CS 124 Assignment 6

Machine Translation

Language selected: Spanish

February 27, 2014

KEY LANGUAGE DIFFERENCES

For this assignment our team chose to translate from Spanish to English. Here are some of the key differences between the two languages:

#	Difference From English	Description
1	Verb Tense	There is a lack of one-to-one equivalencies for verb tense. For example, the translation of Spanish's "yo hablo" can take on several english verb tenses such as "I speak" or "I am speaking".
2	Grammar	While both Spanish and English are Subject-Verb-Object languages, Spanish often places emphasized words at the end of the sentence where they might have occurred earlier in an equivalent English sentence. For example, "es difícil esta clase" places class at the end of the sentence, whereas its equivalent English translation ("this class is hard") puts class at the beginning.
3	False Cognates	Spanish and English have many words that appear and/or sound similar, but have very different meanings. For example, Spanish's "advertencia" might appear to have the same root as English's "advertisement", but actually translates to warning, advice, or reminder. An even more extreme example of this is the Spanish word "colorado", which translates as "colored", not the U.S. state of "Colorado".
4	Dropped Subject Pronouns	As Spanish has verb conjugations that imply the "doer" of the action, subject pronouns (I, you, he/she) are often dropped from Spanish sentences. For example the long version of "I swim" in Spanish is "yo nado", however in this case the "yo" will be frequently dropped to make the sentence simply "nado". Spanish speakers often drop the subject pronoun from "understood" objects that are repeated in subsequent sentences. For example, the Spanish sentence "Tengo un libro de texto. Está azul" drops the subject pronoun in the second sentence, as "Está" implies we're talking about the textbook.
5	Idiomatic Expressions	Both Spanish and English have widely used idiomatic expressions that do not have the same meaning when translated word-for-word. For example Spanish's "tengo 20 años" translates as "I'm 20 years old", not "I have 20 years".
6	Differences In Capitalization	Spanish, unlike English, does not capitalize days of the week, nor months, nor foreign languages.
7	Contextual in, on, and at	English's "in", "on", and "at" can all be translated to Spanish's "en". However, translation from Spanish's "en" to English can take the very different forms of "in", "at", and "on", completely depending on context and subject.
8	Plural Adjectives	Spanish has plural adjectives whereas English does not. For example, "altos" or "rojas" which translate to "high" and "red", respectively.

15 SENTENCE CORPUS*

DEVELOPMENT SET

1.) Los cambios regulatorios castigan a Acciona, que pierde 1.972 millones.

2.) Es la cara y la voz de los socialistas para la cita de las europeas que, dice, marcarán el inicio del cambio político en España.

3.) Marruecos no aceptó la ayuda española para el rescate de cuerpos.

4.) El Foro Económico Mundial presenta los 10 grandes avances que tendrán un papel decisivo en el cambio del mundo moderno.

5.) Intentaron reanimar a los dos más graves pero no fue posible.

6.) En su declaración ante el juez, los verificadores también han precisado que cobran 750 euros por día de trabajo.

7.) Somos profesionales con experiencia y creemos que tenemos una auténtica oportunidad para la paz.

8.) ¿Cuál es su número de teléfono?†

9.) Un juez de Miami estudia si el público tiene derecho a ver el pene de Bieber.

10.) Nadal vuelve a su sitio.

TEST SET

1.) La semana fantástica de Ibra.

2.) Las 30 profesiones que aseguran un trabajo para la próxima década.

3.) El Gobierno tiene que mostrarle a Reding su malestar.

4.) Las españolas saben administrar el Estado.

5.) No me gusta, ni me parece provocador.

---

* Unless noted otherwise all sentences in corpus taken from the headlines of www.elmundo.com on February 23rd, 2014.

† Taken from http://www.enchantedlearning.com/language/spanish/phrases/questions.shtml on February 23rd, 2014.

BASELINE VS. FINAL SENTENCES

DEVELOPMENT SET

Baseline	Final Output
The changes regulatory punish to Acciona, that lost 1.972 million.	The regulatory changes punish to Acciona, that loses 1,972 million.
This is the face and the voice of the socialists for the appointment of the European that, say, scored the initiation of change political in Spain.	This is the face and the voice of socialists for the appointment of the European that, says, scored the initiation of political change in Spain.
Morocco not accepted the help Spanish for the rescue of bodies.	Morocco not accepted the help Spanish for bodies rescue.
The Forum Economic World present the 10 large developments that have a role decisive in the change of world modern.	Forum Economic World presents the 10 large developments that they have a decisive role in the world change modern.
Tried to revive to the two more serious but not was possible.	It tried to revive to the two more serious but it was not possible.
In your statement before the judge, the checkers also have specify that charge 750 euros per day of work.	In his statement before the judge, checkers also have specifying that charges 750 euros per work day.
Are professionals with experience and think that have a authentic opportunity for the peace.	They are professionals with experience and they think that they have an authentic opportunity for the peace.
What this is your number of telephone?	What this is his telephone number?
A judge of Miami is studying if the public have right to view the penis of Bieber.	Miami's judge is studying if the public has right to view Bieber's penis.
Nadal return to your room.	Nadal returns to his room.

