

# A Statistical Machine Translation System for Spanish-to-English Translation

Albert Chu, Brad Huang, and Nicholas P. Moores, Stanford University

Dan Jurafsky (for CS 124)

(Dated: 2 March 2015)

## I. COMMENTS ON *F*, SPANISH

To enumerate the ways we were able to capitalize on or were challenged by building this machine translation system, it is necessary to first discuss the relevant differences between English and Spanish. Key differences are that:

1. Spanish is a much more heavily inflected language than English, with different verb conjugations for *-ir* and *-er* and *-ar* final verbs and different conjugations for mood and tense. English only inflects verbs for number in the third-person present tense, and has largely abandoned marking verbs for number in the past tense.
2. Within noun phrases, Spanish adjectives tend to appear following the nouns they modify instead of before them as in English (though adjective placement in Spanish can be a cue to meaning)
3. Spanish word ordering is much more flexible than is English's. Whereas English rarely deviates from SVO (Subject-Verb-Object) word ordering, Spanish speakers do so whenever they desire to place particular emphasis on a word (and will then place that word toward the end of the sentence).
4. Spanish uses grammatical gender, whereas English abandoned such a system hundreds of years ago.
5. Spanish heavily relies on the use of the subjunctive mood when describing the past, whereas the use of the subjunctive mood in English has become less commonplace.
6. Spanish nouns and adjectives must also agree by both gender and number, as must the particles used at the beginning of noun phrases (for example, the definite articles *el*, *la*, *los*, and *las*).

7. English indefinite articles are marked for whether their noun begins (phonetically) with a vowel, whereas Spanish does not.
8. Spanish will often use a (non-progressive) present tense form to indicate that something is currently happening, whereas English will use primarily the progressive present tense form to indicate this; a non-progressive translation would seem wrong to a native English speaker.

Many of the difficulties which arise from translating Spanish into English rest on the morphosyntactic differences between the two languages. The lack of grammatical gender in English poses the problem that many of the words in a Spanish sentence that have to agree based on the gender of the noun at the head of the noun phrase must instead map to the same English articles and adjectives (but that the same does not hold for number). It is therefore a challenge to train a system to recognize that *traducido* and *traducida* map to the same English word, *translated*. Fortunately, since it is Spanish that is more inflected, it is easier to translate from Spanish to English since one can generalize across or collapse inflections, whereas an English-to-Spanish translation would require the system to correctly output sentences whose words are felicitously matched for Spanish gender, number, tense, and mood.

## II. IBM IMPROVEMENT STRATEGIES

In the course of this project we applied several separate strategies attempting to improve our statistical machine translation system.

### 1. `LanguageModel.py`

We first implemented a trigram language model with stupid backoff on English. For each Spanish word  $f$  to be translated, we select the top two words with the highest  $t(f|e)$  probabilities. However, for a Spanish sentence of length  $m$ , we would have  $2^m$  combinations of possible word-for-word translations. To limit the search space, the variable  $\alpha$  controls how many second-highest probability words can appear in one sentence, and is usually set at  $\alpha = 3$  for development.

### 2. `PhraseTranslation.py`

We then attempted to further improve the model by implementing a phrase table. During training, the IBM Model 1 is run twice to construct probability tables from English to Spanish and from Spanish to English. Word pairs of reciprocal best word-to-word translation probability are then selected, and some other word pairs are selected heuristically. Smaller phrases are then combined into larger phrases and a dictionary mapping Spanish phrases to English phrases is thus constructed.

However, we discovered that the phrase table does not increase either the BLEU-1 or BLEU-2 scores. One common type of errors phrase table makes is that words that appear less frequently in both languages may not be combined into phrases, and thus may be eliminated from the translation of phrase table, thus decreasing the BLEU scores. (Output of this script is named `output.pt` in the output folder for your reference.)

### 3. `nltpos.py`

Next, we used post-processing to improve the translation. One common word order difference between English and Spanish is that Spanish sentences usually have nouns preceding their modifiers while English is the opposite. As a result, we use NLTK's POS tagger to identify noun-adjective pairs and reverse the order. This increases the BLEU-2 score while not affecting the BLEU-1 score as expected.

#### 4. `Utility.py`

We introduced several other post-processing strategies. We observed that because of the unweighted alignments considered in IBM Model 1, some words are translated into stop words for no apparent reasons, so we check for and eliminate consecutive stop words (which should not appear in English) after translation.

