JB Bakst
Zach Ellison
Alex Grover
Zach Saraf

CS 124: Machine Translator

For our Machine Translator, we chose to create a Spanish-English direct translation program. We chose Spanish primarily because we are all familiar with the language, and that was a big factor in our ability to judge our translations as well as our ability to create effective strategies to improve the quality of the generated sentences.

Compared to other languages, Spanish stood out to us as a relatively straightforward language to translate to English. It uses the same character set (this would be a huge challenge in translating from a language such as Japanese to English), and it uses similar grammatical constructs as English. However, translating Spanish to English is not without its own unique challenges. Because the languages can be similar in some aspects, we found that many of the grammatical conversions necessary to improve our translator were somewhat more nuanced than they would have been coming from a language where it is difficult to even convert the words to their appropriate English word.

The biggest challenges that we faced were the ordering of words, the inclusion/deletion of certain words, and general colloquialisms between Spanish and English. For example, Spanish places adjectives after nouns whereas English places adjectives before nouns, and we found a library for POS tagging that simplified the solution to this challenge. But in other cases, there was not as clear of a solution. Spanish includes many articles (such as the words 'el' and 'la') that would in many cases not appear in a correct English translation, but determining the context in which an article should be included or deleted is a major challenge that POS tagging could not sufficiently solve. We ultimately developed a solution that addresses this challenge, but we could not account for all cases as accurately as we would have liked. Lastly, there are many colloquial and idiomatic phrases that simply do not translate directly between the two languages. This was a challenge that we chose not to address, since it would require having an additional dictionary to handle all idiomatic phrases, and we did not feel that developing this dictionary would be a valuable use of our time.

We constructed our corpus from an article we found online, originally taken from a Spanish-language newspaper. It can be found in the Appendix to this document, along with the Google Translate translations of each sentence.

To aid us in the development of our translator, we first implemented a scoring mechanism that pairs our Spanish sentences with their English counterparts as translated by Google Translate. We put these translations in our corpus files and implemented BLEU scoring with a variable length n-gram. While we know that this is not the perfect way to score our sentences as there are various possibilities for a correct translation and Google Translate can make mistakes, it gave us a baseline score to improve on as well as allowing us to concretely see how much improvement each of our strategies made.

The strategies we created include 3 pre-processing strategies and 5 post-processing strategies that we developed iteratively in response to the output of our translator at each stage of development. We developed them as follows:

1. Stripping Spanish punctuation (¡ and ¿) from our parsed sentences. We handle other punctuation by simply passing it through our direct translator as we found that most punctuation in our dev set is shared between English and Spanish.

2. Adjective-noun swapping for incoming Spanish sentences. This preprocessing strategy was one of the most obvious fixes that made the translated sentences seem incorrect at first glance. We used the NLTK library for part-of-speech tagging and then swapped adjectives that were followed by nouns, except in the case that they were followed by a verb. This was the most substantial change that we found improved our scores a lot.

3. Verb negation. In Spanish, the verb is negated as 'no + VERB', but in English we would like this to read 'VERB + not'. For example, the Spanish sentence 'no puede' should be translated as 'can not' rather than 'no can'. We used the NLTK library POS tagging to distinguish when 'no' is followed by a verb.

4. Translations of 'que'. In Spanish, the idiomatic use of this word is highly dependent on the surrounding sentence. This word has many possible translations so rather than strictly using the dictionary, we check for phrases that we recognize and use the correct translation. We wanted to nail down each of these cases as 'que' is possibly the most common word in Spanish. One special case is when 'que' begins a sentence or is preceded by 'lo'. In this case, que has the standard translation to 'what'. If 'que' is preceded by 'persona', 'personas', 'gente' or any name, then 'que' is translated to 'who'. We wanted to generalize out these three words so that it worked on any synonyms, but our POS tagger was not accurate enough to identify this so we had to hard code in these words. Another case was if 'que' was preceded by any verb. In this case, the 'que' is translated to 'to'. In all other cases, 'que' is translated to 'that'.

5. Translating 'una' to 'a/an' depending on the context. Our dictionary has the literal translation but the correct word to use is dependent on the following word (e.g. 'una vez' means 'one time' but 'una chica' means 'a girl'). If we decide on using a/an, we do post processing to change 'a' to 'an' if the following word begins with a vowel. For example, 'a time' versus 'an eagle'.

