CS124 PA6 Translation Writeup

Introduction and our choice of language:

For Assignment 6, we selected Spanish as the language for our translator, given two of our team member's familiarity with the language. The challenges that we faced with translating Spanish to English were comparatively few relative to other languages like Chinese, which offers no unambiguous separation between words, or Arabic, which does not explicitly identify vowels in words. While Spanish uses the Latin alphabet, it also employs a set of accent marks that are not present in English, including: á, é, í, ó, ú, ü, ñ, ¿, ¡, which are distinct from their unaccented counterparts and thus pose minor encoding issues.

Although Spanish grammar is structurally similar to English grammar, there are key distinctions. Some notable differences include: the grammatical role of object pronouns, the noun-adjective ordering, the frequent omission of subject pronouns, the use of definite articles (e.g. 'I like world peace' vs 'I like the world peace'), ambiguity over the use of the conjunction 'que', the use of prepositions in infinitives, and others. In Spanish, adjectives are frequently--but not always--positioned after the nouns they modify (e.g. "elefante rosa" or "elephant pink"), while in English adjectives precede their modified nouns in many cases. Pronouns are frequently omitted because Spanish verbs encode their subjects (for example, the conjugated verb "tengo" communicates "I have"; the pronoun "I" is unneeded). This presented an interesting challenge, as pronouns are necessary for fluency in English, but are often difficult to determine without a larger context or sophisticated parsing of syntactic relations.

One final difference is that sentence word order in Spanish tends to be less rigid than in English, meaning sentence structures can take the form of not only Subject-Verb-Object, but also VSO, OVS, and SOV.

Pre-processing strategies:

We had a basic pre-processing system, consisting of a direct translation combined with part-of-speech tagging, performed with code from the Apache OpenNLP project and an open source Spanish probabilistic tagging model.

Post-processing strategies:

Our post-processing strategies are defined in the methods below. For each method, we will describe the differences between Spanish and English that the strategy was designed to address, in addition to motivating the strategies by pointing to the characteristics of the dev set that led us to design them.

Our post-processing strategies are as follows (see Translator.java for the code):

1. resolveQueAmbiguity(tsentence);
2. detectCommonPhrases(tsentence);
3. switchAdjNouns(tsentence);
4. switchNegation(tsentence);
5. switchObjVerbs(tsentence);
6. flipQuestionWord(tsentence);
7. fixInfinitive(tsentence);
8. porBy(tsentence);
9. paraInOrderTo(tsentence);
10. addPronounToVerb(tsentence);
11. removeReflexive(tsentence);
12. Double-check strategies using counts from bigram corpus (occurs in the swap function in TaggedSentence.java)

1. Resolving 'que' ambiguity:

The word "que" is frequently used in Spanish to mean "that", although there are a number of contexts in which its meaning varies. When "que" is preceded by a comparison word (e.g. "más que"/"more than" or "más difícil que"/"more difficult than"), it should instead translate to the English word "than". We thus detect when "que" is used to make a comparison, and choose its English translation to reflect this.

Another frequent use of "que" that consistently translates to a word other than "that" in English is as part of a preposition describing time, in which "que" should be absorbed into the other part of the preposition in the English translation. For example, the direct translation of "mientras que" would be "while that", but in English the same idea is expressed simply using "while". We thus detect such cases and remove "that" in the final translation.

Without this change, the translation "In the last days, however, the issue that concerned him had not it been more than one: the daughter of the merchant that lived in the city where would arrive within of four days" from the dev set would instead have read "...more that one...".

2. Resolving common phrase issues:

In Spanish, there are several idiomatic phrases that, as phrases, take on meanings in English that differ significantly from the meanings of their constituents. For example, the commonly used phrase "sin embargo" directly translates to "without embargo", but is generally understood to mean "however". For this reason, we detect a set of frequently used idioms and replace them with their accepted English

translations. We are thus able to achieve the fluent “In the last days, however, the issue that concerned him…” from the sentence provided in the previous strategy iinstead of the confusing “In the last days, without embargo…” In a large-scale system, a strategy like this would ideally contain a larger set of such phrases, and chunk these phrases as units during tokenization using an alignment algorithm.

