CS124 Assignment 6

Introduction

For this assignment, we chose to translate from German into English. Both languages are Germanic languages, and there is much commonality in the vocabulary used in German and English. However, there are key differences between the languages that posed challenges for machine translation:

- **Verb order.** Verbs in German are always in the second position of a declarative clause. In a relative clause or a clause involving modal verbs (like “must”, “can”, “should”), the verb comes at the end of the sentence.
    - When parsing a present perfect sentence (the most common way of expressing something that happened in the past) the verb in the second position is either “haben” or “sein”, and the verb that is actually describing the action done in the past is placed at the end of the sentence (or clause if there are multiple). For example, “Ich bin nach San Francisco letztes Wochenende gegangen” would translate to “I went to San Francisco last weekend” where “gegangen” is the conjugation of “gehen” the verb to go in the past.
    - When parsing a sentence with a relative clause, the verb comes at the end of the clause, whereas in english it comes after the subject of the relative clause. An example would be “Der Mann, der ein Arzt ist”, which translated into English is “The man, who is a doctor”.
    - Sometimes verbs and their subjects come later in the sentence, but in English they should come at the beginning of the sentence. For example, “Heute habe ich viele Hausuafgaben gemacht.” Where the subject is “ich” and the verb phrase is “habe gemacht.” The sentence translates to “I did lots of homework today.”
- **Extra/Missing/Incorrect articles.** When literally translating to English, it can be found that there are extra articles, particularly “the”, appear. For example, “im 1901” means “in 1901”, but would be literally translated to “in the 1901.” This comes up in dates and dative clauses, because dative clauses modify articles in such a manner such that they become indistinguishable from words that get literally translated with extraneous articles in English. In the genitive case, articles such as “die” and “der” are used to express possiveness, but are directly translated to “the” rather than “of the”. For example, in the last sentence of our development set, “Die Liste der größten Unternehmen” should be translated to “The list of the largest companies,” but the possession is not carried over when translating “der”. Lastly, articles used in a relative clause are most often improperly translated, as their role serves to carry over the subject of the previous clause into the new clause, which usually requires a *wh-determiner* such as “which” or “whatever”.
- **Adverb positioning.** Adverbs and adjective phrases in German are not always

ordered in the same way that they are in English, specifically in sentences expressing something in the past with the present perfect tense. For example, one of our development sentences “Was haben eigentlich die amerikanischen und britischen Geheimdienste gemacht?” would be translated as “What have actually the american and british secret service agencies made?”, while it would be more correct to say “What actually have the american and british secret service agencies made?”

These are just a few of the differences between German and English, but these are the most common differences that not only German speakers trying to speak English would encounter on a daily basis, but also were encountered in our development set.

Corpora

Dev set

1. Manche haben sich im Laufe der Jahrhunderte aus einem Jahrmarkt entwickelt. [http://de.wikipedia.org/wiki/Volksfest]
2. Seine Schauspielausbildung absolvierte er bis 1972 am Max Reinhardt Seminar in Wien. [http://de.wikipedia.org/wiki/Götz_Kauffmann]
3. Ich muss ganz klar sagen: Von der Existenz und dem Umfang dieses Überwachungssystems wissen wir nur durch Edward Snowden. [http://alles-schallundrauch.blogspot.com/2013/11/gregor-gysis-rede-zum-nsa-skandal.html]
4. Die Stadt wurde 1170 von den Anglonormannen unter der Führung von Richard de Clare und seinem irischen Verbündeten Diarmuid Mac Murchadha Caomhánach eingenommen. [http://de.wikipedia.org/wiki/Dublin]
5. Seit dem Beitritt Irlands zur Europäischen Gemeinschaft 1973 hat sich Dublin zu einer europäischen Metropole entwickelt. [http://de.wikipedia.org/wiki/Dublin]
6. Der erste Teil der Strecke, der Dublin Port Tunnel, wurde im Dezember 2006 nach sechs Jahren Bauzeit eröffnet. [http://de.wikipedia.org/wiki/Dublin]
7. Ihre Schauspielausbildung erhielt sie an der Staatlichen Hochschule für Musik und Darstellende Kunst Stuttgart. [http://de.wikipedia.org/wiki/Ursula_Buschhorn]
8. Sie bilden die bunte Kulisse für die rasanten und erotischen Abenteuer im wilden Agenturgeschehen. [http://de.wikipedia.org/wiki/Gute_Mädchen,_böse_Mädchen]
9. Was haben eigentlich die amerikanischen und britischen Geheimdienste gemacht? [http://alles-schallundrauch.blogspot.com/2013/11/gregor-gysis-rede-zum-nsa-skandal.html]
10. Die Liste der größten Unternehmen in Luxemburg enthält die vom Forbes Magazine in der Liste Forbes Global 2000 veröffentlichten größten börsennotierten Unternehmen in Luxemburg. [http://de.wikipedia.org/wiki/Liste_der_größten_Unternehmen_in_Luxemburg]