6. Correcting capitalization. This boosted our scores as our scorer compares literal words. Thus, it was always incorrectly marking the first word in the result sentence as wrong because we always begin by converting everything to lowercase. It was therefore not cased the same way the Google Translate result was (as all Google Translate sentences begin with a capital).

7. Article Correction: Used a specialized model incorporating Laplace-Smoothed Unigrams and Bigrams to decide whether 'the' should be included in the output English sentence. In many cases, it is not included as an article. The rules as to when this happens are complicated so we decided to use a statistical model to test for the most likely sentence with every possible permutation of 'the's removed. We used the Holbrook Corpus from Assignment 2 to train the Unigram and Bigram Models. If the word following the 'the' in our direct translation appears more often with a 'the' in front of it, we include 'the' in our translation, otherwise we drop the 'the' and just include the word itself.

8. Direct Object / Verb Swapping: In Spanish, many verbs refer to a direct object using a pronoun that precedes the verb (such as 'me', 'te', 'lo', 'la', 'nos', 'os', 'los', and 'las'). During pre-processing, we swapped all direct object - verb pairs so that the object would follow the verb, and we and marked the direct object with a tag showing that it was swapped. Then

during translation, we picked up on the tag indicating that a word was a direct object that had been swapped and translated it as a special case, outside of the dictionary. A good example in the dev set that demonstrates the effect of this strategy is sentence 7, which with a simple direct translation would end with “that the made feel uncomfortable”, but with this improvement is translated to “that made them feel uncomfortable”.

Google Translate Analysis:

We could not find any cases where our system performs better than Google Translate. However, we did find that while some of our translations were very different from Google Translate’s, our translations were not necessarily wrong. Google Translate had more colloquial translations, while our translations were much closer to direct translation.

For sentence one, Google Translate (GT) uses the word ‘ability’ instead of ‘capacity’ which is the better translation here because this word applies better. GT also translates ‘block’ to ‘block out’ which is a better translation in this context. GT also adds a ‘the’ before world which we should have caught with our article correction. Overall, GT translates sentence one better than we did.

For sentence two, GT translates ‘programa de computadora’ to ‘computer program’ while we translate it to ‘program of computer’. GT also correctly translates the direct object ‘la vincula’ to ‘links it’, whereas we did not correctly translate this. This was because our part of speech tagger did not correctly identify ‘vincula’ as a verb, so we did not treat the word ‘la’ as a direct object and rather translated it directly as ‘the’. Overall, GT translates sentence two better than we did.

For sentence three, GT translates ‘ademas’ into ‘furthermore’ while we translate this to ‘further’. GT has the better translation here. Our translation also adds in an unnecessary ‘the’ while GT does not. GT also uses the gerund ‘people suffering’ while we translate this to ‘people who suffer’. These are both completely fine, with neither translation being better. Finally, GT translates ‘de’ to ‘from’ while we translate it to ‘of’ preceding psychosis. GT was correct in this as well. Overall, GT translates sentence three better than we did.

For sentence four, GT translates ‘sentirse’ as ‘feeling’, whereas we simply translate it as ‘feel’, making GT’s translation more correct. However, GT does not do the same thing with ‘experimentar’ (it simply leaves it as ‘experience’). GT also removes the ‘a’ before ‘deep relaxation’, and we keep the ‘a’. This did not provide a significantly better translation but is notable.

For sentence five, GT handles indirect objects and we do not. This leads to GT having a significantly better translation. We also do not handle the idiomatic phrase ‘de pie’ as standing.

Error Analysis on Test Set:

Most of our processing applies in at least one sentence in the test set, with the exception of the Spanish punctuation removal and adjective-noun swapping. Sentence capitalization applied to every sentence. Article correction applied in sentences one, two and three. While it incorrectly removed a ‘the’ in sentence one and incorrectly kept a ‘the’ in sentence three, it correctly kept a ‘the’ in sentence two. Our “que” correction was correct in every sentence it was applied to: one, two and three. Our consonant correction was correctly used on sentences two and four. Verb negation was used correctly in test sentence five; however, it also missed a case

in sentence five that should have been negated because there was an indirect object in between the “no” and the verb.