3. Switching adjective and noun order:

Typically in Spanish, adjectives come after nouns (for example, “the white car” becomes “the car white”). To address this, we applied a simple heuristic rule to switch any adjectives and nouns that are adjacent to each other in the sentence. In our dev set, there were a number of occurrences of adjacent adjectives and nouns that motivated this rule - for example, we wanted “weight healthy”/“peso saludable” to become “healthy weight”.

4. Switching negation order:

In English, we negate verbs by adding “not” after the verb, whereas in Spanish, verbs are negated with a “no” before the verb. We applied a rule to switch “no” with the verb following it. Although the resulting English is not entirely idiomatic (since in English we often negate verbs with “did not” or “does not”), it results in valid and easily read English. In our dev set, an example of incorrect negation order comes in sentence 5, in the phrase “no podía entender,” which directly translates as “not could understand.”

5. Switching objects and verbs:

In Spanish, reflexive and object pronouns come before the verb. In our dev set, some examples include “le preocupaba,” “le pedían,” and “se sentó,” which directly translate to “him worried,” “of him they asked,” and “himself he sat.” Switching the pronouns and the verb in these cases results in readable English (“worried him”, “they asked of him”, “he sat himself”).

6. Switch the verb order for question words:

We had one question in our dev set, whose direct translation started with “We will use in the future…?” while the English translation should read “Will we use in the future….?” We resolved this by identifying subject-verb pairs in questions (which we identified by looking for question marks) and switching the order of these pairs. Although it only occurs once in the set of sentences we selected, questions in Spanish frequently differ from their propositional counterparts only in the inflection, as indicated with “¿” at the beginning of a sentence. In English, however, we frequently invert the verb and its object to indicate that we are asking a question. That is, a question in Spanish could begin with “¿Podemos…”/”We can...”, while in English we would switch the verb and its object (resulting in “Can we...”) in order to indicate that we are asking a question.

7. Fixing the use of prepositions with infinitives:

In English, infinitive verbs are almost always translated directly with the preposition ‘to’ in front of the verb. However, in Spanish, the infinitive of a verb does not contain any prepositions. Therefore, there are a number of instances where prepositions are improperly incorporated within the sentence, such as this example from the dev set: ‘is key to help to to maintain.’ This should instead read ‘is key to help maintain.’

8. Resolving ‘por’ ambiguity

The word ‘por’ in Spanish is very common, but has multiple translations depending on context. The most common translation is “for,” but we identified when it is intended to mean “by” in constructing the passive voice. For example, from our dev set, “estas ciudades están dirigidas por personas que...”). We applied a simple rule where we translated “por” as “by” when “por” followed a participle (e.g. “dirigidas”).

9. Resolving ‘para’ ambiguity

As with ‘por,’ ‘para’ is most frequently translated as ‘for,’ but can also mean ‘in order to.’ We applied a rule to translate ‘para’ as ‘to’ when it preceded an infinitive, and then to remove the “to” from the infinitive, as consistent with rule number 7. In our second dev sentence, this rule applies to the example “es clave para ayudar …”

10. Adding pronouns to verbs without a pronoun

In Spanish, it is frequently acceptable to drop the pronoun that describes the doer of a particular verb. For example, in English, we might say ‘He went to the park’, while in Spanish, it is perfectly okay to drop the ‘he’ and have only ‘Went to the park’. This presents a significant challenge because one must determine first if there is an explicitly named doer, and if so, if the doer is a pronoun, it is necessary to identify the antecedent to determine which pronoun (he, she, them, they etc.) to use. Identifying the doer is challenging because there are numerous instances in grammar where verbs may have a well-defined doer, but the doer is relatively far from the verb, in the case of one dev set sentence: ‘that these lodgings of first, in terms of benefits, offer…” It is quite difficult to see that in this case ‘offer’ has a well defined doer, ‘these lodgings.’ Problems arise when there are sentences such as ‘the daughter of the merchant that lived in the city where would arrive’, where we can see that ‘would arrive’ is missing the right pronoun, ‘she’. Ultimately, we were only able to effectively add the third person singular gender-neutral pronoun, it. In many contexts, this is not appropriate. However, we could not reliably identify the gender and number of the antecedent, so we only add ‘it’, which improves readability.

