Angad Gogia
Eric Hermann
Connor Scherer

Translation Write-up

Introduction

For our translation, we chose to use the direct MT system to translate from Spanish to English. The two languages clearly have a number of features in common, but we had to tackle some key differences to have an effective system.

- One classic example of this is the ordering of nouns and adjectives. In English, we clearly have adjectives preceding the nouns, whereas in Spanish this order is reversed. For instance, one of our development sentences contains the phrase "asuntos personales" – without accounting for the ordering of adjectives/nouns, the translation method would translate this as "issues personal" rather than the correct translation "personal issues."
- Along these same lines, another difference between the two languages is the ordering of the direct objects and verbs. In English the direct object follows the verb, while again this order is reversed in Spanish. A concrete example of this in our corpus is "lo hiciera," which appropriately translates to "did it," rather than the direct translation "it did."
- One major difference between the two languages, and one that we were not able to address as completely as we would have wanted (discussed more in the error analysis section), was structure of infinitive verbs vs. gerunds. English gerunds, for instance, all end in "-ing," while words in Spanish that end in "ar" may either be the infinitive form or a gerund, depending on what comes before. Specific examples in the corpus show this difference: "sin dar cuentas" translates to "without be*ing* accountable" (gerund form of be), while "podían gastar" translates to "they could spend" (infinitive form of spend).
- Another difference between the two languages is how they handle numbers and punctuation within numbers. Specifically, English uses commas to delineate large numbers (e.g. "130,900 euros") while Spanish will use periods in this scenario (e.g. "130.900 euros"). They also have contrasting conventions for the use of decimal points (where obviously English uses periods and Spanish uses commas). While this distinction doesn't add much in terms of semantic understanding of a translation, we found it to be an interesting distinction to analyze for our system. One important aspect to note: this difference in languages only applies for certain Spanish languages and not for all varieties – in Mexico and Central America, for instance, they use the same conventions as English, while Spain and South America uses the switched punctuation. Because our test set was taken from the same article (and therefore the same language) as our development set, we thought it appropriate to address this difference.
- Another distinction in the languages is the use of "a" following a verb. Spanish will often times follow a verb with "a" before the following noun. In

our corpus, for instance, there is a phrase “se convence a consejeros,” which translates to “convince directors.” Clearly, English omits the “a” between the verb and the noun, which was difficult to account for because there is not a specific pattern as to when the “a” would be present or not.

- One of the more difficult differences between the languages is the use of the word “se” before verbs in Spanish. The “se” is difficult to account for because it can have different meanings, depending on if it’s used an indirect object pronoun, as part of a reflexive verb, as a reciprocal pronoun… Again, the same example as just mentioned applies here: “se convence a consejeros.” Clearly English does not use the “se” in its translation at all, so in the translation system you have to determine whether to drop the word altogether or interpret it as a pronoun that needs to be translated.
- One challenge in tackling these two languages are certain idioms that don't translate word for word to English. Some of these phrases come up frequently in translating the languages (such as “como se”), and we looked at multiple different alternatives to solving this issue. This is probably not unique to Spanish, as probably most foreign languages have common idioms, but was definitely a challenge to address in our MT system.

Corpus

Our corpus was taken from a Spanish website called “Noticias Españas,” and the article was titled “Un regalo de plástico envenenado” (http://noticiaes.com/espana-noticias/un-regalo-de-plastico-envenenado/). The first 10 sentences served as our development set and the next five sentences as our test set. The individual word translations for our dictionary were taken from Google Translate, where all translations were included per word.