While our system improves upon the direct translation in 4 of the 5 test sentences according to a BLEU score comparison with the Google Translate translations, we do see significant room for improvement. Our biggest area for improvement is in synonym translation. We did not implement a dictionary with multiple options, which we choose from and ended up with mistranslations. Examples include ‘capicidad’ and ‘de’ in the first sentence (as well as ‘de’ in the third) and ‘si’ and ‘accordar’ in sentence five. We could improve our system design by including multiple definitions for a word and selecting them based on part of speech and fluency as determined by scoring on a Stupid-Backoff Model.

As mentioned in the previous paragraph, our article correction has room for improvement with ‘the’s begin incorrectly removed in sentence one and incorrectly kept in sentence three. This error could probably be improved with a better training model for the correction (i.e. more robust than the Holbrook Corpus). We also are not currently taking into account the inherent difference in meaning that the word ‘the’ brings in terms of specifying an instance of something instead of talking about it more generally. We could also improve the system by modelling a sentence to determine if the ‘the’ is referring to a specific instance of an object.

We also have room for improvement in colloquial translations of phrases such as ‘program de computadora’ in sentence two. While ‘program of computer’ is technically correct, ‘computer program’ is obviously a better colloquial translation. This is difficult to improve in a generalized manner, but we could include a dictionary of phrases for common colloquial phrases and prioritize phrase translations over direct word-by-word translations to improve our performance.

Our verb negation also has room for improvement; in sentence five, there is a verb negation where the negative ‘no’ is separated from the verb by a direct object, which is something we don’t consider. Additionally, although it doesn’t appear in the test set, there is room for improvement in the negation because different verbs (by tense and person) are negated differently and our negation is certainly not comprehensive to all of those situations. We could improve our system by separating negation situations where we simply want to swap ‘no VERB’ in Spanish for ‘VERB not’ in English (e.g. ‘should not’), which is the case we included in our system, with negation of verbs that should be translated as ‘do/does not’ or ‘did not’ (for past).

Appendix

Corpus: Development set

9. ¿Puede una luz tener el efecto de una droga?
10. La meditación es una habilidad que no está al alcance de todos.
11. El video contiene imágenes de una luz intermitente extremadamente rápida y no debe ser visto por personas que tienen epilepsia.
12. La luz, dirigida a la cara de la persona, estimula la glándula pineal del cerebro.
13. Un poco como drogarse, pero sin el bajón y la mandíbula dolorida.
14. Obviamente la gente puede ser escéptica porque es algo nuevo.
15. En una demostración con algunos empleados de la BBC, algunos dijeron sentir picazón en los ojos y una sensación de sacudida que los hizo sentirse incómodos.
16. Sin embargo, le dijo a la BBC que el 99% de las personas que utilizan la luz quedan cautivadas por lo que experimentaron porque es muy diferente de la realidad y lo disfrutan.
17. Un zumbido de colores y formas.
18. Actúa como una especie de vía rápida para las personas que quieren llegar a ese tipo de estado.

Corpus: Test set

1. Esa capacidad de bloquear el mundo que te rodea y relajarse.
2. La luz está emparejada con un programa de computadora que genera la secuencia de luces y la vincula con la música.
3. Además, los inventores advierten contra el uso por personas que sufren de psicosis.
4. Otros, por su parte, describieron sentirse fuera de sus cuerpos y experimentar una profunda relajación.
5. Me sentí ingrávida, no me podía acordar si estaba sentada o de pie porque no podía sentir mi cuerpo.

Corpus: Development set translations (from Google Translate)

1. Can a light have the effect of a drug?
2. Meditation is a skill that is not within everyone's reach.
3. The video contains images of an extremely fast flashing light and should not be seen by people who have epilepsy.
4. The light, directed at the face of the person, stimulates the pineal gland of the brain.
5. A little bit like getting high, but without the crash and the sore jaw.
6. Obviously people can be skeptical because it is something new.
7. In a demonstration with some BBC employees, some said they felt itchy eyes and a sense of shock that made them feel uncomfortable.
8. However, he told the BBC that 99% of the people that use the light are captivated by what they experienced because it is very different from reality and they enjoy it.
9. A buzz of colors and shapes.
10. It acts as a kind of fast track for people who want to get to that type of state.

Corpus: Test set translations (from Google Translate)

1. That ability to block out the world around you and relax.
2. The light is paired with a computer program that generates the sequence of lights and links it to music.
3. Furthermore, the inventors warn against use by people suffering from psychosis.
4. Others, for their part, described feeling out of their bodies and experience deep relaxation.

5. I felt weightless, I could not decide whether I was sitting or standing because I could not feel my body.