11. Remove reflexive pronoun

In Spanish, a reflexive pronoun is frequently used with a verb when a subject is acting upon itself. For example, “I learned” would be “me aprendí”/ “I learned me”. In English, we do not include the reflexive pronoun. However, the pronoun “me” would be used in English if the subject of the verb is different from its object. This is hard to detect in cases other than the first-person, so the remove-reflexive strategy detects when the reflexive pronoun is combined with a verb conjugated in the first person, and removes the reflexive pronoun in this case.

12. Double-check strategies using counts from bigram corpus

A number of the aforementioned strategies depend on switching adjacent words of a defined set of verb forms. However, English has a number of exceptions to these rules. For example, while switching an adjective to precede the noun it modifies almost always produces a more fluent result in English, there are certain cases where this is not the most fluent ordering--for example, in one of the test sentences, switching “espalda erguida”/”back upright” results in the more awkward “upright back”. Thus, any time a strategy includes such a swap, we use a bigram corpus (from http://www.ngrams.info) to check that the bigram produced by the swap occurs more frequently in English, and only perform the swap if this is the case.

Error Analysis:

One source of error comes from the inability of our system to detect macro-level structure in a sentence. For example, instead of switching nouns and adjectives, we really should be switching noun phrases and adjective phrases - however, this would require a fairly sophisticated mechanism for creating a parse tree, and we found that our word-level heuristic rules combined with a simple English bigram model got us far. Such analysis would also allow us to effectively use punctuation (particularly with respect to commas and clause identification) to help us more effectively translate our sentences. While we did not address the higher level SVO vs VSO, OVS, SOV sentence structural differences, we wrote generalized rules for almost all of the aforementioned grammar differences, and accounted for variability in structure by probabilistically choosing reorderings based on bigram counts from a corpus of English words. Thus, although we were generally unable to detect and reorder complicated clauses, the phrases we produced were significantly more fluent than their respective direct translations.

Additionally, our system was lacking in named-entity recognition, as well as more extensive Spanish idiom recognition. For example, our system did not pick up on “con el fin de”/ “with the end of” as meaning “in order to”. This is a pre-processing step that would have solved some of the more glaring

errors in our test set translations, and would be a natural place to start in an attempt to improve our system.

In a number of rules we created, whether the rule should be applied or not depended on certain contextual elements (e.g. recognizing the subject in adding pronouns to verbs, or understanding the role of an adjective in switching adjectives and nouns), which our system was not granular enough to understand. Building out probabilistic models on a case-by-case basis to recognize the gender of the subject or to classify adjectives more precisely than our POS tagging could fix this problem.

One further challenge we encountered was the issue of accurate part of speech tagging. While the Apache OpenNLP tagger that we used was largely effective and immensely useful in helping us develop strategies, it was from time to time inaccurate with its grammatical tagging. In one example, both pronoun and verb were tagged together as verb, which resulted in an error when we ran our strategy to add pronouns to verbs lacking a pronoun (in Spanish, it is frequently okay to omit a verb's pronoun). With certain words such as 'that', which in certain contexts can function as a pronoun, an adverb, or a conjunction, it is often difficult to effectively identify which part of speech it is in a given context, especially when idiomatic rules such as dropping pronouns are applied.

Finally, we sometimes chose the wrong translation for a given word (for example, hacer -> do instead of hacer -> make). This could be improved by choosing a translation based on the bigram language model we constructed.

Comparative analysis with Google Translate:

1. ***SPANISH:*** *Los Estados Unidos no sufren, fronteras adentro, el problema de la explosión de la natalidad, pero se preocupan como nadie por difundir e imponer, en los cuatros puntos cardinales, la planificación familiar.*

   GT: *The United States does not suffer, in borders, the problem of the baby boom, but no worry as to disseminate and enforce, in the four cardinal points, family planning.*

   Ours: *The United States not suffer, borders inside, the problem of the explosion of the birthrate, but care oneself like no one to spread and impose, in the four cardinal points, the family planning.*

   Analysis: Google picks up on the "baby boom" phrase but fails to idiomatically parse the second half of the sentence. Our system has two major issues: first, it fails to recognize "The United States" as a singular entity, and second, it works against itself by having the language

model override the negation rule. A fix for the second problem would be to add "does not/did not" as other alternate English negations and then picking the best fit among "does not X", "X's not" and "not X" according to the language model.

