Language

We decided to work on translating Spanish phrases to English ones using a Statistical Machine Translation system. Although they both theoretically arise from the same tongue, Proto-Indo European, the immediately diverge from there: Spanish is an Italic, Latino-Faliscan language, more directly from Latin; English is a Germanic language and thus more closely related to that family of languages. This does mean that there are unsurmountable differences between them, however—there are many cognates, and the general grammatical structure is more similar than dissimilar between the languages.

Some of the most tangible differences which we noticed between these languages include:

- Spanish sentences almost exclusively have adjectives after nouns, whereas in English adjectives generally are put before nouns. This is a problem which can likely by solved by correctly handling alignments and creating a high-probability valid English sentence from ngrams, but it makes it more likely that an English sentence will have an inappropriate ordering.
- In Spanish, nouns can be masculine or feminine, and a noun's articles and adjectives must match the noun in gender. This means that there are probably multiple Spanish versions of any one given English article or adjective. While this does not present a huge problem when translating Spanish to English (with any luck, both genders of a Spanish adjective will map to the same English equivalent) it adds complexity and would certainly complicate any translation from English back to Spanish.
- Negating sentences in Romance languages, including Spanish, is more complicated than it is in English. Although our parser is able to interpret the Spanish "no" as the English "no", the negation often goes deeper by changing word order, changing verb use, and sometimes even changing the conjugation of a verb, as in the Spanish imperative tense. While the EM algorithm is likely able to gloss over a few of these differences, it by no means can fix them all. The addition of phrase-based translation would likely help because then a Spanish negative phrase could be translated to an English one, but assuming a random sample of sentences, examples of this kind of translation are likely to be scarce, so it may be harder to appropriately train any kind of algorithm to recognize type of phrase.
- In Spanish, word order and other subtle differences are often used just to create emphasis, and there is no excellent equivalent to this in English. To a certain extent, this presents a practical problem because it means that a Spanish sentence will be less likely to translate to a grammatically sound English sentence. But beyond that minor problem with fluency, this paradigm may also create a deeper issue will faithfulness: if a certain word is meant to be emphasized or carry extra meaning in a Spanish sentence, our model does not appropriately emphasize it based on its position (that is, if we see a Spanish sentence in this style in our test set, the emphasized word will be translated just as it would be if it were earlier in the sentence).
- There are some subtle differences in the alphabets and style, such as the addition of slightly modified characters (e.g. ñ, é, etc.) and use of different punctuation (e.g. ¡, the use of , instead of . as a decimal point in numbers, etc.). This doesn't present a particular problem to our system, but it does preclude any naïve interpretations of punctuation and the like.

BLEU Scores

Our final BLEU scores on the test set were as follows:

BLEU_1: 27

BLEU_2: 1.9

Improvement Strategy

We noticed that our bag of words for any translation consisted mostly of common words such as “as” and “the.” Even for a word that does not translate to a common word, such as “casa,” the translation was often such a common word. This was because P(E) was significantly higher for words such as “the” than words such as “house,” so when we calculated our bag of words using the formula t(f|e) * p(E), common words often got a higher result.

To solve this, we hardcoded the values for common words such as “the,” “and” and “as.” We did this by googling “most common English words,” getting a list of the 100 most common English words, and then manually translating each of those words into Spanish using Google translate. This gave us a table of common English words and their Spanish translations.

We then hardcoded a HashMap of common Spanish words to their English translations, so that when we come across a Spanish word, we can simply check if it is one of these common words and know its translation directly from there. Furthermore, we hardcoded a HashSet of common English words. While calculating argmax(p(f|e) * p(e)) for a particular Spanish word, we ignored all such common English words. This made sense since we had already checked if the Spanish word was part of our “common words” list, and thus if its translation was part of the “common English words,” so we don’t have to consider these common English words in our translation for other Spanish words.

Furthermore, we tried a few ways of incorporating language models into our system. Our first attempt was with trigrams; though we couldn’t find a trigram corpus on the internet, we tried to use our training English sentences to make a trigram map, and then use those trigrams to correctly align our bag of words. Although this was a promising strategy, it did not work as well as we hoped since our training data was not big enough. What worked better was a bigram corpus we got off the internet (a link to it had been posted on Piazza); we used this corpus to predict the alignment of our bag of words and that significantly improved our BLEU scores. However, we eventually decided on the following strategy:

For each word in our bag of words, we check the bigram score with the next 2 words in our bag of words; the word with the higher bigram score took the next position. This means that in our final code, we didn’t shuffle our bag of words too much; each word is followed by one of the following 2 words from the original bag.

