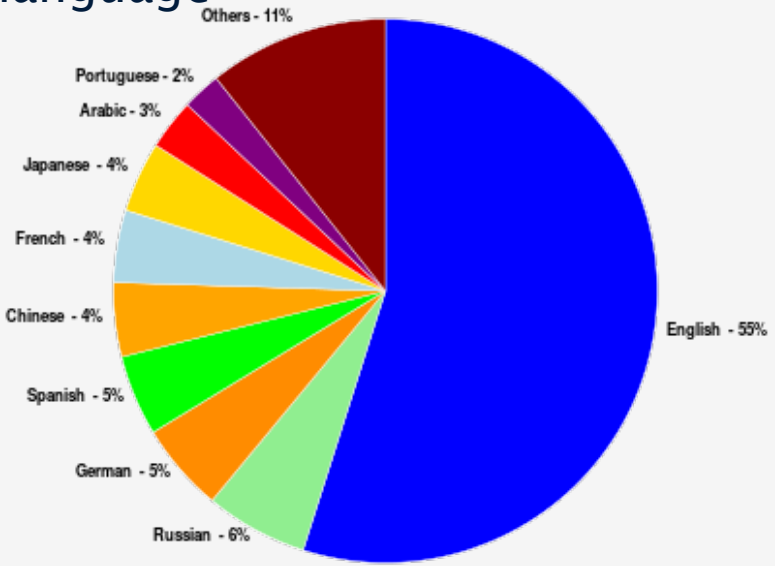
NLP

Introduction to NLP

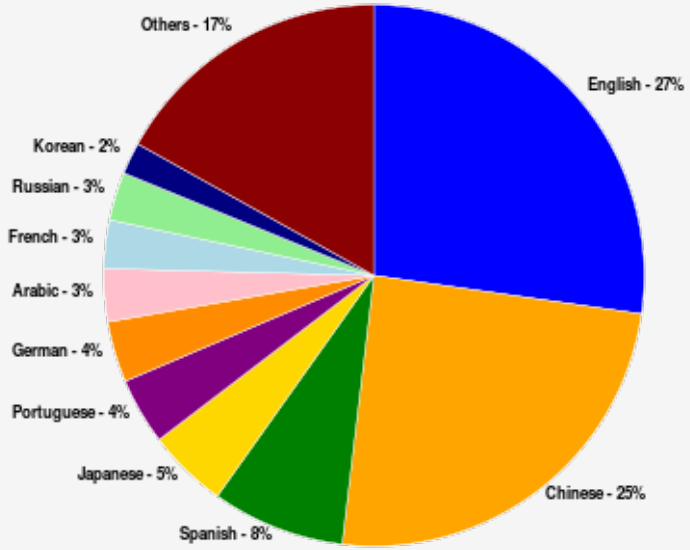
Machine Translation

Multilingual Users

- Content languages for websites language



Percentage of Internet users by

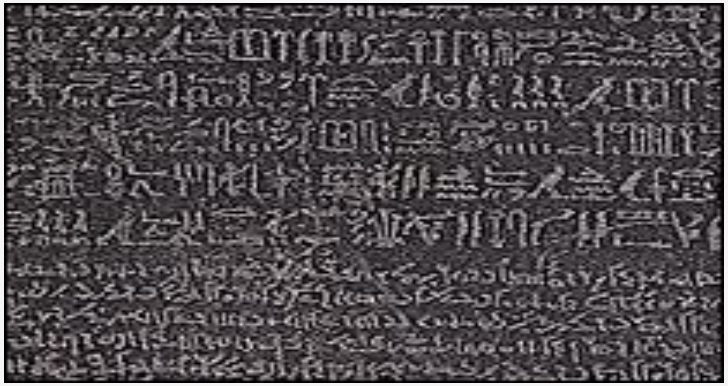


April 2013

http://en.wikipedia.org/wiki/Global_Internet_usage

[The Tower of Babel, by Pieter Bruegel the Elder, 1563]

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

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

Arcturan Problem – 1/4

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	ARCTURAN
ok-voon ororok sprok.	at-voon bichat dat.
ok-drubel ok-voon anok plok sprok.	at-drubel at-voon pippat rrat dat.
erok sprok izok hihok ghirok.	totat dat arrat vat hilat.
ok-voon anok drok brok jok.	at-voon krat pippat sat lat.
wiwok farok izok stok.	totat jjat quat cat.
lalok sprok izok jok stok.	wat dat krat quat cat.
lalok farok ororok lalok sprok izok enemok.	wat jjat bichat wat dat vat eneat.
lalok brok anok plok nok.	iat lat pippat rrat nnat.
wiwok nok izok kantok ok-yurp.	totat nnat quat oloat at-yurp.
lalok mok nok yorok ghirok clok.	wat nnat gat mat bat hilat.
lalok nok crrok hihok yorok zanzanok.	wat nnat arrat mat zanzanat.
lalok rarok nok izok hihok mok.	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.

crrok

kantok

ok-yurp

Arcturan Problem – 4/4

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.

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.

Arcturan Solution – 1/3

C-1 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	ARCTURAN
Ok-voon ororok sprok. Garcia and associates.	At-voon bichat dat. Garcia y asociados.
Ok-drubel ok-voon anok plok sprok. Carlos Garcia has three associates.	At-drubel at-voon pippat rrat dat. Carlos Garcia tiene tres asociados.
Erok sprok izok hihok ghirok. His associates are not strong.	Totat dat arrat vat hilat. Sus asociados no son fuertes.
Ok-voon anok drok brok jok. Garcia has a company also.	At-voon krat pippat sat lat. Garcia tambien tiene una empresa.
Wiwok farok izok stok. Its clients are angry.	Totat jjat quat cat. Sus clientes están enfadodos.
Lalok sprok izok jok stok. The associates are also angry.	Wat dat krat quat cat. Los asociados tambien están enfadados.

