A Direct Translation System from Spanish to English

Contributors: Anonymous

CS124

Introduction:

We implemented a direct translation system from Spanish (F) to English (E). A simple baseline system for direct translation would take each Spanish word in a given sentence and replace them directly with any of the possible English translations. However, this approach is very rudimentary and we wanted to improve upon it by identifying the multiple ways in which Spanish linguistically diverges from English. Through consideration of these differences, we created a translation system that was more faithful to the true meaning and fluent than the baseline approach.

One of the first differences identified was the ordering relation between nouns and adjectives. In Spanish, nouns come before their descriptive adjectives, meaning the object is known and then described. An easy example of this can be found in the phrase “una casa blanca”, where “casa” is the noun for house and “blanca” is the matching adjective. This differs distinctly from English, where adjectives precede the nouns: “a white house”. Another related distinction between the two languages stems from the use of nouns as adjectives. In English, to use a noun as an adjective, we simply prepend it to the noun we intend to describe, as in the example “history teacher”. In Spanish, this is achieved by taking the noun to be described, appending “of” (“de”), and then appending the descriptor noun, as in “profesor de historia”. These two separate, yet closely related and highly used description conventions are important differences between Spanish and English.

Another divergence occurs with the negation of verbs. In Spanish, to negate a verb, one simply prepends “no” in front of the verb, as in “el no grita”. This differs from English, in which one prepends an auxiliary verb such as “does”, as in “he does not yell”. Essentially, this means that English requires the insertion of another word besides the negative itself to negate a verb.

A subtle divergence with substantial translation implications is the cold nature of Spanish. Cold refers to the implicitness of subjects hidden within verbs, or that subjects are often left to be entirely implicit. The lack of an explicit subject is made for in the conjugations of verbs which give much more context in Spanish than in English. This is made obvious with an extreme example of the word “to have”. In English the conjugations are: “I have”, “you have”, “he/she/it has”, “we have”, “they have”. For five different subjects there are only two different conjugations (“have”/”has”). In Spanish the conjugations are: “Yo tengo”, “tu tienes”, “el/ella/usted tiene”, “nosotros tenemos”, “ellos/ustedes tienen”. Spanish has 5 conjugations for 5 different subjects, meaning each conjugated verb has much more implicit meaning built in. This implicit meaning allows Spanish to leave out subjects more readily without a loss of comprehension, unlike English.

An additional difference between the languages is the changing of the word “a” in English when followed by a vowel-begun word. The equivalent word in Spanish (“un”) does not change upon the same condition.

One last important divergence is the tendency of English sentences to drop certain

words such as articles (“the”, “a”, “an”) that are apparent in the corresponding Spanish sentence. For example, the corresponding Spanish phrase to “Last year, I went…” would be “El año pasado, fui…” where the article “El” has no corresponding “the” in the English phrase. Ultimately finding these words that should be dropped and removing them leads to more fluent English sentences.

Methodology / Strategic Reasoning:

Given the observed linguistic differences between Spanish and English, we implemented a series of strategies to effectively translate Spanish to English.

Strategy 1:

The most significant strategy implemented was a bigram language model that incorporated Kneser-Ney Smoothing and backed off to a Laplace smoothed unigram model. This language model was trained on the NLTK Brown Corpus. This model was key in choosing a better translation as it was able to quantify how similar any given translation was to actual fluent English. In order to utilize this language model to its greatest potential we implemented a modified version of Uniform Cost Search to determine which set of translated words best fit the model. More formally, given a lattice structure where each node represents a possible translation and each outgoing edge represents a possible translation of the next word, Uniform Cost Search begins searching at a set of start nodes (corresponding to the set of possible translations for the first word) for the path to a goal node (corresponding to set of possible translations for the last word) while minimizing the cost function as defined by the language model on any two nodes. The result of combining Uniform Cost Search with our language model is a sentence that, according to the language model, is the best fit for English. This strategy was motivated by the fact that we had many possible translations for any given word, and subsequently the space of possible sentences grew exponentially with the size of the sentence. This strategy affected every sentence in both the development set and test set.

