年钢琴家郎朗在今年的新春美国巡演所到之处场场爆满，征服了挑剔的美国观众，好评如潮，当地音乐评论家认为郎朗已达到人琴合一、出神入化的境界，但面对赞誉郎朗依然十分谦逊。
4. 刚过而立之年的郎朗穿着牛仔裤和休闲皮夹克，和舞台上的华丽酷炫相比十分低调，但始终洋溢着自信和真挚。
5. 正是这份对音乐的执着追求让他越来越成熟与理智，这也让他的音乐技巧和舞台表现力更加稳健，受到了业内外的广泛认可和欢迎。
6. 他说，一个现代意义上的音乐大师应该不仅仅停留在用手指来展示演奏技巧的层面上，还应该对音乐理论、教育以及未来发展做出思考。
7. 基于这样的思考，郎朗近年来对商业演出做了较大幅度的调整，把相当一部分精力投入到音乐教育、国际交流以及履行社会责任等领域，并在去年10月被联合国任命为关注全球教育的联合国和平使者。
8. 郎朗1983年出生于沈阳的一个普通家庭，3岁开始学习钢琴，14岁以优异的成绩被美国费城柯蒂斯音乐学院录取，师从院长格拉夫曼。
9. 从他们身上我学到了音乐大师的精湛艺术和理论修为。
10. 除了在已经获得非凡成绩的古典音乐领域，郎朗近年来还积极尝试古典与流行音乐的交集。

Test Set:

1. 他认为，大多数年轻人还是只对欧美的流行音乐比较熟悉。
2. 郎朗此次为期一个月、共20场左右的美国巡演从2月初开始，以肯塔基、俄亥俄、弗吉尼亚等州的中等城市为主。
3. 专访当晚，郎朗的钢琴独奏音乐会在弗吉尼亚夏洛茨维尔市的派拉蒙剧院礼堂举行。
4. 与纽约、波士顿等国际大都会相比，这些地方的文化特色更加突出，演出过程也是一次很好的了解西方历史和文化的学习之旅。
5. 对观众的追捧和专业人士的称道，这位年轻的钢琴家没有自满，依然保持着一份难能可贵的单纯和清醒。

Part Three: The output of your system.

Dev Set:

1. On February 15th, at the US Virginia Central part the City of Charlottesville Paramount theater, China youth pianist Lang Lang played piano.
2. China youth pianist Lang Lang at that night at the US Virginia Central part the City of Charlottesville's Paramount theater hall held New Year solo concert, received warm welcome.
3. Although China youth pianist Lang Lang at this year's New Year the US tour for all visited places was

sold out, won over picky the US audience, was well-received, local music critic has already thought Lang Lang has already achieved person piano has already become one, superb realm, but has already faced compliment Lang Lang still very humble.

4. Just is passing thirties Lang Lang is wearing jeans and leisure leather jacket, and on stage gorgeous cool is being compared to very low-key, but always is being filled with confidence and sincerity.
5. It is this toward music's dedication pursued let he more and more mature as well as calm and collected, this also let he's music skill and stage expressiveness more and more steady, received both inside and outside the circle's wide-spread recognition and welcome.
6. He said, a modern definition aspect's music master should not only stayed at used finger showed played skill's level aspect, also should toward music theory, education and future development contributed thinking.
7. Based on thus has been thinking, Lang Lang in recent years toward business performance has been making relatively big magnitude's adjustment, has been putting sizable portion energy has been putting to music education, international communication and has been fulfilling society responsibility etc field, and at lasted year October was by UN appointed as paid attention to global education UN messenger of peace.
8. Lang Lang in 1983 was born in Shenyang a typical family, at age 3 began studying piano, at age 14 with excellent achievement was by the US Philadelphia Curtis conservatory admitted, following Dean Graffman.
9. From they body I had learned music master superb art and theoretical training.
10. Except at already acquire extraordinary achievement's classical music field, Lang Lang in recent years also actively has been trying classical as well as pop music intersection.

