CS124 PA6 - Machine Translation (1906 words)

# Machine Translation: Spanish - English

## Introduction

In this paper, we explore the key concepts and challenges that accompany the task of translating passages of Spanish text to English. Below we describe several of the key differences between the two languages, many of which proved useful in our development of a machine translation tool.

Note: If you want to examine our translation algorithm, please use `translate2.py` in the `code/` directory.

### Descriptor Placement

Spanish adjectives are more commonly written after the noun, whereas in English it is standard to place adjectives first. For example, "la casa roja" should be translated as "the red house".

### Gendered/Pluralized Adjectives

English adjectives are non-gendered. In Spanish, the adjective is made to agree with the subject it describes. For example, "The banana is red and the apples are red." is translated as "El platano es rojo y las manzanas son rojas."

### Verb Tense

Spanish has a number of verb tenses that do not appear in English, and that are instead achieved through the use of one of the three major tenses (past, present, future). For example, The imperfect tense: "Estaba en Chile por dos años." - "I was in [stayed] Chile for two years."
The preterit tense: "Estuve en Chile por una semana" - "I was in Chile [visited] for a week."

### Possessives

In English, a possessive can be formed by the addition of "'s". This contraction is replaced in Spanish by the passive voice. For example, "Juan's house" - "La casa de Juan"

### The Formal "You"

Spanish has a "vosotros" form that can either mean a pluralized "you" ("you guys"), or a term of respect when addressing an elder or superior. ("¿Vosotros estaís buenos, señores?")

### Adjective-Noun Placement

In Spanish, the positioning of an adjective relative to the noun it describes is important. For example, "Tengo un viejo amigo" (I have a longtime friend) is different from "Tengo un amigo viejo" (I have an elderly friend).

## Working Corpus

Please see the appendix of this write-up for the following information:

1) Development corpus and sources (Spanish) [**DC**]

2) Test corpus and sources (Spanish) [**TC**]
3) Baseline test corpus translations (English) [**BT**]
4) Final test corpus translations (English) [**FT**]
5) Google test corpus translations (English) [**GT**]

*(Sentences from each set are referred to by their two-letter code in the list below [XX] followed by the sentence number. For example, sentence 3 in the Spanish test corpus is **TC3**. This information is also available in the zipped folder.)*

## Translation System

We ran the machine translation system using the following heuristics. The Spanish sentence was run through each of these steps and the output was reassembled at the other end.

### Laplace Unigram Language Model

This was a high-level strategy that dramatically affected the clarity and faithfulness of our results. Having run the input sentence through several strategies of a more limited scope (see above), we then ran the remaining untranslated words through a Laplace-smoothed Language Model [LM] in order to choose the most accurate translation. In almost every sentence in both the dev and test sets, our baseline translations were dramatically improved. This is because many Spanish words in our dictionary had multiple possible definitions, and there were consistently words in the final translation that were not able to be fully interpreted using the previous methods. The LM was able to choose the best translation from among all possible options generated from these strategies, and thereby significantly improved our results.
Affected Sentences: **TC1,2,3,4,5**

### Part-Of-Speech Tagging

We used the NLTK python library to generate a POS tagger based on the CESS-ESP corpus of Spanish words. This allowed us to identify a large number of the words as either verbs, nouns, adjectives, or pronouns. Based on these results, we were able to generate different sets of rules through which we translated each word. However, the results generated from the NLTK library were somewhat sparse, so we then used an edit-distance algorithm to determine the closest dictionary match for all untagged words and tagged based on the English definition. This was useful in every sentence in the test set.
Affected Sentences: **TC1,2,3,4,5**

### Stemming

We used the NLTK library to generate a stemmer into which we were able to input a Spanish word and receive a stemmed, simplified version of that word. This was used for POS tagging, as well as for choosing initial word translations from our dictionary, but was only useful in sentences with a large number of regular, conjugated verbs and nouns.
Affected Sentences: **TC2,3,4**

### Regular Verb Conjugator

There is a set of common rules to verb conjugation in the Spanish language, so we decided to implement a hard-coded translator for the set of "regular" verbs ending in '-ar', '-er', or 'ir'. We did this for the three most common tenses (present, past, and imperfect). For example, in **TC1**, we correctly translate "el buen humor existe en la Rusia" to "the good spirit exists in the Russia".
Affected Sentences: **TC1,2,3**

### Check Stem-Changing or Irregular Words

This strategy applied in only two of our dev/test set examples, but is common in the Spanish language. Certain verbs are "stem-changing", where the spelling of the word changes depending on the conjugation. Our stemmer would misinterpret these words and therefore the words would not show up in our dictionary. To solve this, we isolated all verbs that had "ue" or "ie" in the middle and substituted the regular letter spelling ("o" and "e", respectively). These words could then be treated as regular verbs.
Affected Sentences:
**TC1**:"tener" is conjugated as "tiene".
**DC3**: "mostrar" is conjugated as "muestren".