Strategy 2:

Additionally, we utilized a Spanish NLTK HMM Parts-Of-Speech Tagger trained on the NLTK CESS_ESP Corpus to identify areas where the aforementioned linguistic discrepancies may have occurred. From there, we could apply strategies before and after direct word translation to have the word ordering match that of fluent English.

For example, the parts-of-speech tagger was used to identify and reorder a noun adjective pair. Using the same example as above, “casa blanca” would be tagged as “[noun] [adj]” and upon identification of this ordering of tags, “casa blanca” would be reordered into “blanca casa” before translation. The resulting translation would then be “white house” which is more fluent English than “house white”.

Strategy 3:

The other POS-Tagger related strategy described in the above section is switching nouns in “[noun1] de [noun2]” phrases. Anytime the sequence of tags associated with an untranslated string has this ordering, our system runs both the original three word sequence as

well as a flipped sequence through our language model. The most probable sequence is used in the translation. Since nouns are prevalent in every sentence and often have associated adjectives this strategy will affect almost all sentences.

Strategy 4:

In order to address the difference in verb negation, we implemented a post-strategy that used the POS tagger to add the auxiliary verb “does” to help the fluency of the translation. More specifically, given the word “no” followed by a word tagged as a verb, we would switch the word “no” to “does not”.

Strategy 5:

One small adjustment we made to our system that helped the fluency of our translations was to allow some words to have no translation at all, giving them the option to essentially be dropped from the sentence. We derived a set of words by looking at specific cases in our development set, finding that the words “un”, “se”, and “a” often should be left out for a most accurate translation. In order to implement this, when “un”, “se”, or “a” was translated we also added the option for it to be dropped before consulting UCS and our language model. We let UCS determine whether or not the word should be dropped, but this strategy allowed for that possibility.

Strategy 6:

In order to address the cold nature of Spanish we used a few heuristics derived from the development set regarding the proper pronoun insertion. When a verb that was deemed to have an implicit pronoun was preceded by a noun or a capitalized word, we simply did not insert a pronoun. We found that the most natural translation for a verb preceded by a noun was to drop any implicit pronoun. We applied this to capitalized words as well as a heuristic for proper nouns. If an implicit-pronoun verb was not preceded by a noun or capitalized word, we then searched the sentence for gender indicative words. Using the GenderClassifier that was given to use earlier in the course, we determined whether or not there was a male or female in the sentence. If there was a male, we inserted “he” before the verb- if there was a female, we inserted “she”. If both were identified, we inserted “he” as our development set mostly consisted of male subjects. If no gendered word was identified, we dropped the pronoun. This is due to the fact that our development set had mostly human subjects, meaning inserting another pronoun such as “it” would affect the faithfulness of the translation. We found that this helped both our faithfulness and fluency as we were able to express implicit subjects with a decent degree of accuracy. This strategy affected every sentence in both the development set and test set.

Corpus:

Preprocessing Strategies	**Examples**
[noun] [adj] -> [adj] [noun]	“casa blanca” -> “blanca casa”

[noun1] de [noun2] -> [noun2] [noun]	“profesor de historia” -> “historia profesor”
“a”,”un”,”se” -> “”	“se puede caminar” -> “puede caminar”
Postprocessing Strategies	**Examples**
no [verb] -> does not [verb]	“él no grita” -> “he does not yell
Holistic Strategies	
Language Model	
Implicit Subject Identification	“pienso que” -> “I think that”

Note: The highlighted words are those that the strategy affects, but due to inconsistencies in the POS-Tagger, etc., not all these phrases may be modified by the MT system. Additionally, a strategy may be identified yet not applied as pa