Test Set:

1. He thought, majority young people still only toward Europe and America's pop music relatively familiar.
2. Lang Lang this time lasted one month, in total 20 field, around's the US tour from February the beginning began, with Kentucky, Ohio, Virginia etc state's median city as the focus.
3. Interviewed at that night, Lang Lang's piano solo concert at Virginia the City of Charlottesville's Paramount theater hall held.
4. As well as New York City, Boston etc international metropolitan is compared to, these place's culture characteristic more and more outstanding, performance process also one very good understanding the West history and culture's learning trip.
5. Toward audience's is being highly sought after and professional praise, this young's pianist is not having complacency, still is keeping a very valuable simplicity and clear-headedness.

Part Four: For each post-processing strategy you implement, a description of what differences between Language F and English that strategy was designed to address. Make sure you motivate the strategies by pointing to the characteristics of the dev set that led you to design them.

Processing One: Chinese word identification and tokenization

Problem to address

As mentioned in Part One of the report, as a logographic language, Chinese does not have natural delimiter of white space as compared to English. Nevertheless, a good tokenization strategy to identify words in Chinese is essential to the correctness of the translation at the most basic level, i.e., the vocabulary level. It is hard to determine where to separate a string of atomic Chinese characters, as a

word in Chinese can contain a variable number of characters. For example, “左” and “右” means “left” and “right” in Chinese, but when they are put together, “左右” means “approximately” or “manipulate”. Apparently, tokenizing such composite words differently will lead to drastically different translations. One way to tokenize is to use Stanford NLP Word Segmenter, though the segmenter has a high error rate. Hence, by observing the dev set, we devise our own tokenization strategy that has a significantly better performance.

Insight

Observing our dev set, we notice that if a composite word appears in a sentence, the longest legal word should be treated as a token. By *longest*, we mean the token contains the most number of Chinese characters. By *legal*, we mean the word appears in the Chinese vocabulary. For example, one phrase in the dev set is “钢琴家”, and in fact “钢” (steel), “琴” (zither), “家” (home), “钢琴”(piano), and “钢琴家” (pianist) are all valid entries in the Chinese vocabulary. In this case, the correct translation is to treat the entire token as a whole, i.e. “pianist”. It is possible to have two words, meant to be separate, appear side-by-side in a sentence. Our approach will fail in that case. However, this is extremely rare and it takes a native Chinese speaker some time to conjure up such an example.

Solution and Result

Based on our insight, we run a forward greedy search algorithm for the longest legal token in our Chinese text. This works extremely well, in part because the dictionary size is small (as it only contains legal words present in the 15 sentences) and therefore the probability of encountering a long, legal but incorrect token is small. As expected, this strategy works correctly for *all five* test set sentences, outperforming the Stanford NLP Word Segmenter. This accurate result of word tokenization in all the sentences ensures the fidelity of translation, and provides a solid foundation for subsequent post-processing strategies that achieve a high fidelity in our translation.

Processing Two: Output formatting, punctuation transformation, and number-based phrase fixing

Problem to address

Translating from a logographic language such as Chinese into a phonetic language such as English that is based on alphabets [1], we need to process the formatting of the output sentence, such as capitalization and white space; note that both features are absent in Chinese. In addition, Chinese has a much more elaborate set of punctuations, such as 《》、 that serve special purposes but are absent in English. For example, all titles of books, newspaper, and magazines are enclosed in 《》, such as 《哈利·波特》 (*Harry Potter*). Hence, we need to transform these punctuations into their English counterparts depending on the context as well. Lastly, when digits are intermixed with characters, the traditional Chinese idiom differs from the English counterparts. For example, “3岁” is translated as “three age” in the base line, though in English, “at age 3” is the correct translation. Hence, we need a systematic way to handle these formatting, punctuation and character-digit intermixing.

Insight

From our dev set, we observe that all three issues have standard formatting in English. For example, the first letter of an English sentence is capitalized, and white space is inserted between two consecutive tokens with the exception of punctuation token. To correct for unseen punctuations in Chinese, we find their approximate counterparts in English. For example, “、” is used in Chinese for enumeration (as in the dev set sentence “投入到音乐教育、国际交流以及履行社会责任”), and hence one way is to link “、” to comma in English, that is used in enumeration of items in English.

