Natural Language Processing

In-class Word Alignment Exercise

Anoop Sarkar
http://anoopsarkar.github.io/nlp-class

(1) **Human Translation**

NASA's latest mission to Mars has found some strange tablets. One tablet seems to be a kind of Rosetta stone which has translations from a language we will call MARTIAN-A (sentences 1a to 12a below) to another language we will call MARTIAN-B (sentences 1b to 12b below). The ASCII transcription of the alien script on the Rosetta tablet is given below:

```
1a. ok'sifar zvau hu .

1b. at'sifar somuds geyu .

2a. ok'anko ok'sifar myi pell hu .

2b. at'anko at'sifar ashi erder geyu .

3a. oprashyo hu qebb yuzvo oxloyzo .

3b. diza geyu isvat iwla pown .

4a. ok'sifar myi rig bzayr zu .

4b. at'sifar keerat ashi parq up .

5a. yux druh qebb stovokor .

5b. diza viodaws pai shun .

6a. ked hu qebb zu stovokor .

6b. dimbe geyu keerat pai shun .

7a. ked druh zvau ked hu qebb pnah .

7b. dimbe viodaws somuds dimbe geyu iwla woq .

8a. ked bzayr myi pell eoq .

8b. gakh up ashi erder kvig .

9a. yux eoq qebb zada ok'nefos .

9b. diza kvig pai goli at'nefos .

10a. ked amn eoq kin oxloyzo hom .

10b. dimbe kvig baz iluh ejuo pown .

11a. ked eoq tazih yuzvo kin dabal'ok .

11b. dimbe kvig isvat iluh dabal'at .

12a. ked mina eoq qebb yuzvo amn .

12b. dimbe kvig zeg isvat iwla baz .
```

We would like to create a translation from the source language which we will take to be MARTIAN-B and produce output in the target language which will be MARTIAN-A. Due to severe budget cutbacks at NASA, decryption of these tablets has fallen to people like you. In this question, you should try to solve this task by hand to get some insight into the process of translation.

a. Use the above translations to produce a translation dictionary. For each word in MARTIAN-A provide an equivalent word in MARTIAN-B. If a word in MARTIAN-A has no equivalent in MARTIAN-B then put the entry "(none)" in the dictionary.

b. Using your translation dictionary, provide a word for word translation for the following MARTIAN-B sentences on a new tablet which was found near the Rosetta tablet.

```
13b. gakh up ashi woq pown goli at'nefos .

14b. diza kvig zeg isvat iluh ejuo .

15b. dimbe geyu pai shun hunslob at'anko .
```

The MARTIAN-A sentences you produce will probably appear to be in a different word order from the MARTIAN-A sentences you observed on the Rosetta tablet. Some words might be unseen and so seemingly untranslatable. In those cases insert the word `?` for the unseen word.

c. The word for word translation can be improved with additional knowledge about MARTIAN-A word order. Luckily another tablet containing some MARTIAN-A sentences (untranslated) was found on the dusty plains of Mars. Use these MARTIAN-A sentences in order to find the most plausible word order for the MARTIAN-A sentences translated from MARTIAN-B sentences in (1b).

```
ok'anko myi oxloyzo druh .
yux mina eoq esky oxloyzo pnah .
ok'anko yolk stovokor koos oprashyo pnah zada ok'nefos yun zu kin hom .
ked hom qebb koos ok'anko .
ok'sifar zvau hu .
ok'anko ok'sifar
myi pell hu .
oprashyo hu qebb yuzvo oxloyzo .
ok'sifar myi rig bzayr zu .
yux druh qebb stovokor .
ked hu qebb zu stovokor .
ked bzayr myi pell eoq .
ked druh zvau ked hu qebb pnah .
yux eoq qebb zada ok'nefos .
ked amn eoq kin oxloyzo hom .
ked eoq tazih yuzvo kin dabal'ok .
ked mina eoq qebb yuzvo amn .
```

Using this additional MARTIAN-A text you can even find a translation for words that are missing from the translation dictionary (although this might be hard to implement in a program, cases that were previously translated as `?` can be translated by manual inspection of the above MARTIAN-A text).

(2) The following is a small parallel text (the same text in two different languages). The 1st column contains phrases in Udihe. The 2nd column contains the English equivalent.

bˈata zäːŋini	the boy's money
si bogdoloi	thy shoulder
jaː xabani	the cow's udder
su zäːŋiu	your money
dili tekpuni	the skin of the head
si jaːŋiː	thy cow
bi moːŋiː	my tree
aziga bugdini	the girl's leg
bi nakta diliŋiː	my boar head
nakta igini	the boar's tail
si bˈataŋiː bogdoloni	thy son's shoulder
teŋku bugdini	the leg of the stool
su jaː woːŋiu	your cow thigh
bi woːi	my thigh

ŋ, ˈ are consonants, ä is a vowel. The ː indicates length of preceding vowel (so for example i+i is written as iː). The archaic English *thy* is used to indicate singular and *your* is used to indicate plural.

(3) Translate into English:

a. su bˈataŋiu zäːŋini

b. si teŋku bugdiŋiː

c. si teŋkuŋiː bugdini

(4) Translate into Udihe:

a. the boy's thigh

b. our boar

c. my daughter's tree

Udihe speakers mostly live in the Siberian far east, and the language is classified as belonging to the Tungus-Manchu language family. There are roughly 100 people who still speak this language. The language is almost extinct. Other than the parallel text given above, you do not need any knowledge about the language and its speakers to answer the questions, but if you are curious, here are some web pages on the Udihe language:

`http://www.ethnologue.com/show_language.asp?code=ude`
`http://en.wikipedia.org/wiki/Udege_language`

Thanks to B. Iomdin who originally created the parallel text and the concept behind the question for an international olympiad in computational linguistics. The question has been somewhat simplified to make the computational aspect of the translation more explicit.