Translation System

The iterative order of our translation strategies are presented below:

- *Pre-processing punctuation.* This was a pretty simple fix, but an important one. We wanted to make sure to take out all the beginning punctuation (¿ and ¡ symbols), as these are clearly not an element of the English translation. This was a helpful fix in our development set, as some of our first few sentences in the corpus were questions. We also had to process the sentences to handle quotation marks, commas, and parentheses appropriately. While the leading ¿ and ¡ were not a feature of our test sets, the punctuation mark pre-processing was used in sentence 3 of the test set (“en cualquier cosa”).
- *Post-processing*. After all the translations are taken care of, we handled some of the post-processing work by ensuring all the sentences began with a capital letter (as words could get switched around during the translation process) and that all the words were split up appropriately. The post-processing work had an impact on every sentence in the test set.
- *Unigram language modeling*. Because our dictionary has multiple English translations for every Spanish word, we had to have a method to choose which English word would be the best translation for any given Spanish word. We

started by using a unigram language modeling system. We downloaded a unigram English dictionary, and our first process just chose the most popular English word as the appropriate word to use for the translation. Just by the very basis of a direct translation system, this method affects every sentence in the test set as choosing appropriate English words is the central feature of a direct translation system.

- *Bigram language modeling*. We chose to expand on our language modeling system by including a bigram model. Again, English bigrams that were more common were given a higher score. We chose to weight the bigram model by 9 times the unigram model, as we saw in our development set that this led to the best results. Again, based on the features of direct translation systems, this language model affected all sentences in the test set.
- *Switching nouns and adjectives*. This again addresses one of the issues that was mentioned above, in that the order of Spanish nouns and adjectives are the opposite order as in English (e.g. "bruja verde" vs. "green witch"). This was a very effective strategy in our development set, and even more so in the test set. You see that this strategy comes into play in at least three test sentences: 1 ("actividad sindical" changes to "union movement," where union/sindical is the adjective), 4 ("técnica sutil" changes to "nice technique"), and 5 ("decisiones estratégicas" changes to "strategic decisions").
- *Switching verbs and direct objects*. In our development set, we saw that it was helpful to address the concern about the different orderings of direct objects and verbs between Spanish and English. We saw this, for instance, in our development set when we saw that "lo hiciera" should translate to "do it," rather than "it do," as it would before the change. We think this is an appropriate strategy to apply in terms of the Spanish language in general, as this is a fairly common pattern. That being said, we did not see examples of this pattern in the development set, so it ultimately did not affect our final translation.
- *Fixing numbers*. This strategy points to the difference in how Spanish and English handles the punctuation within numbers, as mentioned in the first section of the write-up, and involved switching commas with periods and vice versa. This was a strategy we thought important to include based on our development set, as large numbers were included frequently in those sentences. It ultimately affected just one sentence in the test set, as there was only one sentence that included numbers ("175,200 euros" vs. "175.200" euros), but helped clarify the meaning of the sentence.
- *Handling common phrases.* This was a strategy we found to be very helpful in our development set, as there were certain phrases we found to be common in our corpus that don't translate well using the normal direct MT system. While this may be viewed as overfitting the data set, we thought it would be helpful in a general Spanish translator, as these phrases are all common Spanish phrases and we would expect an effective translator to have certain phrases hard-coded into its system. The phrases we chose to analyze were "como se" ("how does one"), "no obstante" ("however"), and "de que" ("that"). These

came up multiple times within our small corpus, so would expect them to similarly be frequent in the general language. It just so happens that our test set did not include these phrases, and so it did not affect any of those sentences.

- *Reflexive verbs*. In our development set, we saw many instances of reflexive verbs, where "se" preceded a verb. In these instances of the development set, it was appropriate in our translations to translate the "se" preceding a verb to be "oneself," as this is the general meaning of the reflexive "se" (e.g. "se moleste" to mean "disturb oneself"). There was one instance in the test set where the reflexive verb modification added benefit: in sentence 2, the phrase "se debe" is appropriately translated to "should oneself." Other than this, there were no other reflexive verbs in our test set.
- *Removing verb-following 'a.'* This was again a specific challenge to address unique to Spanish, in that there is often an 'a' that follows a verb before a noun. We saw this in our development set multiple times, and knew it to be a common Spanish pattern, so we made the simple fix to remove any 'a' following a verb, as English omits this 'a.' In the test set, this strategy affects sentence 2, which contains the phrase "debe a," where it is appropriate to remove the following 'a.' Again, we expect that in a larger test set, this strategy would affect a larger number of sentences as this is a common pattern.

