NLP
Introduction to NLP

*Machine Translation*
Multilingual Users

- Content languages for websites
  - Percentage of Internet users by language

http://en.wikipedia.org/wiki/Global_Internet_usage
The Rosetta Stone

Carved in 196 BC in Egypt
Deciphered by Champollion in 1822
Mixture of Egyptian (hieroglyphs and Demotic) and Greek

http://www.ancientegypt.co.uk/writing/rosetta.html
English-Cebuano Bible Example

In the beginning God created the heaven and the earth.
Sa sinugdan gibuhat sa Dios ang mga langit ug ang yuta.

And God called the firmament Heaven.
Ug gihinganlan sa Dios ang hawan nga Langit.

And God called the dry land Earth
Ug ang mamala nga dapit gihinganlan sa Dios nga Yuta

• use: co–occurrence, word order, cognates
• corpora are needed
• sentence alignment needs to be done first

http://en.wikipedia.org/wiki/Bible_translations_by_language
NACLO Problem

• http://nacloweb.org/resources/problems/2012/N2012-C.pdf
• http://nacloweb.org/resources/problems/2012/N2012-CS.pdf
• Problem by Simon Zwarts, based on work by Kevin Knight
It's hard enough to translate between languages when you understand both languages. It's harder still when you only understand one. But what do computers do? They don't truly understand either language. To illustrate the challenge that computers face, Kevin Knight posed this classic puzzle (Knight 1997): given two equivalent texts in two unknown alien languages, how would you go about translating one to another?

It is the year 2354 AD. Our scientists have been eavesdropping on messages between two alien civilizations for a very long time, but we have never met either. The closest aliens, the Centauri, have finally begun to communicate with us. Their first message was a message of peace, “Farok crrok hihok yorok clok kantok ok -yurp.”

Now, we know that the Centauri have been in contact for some time with the Arcturan race, who live in another solar system. We have never had contact with the Arcturans, but newly developed technology makes it possible for us to send them a message. We would like to send them, first, a message of peace, but because we do not understand their language, this is not an easy task.

Luckily, we have intercepted communications from the Centauri that include both languages. Here are 12 sentences in Centauri and their 12 translations in Arcturan. Unfortunately, because we have only been eavesdropping, their meaning is unknown. However, we do know that the sentence pairs on each line are translations of each other. We want to use this information to translate the original peace message from the Centauri and then send this to the Arcturans. Your assignment will be to do this translation.
Arcturan Problem – 2/4

CENTAURI
ok-voon ororok sprok.
ok-drubel ok-voon anok plok sprok.
erok sprok izok hihok ghiero.
ok-voon anok drok brok jok.
wiwok farok izok stok.
lalok sprok izok jok stok.
lalok farok ororok lalok sprok izok enemok.
lalok brok anok plok nok.
wiwok nok izok kantok ok-yurp.
lalok mok nok yorok ghiero clok.
lalok nok crrook hihok yorok zanzanok.
lalok rarok nok izok hihok mok.

ARCTURAN
at-voon bichat dat.
at-drubel at-voon pippat rrat dat.
totat dat arrat vat hilat.
at-voon krat pippat sat lat.
totat jat quat cat.
wat dat krat quat cat.
wat jat bichat wat dat vat eneat.
iat lat pippat rrat nnat.
totat nnat quat oloat at-yurp.
wat nnat gat mat bat hilat.
wat nnat arrat mat zanzanat.
wat nnat forat arrat vat gat.
Arcturan Problem – 3/4

C-1 Let’s start with the first Centauri word: “farok”. This word occurs in two of our Centauri sentences. Given that these sentences’ Arcturan translations only have one word in common with each other, we can assume that this word is the translation for “farok”. Which word it is?

farok

C-2 Do the same thing for “hihok” and “yorok”. For “yorok” you will need to make some assumptions about word ordering.

hihok
yorok

C-3 The Centauri word “clok” only occurs once. However, you can figure out its Arcturan translation in another way.

clok

C-4 Try to use the processes from the previous assignments to complete as much as possible of the following table.

crrrok
kantok
ok-yurp
C-5 Complete the translation of “farok crrok hihok yorok clok kantok ok-yurp.” Keep in mind that Centauri and Arcturan sentences can have a different word order. There may be more than one correct reply.

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C-6 After some years a reply message is received in Arcturan. It reads, “Totat nnat forat arrat mat bat.” Translate this message into Centauri. There may be more than one correct reply.

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The questions in this assignment are based on examples in Knight (1997). In fact, both Centauri and Arcturan have underlying real world languages, as it turns out Centauri is English and Arcturan is Spanish. The languages are obfuscated to Centauri and Arcturan in order to illustrate how a Statistical Machine Translation (SMT) system must start from scratch, since it has no prior knowledge of how the languages work.

**CENTAURI**

Ok-voon ororok sprok.
Garcia and associates.

Ok-drubel ok-voon anok plok sprok.
Carlos Garcia has three associates.

Erok sprok izok hihok ghirok.
His associates are not strong.

Ok-voon anok drok brok jok.
Garcia has a company also.

Wiwok farok izok stok.
Its clients are angry.

Lalok sprok izok jok stok.
The associates are also angry.

**ARCTURAN**

At-voon bichat dat.
Garcia y asociados.