1	**Vivas dice que no incita la violencia sino que brinda "consejos de autodefensa" y asegura que no piensa entregarse.**
2	**En Washington un alto funcionario del gobierno describió el arresto como "enorme" y dijo que era "una operación dirigida por la policía mexicana, pero con un fuerte apoyo del gobierno de EE.UU".**
3	**Phil Jordan, que pasó tres décadas con la DEA y dirigió El Paso Intelligence Center de la agencia, dijo que el arresto representa un duro golpe para capo más poderoso del mundo.**
4	**El ambiente en la plaza de la Independencia en Kiev, la capital de Ucrania, era solemne este domingo mientras miles de personas se reunían para recordar a las víctimas de las manifestaciones y en un momento de dudas sobre quién tomará las riendas del país.**
5	**Jason Collins se convirtió en el primer hombre abiertamente homosexual que jugará en la NBA al firmar un contrato con los Nets de Brooklyn este domingo.**
6	**Aparentemente, dos fuerzas políticas intentan tener el control del país: la oposición y Yanukovych, quien dice que aún es el presidente y que está a cargo pese a que viajó a la parte este de Ucrania.**
7	**Las patinadoras artísticas que estuvieron en el hielo completaron dos presentaciones en los Olímpicos de Sochi 2014; un programa corto de dos minutos y 50 segundos y un patinaje libre de entre cuatro y cinco minuto.**
8	**Koum creció en Ucrania, migrando a principios de los años noventa a Estados Unidos cuando era un adolescente.**

9	In 2009, fundó WhatsApp, el servicio que ha acumulado 450 millones de usuarios en poco más de cuatro años y crece su base de usuarios a una tasa de un millón de personas al día.
10	Guzmán, a quien las autoridades dicen que ha eludido la captura durante años, es buscado en Estados Unidos por múltiples cargos federales de narcotráfico y el año pasado fue nombrado el enemigo público número 1 por la Comisión del Crimen de Chicago.
11	Joaquín "Chapo" Guzmán fue detenido en un hotel de Mazatlán, durante la noche del viernes al sábado y que estaba en compañía de una mujer, según confirmó un funcionario de EEUU.
12	Esto ha sido esperado por mucho tiempo y esperamos que ponga un alto a los señalamientos sin sentido de que este gobierno mexicano no está enfocándose en la seguridad y de que Estados Unidos y México no están trabajando bien juntos.
13	Según las autoridades mexicanas Guzmán Loera logró reestablecerse como uno de los narcotraficantes más importantes desde su fuga y desde 2009 la revista Forbes lo incluyó en su lista de las personas más poderosas del mundo.
14	Este domingo se realizó una concentración en respaldo al gobierno del presidente Nicolás Maduro.
15	Durante una conferencia de prensa transmitida por la televisión estatal, Maduro dio marcha atrás diciendo que CNN podía quedarse.

Sources: *sentence number in []

[2, 3, 10, 11, 12, 13] http://cnnespanol.cnn.com/2014/02/22/capturan-al-chapo-guzman/
[15]http://cnnespanol.cnn.com/2014/02/22/maduro-dice-ahora-que-cnn-puede-quedarse-en-venezuela/
[1,14]http://cnnespanol.cnn.com/2014/02/23/las-calles-de-la-venezuela-entre-la-polarizacion-y-el-llamado-a-la-paz/
[9]http://cnnespanol.cnn.com/2014/02/23/el-creador-de-whatsapp-de-humilde-inmigrante-a-multimillonario/
[7]http://cnnespanol.cnn.com/2014/02/23/cuantas-calorias-quemaron-los-atletas-olimpicos-de-sochi/
[5]http://cnnespanol.cnn.com/2014/02/23/jason-collins-se-convierte-en-el-primer-basquetbolista-abiertamente-gay-de-la-nba/
[4,6,8]http://cnnespanol.cnn.com/2014/02/23/quien-manda-en-ucrania-reina-la-incertidumbre-politica-en-un-pais-dividido/