Solution and Result

We hence use a series of regular expressions in our post-processing step to correct for the sentence

formatting, punctuation handling and digit-character intermixing. The result in our test is very satisfactory, partly because our dev set has a quite comprehensive coverage of different scenarios, and partly because the handling of these three areas are very standard. We can see in our translated test set that *all five* sentences have the correct format, punctuation and digit handling (e.g. "2 month" is correctly transformed into "February") after this step of post-processing. Besides, notice that in Chinese, people often omit propositions such as "at" and "on", but they have been correctly added in the test set translation, which improves both the fluency and fidelity of the result.

Processing Three: Resolving polysemous ambiguities

Problem to address

As mentioned in Part One of the report, a single word in Chinese can have multiple different meanings. To complicate the issue, these different meanings serve very different part-of-speech functions, and can only be determined through the context and the structure of the sentence. For example, in our dev set, "自信" in "洋溢着自信和真挚" means both "confident" (adjective) and "confidence" (noun), but in this context, the latter choice is correct. Hence, in order to provide a both a faithful translation and to ensure the grammatical correctness (which is related to fluency), we need a mechanism to select the correct meaning of the Chinese word.

Insight

The key observation is that in most cases in our dev set, the part-of-speech determines the meaning of a given Chinese word. Using the same example of "洋溢着自信和真挚", "自信" in this case is an object following the main verb of the subclause, "洋溢" (be filled with). Hence, we deduce that the noun option (confidence) is a better choice than the adjective version (confident). There are many other similar examples, such as "认可", "欢迎", etc. Thus, by decomposing the sentence into its constituent POS parts and analyzing the individual structure and function, we are able to accurately resolve the ambiguities for polysemous words. Nevertheless, we note that there are exceptions where our method does not apply. For example, sometimes, as in English, the word is used sarcastically to imply the opposite meaning, such as "你真够意思" which is often used to imply the negated version. Correctly determining the meaning in these cases can involve sentiment analysis and other higher level language processing, which is not covered in our project.

Solution and Result

To best identify the part-of-speech of tokenized Chinese text, we employ the Stanford NLP POS tagger, based on the maximum entropy and cyclic dependency network approaches [5, 6]. The model is trained on a combination of CTB7 texts from Chinese and Hong Kong sources with distributional similarity clusters, which has a 93.99% accuracy on a combination of Chinese and Hong Kong texts and 84.60% accuracy on unknown words [4]. On the other hand, we annotate our dictionary with the Penn Chinese Treebank (3.0) POS tags [2] to resolve words that have multiple meanings and POS functions. To resolve polysemous ambiguities, our algorithm first search for matching tags. If no matching is found due to the small error in POS tagging of the source Chinese text, we relax the tags by matching a broader categories of tags (for example, instead of searching for temporal noun (NT), we search for all nouns (NT, NR, NN)). In the worst case of no matching even in the broader category, we probabilistically select a tag according to a uniform distribution.

We see that this method successfully resolves many polysemous ambiguities in the test set for *all five* sentences. For example, multiple meanings or POS functions in "自信", "巡演", "举行", "演出", "单纯", and "清醒" are all correctly resolved in the translation output. In fact, as later we will compare our result to that from Google Translate later, we do a better job in resolving these ambiguities,

which make the translation faithful to the original meaning and fluent in terms of grammatical correctness.

Processing Four: Determine verb morphology: tense and grammatical person

Problem to address

One important difference between Chinese and English, as elaborated in Part One, is that the Chinese language does not differentiate verb morphology for tenses and grammatical persons. In other words, the same verb is used in all tenses, for all persons. For example, the same verb "做" is used in "我昨天做实验" (I did the experiment yesterday), "我正在做实验" (I am doing the experiment right now), "我刚做完实验" (I have done the experiment), and "我明天做实验" (I will do the experiment tomorrow), and even in the third person "他做实验" (he does experiment); on the other hand, the verb "do" has morphed into "did", "doing", "done", and "does". As there is no way to determine the tense and grammatical person from the Chinese verbs themselves, we have to look at the context of the sentence.

