

Assignment 5

N-grams Machine Translation

DUE: Wednesday, February 18

1. We talked about n-grams in both the spam and spelling units, but n-grams are used in a lot of applications, including **part-of-speech tagging**—i.e. automatically assigning parts of speech to a text. So, now, you’re going to get a chance to explore the notion of an n-gram, while learning a bit about part-of-speech in the process.

Go to <http://pie.usna.edu>. Read this page [it will probably load slowly, as it is under development], to get an idea of what the BNC is and what this webpage does. We’re going to explore the notions of *n-grams* and *phrase frames*, so you’ll want to click on the links where it says, “Via two basic query pages users can explore n-grams [<http://pie.usna.edu/explore.html>] and phrase-frames [<http://pie.usna.edu/explorep.html>].”

I recommend that you play around a bit with both pages before trying the homework questions, to get a feel for what each page does. Just pick some random words and different lengths of n-grams and make sure you understand the results you get. You may want to try the example of *play* in part (b) to make sure you understand what I say there. The page <http://pie.usna.edu/gettingstarted.html> gives explanations on how it works.

Let’s use the word *round*, and for all the questions, keep the default settings (for “minimum frequency,” “maximum frequency” and so on).

- (a) What is the BNC and what is it used for?
- (b) How many different parts of speech does *round* have? What are they? Give the three-letter codes and the official names.
- (c) Find me a trigram of the following form: “POS-tag round POS-tag” where *round* only has one part of speech (POS). For example, if the word were *play* instead of *round*: *play* can be an infinitive verb (VVI), a common noun (NN1), or a finite verb (VVB). However, in “AT0 play PRP” (e.g. *a play on*), *play* is **always** a common noun (NN1)
- (d) On the frames page, look at the 4-grams where *round* is the third word.

- i. Take an example where the asterisk is the second word [pick from one of the three].
 - ii. Provide all the examples of words in the BNC which appear where the asterisk does. Be sure to include POS information, along with the words.
 - iii. Imagine that you are trying to figure out what the part of speech of *round* is. From part (b), we know that it has many parts of speech. In your own words, describe how this frame information is useful.
2. (a) The Japanese sentence in (1a) is most literally translated as the “English” sentence in (1b). Give a more natural translation.
 - (1) a. Sinda node minna ga kanasinda hito wa ymei desita.
 b. The man who everyone was distressed because (he) died was famous.
 - (b) The sentence below was found in an Austrian skiing hotel. How should it actually read?
 - (2) Not to perambulate the corridors in the hours of repose in the boots of ascension.
 3. When translating from English into the Native American language Mam (in Guatemala), a translator reported the following:
 - *ntz?ica* = ‘older sibling’
 - *witzin* = ‘younger sibling’

Both words are used for males and females.

- (a) Draw a Venn diagram showing how the English words *brother* and *sister* overlap with the Mam words *ntz?ica* and *witzin*.
 - (b) You come across the text: *Maxwell is the brother of Santiago*, but it gives no indication of who is older. How would you preserve this age ambiguity in Mam?
4. At altavista’s babelfish
<http://www.babelfish.altavista.com/babelfish/tr>
 you can type in a sentence in English, translate it into Spanish, and then “backtranslate” that sentence into English. The following are a list of such “triads.” The (a) examples show the original English. The (b) show the Spanish, along with annotations indicating English equivalents. The (c) examples show how that Spanish sentence gets backtranslated into English. Read the examples and answer the questions about them.

- A. (3) a. There burst into the room three men with guns.

- b. Estalló en los hombres del sitio tres con
He/she/it exploded in the men of the place three with
los armas.
the arms
- c. It exploded in the men of site three with the arms.
- i. English always has to have what we call a subject. For example, *John speaks Spanish* is a good sentence, whereas *Speaks Spanish* is not. Based on (3b), do you think Spanish always has to have a subject? *Yo hablo* is literally 'I speak'; what is another way to say this in Spanish?
- ii. Is the meaning of *there* in (3a) more like the *it* in (4a) or more like the *there* in (4b)?
- (4) a. It is raining.
b. Put the books there.
- iii. I want to know how this use of *burst* is different from *burst* in *The balloon burst in my hands*. Give (partial) synonyms for both cases – i.e. what word(s) could replace *burst* in (3a) and what word(s) could replace *burst* in *The balloon burst in my hands*?
- iv. People who work at Bank One could be called *Bank One people*. Likewise, *the room three men* gets interpreted as *the men of site three*. What kind of ambiguity is this? Take the original sentence and put brackets around the prepositional phrases.
- v. It is difficult to find an exact translation of *gun* in Spanish. We have: *fusil* = 'rifle', *escopeta* = 'shotgun', *cañon* = 'cannon', *revólver* = 'revolver'. Describe this in terms of hyponymy/hypernymy.
- vi. Generate a "triad" at babelish using Spanish and using the English word *girlfriend*. Describe your results in terms of hyponymy/hypernymy.
- B. (5) a. It's 7:30, and I'm wearing pajamas.
b. Es 7:30, y estoy usando los pijamas.
He/she/it is 7:30 and I am using/wearing the pajamas
c. He is 7:30, and I am using the pajamas.
- i. Briefly describe three problems with this example.
- ii. Does the problem seem to be in going from English to Spanish or in going from Spanish to English for this example?
- iii. *using* is probably a more frequent translation of *usando* than *wearing* is. But we have to wonder how we got *usando* in the first place. Go to www.wordreference.com, and look up both *wear* and *wearing* in the English-to-Spanish dictionary. Give an entry where *usando* (or some related form such as *uso* or *usar*) is listed as the meaning for *wear* or *wearing*.

iv. A better translation is in (6). Give two difficulties an automatic translation into English would face.

(6) Son las 7:30, yo llevo una pijama.
They are the 7:30 and I wear a pajama.
'It is 7:30, and I am wearing pajamas'

v. Go to babelfish (or some other translator) and find an English translation of (6). Is this better or worse than (5c)?

5. Go to: <http://www.tashian.com/multibabel/>

This site generates triads from babelfish automatically. For the first four parts of this question, ignore the Chinese, Japanese, and Korean options.

- A. Come up with an example sentence that you're going to translate and backtranslate and write it down. Be funny; be creative; pick a song lyric or movie quote; whatever. Just make sure that the sentence is sufficiently interesting, so that you are able to answer all of the following questions.
- B. Enter your sentence, and examine all the (English) backtranslations. Write down all the backtranslations and for each backtranslation, give me its score (1-4) on the intelligibility scale (introduced in class).
- C. In terms of quality, pick the best and worst backtranslations. Explain how you arrived at the best and worst – i.e. think about intelligibility, accuracy, error analysis. (For error analysis, think of criteria you can use for determining quality: meaning change, tense change [present, past, future], word choice, missing/added words, word order, “word salad,” etc.)
- D. Take the worst backtranslation and its corresponding translation (German, French, etc.). Using an online dictionary¹, or other sources², give a word-by-word transliteration of the translation – include multiple meanings where relevant. Indicate your source(s).
- E. Now, turn on the Chinese, Japanese, and Korean option. Are these backtranslations generally better or worse than the others? Why do you think that is?

¹www.wordreference.com has dictionaries for a variety of languages

²e.g. paper dictionaries, your own knowledge if you know one of these languages, or native speaker consultation