TEST SET

Baseline	Final Output
The week great of Ibra.	The week Ibra's great.
The 30 professions that ensure a work for the next decade.	The 30 professions that ensure a work for the next decade.
Government have to show to Reding your discomfort.	Government have to show to Reding your discomfort.
Spanish know manage State.	Spanish know manage State.
Me do not like, or me seems provocative.	Me do not like, or me seems provocative.

PROCESSING STRATEGIES

#	Strategy	Description	Application In Development & Test Sets
1	Swap adjacent adjectives and nouns	In Spanish, adjectives generally come after nouns whereas in English adjectives generally come before nouns. For example, “casa verde” would translate to “green house.” This shows up in our corpus: Notice that “change political” is switched to “political change. This strategy is almost always true but there are a few exceptions like the English word “galore.” We implemented this using a part of speech tagger.	Baseline: this is the face and the voice of the socialists for the appointment of the European that, say, scored the initiation of change political in Spain. Translation: This is the face and the voice of socialists for the appointment of the European that, says, scored the initiation of political change in Spain.
2	Swap position of negation	In Spanish, negations generally come before the verb, which is generally not the case in English. For example, “No estoy cansado” translates to “I am not tired” as opposed to “no am tired.” “not was” is correctly swapped to “was not.” One other fine point is the case where a pronoun occurs in between the negation and verb. This showed up in our test set as “No me gusta.” In cases where we had a pronoun after the negation, we checked one word ahead to see if it was a verb, and then handled the situation similarly. We implemented this using a part of speech tagger.	Baseline: tried to revive to the two more serious but not was possible. Translation: It tried to revive to the two more serious but it was not possible. Baseline: not me like, or me seems provocative Translation: Me do not like, or me seems provocative.
3	Swap nouns when separated by “of”	In Spanish often times a construction such as “[Noun1] de [Noun2]” will be used. When “de” translates to “of” this often should be represented as “[Noun2] [Noun1]” in English. “día de trabajo” should translate to “workday” as opposed to “day of work”. Likewise “número de teléfono” should translate to “telephone number” as opposed to “number of telephone”. We used a language model for word disambiguation as described later to help with this strategy. In addition, we used a part of speech tagger.	Baseline: in your statement before the judge, the checkers also have specify that charge 750 euros per day of work. Translation: In his statement before the judge, checkers also have specifying that charges 750 euros per work day. Baseline: what this is your number of telephone? Translation: What this is his telephone number?

#	Strategy	Description	Application In Development & Test Sets
4	Possessives with apostrophes in English	Spanish does not use the apostrophe “s” construction to denote possessives. So “El carro de David” should be translated to “David’s car” as opposed to “The car of David.” “de” can be translated to a number of different words, but it is only when it translates to “of” can it denote a possessive. Also, to disambiguate with the strategy above, possessives generally occur when the second argument is a proper noun. This doesn’t capture every case, but it has high precision. “de” is disambiguated to “of” and “Bieber” is a proper noun, so we translated “el pene de Bieber” to “Bieber’s penis.” Similarly, “Ibra” is a proper noun, so we do a similar translation. We used word disambiguation (as discussed in a later strategy) and a part of speech tagger.	Baseline: a judge of Miami is studying if the public have right to view the penis of Bieber. Translation: Miami's judge is studying if the public has right to view Bieber's penis. Baseline: The week Ibra's great. Translation: the week great of Ibra.
5	Select relevant translation based on part of speech	Many Spanish words in our dictionary had a many-to-one relationship. We included all translations and labeled the parts of speech. To choose a translation, we used a Spanish part of speech tagger, then chose the corresponding English translation. In cases where we could not find a match, we choose the highest probability translation. This significantly improved all of our sentences. We just present a few for the sake of space but please note that this strategy applies to all sentences and to all parts of speech in the test set.	Baseline: Nadal return to your room. Translation: Nadal returns to his room. The model correctly chooses “his” for the translation for “su.” Baseline: the government have that show to Reding your discomfort Translation: Government have that show to Reding his discomfort “su” can either be a pronoun or an adjective. Our part of speech tagger correctly identifies that a pronoun is needed and chooses the translation accordingly. In the case of pronouns for “su” there are multiple possibilities in which case we use the language model as described later to disambiguate.
6	Add pronoun when neglected in Spanish (may not be correct pronoun)	Spanish is a pro-drop language and frequently omits pronouns. For example “estamos cansados” should translate to “we are tired” instead of “are tired”. This strategy aimed to identify where pronouns should be included. We looked at the word right before all verbs. At a high level, if it wasn’t a noun or pronoun, we included a pronoun. There were also cases that specifically cancelled the need to add a pronoun such as when the word before the verb was a determiner, “wh” word, adverb, preposition, among a few others. In all three cases, the added pronouns are underlined. This part was implemented with a part of speech tagger.	Baseline: the Forum Economic World present the 10 large developments that have a role decisive in the change of world modern. Translation: Forum Economic World presents the 10 large developments that they have a decisive role in the world change modern. Baseline: tried to revive to the two more serious but not was possible. Translation: It tried to revive to the two more serious but it was not possible. Baseline: are professionals with experience and think that have a authentic opportunity for the peace. Translation: They are professionals with experience and they think that they have an authentic opportunity for the peace.

