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: 27 February 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.

- Spanish: *el hombre corre* and *yo corro* and *los estudiantes corren*
- English: *the man runs* and *I run* and *the students run*

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)

- Spanish: *un pobre hombre* and *un hombre pobre*
- English: *a poor (unfortunate) man* and *a poor (destitute) man*

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).

- Spanish: *Masoud trabaja en el mercado.* and *Trabaja Masoud en el mercado.* and *En el mercado trabaja Masoud.*
- English: *Masoud works in the market.*

4. Spanish uses grammatical gender, whereas English abandoned such a system hundreds of years ago.

•Spanish: *la abuela hermosa* and *el abuelo hermoso* and *los estudiantes hermosos*

•English: *the beautiful grandmother* and *the beautiful grandfather* and *the beautiful students*

5. Spanish heavily relies on the use of the subjunctive mood when describing the past, whereas the use of the subjunctive mood in English hash become less commonplace.

•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)*

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*).

•Spanish: *una abuela* and *una computadora*

•English: *a question* and *an answer*

7. English indefinite articles are marked for whether their noun begins (phonetically) with a vowel, whereas Spanish does not.

•Spanish: *una abuela* and *una computadora*

•English: *a question* and *an answer*

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.

•Spanish: *Ellos comen* but *Ellos son comiendo.**

•English: *They eat** and *They are eating*

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 two separate strategies to improve our statistical machine translation system.

1. We first improved the baseline IBM algorithm by implementing a trigram language model. This was done mostly to improve performance surrounding errors caused by differences in word order between Spanish and English. For a given trigram, we permute the trigram and select the trigram with the highest probability. We implemented this language model in the included script `LanguageModel.py`.

2. We then

III. ERROR ANALYSIS

Infrequent words become 'fawn'. Infrequent inflected forms are also translated as 'fawn'; if we stemmed we'd have to try to unstem the morphology back.

Could use a gazeteer to help with proper nouns.

Baseline is to weight all alignments are equal, though this way you can easily calculate the probability for 'the' and 'la'. One way to improve that is to limit the alignment to maybe a few words before and after, so voy could align to +- 1 precition (only align to I and go and to) in the mapping "yo voy a la tienda" to "I go to the store".

Use stop word filtering but then we have to deal with the fact that the stop word in one doesn't always map to a word in the other language's sentence.

Use POS tag on the document, and align words of the same POS tag (ex: only aligning noun to noun).

Mentioned inflection because 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 way to do that is stemming, but then you'd have to deal with putting the right stem back.

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. Possible method of reducing includes restricting the length of the target sentence and eliminating 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 repeating words to words in another language is abnormally high. As a result, the translations from the baseline contains unrelated spurious words that occurs only because they are repeated several times in many 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 'Military'... Several ways to improve this: (1) make sure that a word does not align to words far from its corresponding index in the other sentence, (2) use stop word filter to filter out abundant words like 'el', 'la' ... and 'the', 'a', ... and 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 it increases the runtime a lot (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, because we consider all alignments equal, the English word with maximum p(F—E) is not always the actual corresponding word, but the word that appears quite frequently next to the actual word. This can be improved by (1) phrase translation: we could translate in phrases rather than words to avoid possible examples like 'San Francisco' being translated as 'Francisco San', and (2) language model: we could evaluate the English words with high p(F—E) and select the combinations based on a language model p(E).

IV. COMPARATIVE ANALYSIS WITH GOOGLE TRANSLATE

1. •Spanish: Es bastante posible que no , pero puede ser que finalmente Occidente pueda presumir de proteger los derechos humanos .

   •Output: is quite possible that not , but can be that finally west can fawn of protect the rights human .

   •Google: It is quite possible that no, but it may be that the West can finally boast of protecting human rights.

2. •Spanish: Ya era la segunda edicin del programa de &quot; La obesidad de no es una casualidad &quot; , que apoya la Compaa de Seguro Sanitario General y la empresa Unilever .

   •Output: already was the second edition the programme of &quot; the obesity of not is a fawn &quot; , that supports the company of sure health general and the company fawn .