2. ***SPANISH:*** *Siéntate en calma quince minutos cada mañana en un lugar tranquilo, confortable, con la espalda erguida y las manos cruzadas.*

   GT: *Sit quietly fifteen minutes every morning in a quiet, comfortable, with your back straight and hands clasped place.*

   Ours: *Sit down in calm fifteen minutes each morning in a quiet place, comfortable, with the back upright and the hands crossed.*

   Analysis: Google thinks the entire end of the sentence is an adjective, and so it moves the word 'place' far away from where it should be. Our system does not make this mistake, but has a less fluent translation in a couple points ("in calm" vs "quietly," "the back upright" vs. "straight" and "the hands crossed" vs "clasped".)

3. ***SPANISH:*** *Desde entonces no paró de hablarme una y otra vez hasta que me aprendí la historia de memoria.*

   GT: *Since then he stopped talking to me again and again until I learned the story from memory.*

   Ours: *Since then not it stopped to talk to me one and other time until I learned the history of memory.*

   Analysis: Google's translation misses that "stopped" should be negated, but figures out in context "historia de memoria" means "story from memory" rather than "history of memory." Our translation has the same negation issue as in the first test sentence, and misses the idiom of "una y otra vez" as "again and again." In addition, our translation inserts the pronoun 'it' to match with the verb 'stopped', instead of the correct 'he', which Google uses. An improved system would determine the gender and number of the pronoun's subject, and choose a pronoun to reflect this.

4. ***SPANISH:*** *¿Qué haces tú por aquí?*

GT: *What are you doing here?*

Ours: *What you do you for here?*

Analysis: Our system failed to recognize the Verb-Subject structure in this sentence and so added "you" as a subject pronoun before the verb, while retaining the "you" after the verb. Our system also missed that "por aquí" is a particular usage of "por" that can be reduced to "around here." Since this is a very short and common sentence, Google translate likely picked out the idiomatic translation from an aligned corpus.

5. ***SPANISH:*** *Lo interesante de éste y otros programas similares es que van aprendiendo del usuario conforme pasa el tiempo con el fin de hacer más útiles sus sugerencias.*

   GT: *The interesting thing about this and other similar programs is that they learn the user as time passes in order to make useful suggestions.*

   Ours: *It interesting of this and others similar programs is that it go learning of the according user pass the time with the end to do more useful their suggestions.*

   Analysis: Google knows to parse "Lo interesante" as "The interesting thing," as well as to reduce "van aprendiendo" (directly "they go learning") to "they learn." It also picks up on "con el fin de" as meaning "in order to." Our translation does not realize that "other" should be singular, decides the pronoun for "van" is "it" instead of "they," has a lot of confusion surrounding "conforme pasa el tiempo," and mis-translates "hacer" as "do" rather than "make."

Conclusion:

Ultimately, though imperfect, our approach of directly translating the Spanish sentences, then applying a series of simple strategies based on general differences between English and Spanish to improve the structure of the translation, then finally applying a bigram model to check the accuracy of our changes, was relatively effective in improving the quality of the translation. The biggest failings of our translator were a lack of a high-level grammatical model, as well as a failure to seriously leverage a large corpus of data to effectively translate on the phrase level. In retrospect, some of our rules appear to have been too specifically tailored to our specific development set, and thus were brittle when applied to the test set. Instead of focusing purely on making rules that could be easily generalized, it seems that we should have focused more on making rules that were highly robust, with more thought to edge cases and more care in ensuring that our code reflected our intent.

Appendix -- Dev Sentence Translations

1. **SPANISH:** Debido al precio, es importante que estas habitaciones de primera, en términos de beneficios, ofrezcan algo más que botellas de champú genérico o un chocolate en la almohada.