Google Translate

Index	Spanish sentence	Google's translation	Our translation	Comment
808	Si hablamos de su trabajo con los sensores de Xbox Kinect , ¿ qué quejas tiene usted de las cámaras modernas ?	If we talk about his work with Xbox Kinect sensors, what complaints you have modern cameras?	if standard of their work with policy sensors of files kinect wait why complaints has you of are cameras modern xbox	Google clearly has the better translation here, and one of the main reasons is its recognition that this ends with a question, so the translation is phrased accordingly with a question mark at the end. Google's translation understands that Kinect and Xbox are supposed to go together, whereas our translation has not understood that.
31	Parece ser que dos de cada tres electores hispanos apoyan al partido demócrata .	It seems that two out of three Hispanic voters support the Democratic Party .	be which two of each topic voters hispanos support style party democrat three	Google translate has the better translation again; it understands the concept of adjectives such as Democratic, so it's able to understand that "democratic party" should be a term, as opposed to party democrat, which is what our translation has. It also seems to have a better language model that has been able to put together a well-formed English sentence, unlike our MT system.
2012	" Se come crudo y luego se bebe un vaso de aquavit de un trago ", dice Havard Halvarsen , bombero de la ciudad que es también lo que se conoce como el " General del Rakfisk ", que se encarga de organizar el festival .	"It is eaten raw and then a glass of aquavit drink in one gulp," says Havard Halvarsen, City firefighter who is also what is known as the "General rakfisk" which organizes the festival.	are eating crude and then are drunk a glass of domain of a bitter domain says havard halvarsen am of whisky city the is also submission which are gifts like he domain general index rakfisk domain which are body of which he festival everything	Google translate has done a better job again. One good reason is that it understands quotes and how they affect the layout of a sentence, which our MT system doesn't do yet. It also is able to identify that General rakfisk is the organizer of the festival, which our MT system didn't do. I think in general, this is because our MT system isn't yet trained to deal with proper nouns, which complicate the layout of a sentence.
2045	" ¿ Cuáles son vuestras puntuaciones para el mejor	"What are your scores for the best fish up there	domain library are release points for he best fish are up at are stories thor-	Again, our MT system hasn't dealt with quotes properly. Furthermore, our language model hasn't put together the bag of words in a way

	pescado allí arriba en las montañas , Thor-Juergen ? "	in the mountains, Thor-Juergen?"	juergen domain wait	that it should, and as in above, our MT system hasn't dealt with proper nouns properly by capitalizing them.
903	Ahora , cada héroe de culto tiene sus 100 fans .	Now every cult hero has his 100 fans.	each hero of worship has now your fan 100	These translations are quite similar in their meaning; cult and worship have similar meanings, and other important words such as fan and 100 were picked up. However, Google translate did a better job of putting the sentence together, as it understood that 100 was associated with fan, which meant that fans should be plural.

Error Analysis

One big error we ran into was that words with similar meanings tended to have competing t values, so one would often show up instead of the other. For instance, in the following example, the word "release" showed up instead of the word "freeing" since they have similar meanings:

Goal translation: Parliament does not support amendment freeing tymoshenko

Our translation: he no supported an amendment for release to parliament tymoshenko

A good way to avoid this error would have been to do more iterations so eventually t values would have converged to the correct ones.

Another error we ran into was that we chose a 1:1 ratio of English words to Spanish words; we chose J = i. In practice, of course, a good MT system should play around with different word counts to ensure that it get the best translation, since some long Spanish sentences can be translated into much shorter English sentences, and vice versa. This caused many errors for our translation, for instance:

Spanish sentence: Una buena idea , siempre que , esta vez , sea realista .

English translation: A good idea , provided that it will , prove realistic this time .

English translation on our MT system: an idea always good requests time which might realistic

Here, we see that the gold translation is 2 words longer than the Spanish sentence, so our MT system could not have achieved that since it assumed that the translation should be just as long as the Spanish sentence.

A good way to avoid this problem would be to allow for J +- 5 I, so J could be anywhere within 5 words of I, and check over all such Js.

Another error our MT system is running into is it's lack of understanding of punctuation. Right now, our system does not account for proper nouns and capitalizing them. Furthermore, it gets rid of all punctuation such as quotes, commas and full stops. This means that outputs from our MT system are often not understood since they haven't been punctuated properly. We also haven't accounted for possessive nouns and plurals.