Linguistics 384
Homework 4

Spelling Correction and Machine Translation

DUE: Wednesday, November 9, 2005

1. (15 points) Go to my online spell checker at

   `http://www.ling.ohio-state.edu/~adriane/384/ispell/`

   and enter the text below. For each of the 5 misspellings (shown in bold), answer the following questions.

   > **Bobb** and his friend Abraham, or "**bae**" for short, were **acberbated bay** their other friend Arbuckle's **ccat**.

   (a) How would you classify this spelling mistake?

   (b) Was the misspelling caught by the spell checker?

   (c) How many potential corrections does it give?

   (d) Is the correct spelling listed among the options?

2. (10 points) Re-read slide 31 in the lecture notes on spelling correction. We discussed a few rules which can be used to correct certain misspellings. In addition to the ones given on this page, think of two other useful rules based on your knowledge of English spelling.

3. (20 points) Create a dictionary entry for an English to Hungarian MT dictionary for each of the following two words: *book* (noun) and *exhale* (verb). To receive full credit, it is crucial that your entries contain adequate information to rule out the possibility for *book* to be the subject of *exhale* (since books do not exhale). You may use the examples on the lecture notes and discussed in class as a basis for this, but feel free to add features wherever necessary. You can use http://dict.sztaki.hu/ or any other online/paper dictionary for the Hungarian translation.

4. (20 points) Go to: http://babelfish.altavista.com

   This site allows you to type in text and translate it into another language. Notice that you can also *backtranslate*, i.e., translate back into the original language, by copying and pasting. To answer the following questions, you will need to play with the MT system with a few interesting examples and use them to defend your answers.

(a) Describe the features of and differences between transformer systems and transfer systems.

(b) Based on your discussion and your experience with the system, do you think this site uses a transformer system or a transfer system? What facts about the system made you think so? Be sure to give example sentences along with their backtranslations to defend your answer.

5. (a) (20 points)
In (1), (2), and (3) below, align the words in the English (a) examples with the words in the Hungarian (b) examples. Note that several English words may correspond with one Hungarian word (many-to-one), one English word may correspond with several Hungarian words (one-to-many), and some English words may correspond with no Hungarian word at all (one-to-null). I have provided a word-by-word translation underneath the Hungarian (b) examples—this is just to let you know what each Hungarian word roughly means.

(1) a. That cat is friendly.
b. Az a macska barátságos.
that the cat friendly

(2) a. I have no money.
b. Nekem nincs pénzem.
to me is not money

(3) a. I think that Peter is going by train.
b. Én azt hiszem, hogy Péter vonattal megy.
I that think that Peter by train go

(b) Now pick one English word that can be translated into at least two different Hungarian words based on your alignments. Describe how you would derive probabilities of translating this word into each of the candidate Hungarian words from the alignments.

(c) If you didn't have word alignments, you could use a bag of words model. For the same word you picked, how would the candidate Hungarian words and their associated probabilities differ from those in part (b)?

(d) The bag of words model, of course, gets better as it sees more data. Describe how the following extra sentences may help you translate certain words better if you're using a bag of words model. Which words get easier to translate and why? Illustrate with at least one specific English word.

(4) a. Peter saw the cat.
b. Péter látta a macskát.
Peter saw the cat

(5) a. I believe that this book is interesting.

b. Én azt gondolom, hogy ez a könyv érdekes.
I that believe that this the book interesting

6. (15 points) The following tree was taken from the Wall Street Journal corpus (with some modifications). Given this tree, give me the phrase structure rules. Here are two rules to get you started:

S → NP VP
NNP → Vinken