In addition, since English requires phonological agreement between articles and nouns (i.e. 'a dog' vs 'an elephant'), we also check for this agreement and change the articles to 'a' or 'an' accordingly.

### III. ERROR ANALYSIS

*Please see the appendix for additional comments on what causes our errors to occur.*

#### A. General Error Analysis

In general, most of our errors appear to occur due to the following nuances of the difference between English and Spanish:

1. Length: Spanish sentences are in general longer than that of English, so it is beneficial to reduce the length of the sentence when translating. One possible means by which to reduce errors related to difference in length would be to restrict the length of the target sentence and eliminate words in the source sentence based on rankings like the Moldovan et al. criteria.
2. Alignment: All of the possible alignments of a given sentence pairs are treated equally. This means that the probability of aligning to repeating words is abnormally high. As a result, translations from IBM Model 1 contain unrelated spurious words that occurs only because they are repeated several times in several sentences of the corpus. For example, when we train on the first 300 sentences of the dev set, the output translation contains a lot of instances of *Military*. Some ways to improve this would be to: (1) make sure that a word does not align to words far from its corresponding index in the other sentence, (2) use stop word filtering to filter out abundant words like *el*, *la*, *the*, *a*, et cetera, and to assign hard rules to translate these stop words, (3) use a part-of-speech tagger to tag the training documents, and only consider alignments that align words of the same POS. Methods (1) and (3) are difficult because they increase the runtime a lot due to increased computational complexity (we have to actually look into each alignment), while method (2) would create some issues because there may not be 1-some correspondences between all Spanish and English stop words.
3. Word Order: Spanish and English have in general the same word orders; however, Spanish word order is more flexible than word order in English. To properly account for this, we would need a syntactic language model. See the second Google translate output for an example.

## B. Language Model:

The errors we encountered using the language model approach relate to improperly ordered translation of noun phrases. Since Spanish nouns usually precede the adjectives that modify them, the English translations given by this model mistakenly keep the original order of the Spanish words, failing to correct for the different ordering of noun phrases. Our system correctly translates the Spanish *de* to English *of*, but fails to take the adjective which follows *de* in the Spanish source sentence and place it directly before the English noun in the translated sentence. In these instances, the output is semantically correct but syntactically uncommon. The system also often fails to distinguish between words in Spanish that function as either an adjective or a noun, such as the translation of *tecnico* into (*technical*, adjective) when it should have been (*technician*, noun). The English translated counterparts for Spanish stop words also often end up being erroneously removed as our model has trouble correctly reinstating the stop words in the English translation. We could address these errors by .

#### IV. COMPARATIVE ANALYSIS WITH GOOGLE TRANSLATE

*Please see the Appendix for a full enumeration of the source sentence, our output, and the gold translations.*

1. **Google Translation:** This attorney for 48 years specializing in human rights and president of the Egyptian Organization for Human Rights (EOHR) defended the Muslim brothers when they were held in prison or had to deal with a trial during the dictatorship of Mubarak.

Analysis:

- It's funny how both translations did not understand that 'de 48 años' actually refers to the age of the lawyer..
  - Spanish omits subjects whenever possible and uses inflections to identify subjects mentioned before. Our translation did not account for verb inflections, so we weren't able to identify that the subject for '... se encontraban...' is 'hermanos musulmanes' as mentioned right before that verb.
  - From this translation we can observe that our model is poor at translating verbs. As mentioned above, this is a result of the various inflected forms possible in Spanish. Google probably encodes Spanish inflections in their translation model, so that they can identify subjects and translate words more successfully.
2. **Google Translation:** "But oil is not the only reason why it is going so well" Says Anna, our waitress, while passing trays rakfisk and, with her long blonde hair and striking blue eyes, represents the image of the Norwegian welfare.

Analysis:

- In the quote section, the latter part is actually 'why we are doing so well', and neither translations are perfect in this case. Google translate changes 'nos' into 'it' because of the following 'va' (third person of 'ir'), while our translation just did word-to-word translation to get 'we is'. Not clear whether that is a typo or

spoken error made, but it is difficult to tell the the correct subject-verb based on 'nos va' as presented without the environment.

- Observe that there are several deletions in our translation that should not occur - 'our (waitress)' and 'long (blonde hair)'. This usually occurs to words with low frequency in the corpus. Because their appearances are so infrequent in the corpus, they are mapped to 'NULL' during IBM Model 1's training process, and is thus eliminated from the final translation.
- It is also interesting how although the Spanish sentence aligns pretty well to the actual translation in terms of word order, both Google and our translation translated the word 'representa' while that word is not in the actual translation.