Test set

1. Volksfeste sind im Brauchtum verankerte regional typische Feste, die oft eine

lange Tradition besitzen. [http://de.wikipedia.org/wiki/Volksfest]

2. Sie haben aber in Wirklichkeit das Gegenteil betrieben. [http://www.linksfraktion.de/reden/edward-snowden-asyl-gewaehren/]
3. Heute ist die Stadt im Norden von einem 5 bis 6 Meter hohen Deich sowie im Süden von einem 9 Meter hohen Deich gegen Wassereinlauf geschützt. [http://de.wikipedia.org/wiki/New_Orleans]
4. Der Name ist von der römischen Goldmünze Aureus abgeleitet. [http://de.wikipedia.org/wiki/Øre]
5. Der zweite Audioflyer wurde im November 2008 in einigen Großstädten Deutschlands verteilt. [http://de.wikipedia.org/wiki/Herr_von_Grau]

Translation system

We implemented a direct translation system along with preprocessing and postprocessing strategies to improve the translation. In implementing our strategies, we used the following tools:

1. **Clause parser** that we implemented ourselves to divide a sentence into clauses.
2. **Google Translate** to create our bilingual dictionary. For each German word encountered in our corpora, we stored its part of speech and a list of possible English translations in our dictionary. We modified a python google translate scraper found online to do this (see citation at end).
3. **pattern.de** (Python module) for part of speech tagging (according to the Penn Treebank II tagset) and sentence chunking in German.
4. **nltk** (Python module) for word tokenization utility methods.
5. **The Brown Corpus** for training our homemade bigram/trigram language models.

Translations of the dev set sentences

Raw direct translations	**Our system's final translations**
some have itself in the over the centuries from a fair developed .	Some have developed over in the centuries from a fair .
its acting training completed he to 1972 at the Max Reinhardt seminar in Vienna .	He completed its acting training to 1972 at the Max Reinhardt seminar in Vienna .
I must all clear say : from the existence and the scope this monitoring system know we only through Edward Snowden .	I all clear must say : we only know from the existence and the scope of this monitoring system through Edward Snowden .
the city was 1170 from the Anglo-Normans under the leadership from Richard de Clare and his Irish ally Diarmuid Mac	The city was taken 1170 from the Anglo-Normans under the leadership of Richard de Clare and his Irish ally Diarmuid

Murchadha Caomhanach taken .	Mac Murchadha Caomhanach .
since the accession Ireland to European community 1973 has itself Dublin to a European metropolis developed .	Since the accession Ireland to European community 1973 has developed Dublin to a European metropolis .
the first part the route , the Dublin port tunnel , was in the December 2006 after six years construction open .	The first part of the route , the Dublin port tunnel , in December was open 2006 after six years construction .
their acting training received they to the state college for music and Performing art Stuttgart .	They received their acting training to the state college for music and Performing art Stuttgart .
they form the colorful backdrop for the rapid and sexy adventure in the wild agency events .	They form the colorful backdrop for the rapid and sexy adventure in the wild agency events .
What have actually the American and British secret services made ?	What actually have the American and British secret services made ?
the list the largest business in Luxembourg contains the from Forbes Magazine in the list Forbes global 2000 published largest listed business in Luxembourg .	The list of the largest business in Luxembourg contains the published from Forbes Magazine in the list Forbes global 2000 largest listed business in Luxembourg .