#	Strategy	Description	Application In Development & Test Sets
7	Choose correct pronouns	After detecting where pronouns should be included, the next step was to choose the proper pronoun. We chose the pronoun based on the verb it was associated with. Using a part of speech tagger, we determined whether the word was singular or plural and chose the pronoun accordingly.	Baseline: the Forum Economic World present the 10 large developments that have a role decisive in the change of world modern. Translation: Forum Economic World presents the 10 large developments that they have a decisive role in the world change modern. Baseline: tried to revive to the two more serious but not was possible. Translation: It tried to revive to the two more serious but it was not possible. Baseline: are professionals with experience and think that have a authentic opportunity for the peace. Translation: They are professionals with experience and they think that they have an authentic opportunity for the peace.
8	Conjugate verbs based on present/ past, singular/ plural, present/past participle	Our strategy here was to correctly conjugate the English word, given that many of our translations in our dictionary only provided the infinitive form. This was a non-trivial improvement since some words are exceptions to the general conjugation rules in English like the word “is”. Using a part of speech tagger, we identified tense, person, and participle. This positively affected almost all of our sentences. Here are just a few in the interest of space:	Baseline: the changes regulatory punish to Acciona, that lost 1.972 million. Translation: The regulatory changes punish to Acciona, that loses 1,972 million. Baseline: a judge of Miami is studying if the public have right to view the penis of Bieber. Translation: Miami's judge is studying if the public has right to view Bieber's penis. Baseline: Nadal return to your room. Translation: Nadal returns to his room.

#	Strategy	Description	Application In Development & Test Sets
9	Remove Determiners	Spanish frequently uses articles such as “lo” and “las”. These roughly translate to “the” in English. However, not all of these occurrences of articles need to be included in the English translation. English has pretty general rules for dealing with this. Proper nouns, plural nouns, and mass nouns do not need determiners. Countable singular nouns do. We applied these rules and it affected the following sentences:	Baseline: a judge of Miami is studying if the public have right to view the penis of Bieber. Translation: Miami's judge is studying if the public has right to view Bieber's penis. Baseline: the Spanish know manage the state . Translation: Spanish know manage State. Notice that we removed “the” in front of “Bieber’s penis” and “the” in front of “Spanish.” This is required for proper English. There were a number of other cases that we removed the “the” for better readability, although it is not strictly necessary to do so. Baseline: this is the face and the voice of the socialists for the appointment of the European that, say, scored the initiation of change political in Spain. Translation: This is the face and the voice of socialists for the appointment of the European that, says, scored the initiation of political change in Spain. Baseline: Morocco not accepted the help Spanish for the rescue of bodies. Translation: Morocco not accepted the help Spanish for bodies rescue. Baseline: the Forum Economic World present the 10 large developments that have a role decisive in the change of world modern. Translation: Forum Economic World presents the 10 large developments that they have a decisive role in the world change modern. Baseline: in your statement before the judge, the checkers also have specify that charge 750 euros per day of work. Translation: In his statement before the judge, checkers also have specifying that charges 750 euros per work day.