Arcturan Solution – 2/3

Lalok farok ororok lalok sprok izok enemok. The clients and the associates are enemies.	Wat jjat bichat wat dat vat eneat. Los clientes y los asociados son enemigos.
Lalok brok anok plok nok. The company has three groups.	Iat lat pippat rrat nnat. La empresa tiene tres grupos.
Wiwok nok izok kantok ok-yurp. Its groups are in Europe.	Totat nnat quat oloat at-yurp. Sus grupos están en Europa.
Lalok mok nok yorok ghirok clok. The modern groups sell strong pharmaceuticals.	Wat nnat gat mat bat hilat. Los grupos modernos venden medicinas fuertes.
Lalok nok crrrok hihok yorok zanzanok. The groups do not sell zanzanine.	Wat nnat arrat mat zanzanat. Los grupos no venden zanzania.
Lalok rarok nok izok hihok mok. The small groups are not modern.	Wat nnat forat arrat vat gat. Los grupos pequeños no son modernos.

Arcturan Solution – 3/3

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:

crrok	(empty)
kantok	oloat
ok-yurp	at-yurp

"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 *crrrok* **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 *crrrok*, which had no translation. *Crrok*, however, turns out to be a natural bridge between *nok* and *hihok*, giving you the translation:

wiwok rarok nok crrrok hihok yorok clok.

Parallel Corpora

- The Rosetta Stone
- The Hansards Corpus
- The Bible

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.

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-D.pdf
- http://nacloweb.org/resources/problems/2012/N2012-DS.pdf
- Problem by Dragomir Radev

Many languages are related to each other for historical reasons. They may have a common ancestor or they may have borrowed words from each other. Linguists group languages into families and branches, based on their common ancestry.
Here is a list of translations of the first article of the Universal Declaration of Human Rights in 17 languages:

Your task it to identify similarities among these languages and group them into seven clusters (groups) of related languages as sketched in the diagram below:

Here is a list of translations of the first article of the Universal Declaration of Human Rights in 17 languages:

A. (English) All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

B. (Latin) Omnes homines dignitate et iure liberi et pares nascuntur, rationis et conscientiae participes sunt, quibus inter se concordiae studio est agendum.

C. Vsi ljudje se rodijo svobodni in imajo enako dostojanstvo in enake pravice. Obdarjeni so z razumom in vestjo in bi morali ravnati drug z drugim kakor bratje.

D. Dieub ha par en o dellezegezh hag o gwirioù eo ganet an holl dud. Poell ha skiant zo dezho ha dleout a reont bevañ an eil gant egile en ur spered a genvreudeuriezh.

E. Tuots umans naschan libers ed eguals in dignità e drets. Els sun dotats cun intellet e conscienza e dessan agir tanter per in uin spiert da fraternità.

F. Toate ființele umane se nasc libere și egale în demnitate și în drepturi. Ele sunt înzestrate cu rațiune și conștiință și trebuie să se comporte unii față de altele în spiritul fraternității.

G. Genir pawb yn rhydd ac yn gydradd â'i gilydd mewn urddas a hawliau. Fe'u cynysgaeddir â rheswm a chydwybod, a dylai pawb ymddwyn y naill at y llall mewn ysbryd cymodlon.

H. Visi žmonės gimsta laisvi ir lygūs savo orumu ir teisėmis. Jiems suteiktas protas ir sąžinė ir jie turi elgtis vienas kito atžvilgiu kaip broliai.

I. Totu sos èsseres umanos naschint lìberos e eguales in dinnidade e in deretos. Issos tenent sa resone e sa cussèntzia e depent operare s'unu cun s'àteru cun ispìritu de fraternidade.

J. Gizon-emakume guztiak aske jaiotzen dira, duintasun eta eskubide berberak dituztela; eta ezaguera eta kontzientzia dutenez gero, elkarren artean senide legez jokatu beharra dute.

K. Kai rahvas roittahes vällinny da taza-arvozinnu omas arvos da oigevuksis. Jogahizele heis on annettu mieli da omatundo da heil vältämättäh pidäy olla keskenäh, kui vellil.

L. Všetci ľudia sa rodia slobodní a sebe rovní , čo sa týka ich dostôjnosti a práv. Sú obdarení rozumom a majú navzájom jednať v bratskom duchu.

M. Nascinu tutti l'omi libari è pari di dignità è di diritti. Pussedinu a raghjoni è a cuscenza è li tocca ad agis-cia trà elli di modu fraternu.

N. Saoláitear na daoine uile saor agus comhionann ina ndínit agus ina gcearta. Tá bauidh an réasúin agus an choinsiasa acu agus dlíd iad féin d'iompar de mheon bhrthreachais i leith a chéile.

O. Visi cilvēki piedzimst brīvi un vienlīdzīgi savā pašcieņā un tiesībās. Viņi ir apveltīti ar saprātu un sirdsapziņu, un viņiem jāizturas citam pret citu brālības garā.

P. Kaikki ihmiset syntyvät vapaina ja tasavertaisina arvoltaan ja oikeuksiltaan. Heille on annettu järki ja omatunto, ja heidän on toimittava toisiaan kohtaan veljeyden hengessä.

Q. Wszyscy ludzie rodzą się wolni i równi pod względem swej godności i swych praw. Są oni obdarzeni rozumem i sumieniem i powinni postępować wobec innych w duchu braterstwa.

Solution

1.	CLQ	Slavic
2.	BEFIM	Romance
3.	J	Basque
4.	HO	Baltic
5.	DGN	Celtic
6.	KP	Finno-Ugric
7.	A	English

NLP