3. **Google Translation:** Men and women mention his career, stress and cost of property and education as reasons that prevent them from having children.

Analysis:

- 'su' in Spanish gets translated into 'his' by Google (true in a word-to-word scenario), but 'their' by our model (better here, since it refers to both 'Mena and women'). It is an interesting case because it should actually be 'sus' here since the phrase actually refers to 'their careers'.
- Since we are not eliminating stop words when necessary, our model does place the multiple 'the's in the final translation. Although we made attempts during development, we could not think of a good statistical strategy to eliminate stop words this way.
- Similarly, when the Spanish has '... el costo **de** las preiodades y **de** la educacin ...', our model translated both 'de' in the sentece, while Google identified that the sentence does not require both (both referring to 'cost of').
- Again, notice that 'motivos' and 'impiden' are translated incorrectly, one into a stop word and another eliminated, in our translation. As mentioned above, it is likely that these inflected forms appeared too few times in the corpus, which led to incorrect word-to-word mappings.



4. **Google Translation:** The measures taken since 2009 ensured that the rate of registration of new voters fell in 2010 compared with 2006.

Analysis:

- Google translated 'lograron' into 'ensured' while our model translated it into 'succeed'. 'Lograr' usually means 'achieve, succeed', and 'lograron' here actually means 'lead to' in this case. Although Google's translation is closer to the actual meaning, it is still debatable given the negative connotation ('17 % drop') afterwards
- Notice again the incorrect translations of 'inscripcin' and 'cayera'. This is possibly due to the same reason why some words in previous examples are incorrectly translated by our model.

5. **Google Translation:** The Prime Minister recently said the ODS (Civic Democratic Party) would not bother entrepreneurs with inspectors. So what is prohibited or permitted?

Analysis:

- A clear difference between our translation and Google's translation is the negation part. Google identified that there is negation in the sentence and translated accordingly, while our model did not.
- In the same vein, Google probably identified the inflected form of the verb and translated it into 'would not bother', while our model does nothing about Spanish inflection, so our translation did not align to Spanish that well.
- It is interesting, though, to observe that the last part of the sentence, where the real translation is '- so is it forbidden or allowed ?'. Both Google and our model failed to identify that the subject of the question is 'ODS' as mentioned in the previous part of the sentence.

## V. APPENDIX

### A. Illustrating Examples of Language Differences

- Spanish: *el hombre corre* and *yo corro* and *los estudiantes corren*
- English: *the man runs* and *I run* and *the students run*
- Spanish: *un pobre hombre* and *un hombre pobre*
- English: *a poor (unfortunate) man* and *a poor (destitute) man*
- Spanish: *Masoud trabaja en el mercado.* and *Trabaja Masoud en el mercado.* and *En el mercado trabaja Masoud.*
- English: *Masoud works in the market.*
- Spanish: *la abuela hermosa* and *el abuelo hermoso* and *los estudiantes hermosos*
- English: *the beautiful grandmother* and *the beautiful grandfather* and *the beautiful students*
- Spanish: *Cuando eras estudiante, tu no trabajabas mucho* and *Cuando eras estudiante, no trabajaste mucho*
- English: *When you were a student (for a long time), you didn't work much (habitually)* and *When you were a student (for a long time), you didn't work much (for a set time)*
- Spanish: *una abuela* and *una computadora*
- English: *a question* and *an answer*
- Spanish: *una abuela* and *una computadora*
- English: *a question* and *an answer*
- Spanish: *Ellos comen* but *Ellos son comiendo.\**
- English: *They eat\** and *They are eating*

### B. System Output as Compared to Google Translate Output Examples

1. **Spanish Source:** Este abogado de 48 años especializado en derechos humanos y presidente de la Organizacin Egipcia de Derechos Humanos ( EOHR ) defendi a los

hermanos musulmanes cuando se encontraban retenidos en prisin o tuvieron que hacer frente a un juicio durante la dictadura de Mubarak .

**Google Translation:** This attorney for 48 years specializing in human rights and president of the Egyptian Organization for Human Rights (EOHR) defended the Muslim brothers when they were held in prison or had to deal with a trial during the dictatorship of Mubarak.

**Our Translation:** This lawyer of 48 years specialises in rights human and President of the organisation Egyptian of human rights ( EOHR ) for to the Muslim Muslims when who held in prison or had that to with to a trial for the dictatorship of Mubarak .

