# 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.

# Improvement Strategy

We used trigrams to improve the baseline IBM model 1 system…

# Google Translate

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Spanish sentence | Google’s translation | Our translation |
| 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? |  |
| 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 . |  |
| 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. |  |
| 2045 | " ¿ Cuáles son vuestras puntuaciones para el mejor pescado allí arriba en las montañas , Thor-Juergen ? " | "What are your scores for the best fish up there in the mountains, Thor-Juergen?" |  |
| 903 | Ahora , cada héroe de culto tiene sus 100 fans . | Now every cult hero has his 100 fans. |  |

Analysis…

# Error Analysis

Our error analysis on the test set, including specific reference to what the code does and ideas for how further work might fix the remaining errors…