Raw direct translations (test set)	**Our test set translations**
festivals are in the Customs anchored regional typical Festivals , the often a long tradition have .	Festivals in the Customs are anchored regional typical Festivals , often have a long tradition .
they have but in reality the contrary operated .	They but have operated in reality the contrary .
today is the city in the north from a 5 to 6 meter high dike as well as in the south from a 9 meter high dike against water inlet protected .	The city as well as is protected in the north from a 5 to 6 meter high dike in the south from a 9 meter high dike against water inlet .
the name is from the Roman gold coin	The name is derived from the Roman gold

Aureus derived .	coin Aureus .
the second audio flyer was in the November 2008 in some cities Germany distributed .	The second audio flyer was distributed in November 2008 in some cities Germany .

We implemented the following strategies:

Reorder the subject of a sentence
(~1 points: affected 3 sentences in dev set and 1 in test set)
In German, when the subject and verb are not the first element of a sentence, then the direct translation is not fluent in English. For example in our development set we have “Ihre Schauspielausbildung erhielt sie an der Staatlichen Hochschule für Musik und Darstellende Kunst Stuttgart.” The subject of this sentence is “sie”, and the verb is “erhielt”. With a direct translation, it would seem that “Ihre Schauspielausbildung” is the subject (“Her acting education”). This is not the case, as the real meaning of this sentence is “she received her acting education…” In the case where the sentence contained a pronoun immediately after a verb, the pronoun is the subject of the sentence and this subject-verb pair can be rotated to the beginning of the sentence. If there is no pronoun, then this case becomes much more difficult to handle, as the sentence may start with a noun and the verb may be followed by a noun. For example, “Mein Freund hat ein Hund” which translates to “My friend has a dog”, but using this rule may be incorrectly interpreted as “A dog has my friend”. In this case, it is most likely that the initial noun is the subject, so we decided to leave it as so.

Fix adverb/adjective phrase order
(~1 points: affected 2 sentences in dev set and 1 in test set)
In German, the positioning of adverbs and adjective phrases typically align with English, but there are exceptions. In our development set, you can see that “Was haben eigentlich die amerikanischen und britischen Geheimdienste gemacht?” was directly translated to “What have actually the American and British secret services made?” where “eigentlich” means “actually.” This translation is a bit awkward, and would be more fluent if it was phrased as “What actually have…” In order to conform to English norms, we had to move the phrases around. We determined the part of speech using the pattern.de part of speech tagger. Whenever we see adverbs or an adjective phrase, we look back for what it modifies, and potentially do reordering

Fix order of present perfect verbs
(~5 points: affected 4 sentences in dev set and 5 in test set)
In German, if there is a verb phrase in a perfect tense (have done, has become, etc.) the second part of the verb phrase is moved to the end of the clause. For example, in our

dev-set, one sentence is: *“Seit dem Beitritt Irlands zur Europäischen Gemeinschaft 1973 hat sich Dublin zu einer europäischen Metropole entwickelt”.* In order to do this, we stored a text file of all perfect german verbs. If we ever encounter one, we look ahead for the second part of the verb phrase, and move it back. However, as some sentences in our development set have multiple clauses, it was first necessary to split our sentence into clauses, and then apply this rule on each clause. As we pulled most of our sentences from articles on the internet, past perfect verbs are very common, and thus are strongly represented in both our development and test sets.

Fix word order in relative clauses

(~1 points: affected 1 sentences in dev set and 1 in test set)
German grammar contains “relative clauses” which are delimited by a relative pronoun. An example of this is the first sentence of our test set *“Volksfeste sind im Brauchtum verankerte regional typische Feste, die oft eine lange”*. Typically, “*die*” means “the”; however, because the underlined clause has no subject, “*die*” refers to the subject of the previous clause (“*Volksfeste”*). In english, relative clauses begin with a word such as who/which/that. In German, in these relative clauses, the predicate is pushed to the end of the sentence. In order to fix this, we first found a list of potential relative pronouns. In our clause parser, anytime we see a relative pronoun, we look forward for any nouns. If none exist, this must be a relative phrase, and we split it off. Once we do this, we look to the end of the clause for a predicate, and move it in front of the relative pronoun.