Insight

Studying the dev set, we find that although the verbs in Chinese do not carry tense information, we would be able to infer the tense from aspect particles such as "着", "的", "过", and "了". For example, in "征服了挑剔的美国观众", the aspect particle "了" following the verb "征服" indicated a past tense. In another example in the dev set, "但始终洋溢着自信和真挚", the aspect particle "着" indicates a progressive tense. In addition, time signals such as a specific date, e.g. "2月15日" (February 15th), temporal modifiers, e.g. "今年" (this year), and other adverbs that describe time relativity, e.g. "刚刚" (just now) also help to determine the tense of the sentence. As for grammatical persons, we made a similar observation that certain key words, such as proper noun and pronoun, immediately reveal the person of a sentence. For example, in "他说", "他" (he) is immediately indicative of a third person. We saw that these heuristics are very accurate, and apply to all sentences in the dev set.

Solution and Result

When implementing the above insight, we also take into account of the added complexity that not all parts of the sentence share the same tense. For example, in the dev sentence "基于这样的...联合国和平使者", the first half is present perfect progressive, and the last clause is in past tense. This variation of tenses within the sentence is largely caused by the fact that a single Chinese sentence usually express a few separate ideas; in English, that would usually be broken down into separate sentences. Hence, in our implementation, we first break down each sentence into its clauses, usually delimited by "、", "；", "：" or "，" in Chinese. The tense/person of each subsentence is then determined by identifying time signals, aspect particles, and other aforementioned contextual signals. Since very often, these signals are absent, we hence infer the tense/person for an undetermined clause by performing a nearest-neighbor search. In other words, the tense/person of a clause follows its nearest neighbor.

After the tense/grammatical person is determined, we are able to change the verb morphology accordingly. A verb morphology dictionary is built for all verbs occurring in all the sentences. It includes the infinitive form (i.e. the original form) as the key, and past, present particle, past particle and the third person form as values. For example, the verb "begin" has "began", "beginning", and "begun" in its entry. For tenses like simple past and simple future, a straightforward substitution works. For other slightly more complicated cases, for example, present progressive, a "*be* present-particle" is substituted and the exact form of "be" is determined by the grammatical person of the clause. Hence, it is important to fix tense-related morphology before changing person-based morphology.

A special case of verb morphology happens when two verbs are located consecutively in a

sentence. For example, "开始 演讲" (begin to deliver a speech), where two verbs are underlined separately. In this case, we use POS tagger to identify the pattern, and either use an infinitive marker (e.g., "to") or use the verb morphology dictionary we built to transform the verb into its present-particle form (e.g. begin delivering a speech).

The above approach correctly identified the tenses and transformed the verb forms in *all five* test set sentences. As we will discuss later, our accounting for tenses give us a major advantage over the Google Translate output, where tenses are largely incorrect. For example, present progressive tense is assigned to the test sentence "对观众...和清醒" to indicate the ongoing present state of "朗朗" (Lang Lang). As another example, "专访当晚...礼堂举行" was assigned to past tense, which matches well with the time signal "当晚" (at that night) in the first subclause of the sentence. As for grammatical person, we also see that the *be*-verb in "与纽约...学习之旅" is correctly assigned to the form in the third person present tense, i.e., "is". These correct handlings of word morphology indicate the effectiveness of our discussed procedures above.

Processing Five: Improve fluency by an English bigram language model

Problem to address

The verbatim translation results in some awkward structure in English. As discussed in the introduction, for example, the location of spatiotemporal modifiers in the two languages are very different. For instance, "郎朗1983年出生" is translated as "Lang Lang in 1983 was born" based on a word-by-word approach, while in English, the temporal modifier "in 1983" should be placed either at the start of the sentence with a comma separation ("In 1983, Lang Lang was born"), or be inserted after the verb ("Lang Lang was born in 1983"). This difference in phrasal and POS sequence does not hinder fidelity, but greatly reduces the fluency of the translation.

Insight

By running manual translation of the dev set sentences, we notice that the most fluent way of translating a sentence usually has a highest frequency in an English corpus. Hence, we adopt a statistical machine translation approach as another post-processing step. First, a bigram model is trained using a corpus (in this case, we used the New York Time corpus distributed during the class activity, as our dev/test set is also from a news article source). Second, we would find the word-permutation corresponding to the maximum probability of a translated output. In this case, our assumption is that a more likely permutation should correspond to a more natural English translation. In this example, "in [*year*] was born" is less likely to occur than "was born in [*year*]".