### Fix Pluralization

After classifying a word as a noun and looking up the stemmed word in our dictionary, we were able to reexamine the original Spanish word and append an "s" to the translation if the noun was pluralized (ended in "s" or "es"). This yielded false positives on some misclassified words (see "motorizeds" in **TC3**), but was otherwise largely effective.
Affected Sentences: **TC1,2,3,4**

### Noun-Adjective Sentence Realignment

One of the major differences between Spanish and English is the ordering of nouns and adjectives (see"Descriptor Placement" in the Introduction). To solve this issue we relied on the POS tagger. For each Spanish sentence, whenever we found a noun followed by an adjective, we swapped the translated words in our result sentence. For example, in **TC4**, this correctly translated "impacto legal" as "legal hit".
Affected Sentences: **TC1,3,4**

## Comparison with Google Translate

See the Appendix (section GT) for the Google-translated sentences.

1. The beginning and end of our translated sentence match up well with Google's output. However, in our translation we were not able to determine the superlative describing Sochi as "the most expensive games in history". Although the word "cara" in Spanish can mean look, in this case it is the feminine word for expensive. We missed this conversion, and therefore missed that meaning. Therefore in this case Google provided the better translation.

2. We immediately have a small error translating the phrasal expression “al menos” to “at least”. Our dictionary only contained single spanish words, and so we first translated “al”, then “menos”, without treating the expression as a single phrase. We also are unable to explain that the policemen were killed. This is because morir has an irregular stem-change in the 3rd person preterite form. The rest of the sentence matches up nicely with Google Translate’s output. In this case, Google was once again the better translation.
3. This sentence did not translate well in either model. The sentence’s true meaning is somewhat visible in both cases, but several complex noun-adjective constructions and subject-verb-object orderings caused a variety of issues. One key difference is in our translation of "kidnapping of the officers working in…", and Google’s "kidnapping of official workers in...". In this case, we have two separate meanings and while our translation is perhaps more fluent, Google’s is more faithful. Furthermore, we missed the noun-adjective realignment when translating “presidential former candidate”, which Google did not.
4. Our main problem in this sentence was with verb conjugation. We have "to encrypt" instead of "encrypted", and that is something we would have liked to catch. However, the choice of meaning in the verb itself seems flawed, and Google also fell prey to this error. However, two minor mistakes that our system made - failing to correctly pluralize “taxes”, and writing "millions of Euro" instead of "million Euros", were both caught by Google.
5. This is our best sentence, probably due to the fact it is short and most words are relatively unambiguous. However, we fail to remove “disappearing words” (see below) which negatively affects the sentence fluency. Specifically, the word “of” appears twice in instances where simply deleting it would aid fluency. However, the faithfulness of our translation is preserved, since the meaning remains the same. We also add an "s" to the end of add, since the word is incorrectly tagged as a verb and conjugated.

## Error Analysis

There are several pieces of our translator that could be improved in future iterations.

### Design Flaw #1: Idiomatic Misses

Our translator missed several common idiomatic expressions in translating both our development and test sets. For example, in **TS2**, “Al menos…” was translated as “To the less …”. This translation, while literally accurate, is relatively far from the correct translation of “At least”. Furthermore in **TS4**, “El FC Barcelona” is translated as “The FC Barcelona”. The English translation of a named entity such as a team almost always omits the leading “The”, so this phrase should be translated simply as “FC Barcelona”. To correct these misses, it would be relatively straightforward to add a list of common idioms in the Spanish language and translate them to English by performing a simple lookup. An alternative method would be to train a parallelized language model on a large already-translated corpus of text. Then we train a simple classifier of bi/trigrams (since most of these expressions are only 2-3 words long) that would identify idiomatic expressions.

### Design Flaw #2: Disappearing Words

In Spanish, there are several cases where words in an expression have no corresponding English translation. This most frequently occurs with short, conjunctive words such as “de” or “para”. As a result, many of our translations have extra words in them, including “of”, “to”, and “for”. In **TC1**, “para confirmar” is translated as “of to confirm”, instead of simply “to confirm”. Similarly, in **TC5**, “un poco de mantequilla” is parsed as “a little of butter” rather than “a little butter” or even “a little bit of butter”. A possible solution would entail including the empty string as a possible definition of such “disappearing words” in our dictionary and then using a language model to parse candidate expressions or sentences in order to choose the most likely option.