Output

1	Vivas says that does not incite the violence but that provides "council of self-defense" and assures that does not think to surrender.
2	In Washington high official of the government described the arrest as "huge" and said that was "a operation directed by the Mexican, police but with strong support of the government of United States".
3	Phil Jordan, that happened three decades with the Dea and directed The Step Intelligence Center of the agency, said that the arrest represents hard blow for head more powerful of the world.
4	The atmosphere in the square of the Independence in Kiev, the capital of Ukraine, was solemn this Sunday as thousands of people they met for to remember the victims of the manifestations and on moment of doubts about who will have the reins of the country.
5	Jason Collins he became in the first man openly homosexual that he will play in the Nba to the to sign agreement with the Nets of Brooklyn this Sunday.
6	Apparently, two forces political they attempt to have the control of the country: the opposition and Yanukovych, who says that even is the president and that is care despite that travelled a part this with Ukraine.
7	The skaters artistic that they were in the ice they completed two presentations in the Olympic of Sochi 2014; platform short of two minutes and 50 seconds and skating free from among four and five minute.
8	Koum rose in Ukraine, migrating beginnings of the years ninety States United as was teen.
9	At 2009, founded Whatsapp, the service that has accumulated 450 million of users in short most of four years and grows its base of users a rate of million of people to the day.
10	Guzmán, whoever the authorities they say that has avoided the catch during years, is searched on States United by multiple charges federal of drug trafficking and the year last was appointed the enemy public number 1 to the Commission of the Offense of Chicago.
11	Joaquín "Chapo" Guzmán was arrested on hotel of Mazatlán, during the night of the Friday to the Saturday and that was in company of a woman, according to confirmed official of United States.
12	This has been expected as much time and <WE> hope that puts tall them signs without sense of that this government Mexican no is focusing in the security and from that States United and México not they are working well together.
13	By the authorities mexicans Guzmán Loera achieved to recover as one of the

	drug traffickers most important from their escape and from 2009 the journal Forbes the included in his list of the people most powerful of the world.
14	This Sunday performed a concentration on back to the government of the president Nicolás Maduro.
15	During a conference of press transmitted to television the state, Maduro gives march back saying that Cnn could to stay.

Comparison to Google Translate

	Our MT	Google Translate
11	Joaquín "Chapo" Guzmán was arrested on hotel of Mazatlán, during the night of the Friday to the Saturday and that was in company of a woman, according to confirmed official of United States.	Joaquin "Chapo" Guzman was arrested at a hotel in Mazatlan, on the night of Friday to Saturday and was accompanied by a woman, as confirmed by an official of the U.S..
12	This has been expected as much time and <WE> hope that puts tall them signs without sense of that this government Mexican no is focusing in the security and from that States United and México not they are working well together.	This has been long awaited and we hope to put a stop to the nonsense signs that this Mexican government is not focusing on security and that the United States and Mexico are not working well together.
13	By the authorities mexicans Guzmán Loera achieved to recover as one of the drug traffickers most important from their escape and from 2009 the journal Forbes the included in his list of the people most powerful of the world.	According to Mexican authorities succeeded Guzman Loera reestablished as one of the most important drug traffickers and their escape from Forbes magazine since 2009 included it in its list of the most powerful people in the world.
14	This Sunday performed a concentration on back to the government of the president Nicolás Maduro.	This Sunday a concentration was performed in support of the government of President Nicolas Maduro.
15	During a conference of press transmitted to television the state, Maduro gives march back saying that Cnn could to stay.	During a news conference broadcast on state television, Maduro backtracked saying that CNN could stay.