Solution and Result

The bigram model (with Laplace smoothing) was trained using the aforementioned corpus and the code from Programming Assignment 2. As we attempted to find the word-permutation corresponding to the maximum probability, we experimented the following ideas:

1. First, we searched through all permutations of words in a *clause* to find the most likely order based on the trained bigram model. However it did not work well for three reasons:
    a. The search space is huge, as it is factorial in the number of words in clause. For long clauses, the search is never terminated.
    b. Even for short clause, the language model sometimes break up well-formed phrases and produce an awkward expression instead.
    c. If the phrase or idiom is preserved (presumably due to their high frequency in the training corpus), reordering sometimes result in a changed meaning.
2. Second, instead of permuting at the level of words, we permuted in the unit of phrases. By

phrases, we mean a word or a sequence of words which are taken directly out of the dictionary as part of baseline translation. This would alleviate problem 1(a) and solve problem 1(b). Nonetheless, the search space can still be arbitrarily large in certain cases as a clause contains many phrases. As mentioned, this results in a very long processing time.

3. Last, to make our algorithm run within a reasonable time frame, we adopted the strategy to only permute phrases locally. We look at five consecutive phrases at a time and permute them, finding the sequence with maximum probability. This runs much faster. Also, it tends to favor local reordering than global ones, and thus more likely to retain origin meaning of the source.

The bigram language model approach attained a limited success. For example, the temporal modifier "in 1983" is correctly placed after the verb "was born", which conforms to the natural sequence in English. Nevertheless, it also scrambles the position of other phrases in the sentence, leading to low fidelity of the output. One way to improve is to implement a "conditional lazy permutation", i.e., attaching a cost for each permutation of phrases. This helps to preserve the meaning of the original sentence. For further discussion of errors in the language model approach, please see the section on error analysis.

Part Five: Your error analysis, including specific reference to what your code does and ideas for how further work might fix your remaining errors.

As alluded to early, here is the error analysis for our statistical approach in building the bigram (with smoothing) language model:

1. Using language model alone throws away information of source text because when we permutes the phrases, we implicitly assume that the probability of each phrase at each location is entirely determined by the local probability of a bigram. In other words, we lose the long-range order information of the source text. A more comprehensive approach would be to use a noisy channel model with distortion penalty. This would allow a modifier to move from the beginning of a clause to the end, if that is sufficiently likely according to our model, but reduce spurious reordering.
2. The Chinese language does not have determiner such as "the", "a" and "an" preceding nouns (instead, as discussed in the introduction, there are measure words/classifier attached before nouns or noun phrases). In addition, it is very common in Chinese language to omit prepositions such as "on", "in", and "at". Therefore, direct translation often misses some prepositions and determiners, which severely distorts the likelihood produced by our bigram model. Consider the clause "born in a typical family", where "born in" and "in a" have high bigram probability. If "in" is missing, "born a" will be of very low probability. Thus, missing a single preposition could drastically hinder the ability of our bigram language model to produce smooth English. To tackle this problem, we could assign a conditional probability for inserting additional prepositions or determines to the baseline sentence, according to the appearance of noun, noun phrases, or spatio-temporal modifiers. For example, "in" or "on" could be inserted if it is missing in a temporal modifier, such as a specific year, month, or day. We could then choose to keep these inserted preposition/determiner if they improve the normalized likelihood of the sentence by a threshold. By normalized, we mean the total likelihood of a given sentence divided by the number of bigrams (remember that when we insert new words, there will be additional bigrams in the sentence, and we can only fairly compare two likelihood score if they have the same number of bigrams).

Due to the above errors in our statistical approach, we did not process our test set through this post-processing step when presenting our system output in Part Three of the report.