#	Strategy	Description	Application In Development & Test Sets
10	Use commas instead of periods for numbers, don't capitalize day of week or names of months, remove question marks and exclamation points at beginning of spanish sentence, choose between "a" and "an", capitalize proper nouns and first words	This strategy comprised multiple minor, yet general differences between Spanish and English. Spanish uses periods in numbers such as "5.000" for "5,000". It does not capitalize days of the week and months such as "lunes" for "Monday". It has exclamation points and question marks at not just the end of a sentence but also the beginning. In English, we disambiguate between "a" and "an" depending on whether or not it is followed by a vowel sound. Finally, as with English, we capitalize proper nouns and first words in sentences.	Baseline: **t**he changes regulatory punish to **A**cciona, that lost 1.972 million. Translation: **T**he regulatory changes punish to **A**cciona, that loses 1,972 million. Baseline: are professionals with experience and think that have a authentic opportunity for the peace. Translation: They are professionals with experience and they think that they have an authentic opportunity for the peace. Other sentences in the corpus that included proper nouns were also capitalized, but these were not included here in an effort to save space. Test set: Baseline: the week great of Ibra Translation: The week Ibra's great Baseline: the 30 professions that ensure a work for the next decade Translation: The 30 professions that ensure a work for the next decade
11	Language model for word disambiguation	We implemented a n-gram model for word disambiguation. Given a target word within a sentence, our n-gram model computes the probability of all possibilities and chooses the highest probability translation. A great example of this in action is the word "su" which can translate with nearly equal probability to "your", "his", "her", "its", or "their". Our n-gram model chooses the highest probability candidate. Our n-gram model is currently trained on the NLTK Brown corpus with trigrams.	Baseline: Nadal return to your room. Translation: Nadal returns to his room. The model correctly chooses "his" for the translation for "su." Baseline: the government have that show to Reding your discomfort Translation: Government have that show to Reding his discomfort
12	No Plural Adjectives in English	Spanish has plural adjectives but English does not. Generally the dictionary captures these cases but sometimes it confuses the plural adjectives and returns a plural noun. An example of this is "blancos" which translates to "whites." Depending on whether "blancos" is an adjective or noun, we translated all plural adjectives to singular adjectives.	Relevant throughout the Spanish language.

GOOGLE TRANSLATE COMPARISON

Superior translation notated in blue

#	Our Test Set	Google Translate
1	The week Ibra's great.	Fantastic week of Ibra.
2	The 30 professions that ensure a work for the next decade.	The 30 occupations securing a job for the next decade.
3	Government have to show to Reding your discomfort.	The government has to show his displeasure Reding.
4	Spanish know manage State.	The Spanish know how to manage the state.
5	Me do not like, or me seems provocative.	I do not like, and I find provocative

#	Comparison Analysis By Sentence
1	Our translation is better since captures the possessive. The correct translation is "Ibra's fantastic week." Our translation correctly identifies the possessive between "Ibra" and "great" but only fails on the ordering of the sentence and the translation to "great" instead of "fantastic.
2	Google's system does better, although they are very close. The key difference is that Google correctly translates "trabajo" to "job" instead of "work". "Work" is the literal translation. To know that it corresponds to "job" instead, we would need to incorporate larger amounts of context and training data for our language model. Given our small scale and given Google's access to large data sets and compute power, Google can do better on translations like these. Another difference is "professions" instead of "occupations." This is pretty similar and either translation works equally well. The last difference is between "ensure" instead of "securing". Google incorrectly uses a gerund.
3	Google's system does much better on this one. Both are difficult to read and understand, but our translation especially so. Our system removes the leading "The" and conjugates the verb to "have" instead of "has" due to incorrect part of speech tagging. Our part of speech tagger thinks that government is a plural noun, but in this context it is a singular noun. We would need a larger data set and use higher order n-grams with backoff to get a better tagger. Another difference is between "that show to" and "to show". Our translation does a word-by-word translation, but we would need a rule to translate to a more idiomatic English expression. Both incorrectly translate the pronoun to "your" and "his". Both are referring to government, but cannot identify it in the long context. The next difference is between "displeasure" and "discomfort." As explained before, this is due to an issue in the language model. We need a larger data set and higher order n-grams to produce more idiomatic expressions. The last difference is between "Reding your discomfort" and "displeasure Reding." Both are incorrect as they cannot correctly link that phrase with the first part.
4	Google's system is nearly flawless in this case. The leading "the" is optional since "Spanish" is a proper noun. Our systems omits the connecting words "how to." This could be due to the fact that Google's system uses a phrasal or idiom dictionary. One thing our system does better than Google's is picking out proper nouns. State is capitalized in our version, whereas state is not detected as a proper noun in Google Translate. Lastly, the "the" in front of "State" is in Google's version but not ours. Since "State" is a proper noun it is not needed. Overall, Google's translation sounds more fluent, but ours is totally understandable as well.
5	Google's system works better because it captures the correct pronoun and selects more fitting verbs. Our system uses "me" as the subject instead of "I". Google Translate also captures a word better suited to the context: "find", instead of "seems", which we translated. Our attempts to tackle the word choice problems fell under the n-gram model trained under the NTLK Brown Corpus. We could do better with a parser to detect that we need a subject instead of an object for "me". In addition, we would need a larger training corpus to select better verbs.