### Design Flaw #3: Nouns-As-Adjectives

In our translation software, it was common for the POS tagger to classify adjectives as nouns if the adjective was formed from the stem of a noun (for example “manly”). This was particularly obvious in **TC2**, where “policías iraquíes” was incorrectly translated as “polices Iraqis” instead of “Iraqi police” and in **TC4**, where “al caso Neymar” was not translated as “the Neymar case” [the case relating to Neymar], but “the case Neymar". This could have been resolved by creating an improved POS tagger that used a large set of sources that draws from material that is similar to our corpus.

# Appendix - Corpus and Results

## Development corpus and sources [DC]

1. El movimiento del voto en Blanco encabezado por Gustavo Bolívar sigue recorriendo todo el país invitando a que las personas dejen de apoyar a los candidatos que están en el tarjetón para las próximas elecciones para Congreso y que en lugar de ello muestren su insatisfacción con la clase política tradicional votando por 'Nadie'.
   ***El País****: elpais.com.co/elpais/elecciones/noticias/voto-blanco-avanza-su-campana-por-todo-pais*
2. Después de 13 años prófugo, de evadir por un pelo a los militares, a la policía y a sus rivales, Joaquín "El Chapo" Guzmán está de regreso tras las rejas en México.
   ***Yahoo****: news.yahoo.com/quot-el-chapo-quot-guzm-n-nos-ser-202928187.html*
3. Y ahora comienza lo que probablemente sea un largo y complicado proceso jurídico para decidir cuál país lo enjuicia primero.
   ***Yahoo****: news.yahoo.com/quot-el-chapo-quot-guzm-n-nos-ser-202928187.html*
4. En México, es probable que Guzmán enfrente una gama de acusaciones relacionadas con condición de jefe del Cártel de Sinaloa, la organización de narcotráfico más poderosa del país, y su condición de participante clave en la violencia que ha reclamado miles de vidas desde 2006.
   ***Yahoo****: news.yahoo.com/quot-el-chapo-quot-guzm-n-nos-ser-202928187.html*
5. La caravana en la que se movilizaban por el departamento de Arauca los candidatos a la Presidencia de la República y al Senado, Aida Avella y Carlos Lozano, respectivamente, fue atacada por desconocidos que impactaron con armas largas a uno de los siete vehículos en que se movilizaban con sus asesores y escoltas.
   ***El País****: elpais.com.co/elpais/elecciones/noticias/candidata-presidencial-aida-avella-sufre-atentado-arauca*
6. Hasta el momento se desconoce quiénes fueron los perpetradores de este ataque, las autoridades todavía no señalan a ningún responsable del hecho.
   ***Universal****: eluniversal.com.co/politica/aida-abella-relato-ataque-uno-de-sus-carros-escolta-en-arauca*
7. En ese capítulo, el número uno del mundo anotó tres saques directos en todo el partido, algunos de ellos rozando los 200 kilómetros por hora, demostrando que su espalda está en mucho mejor estado que en días anteriores.
   ***Hidalgo Sport****: hidalgosport.com/2014/02/23/page/3/*
8. El líder de 24 años asegura que este es el peor momento del Gobierno chavista, ahora encabezado por Nicolás Maduro.
   ***El País****: elpais.com.co/elpais/internacional/noticias/vamos-seguir-luchando-con-coraje-calle*
9. En efecto, fueron tantas las voces que don Quijote dio, que, abriendo de presto las puertas de la venta, salió el ventero, despavorido, a ver quién tales gritos daba, y los que estaban fuera hicieron lo mismo.
   ***El ingenioso hidalgo Don Quijote de la Mancha***
10. Maritornes, que ya había despertado a las mismas voces, imaginando lo que podía ser, se fue al pajar y desató, sin que nadie lo viese, el cabestro que a Don Quijote sostenía, y él dio luego en el suelo, a vista del ventero y de los caminantes, que, llegándose a él, le preguntaron qué tenía, que tales voces daba.
    ***El ingenioso hidalgo Don Quijote de la Mancha***