Besides, we observed the following errors in our test set, due to our inadequacies when conforming to all English grammar rules. The errors include:

1. Lack of plural form of nouns. For example, "纽约、波士顿等国际大都会" is translated to "New York City, Boston etc international metropolitan" where "metropolitan" should be "metropolitans". This also points to another major difference between Chinese and English: nouns do not have the distinction between singular or plural forms.
2. Lack of inflected form for pronouns. For example, "他们" could be translated to "they", "them" and "their". Instead of "from they body", a better translation for "从他们身上" would be "from their body". Again, this is related to the fact that in Chinese, pronouns have the same form when serving as subject and object, and sometimes as possessive pronouns in the above example.
3. Lack of a main clause. English grammar requires a sentence to have one and only one main clause. For example, in "Although I am sick, I am still working." there is only one main clause "I am still working". But Chinese does not have such requirement and it is perfectly legal and in fact preferred to say "虽然我生病了，但我还在工作". This would be directly translated to "Although I am sick, but I am still working." Our sentence "尽管...但...谦逊" in the training set has exactly this issue.
4. Lack of conjunction. It is perfectly fine in Chinese to have several clauses in a sentence that are juxtaposed without being linked by conjunctions. For example, our translation of "郎朗...格拉夫曼" in the training set is essentially "Yang Yang was born, learned piano, was admitted to conservatory". To be grammatically correct, it needs an "and" before "was".

To resolve these errors, we can program these grammar rules into our model, just as we did for verb morphology transformation for tenses and grammatical persons. Again, the POS tagger will be helpful in determining the grammatical function of phrases. We specifically propose the following approach for each of the aforementioned errors:

1. The simple case involves the identification of cardinal and ordinal numbers [2], which can be accurately handled by the Stanford NLP POS tagger. With these tags, we would readily apply the regular expression to pluralize the nouns if necessary. In more intricate cases, for example, if we see "等", which means "etc." in English, the next noun would be plural. We might even be able to recognize phrases in the form "纽约、波士顿等国际大都会", which should be translated to "metropolitans such as New York and Boston". This can achieved by recognizing expressions in the form "... 、...等...". Besides keyword "等", unique punctuation "、" is very helpful here.
2. There are various heuristics to identify which inflected form the pronouns should take. For example, if a noun follows the pronoun or "的" follows the pronoun, the pronoun should take the possessive form. To be specific, "从他们身上" and "从他们的身上" should be translated to "from their body" instead of "from they body". As another example, a pronoun after prepositions, such as "对", should take the passive form. To be specific, "对他影响很大" should be translated to "have great influence on him".
3. These can be relatively easily handled by including regular expressions to recognize conjunctive phrases in Chinese, such as "虽然...但是..." ("although… but…"), "因为...所以..." ("because… so…"), and remove the latter conjunction to ensure the integrity of the main clause in a sentence.
4. We can solve this problem by generalizing the problem into a higher-level one: identifying clauses in sentence. If we could accurately identify the clauses, we could then analyze the relative relation between clauses and insert the appropriate conjunctions. We can rely on the basic component of a clause as our heuristic for clause identification, for example, an optional subject, a required main verb, and an optional object (and of course, optional adverbial modifiers). Again, by using POS taggers and hints from punctuations (such as commas), we would be able to deduce the clausal decomposition of a

sentence.

Part Six: The output of Google Translate and a comparative analysis commenting on your system's performance compared to Google Translate's. Show where the systems agree, what your system does better than Google Translate, and what Google Translate does better than your system.

Here is the output of and the comment on the result of Google Translate (**GT** for short) for our test set:

1. He believes that most young people or just for the U.S. and Europe are more familiar with pop music.
   Comment: GT outperforms in terms of fluency and it gracefully handles a subclause after a main clause (even though the tense is incorrect in this context). On the other hand, it fails in faithfulness as compared to our output; specifically, GT unfortunately completely alters the meaning of the subclause (a better translation would be "the majority of young people are still only familiar with European or American pop music"). GT's failure could be possibly explained by its poor treatment of genitive/associative marker "**的**" [2], which is totally ignored. Hence, the correct handling of "**的**" in our system (by using a POS tagger) increases the faithfulness of our translation.