ERROR ANALYSIS

#	Error	Proposed Fix
1	Verb not conjugated properly *Spanish*: En su declaración ante el juez, los verificadores también han precisado que cobran 750 euros por día de trabajo. *Translation*: In his statement before the judge, checkers also have specifying that charges 750 euros per work day. *Correct*: In his statement before the judge, checkers also have specified that charges 750 euros per work day. The root of the problem is that our Spanish part of speech tagger incorrectly tags the verb “precisado”. Our system correctly identifies “to specify” is in its participle form but incorrectly identifies it as present tense. Thus, our verb conjugation strategy is working as intended but our Spanish part of speech tagger is causing the error.	Our part of speech tagger is trained on a small dataset and only uses unigrams for tagging parts of speech. To improve our tagger we would need a larger dataset such as the web and use higher order n-grams with a backoff strategy.
2	Pronouns are not the right gender *Spanish*: ¿Cuál es su número de teléfono? *Translation*: What this is his telephone number? *Correct*: What this is your telephone number? Our language model translates “su” to “his” as opposed to “your”.	There is no way for our language model to distinguish between “his” and “your” based on just trigrams because those would produce equally valid trigrams. To fix this, we would need a gender classifier. This would disambiguate between masculine, feminine, and neutral genders.
3	Pronouns are not the right plurality *Spanish*: En su declaración ante el juez, los verificadores también han precisado que cobran 750 euros por día de trabajo. *Translation*: In his statement before the judge, checkers also have specifying that charges 750 euros per work day. *Correct*: In their statement before the judge, checkers also have specifying that charges 750 euros per work day. Our language model translates “su” to “his” as opposed to “their”.	Due to computational constraints and relatively small test set, we are using trigrams for our language model. To correct this problem, we would need to either user higher order n-grams, which consequently which require a larger dataset like the web, or we would need to run a parser to link the pronoun with the object. With a parse, we can link “checkers” with the pronoun and properly select the plural pronoun.

#	Error	Proposed Fix
4	Not translating possessives as <owner's object>, but rather <object of owner> *Spanish*: Un juez de Miami estudia si el público tiene derecho a ver el pene de Bieber *Translation*: Miami's judge is studying if the public has right to view Bieber's penis. *Correct*: A judge from Miami is studying if the public has right to view Bieber's penis. The issue roots from the fact that our language model translates "de" to "of" as opposed to "from". Since it translates "de" to "of", our system then incorrectly thinks we have a possessive.	The error roots from our language model. Unfortunately due to limited computation power our n-gram model has to be trained on small datasets and small values of n. To get a more accurate language model, we would need to train on a larger dataset such as the web and use higher n-grams.
5	Poor word choices given multiple translations *Spanish*: Las 30 profesiones que aseguran un trabajo para la próxima década. *Translation*: The 30 professions that ensure a work for the next decade. *Correct*: The 30 professions that ensure a job for the next decade. The literal translation of "trabajo" is "work" instead of "job". Our system cannot capture the right translation based on the context.	Similar to some of the errors above, the error roots from our language model. Unfortunately due to limited computation power our n-gram model has to be trained on small datasets and small values of n. To get a more accurate language model, we would need to train on a larger dataset such as the web and use higher n-grams.
6	Wrong plurality in translation *Spanish*: El Gobierno tiene que mostrarle a Reding su malestar. *Translation*: Government have that show to Reding his discomfort. *Correct*: The government has that show to Reding its discomfort. Our system incorrectly detects "El Gobierno" as a plural noun as opposed to a singular noun.	The error is due to our part of speech tagger. If we wanted to improve our part of speech tagger, we would need a larger data set along with higher order n-grams with backoff. Another solution is to use a named entity recognizer. This would tell us that "government" is an entity which we could use to conclude that its singular.

WORK CITED

Cover photo acquired from http://bit.ly/1huCZNZ