## Test corpus and sources [TC]

1. Pues la ceremonia de clausura de los Juegos de Sochi, la olimpiada más cara de la historia, sirvió para confirmar que el buen humor existe en la Rusia de hoy.
   ***Libre Prensa***: *libreprensa.com/k/rusia/87202#s/1884646*
2. Al menos 26 policías iraquíes murieron este domingo y diez resultaron heridos en un ataque terrorista contra una sede de la Policía cerca de la ciudad de Mosul, 400 kilómetros al norte de Bagdad.
   ***El País****: elpais.com.co/elpais/internacional/noticias/cerca-26-policias-mueren-ataque-irak*
3. Todo ese poder intimidante lo pusieron a prueba en los ataques contra el canal Globovisión, con el secuestro de los trabajadores oficiales en el Ateneo de Caracas, con el atentado en el 2012 contra el excandidato presidencial Henrique Capriles en la Parroquia de San José de Cotiza y con los ataques de bandas motorizadas contra quienes votan en zonas de influencia de los candidatos opositores.
   ***El País****: elpais.com.co/elpais/internacional/noticias/asi-operan-colectivos-fuerzas-paramilitares-chavistas-venezuela*
4. El FC Barcelona, en un intento de minimizar el impacto legal que rodea al caso Neymar, tiene la intención de depositar mañana lunes en Hacienda los 9.1 millones de euros cifrados por el juez Ruz como cantidad defraudada en el fichaje del delantero brasileño en concepto de impuestos.
   ***ESPN Deportes****: espndeportes.espn.go.com/news/story?id=2028977&s=esp&type=story*
5. En una sartén pon un poco de mantequilla o aceite de tu preferencia y agrega cebolla a caramelizar.
   ***Univision****: cocina.univision.com/cocina-con-sabor/article/2014-02-21/ricas-calabacitas-rellenas-de-queso*

## Baseline test corpus translations [BT]

1. well the ceremony at closing with the sports off Sochi, the Olympics other look of the tale, sirvió at as to uphold what the good spirit to exist into the Russia from hoy.
2. at the least 26 policías iraquíes murieron east sunday and ten to proce injured in a onset terrorist to one headquarters off the Policía nearby of the city by Mosul, 400 kilometers at the north with Bagdad.
3. everything that power intimidating it pusieron of evidence by them onset against he channel Globovision, with the kidnapping off the working officer into the Athenaeum at Caracas, by the assualt by the 2012 against the former candidate presidential Henrique Capriles by the Parish by San José out listed and with them raid out side motorized against quienes vote in zone than clout with them nominee opponents.
4. he FC Barcelona, in an intention out to play down the hit legal what rodea to the instance Neymar, tiene the purpose with to put down mañana Monday into ranch the 9.1 million by Euro to encode because of the magistrate Ruz as amount to deceive into he signing of the leading brasileño into concept with laid.
5. by an sartén to put on one few with butter or oil out your choice and to incorporate onion at to caramelize.

## Final test corpus translations [FT]

1. Then the ritual of closing of the games of Sochi, the Olympics mores look of the history, do of to confirm what the good spirit exists in the Russia of time.
2. To the less 26 polices Iraqis they went this sunday and ten resulted woundeds in a terrorist attack to a headquarters of the police by of the city of Mosul, 400 kilometers to the north of Bagdad.

3. All that power intimidating the people of evidence in the attacks to the channel Globovision, with the kidnapping of the officers working in the Athenaeum of Caracas, with the assault in the 2012 to the presidential former candidate Henrique Carlos in the Parish of without the of listed and with the attacks of sides motorizeds to who vote in areas of influence of the candidate's opponents.
4. The FC Barcelona, in a try of to minimize the legal hit what surrounds to the case Neymar, to have the purpose of to deposit morning Monday in estate the 9.1 millions of Euro to encrypt for the judge Ruz as total to defraud in the signing of the front Brazilian in opinion of taxs.
5. In a pan to put a little of butter or oil of you choice and adds onion of to caramelize.

## **Google test corpus translations [GT]**

1. For the closing ceremony of the Games in Sochi, the most expensive Olympics in history, served to confirm that humor exists in Russia today.
2. At least 26 Iraqi policemen were killed Sunday and ten were injured in a terrorist attack on a police headquarters near the city of Mosul, 400 kilometers north of Baghdad.
3. All that power is intimidating tested in attacks against Globovision, the kidnapping of official workers in the Ateneo de Caracas, with the attack in 2012 against former presidential candidate Henrique Capriles in the Parish of St. Joseph of Price & with motorized mob attacks against those who vote in areas of influence of the opposition candidates.
4. The FC Barcelona in an attempt to minimize the impact legal case surrounding Neymar intends to deposit Hacienda Monday in the 9.1 million euros encrypted by Judge Ruz as disappointed in the signing of Brazilian striker amount in taxes.
5. In a pan add a little butter or oil of your choice and add onions to caramelize.