   GT: Due to the price, it is important that these rooms first, in terms of benefits, offer more than bottles of generic shampoo or a chocolate on the pillow.

   Ours: Due to the price, is important that these lodgings of first, in terms of benefits, offer something more than bottles of generic shampoo or a chocolate in the pillow.

2. **SPANISH:** Limitar la ingesta de alimentos y bebidas con muchas calorías es clave para ayudar a mantener un peso saludable.

   GT: Limit your intake of foods and beverages high in calories is key to help maintain a healthy weight.

   Ours: To limit the intake of food and beverages with many calories is key to help to maintain a healthy weight.

3. **SPANISH:** ¿Usaremos en el futuro los coches de la misma manera que actualmente?

   GT: Will we use in future cars the same way today?

   Ours: Will we use in the future the cars of the same way that at present?

4. **SPANISH:** En los últimos días, sin embargo, el asunto que le preocupaba no había sido más que uno: la hija del comerciante que vivía en la ciudád adonde llegarían dentro de cuatro días.

   GT: In recent days , however , the issue that concerned him was no more than one: the daughter of the merchant who lived in the city where he would arrive in four days .

   Ours: In the last days, however, the issue that concerned him had not it been more than one: the daughter of the merchant that lived in the city where would arrive within of four days.

5. **SPANISH:** Gracias a su correspondencia, podemos descubrir que Lewis Carroll detestaba la celebridad de Alicia y no podía entender, ni soportar, los autógrafos que le pedían por la calle.

   GT: Through correspondence, we may discover that Lewis Carroll Alice hated celebrity and could not understand , nor support , asking him autographs on the street.

   Ours: Thanks to his correspondence, we can to discover that Lewis Carroll detested the celebrity of Alice and could not understand, nor support, the autographs that requested him for the street.

6. **SPANISH:** "Estas ciudades están dirigidas por personas que estuvieron trabajando arduamente para llevar velocidades de internet más rápidas y las últimas tecnologías a sus residentes", declaró la compañía en su publicación de blog.

   GT: "These cities are run by people who were working hard to bring faster Internet speeds and the latest technology to its residents," the company said in its blog post .

   Ours: "These cities are directed by people that were working hard to carry speeds of internet more fast and the last technologies to their residents", declared the company in his publication of blog.

7. **SPANISH:** De gran belleza y vistosos colores, las mariposas monarca son de vital importancia en el ciclo de la vida como agente polinizador y factor de equilibrio ecológico en los bosques que habitan.

GT: Of great beauty and bright colors , monarch butterflies are of vital importance in the life cycle as pollinator and ecological balance factor in the forests they inhabit.

Ours: Of great beauty and showy colors, the butterflies monarch are of vital importance in the cycle of the life like pollinator agent and factor of ecological balance in the forests that inhabit.

8. **SPANISH:** Aquel espíritu de iniciativa social desapareció en poco tiempo, arrastrado por el fiebre de los imanes, los cálculos astronómicos, los sueños de transmutación y las ansias de conocer las maravillas del mundo.

   GT: That spirit of social initiative disappeared in a short time , driven by the fever of the magnets , the astronomical calculations , the dreams of transmutation , and the urge to see the wonders of the world.

   Ours: That spirit of social initiative disappeared in little time, drawn by the fever of the magnets, the astronomical calculations, the dreams of transmutation and the desires to know the wonders of the world.

9. **SPANISH:** Esta información no está verificada, pero da idea de cuán caldeado está el ambiente en la capital de Ucrania.

   GT: This information is not verified, but gives an idea of how the atmosphere is heated in the Ukrainian capital .

   Ours: This information is not verified, but it gives idea of heated how is the environment in the capital of Ukraine.

10. **SPANISH:** Cuando se sentó comenzó a sentirse más tranquila, no estaba sola en la sala y le apetecía mucho ver la película.

    GT: When he sat down he began to feel more calm, was not alone in the room and I really wanted to see the movie.

    Ours: When it sat oneself began to feel more quiet, it was not alone in the room and fancied him much to see the movie.