Comparing our translations with those from Google Translate, one can see that our translations lack fluency while upholding a similar degree of faithfulness. Here is an analysis of each translation:

11. Gives best translation: Google. Both translations uphold a faithful meaning of the sentence. However, Google's translation is more fluent English. For example we translated to "on hotel" when "at a hotel" is the more fluent English phrase. Our translations phrasing of "night of the Friday" and "according to confirmed official of…" are more awkward translations. Google translates more fluent versions of these phrases. Our sentence faithfully keeps the accents while Google's does not.
12. Gives best translation: Google. It is a complex sentence that is not handled well by our direct translation system. The very poor fluency causes our translation to have very little meaning. While Google's translation is not great, translating what probably should be "senseless accusations" as "nonsense signs," it still effectively carries the faithful meaning of the sentence.
13. Gives best translation: Google. Although the general meaning of the sentence can be inferred from our translation, its fluency is quite poor. For example "authorities mexicans" should be "mexican authorities". Google also struggles to form a great fluent translation. It falsely attributes "su" as the plural possessive pronoun "their" when it should be referring to Guzman and be "his" and mentions "escape from Forbes magazine" which is incorrect. Our sentence faithfully keeps the accents while Google's does not.
14. Gives best translation: Google. Our direct translation strategy is certainly apparent in this translation. "En respaldo" should be translated to "in support" while our individual translation of each word led to "on back". Although, Google's "a concentration was performed" could more accurately be "a gathering formed", the meaning is still somewhat clear.
15. Gives best translation: Google. Our direct translation certainly obfuscates the meaning of this sentence. The phrase for "backtracked" is directly translated to "gives march back". Google's system captures this phrase as well as the entire translation fluently and faithfully.

Error Analysis:

One of the most interesting and indicative failures of our system was in sentence 15 with the Spanish phrase "dio marcha atrás". The best translation of this phrase, as given by Google Translate, is "backtracked" but our system translated the phrase as "gives march back" which is a more direct translation as individually the words do mean "gave a backwards march". This result is very indicative of the limitations of our entire approach and system. Direct translation systems have difficulty identifying complex phrases and associating meaning with them to give more meaningful and faithful translations, as demoed here. "Gave a backwards march" means nothing to English speakers, but a human translator would be able to decipher meaning from such a Spanish phrase even without explicitly hearing it priorly and deduce the meaning of "backtracked". In order for our system to be able to properly translate a phrase such as this one, statistical translation would be necessary. An Expectation-Maximization algorithm such as IBM Model 1 would be able to associate the phrase with the correct translation, where as a direct

translation system such as ours would have immense difficulty doing so, even using other statistical methods as peripheral tools.

A few simple failures on our part are demoed in sentence 12, the most apparent being the <WE> tag within the translation. The <WE> tag served as an indicator within our dictionary to denote a verb that potentially had an implicit pronoun, but our development set did not include any verbs that would have an implicit <WE> so we never wrote a case to handle it. A better approach would to have written a case-generalizable implicit pronoun insertion technique or to simply have every case enumerated. Another example of a simple failure on our part is the word “alto”. In sentence 12 “alto” serves as the word “stop” but in our translation says “tall”. This is due to the fact that our dictionary entry of direct translations for “alto” only contains “high” and “tall”. This shows a severe yet simple limitation of our system, as it can only translate Spanish words to the English words we selected for it in our hand-created dictionary. A statistical method does not constrain words to such limited definitions, but rather assigns probabilities to translations. Expectation-Maximization Algorithms give non-zero probabilities to all possible alignments in translated sentences, making the pool of words much larger. A statistical method’s larger translation pool would have served to help our translation in this case of “alto”.

Our system is also very dependent upon proper POS tagging, which is a very difficult task. Subsequently, many of the tags are incorrect, and some of our strategies are not executed when expected to, leading to faulty translations. In sentence 4’s translation the phrase “thousands of people they met”. The implicit noun insertion of “they” was not supposed to happen as people is a noun, yet the POS tagger did not indicate as such and therefore “they” was inserted. This could have been addressed via a more robust heuristic or less reliance on the POS tagger.

Overall, our system produced translations that for the most part carried the faithful meaning with okay to poor fluency. There were times however where the fluency was so flawed that the true meaning was hard to determine. A proposition for improvement would to be to use a more robust method, such the mentioned IBM Model 1, that would increase fluency and ultimately success of our machine translator.