   •Google: It was the second edition of the program & quot; Obesity is not a coincidence & quot; Which supports the General Health Insurance Company and Unilever.

3. •Spanish: De juegos de este tipo no cabe esperar que recree deformaciones y colisiones complicadas , pero de hecho antes de un golpe contra cualquier objeto , no podis predecir cmo reaccionar vuestro auto , con lo cual no todo est en orden .

   •Output: of games of this kind not is wait that fawn fawn and collisions complicated , but of fact before of a coup against any subject , not tell predict how fawn fawn prosecution , with what which not all is in order .

   •Google: Games of this type are not expected to recreate complicated deformations and collisions, but in fact before a coup against any object, you can not predict how your car will react, which not everything is in order.

4. •Spanish: Tenis que estar atentos a los surtidores de gasolina al lado de la carretera , porque justamente ah tenis que cambiar de auto , aunque en la gran mayora de los casos os bastar lo que tenis .

- •Output: you that be to to the fawn of petrol the on of the road , because exactly there you that change of prosecution , although in the great majority of the cases you suffice what that you .
- •Google: You must be attentive to the pumps off the road because right there you have to change cars, although in most cases will suffice what you have.

5. •Spanish: La campaa del defensor del consumidor se dirige en concreto a Josef Ackermann , director del Deutsche Bank , aunque los grandes bancos como Goldman Sachs o Morgan Stanley actan de manera parecida .
   - •Output: the campaign the ombudsman the consumer at in particular to fawn chair , director the deutsche bank , although the large banks as sink sachs or unhesitatingly fawn act of way analogous .
   - •Google: The consumer advocate campaign is directed specifically to Josef Ackermann, head of Deutsche Bank, although the big banks like Goldman Sachs or Morgan Stanley act similarly.

V. COMPARATIVE ANALYSIS

[1] Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., & Reitboeck, H. J. (1988). Coherent oscillations: A mechanism of feature linking in the visual cortex?. *Biological cybernetics, 60(2)*, 121-130.

[2] Gray, C. M., & Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proceedings of the National Academy of Sciences, 86(5)*, 1698-1702.

[3] Gray, C. M., Knig, P., Engel, A. K., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature, 338(6213)*, 334-337.

[4] Grossberg, S., & Somers, D. (1991). Synchronized oscillations during cooperative feature linking in a cortical model of visual perception. *Neural Networks, 4(4)*, 453-466.

[5] Kanizsa, G. (1976). Subjective contours. *Scientific American, 234(4)*, 48-52.

[6] Lee, T. S. (2001). Dynamics of subjective contour formation in the early visual cortex. *Proceedings of the National Academy of Sciences, 98(4)*, 1907-1911.

[7] Nieder, A. (2002). Seeing more than meets the eye: processing of illusory contours in animals. *Journal of Comparative Physiology A*, 188(4), 249-260.

[8] Orban, G. A., Dupont, P., De Bruyn, B., Vogels, R., Vandenberghe, R., & Mortelmans, L. (1995). A motion area in human visual cortex. *Proceedings of the National Academy of Sciences, 92(4)*, 993-997.

[9] Samonds, J. M., Zhou, Z., Bernard, M. R., & Bonds, A. B. (2006). Synchronous activity in cat visual cortex encodes collinear and cocircular contours. *Journal of Neurophysiology, 95(4)*, 2602-2616.

[10] Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience, 18(1)*, 555-586.

[11] Somers, D., & Kopell, N. (1993). Rapid synchronization through fast threshold modulation. *Biological cybernetics, 68(5)*, 393-407.

[12] Traub, R. D., Whittington, M. A., Stanford, I. M., & Jefferys, J. G. (1996). A mechanism for generation of long-range synchronous fast oscillations in the cortex. Nature, 383(6601), 621-624.

[13] von der Heydt, R., & Peterhans, E. (1989). Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *The Journal of neuroscience, 9(5)*, 1731-1748.

[14] von der Heydt, R., & Peterhans, E. (1989). Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *The Journal of neuroscience, 9(5)*, 1731-1748.

VI. APPENDIX FOR FIGURES

Figures are listed below by the experiment they are associated with. See Results and Analysis for other results pertaining to these figures.