Google Translate:

1. Google's translation is a more effective one in this scenario. While with our translation you can understand the sentiment of the sentence, it takes some rearranging of words and general intuition. For instance, one can infer that "The explained four" in our translation really should be "The four explained" or that "your expenses" really should be "their expenses," but these are still errors worth considering. Google's translation handles the sentence very appropriately, matching up identically with what the correct translation would be. Furthermore, "their trade union and professional activity" is clearly a better translation than "your movement and pro," again showing that the pronouns are a little bit off.
2. Google's translation is still more effective, but not as much so as in sentence 1. Our translation correctly translates the beginning part of the sentence ("Neither" instead of "Nor"), but it is easier to understand the general sentiment of Google's translation than ours, as it handles the second clause very well, while our translation has some inaccurate verb tenses and omitted pronouns ("was member" vs. "he was a member").
3. Both translations are pretty equivalent here, though Google's is slightly more so. The sentence is clearly a short one, so there is less room for error, but our translation again omits a pronoun that is necessary to the meaning of the sentence and the chosen word for "declaró" is better handled in the Google translation ("he said" is more accurate than "registered" as a translation).

4. In this sentence, we found that our translation does well in similar places and makes similar mistakes as well. For instance, both translations pretty accurately handle the first part of the first clause ("though subtle and deceitful technique" is a better translation than "nice technique and trickster") but then gets caught up with a difficult to translate phrase towards the end of the first clause, as neither translation handles it well. Google's translation has a better flow in the second clause, and our sentence suffers from a confusing double negative ("without no").
5. Both translations have a difficult time with this sentence. Google's translation does a good job in identifying the subject of the sentence and the positions he has held, while our translation makes the mistake of translating his last name, Rato, as "time," making the translation harder to understand. Both sentences have a difficult time with "vivieron unos plácidos mandatos," as neither translation handles this phrase appropriately. Unlike Google's translation, ours suffers from incorrect pronouns ("your" instead of "their") once again, which bungles the meaning of the sentence.

Error Analysis:

One major challenge in the direct MT system was implementing verbs tenses and conjugations. Our system's design took in all of the different verbs in the tenses they were used, and directly translated those tenses into our translation dictionary (for instance, where we saw "explicaron," the past tense of "explain," we would input "explained" directly into our translation dictionary). In general, this method would produce pretty accurate results in both our development and test set, but a more sophisticated system could cut down on errors. One particular method to address this could be to stem all of the inputted verbs, store the infinitive translation in the dictionary, and then have specific rules as to which verb conjugation is most appropriate given the context of the sentence and the original word. This would ultimately be a much more involved system, as the correct verb tense depends not only on the word itself, but on which words precede it as well (mentioned when discussing the "-ar" verbs in the first section of the report). This update would solve multiple errors in our development set (as there were frequent instances where the verb tense or noun-verb agreement were incorrect). This ultimately only really affected one sentence in our development set, but it led to a very noticeable error, and would certainly affect more sentences in a larger test set. The first sentence in our test set starts with "Los cuatro explicaron," which should translate to "The four explained." Our system outputs "explained" appropriately for "explicaron," but then interprets it as an adjective and switches the noun and verb to get "The explained four." While the verb tense is in fact correct here, the more sophisticated translation system could avoid this error, as if you stem this word into its infinitive form, the POS tagger would not tag this as an adjective form.

A common error we saw in many of our test translations was simply the handling of the word "sus." While this may seem like a trivial error, as it is just one word, it led to many errors in our test set (in sentences 1, 4, and 5), as our test translations often have the word "your" instead of "their" in this place. The reason this happens is that “sus” can translate to either word depending on the context, and our unigram and bigram language models give a higher weighting to the English translation “your” instead of “their.” There are a couple ways to handle this error, and perhaps the most sophisticated system would choose the appropriate word based on the total context of the sentence, not just the word preceding or following the word. For instance, in the first sentence, “Los cuatro explicaron que sus gastos” (“The four explained that *their* expenses”), the “sus” clearly depends on the initial object “The four” rather than the words before or after. If we could condition the correct word perhaps on the subject of the sentence, this problem could have been avoided and more accurate translations been outputted.