Although this only affects one sentence in our dev-set, the structure of a relative clause is something very representative of the difference between the English and German languages. Relative clauses are quite common, and according to our sources (at end of document) show the difference between fluent and non-fluent german speakers.

Remove extraneous reflexive pronouns

(1 point: affected 2 sentences in dev set and 0 in test set)
German grammar allows for reflexive verbs along with reflexive pronouns (e.g. *“Manche haben sich... entwickelt…”*). When these are translated to English, the reflexive pronoun can likely be dropped. In preprocessing, we implemented a check for reflexive pronouns and removed them as necessary.

Remove extra article: “the”

(0-1 points: affected 1 sentence in dev set and 2 in test set)
Our English translation had extraneous instances of the word *“the”* (from German *das*, *dem*, *den*, or *der*). (e.g. *the over the centuries...*) We removed these extra articles to increase translation fluency. In deciding which articles were needed and which were extraneous, we used a language model, applying bigram and trigram counts from the Brown Corpus. If removing “the” gave the phrase a higher probability by some constant factor, then we removed “the” from the sentence.

Although this strategy only affected 1 sentence in the dev set, we argue that it is a general strategy and should be awarded 1 point -- in fact, it improved 2 sentences in the test set!

Add missing preposition: “of”
(1 point: affected 2 sentences in dev set and 0 in test set)
In translating to English, the word “of” was missing in certain places (i.e. before articles such as “the” or “this”. e.g. *“the first part [of] the route”*) To improve fluency, we used our bigram and trigram-trained language model to assess the probability of these phrases with and without the word “of” and to insert the word “of” accordingly.

Select among candidate translations for prepositions: “of” or “from”
(1 point: affected 2 sentences in dev set and 0 in test set)
Many of the words in our German corpora had multiple candidate translations in English. We found that the top candidate was reasonable for most words. However, the German word *“von”* in particular gave rise to errors because it should be translated to *“of”* in some instances and *“from”* in others. (e.g. *“under the leadership from Richard de Clare...”*) To disambiguate the translation of *“von”*, we used our bigram and trigram-trained language model to see whether “of” or “from” was a more probable translation.

Comparison with Google Translate

	Our test set translations	**Google Translate’s translations**
1	Festivals in the Customs are anchored regional typical Festivals , often have a long tradition .	Folk festivals are anchored in the traditions typical regional festivals which often have a long tradition.
2	They but have operated in reality the contrary .	But they have operated in fact the opposite.
3	The city as well as is protected in the north from a 5 to 6 meter high dike in the south from a 9 meter high dike against water inlet .	Today, the city in the north of a 5 to 6 meter high dike and in the south by a 9-meter-high dike is protected against water enema.
4	The name is derived from the Roman gold coin Aureus .	The name is derived from the Roman gold coin aureus.
5	The second audio flyer was distributed in November 2008 in some cities Germany .	The second audio flyer was distributed in November 2008 in some cities in Germany.

Comments on comparison to Google Translate

1. The main difference between the two are the word translations. For example our dictionary translated “Volksfeste sind im Brauchtum” as “Festivals in the Customs” rather than “Folk festivals”. We also missed the preposition “in” and the reflective pronoun “often”. Google’s translation is better for these reasons, but it is still not perfect and is missing a few prepositions.
2. The main difference is the placement of “but” is at the front of the second sentence. Our translation also misses the preposition “to”: “[to] the contrary.” Google’s translation is more readable for these reasons and because of its choice of word translations (e.g. “in fact” compared to “in reality”).
3. The most salient difference between the translations is that ours misplaces the phrase “as well as” near the beginning of the sentence. Also, out translation wrongly omits the word “today” from the sentence. Apart from these errors, our translation is comparable in fluency and accuracy with Google’s. Our system places the verb phrase “is protected” next to the subject “The city”, increasing fluency; Google Translate does not do this. Both systems use the wrong proposition in saying that the city is protected in the north by a 5 to 6 meter high dike (the two translations instead use “from” and “of”).
4. Apart from a capitalization discrepancy in “[a]ureus”, the two translations are identical.
5. Google’s translation is better because ours misses the preposition “in” (“in some cities [in] Germany”). The two translations are identical.