**Actual Translation:** The 48-year-old human rights lawyer and Chairman of the Egyptian Organisation for Human Rights ( EOHR ) defended the Muslim Brotherhood , when imprisoned or in court under Mubarak .

2. **Spanish Source:** “ Pero el petróleo no es la única razón por la que nos va tan bien ” , afirma Anna , nuestra camarera , mientras va pasando bandejas de rakfisk y que , con su larga melena rubia y sus sorprendentes ojos azules , representa la imagen del bienestar noruego .

**Google Translation:** “ But oil is not the only reason why it is going so well ” Says Anna, our waitress, while passing trays rakfisk and, with her long blonde hair and striking blue eyes, represents the image of the Norwegian welfare.

**Our Translation:** “ but the oil is not the only reason the we is so well ” , says Anna , our , while is happening explained of and that , with their long and their surprising blue eyes , represents the image the welfare Norwegian .

**Gold Translation:** “ But oil is not the only reason we are doing so well , ” says Anna our waitress , handing round trays of maturing rakfisk and , with her long blond hair and startlingly blue eyes , the image of Nordic well-being .

**Actual Translation:** “ But oil is not the only reason we are doing so well , ” says Anna our waitress , handing round trays of maturing rakfisk and , with her long blond hair and startlingly blue eyes , the image of Nordic well-being .

3. **Spanish Source:** Hombres y mujeres mencionan su carrera , el estrés y el costo de las propiedades y de la educación como los motivos que les impiden tener hijos .

**Google Translation:** Men and women mention his career, stress and cost of property and education as reasons that prevent them from having children.

**Our Translation:** Men and women mention their career , the stress and the costs of the properties and of the education as the for that them from have children .

**Actual Translation:** Men and women alike mention their careers , stress and the cost of property and education as the reasons preventing them from having children .

4. **Spanish Source:** Las medidas adoptadas desde 2009 lograron que la tasa de inscripción de nuevos electores cayera en 2010 en comparación con el año 2006 .

**Google Translation:** The measures taken since 2009 ensured that the rate of registration of new voters fell in 2010 compared with 2006.

**Our Translation:** The measures taken since 2009 succeed that rate of for of new electorate out in 2010 in compared with the year 2006 .

**Gold Translation:** However , the measures adopted since 2009 have led to a 17 % drop in the registration rate of new voters in 2010 compared to 2006 .

**Actual Translation:** However , the measures adopted since 2009 have led to a 17 % drop in the registration rate of new voters in 2010 compared to 2006 .

5. **Spanish Source:** El primer ministro dijo recientemente que la ODS ( Partido Democrático Cívico ) no molestaría a los empresarios con inspectores . Entonces , est prohibido o permitido ?

**Google Translation:** The Prime Minister recently said the ODS (Civic Democratic Party) would not bother entrepreneurs with inspectors. So what is prohibited or permitted?

**Our Translation:** The first Minister said recently that ODS ( democratic civic party ) not to the employers with inspectors . then , is ban or allowed ?

**Actual Translation:** The Prime Minister recently claimed that the ODS will not be burdening business owners with its checks - so is it forbidden or allowed ?

### C. Version-specific Error Analysis

Through the course of the project, we implemented two improvements to the baseline model: A trigram language model with stupid backoff on English, and implemented phrase translation. We report the characteristics of our errors as they occurred in the baseline case, in the case where the system utilized the language model, and in the case where the system utilized phrase translation.

- **Baseline:** Both infrequent Spanish words and infrequent inflected forms are often translated as *fawn*. In addition, proper nouns are encountered so infrequently that they are often replaced with *fawn* in the translation or are omitted altogether. One strategy we could utilize in the future is to use a gazeteer as additional input to the system to help correct errors in handling proper nouns.

This behavior can be understood based on the fact that in our baseline implementation, all alignments are weighted equally. One way to improve this would be to limit the alignment to maybe only a few words before and after each word, so that in the sentence *yo voy a la tienda*. (translation: *I go to the store.*), *voy* could align to  $\pm 1$  precision (only aligning to *I* and *go* and *to*).

One of the biggest errors was that since there are so many inflected forms, some of them are not identified as the same word as the others. One method we could use in the future is to try stemming the source text.

- **Phrase Translation:** Longer sequences of words are correct when we use phrase translation, but the English translations are often missing stop words or entire dependent clauses. As a whole, our system with phrase translation maintains proper word order but omits words that belong in a correct translation. We could address these errors by encouraging the system not to eliminate infrequent words from the translation of the phrase table.