Linguistics 384
Homework 4

Machine Translation

DUE: Wednesday, February 22, 2006

Problem Solving (to prepare for the quiz on February 22, 2006)

1. Create a dictionary entry for an English to Hungarian MT dictionary for each of the following two words: *book* (noun) and *exhale* (verb).

   Here are Hungarian translations you can use in your entries:

   - *book – könyv*
   - *exhale – kilehel*

   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 from the class notes (or handout) as a basis for this, but feel free to add features wherever necessary. (Hint: nouns have features like HUMAN or CONCRETE. What kinds of features should verbs have? Remember to think about the fact that verbs often have subjects and objects that are nouns. Part of the information you need about a verb relates to what kinds of subjects and objects it can have.)

2. I'm trying to design an MT system so I can translate between four different languages (Portuguese, Quechua, Romanian, and Slovenian), and I want to know how many *transfer* components I'm going to have to build if I: a) use a transformer system, and b) use an interlingua. Portuguese → Interlingua counts as one component; Interlingua → Portugese is another component. Please show your work and explain your answer.

3. (a) 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

Essay (to hand in on February 22, 2006)

1. When translating from English into the Native American language Mam (spoken in Guatemala), a translator reported the following terms used among siblings:

- *ntz?ica* = 'older sibling'
- *witzin* = 'younger sibling'

Both words are used for males and females.

(a) In terms of hyponymy/hypernymy, describe the relationship between the English word *sibling* and these words. (For *ntz?ica*, you would fill in the following: *sibling* is a ______ of *ntz?ica*.)

(b) Draw a Venn diagram (see page 15 in Handout 5) showing how the English words *brother* and *sister* overlap with the Mam words *ntz?ica* and *witzin*.

(c) You come across the text: *Maxwell is the brother of Santiago*, but it gives no indication of who is older. If you were forced to translate this sentence into Mam and you had to preserve this age ambiguity, how would you do it? Describe in English what you would write in Mam, knowing that *ntz?ica* and *witzin* are the only two words in Mam that you can use to describe sibling relationships.

2. System evaluation exercise:

   Go to: `http://www.tashian.com/multibabel/`

   (This website is sometimes a little slow. Please be patient while waiting for the translation page to load.)

   For the first three 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 (there are 5 languages, so make sure you give me all 5 backtranslations), give me its score (1-4) on the intelligibility scale (given on page 18 of Handout 5).

   (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) 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?