Error Analysis

Error: missing word “which” in sentence 1

Another error can be found in sentence 1, where there is a missing “which” in the relative clause. In the original German, the word “die” should have been translated to “which”; however, our system translates it to “the” because the alternate translation “which” is not present in our dictionary. In post-processing, we remove this extraneous/awkward “the”. To fix this sort of error, we would need to do two things:

1. (Consistently) include all possible translations for a German word in our dictionary. We could have used multiple sources to cross-validate definitions or accommodate for words belonging to multiple parts of speech.
2. Properly disambiguate the correct translation of the word (e.g. the translation of “die” to “the” or “which”). To do this, we could have used the parse tree to find out that “die” was tagged as a *wh-determiner* in the original sentence. Then, we could replace it with an appropriate English *wh-determiner* (which, whatever, whichever).

This error brings up a general fact about the difficulty of machine translations in dealing with ambiguous translations of words from the source language.

Error: wrong order of the word “but” in sentence 2
In sentence 2, our system does not move the word “but” (translated from the German “aber”) to the beginning of the sentence, as it should. This error can be traced back to an incorrect POS tag. The POS tagger we used improperly tagged “aber” as an adverb rather than a conjunction. This error may be due to a lack of context, as this sentence was contradicting the previous sentence in the article. This problem is difficult to combat, as the task at hand is to translate a given sentence, not an entire paragraph.

Error: poor word choice of “contrary” in sentence 2
The other error in the translation of sentence 2 is word choice—namely, the translation of the word “Gegenteil” to “contrary” rather than the more fluent translation to “opposite”. Our system uses “contrary” because this was the only translation in our dictionary; however, Google Translate uses “opposite” and hence produces a better translation. This can be fixed by using a dictionary that cross-validates word definitions from multiple sources to generate multiple candidate translations for German words, and using a language model to statistically determine which candidate translation word fits better.

Error: wrong phrase order of “as well as” in sentence 3
In sentence 3 of our test set, “sowie” is improperly tagged as an adverb phrase, when it should have been tagged as a conjunction. This results in the phrase “as well as” being wrongly displaced to the beginning of the sentence (as per our *“Fix adjective phrase verb order”* rule). This error can be attributed to the POS tagger, and can be fixed by using a more robust POS tagger or cross-checking POS tags with multiple sources to validate answers. For example, Google Translate suggests that “sowie” is in fact a conjunction; had we cross-checked our POS tags with Google Translate’s tags, we could have avoided errors like these.

Error: missing word “today” in sentence 3
In sentence 3 again, “heute” (meaning “today”) is being unnecessarily removed by our system. This can be attributed to our *“Fix adverb order”* rule that moves adverbs around based on the verb that they are modifying. “Heute” is tagged (correctly) as an adverb, but our system does not handle it correctly because it miscalculates the clause that “heute” is modifying. “Heute” is in fact modifying the entire sentence, so it should be part of its own clause, as Google Translate accurately identifies. One way that we could approach this error is by building an adverb function that handles adverbs separately by functional category: e.g. time, manner, and place. This would give our system increased understanding of the function of particular adverbs in sentences and lead to better translations.

Sources

List of German Conjugations
http://german.about.com/library/weekly/aa010910b.htm

List of German Relative Pronouns:
http://www.dartmouth.edu/~german/Grammatik/RelativeClauses/relatives.html
Python pattern.de module
http://www.clips.ua.ac.be/pages/pattern-de
Google translate scraper
https://github.com/terryyin/google-translate-python