2. Lang Lang The two-month total of 20 games from around the U.S. tour beginning early February to Kentucky, Ohio, Virginia and other states of the medium-sized city based.
   Comment: In addition to semantic errors such as "games" and "based" that alter the intended meaning (and the wrong capitalization of "The"), again, GT sacrifices faithfulness to fluency. One example is its wrong application of the "from...to..." structure that wrongly juxtaposes temporal and spatial modifiers. In contrast, our system has a relatively higher fidelity and it is fluent in the first clause ("lasted one month"). Our model pales in the second half of the sentence, as it misinterprets the particle "**的**" ('s) based on a wrong POS tag and failed in transforming the noun "**初**" as a noun modifier for "2**月**" and hence the awkwardness in the translated phrase "February the beginning began". Nonetheless, our model captures the correct tense, which is apparently missing in GT (in fact, it lacks a verb all together).

3. Interview night, Lang Lang's piano recital was held at the Paramount Theater in Charlottesville, VA auditorium city.
   Comment: GT outperforms our system in both fidelity and fluency except for the wrong subclause at the beginning of the sentence as GT failed to transform the temporal noun "**当晚**" into either a temporal modifier or a passive subclause as our system did. For the main clause, our system almost matches GT's performance, though it does not gracefully handle the passive voice as well as the location of spatial adverbial modifier ("**弗吉尼亚夏洛茨维尔市的派拉蒙剧院礼堂**") due to imperfection in our language model that deals with phrasal sequence.

4. Compared with New York, Boston and other cosmopolitan culture of these places more prominent during the show but also a good understanding of Western history and culture learning journey.
   Comment: both translations show poor result and it seems GT's aggressive pursuit of fluency further jeopardizes the original meaning. As a reference, our suggested human translation is: "As compared to New York City, Boston and other international metropolitans, these places have a more characteristic culture; thus, this performance trip is also a good learning journey for better understanding the Western history and culture." It is clear that both systems failed to recognize the two causally related clauses in the sentence, and GT tries to conform it to a single-clause by wrongly transforming a subject into a temporal modifier "during the show". On the other hand, our system preserves a larger extent of the

original meaning by being less aggressive in POS transformation. It is an example of the fluency-fidelity trade-off.

5. Pursuit and professionals on the audience's praise, the young pianist is no room for complacency, still maintained a commendable simple and sober.
   Comment: GT failed to identify the correct tense that is correctly used in our system's output. Again, GT's aggressive transformation of the first subclause "对观众的追捧和专业人士的称道" leads to a complete distortion of meaning, which is almost all preserved in our system. On the other hand, our translation wrongly handles the particle "的" after the adjective "年轻" (young) and results in "young's"; to improve the performance, we could have built a simple correction of merging "的" to an adjective (thought the reality is more complicated because as mentioned before, adjectives in Chinese can also plays the role of noun, and in that case "的" still plays the role of a genitive or associative marker). Besides, GT wrongly handles the verb ("... is no room for..."), as it probably wrongly inserts the *be*-verb for the adjective/noun "自满" though in this case (recall in Chinese *be*-verb is not used in this context); however, classifying "自满" as a noun is a better choice, as reflected in our translation.

References:

[1] Sun, Chaofen. *Chinese : a Linguistic Introduction.* Cambridge: Cambridge University Press, 2006.

[2] Xia, Fei. *The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0)*. Institute for Research in Cognitive Science. University of Pennsylvania, 2000. Web. Accessed on Feb. 21, 2014.

[3] Liu, Xiaopeng. *专访：愿做火中凤凰云中鹏 实现从演奏家到音乐大师的嬗变—访中国青年钢琴家郎朗*. Xinhua Net (Feb. 17, 2014). Web. Accessed on Feb. 21, 2014.

[4] Manning, Christopher, Dan Klein, William Morgan, Huihsin Tseng, Anna Rafferty, and John Bauer. Stanford POS Tagger Documentation, v3.3.1 - 2014-01-04. Web. Accessed on Feb. 18, 2014.

[5] Toutanova, Kristina, and Christopher D. Manning. 2000. *Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger*. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[6] Toutanova, Kristina, Dan Klein, Christopher Manning, and Yoram Singer. 2003. *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*. In Proceedings of HLT-NAACL 2003, pp. 252-259.