At-drubel at-voon pippat rrat dat.
Carlos Garcia tiene tres asociados.

Totat dat arrat vat hilat.
Sus asociados no son fuertes.

At-voon krat pippat sat lat.
Garcia también tiene una empresa.

Totat jat quart cat.
Sus clientes están enfadados.

Wat dat krat quart cat.
Los asociados también están enfadados.
Arcturan Solution – 2/3

Lalok farok ororok lalok sprok izok enemok. The clients and the associates are enemies.

Lalok brok anok plok nok. The company has three groups.

Wiwok nok izok kantok ok-yurp. Its groups are in Europe.

Lalok mok nok yorok ghirok clok. The modern groups sell strong pharmaceuticals.

Lalok nok crrrok hihok yorok zanzanok. The groups do not sell zanzanine.

Lalok rarok nok izok hihok mok. The small groups are not modern.

Wat jiat bichat wat dat vat eneat. Los clientes y los asociados son enemigos.

Lat lat pippat rrat nnat. La empresa tiene tres grupos.

Totat nnatquat oloat at-yurp. Sus grupos están en Europa.

Wat nnat gat mat bat hilat. Los grupos modernos venden medicinas fuertes.

Wat nnat arrat mat zanzanat. Los grupos no venden zanzania.

Wat nnat forat arrat vat gat. Los grupos pequeños no son modernos.
The novel sentence which was offered for translation in English is: "clients do not sell pharmaceuticals in Europe."

Answers

C-1 jjat

C-2 hihok = arrat, yorok = mat

C-3 We need to use the process of elimination, when mapping all the words between the two sentences two words are unaligned, we assume these are translations of each other. Thus, clok = bat.

C-4 Here are the new matches:

<table>
<thead>
<tr>
<th>crrok</th>
<th>(empty)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kantok</td>
<td>oloat</td>
</tr>
<tr>
<td>ok-yurp</td>
<td>at-yurp</td>
</tr>
</tbody>
</table>

"crrok" does not seem to have a Arcturan equivalent, like in English the word "do" is not translated in "do not sell" which simply becomes "not sells" in Spanish. (Or to put it another way, the Centauri word crrok has a translation, but it's the "empty" word.)

C-5 jjat arrat mat bat oloat at-yurp

Since you cannot deduce with certainty the exact order of the Arcturan sentence, various orders of these words will be accepted.

C-6 Immediately, you are faced with a dilemma: should you translate totat as erok or wiwok? Because wiwok occurs more frequently and because you’ve never seen erok followed by any of the other words you’re considering, wiwok seems more likely. (However, admittedly, this is only a best guess, and erok will also be accepted.) Next, you consider various word orders. There appears to be no grammatical path through these words. Suddenly, you remember that curious Centauri word crrok, which had no translation. Crrok, however, turns out to be a natural bridge between nok and hihok, giving you the translation:

wiwok rarok nok crrok hihok yorok clok.
Parallel Corpora

- The Rosetta Stone
- The Hansards Corpus
- The Bible
42nd PARLIAMENT, 1st SESSION
EDITED HANSARD • NUMBER 095

CONTENTS

Friday, October 21, 2016

House of Commons Debates

VOLUME 148 • NUMBER 095 • 1st SESSION • 42nd PARLIAMENT

OFFICIAL REPORT (HANSARD)

Friday, October 21, 2016
Speaker: The Honourable Geoff Regan

The House met at 10 a.m.

Prayer

GOVERNMENT ORDERS
Hansards Example

- **English**
  
  - <s id=960001> I would like the government and the Postmaster General to agree that we place the union and the Postmaster General under trusteeship so that we can look at his books and records, including those of his management people and all the memos he has received from them, some of which must have shocked him rigid.
  
  - <s id=960002> If the minister would like to propose that, I for one would be prepared to support him.

- **French**
  
  - <s id=960001> Je voudrais que le gouvernement et le ministre des Postes conviennent de placer le syndicat et le ministre des Postes sous tutelle afin que nous puissions examiner ses livres et ses dossiers, y compris ceux de ses collaborateurs, et tous les mémoires qu'il a reçus d'eux, dont certains l'ont sidéré.
  
  - <s id=960002> Si le ministre voulait proposer cela, je serais pour ma part disposé à l'appuyer.
Language Differences (1/3)

[Example from Jurafsky and Martin]
Language Differences (2/3)

• Word order
  – SVO: English, Mandarin
  – VSO: Irish, Classical Arabic
  – SOV: Hindi, Japanese

• Prepositions (Jap.)
  – to Mariko, Mariko-ni

• Inflection (Sp.)
  – have: tengo, tienes, tenemos, tienen, tener

• Lexical distinctions (Sp.):
  – the bottle floated out – la botella salió flotando

• Brother (Jap.)
  – ootoo (younger), oniisan (older)

• They (Fr.)
  – ellas (feminine), ils (masculine)
Language Differences (3/3)

• Word order in phrases (Fr.)
  – la maison bleue, the blue house

• Word order in sentences (Jap.)
  – I like to drink coffee
  – watashi wa kohii o nomu no ga suki desu
  – I–subj coffee–obj drink–dat–rheme like

• vocabulary (Sp.)
  – wall
  – pared, muro

• phrases (Fr.)
  – play
  – pièce de théâtre
NLP