

General motivation

- Measure the importance of the “language-dependent” part of a system
 - as a criterion for evaluation
 - as a case study for the alleged multilinguality of programs (what is the real gain/loss of adapting systems designed for other language(pair)s?)

An evaluation of the Translation
Corpus Aligner
with special reference to the language
pair English-Portuguese

Diana Santos & Signe Oksefjell

Presentation structure

- Goals of the study
- Description of the study
- What was learned from the study

The context

- The ENPC (the English-Norwegian Parallel Corpus) was created in a project led by Stig Johansson and Knut Hofland (Univ. of Oslo and Bergen) <http://iba.hf.uio.no/prosjekt/>
- The ENPC has been extended to other language pairs, including Portuguese
- The Translation Corpus Aligner (TCA) was developed by K.H. for use in the project

The TCA

- Uses several kinds of information
- Expects human revision (in the cases of sentence alignment different from 1:1)
 - Builds 10x10 matrices
- Employs a word anchor list
 - source language items - target language items
- Has been customized for several language pairs

Particular motivation

- Investigate in detail the particular anchor word list used in the TCA, with a view to
 - improving it
 - gathering information for contrastive tasks
 - automatic creation of anchor word lists
 - word alignment
 - bilingual terminology work
- Look thoroughly at Portuguese language technology in the context of the Computational Processing of Portuguese project

Why?

- Expectations are that patterns occurring for items selected as good translation indicators may
 - at least indicate an upper limit of optimal translation behaviour
 - help generalizing different kinds of translation behaviour asking for different kinds of interpreting reasoning (and consequently processing)

How?

- To evaluate the import of a component if one has **no access to the source code?**
(knowing that many things are taken into consideration to produce the final output)

Many issues related to language-dependence cannot be measured: punctuation, proper names, cognates, sentence size...

- There is an interesting part which can be evaluated: **the anchor list**

Forms of evaluation

1. Estimate the revision and correction involved for a given set of E-P texts
2. Estimate the contribution of the anchor word list to the above process
3. Look at the content of the anchor word list, for each entry and for each pair
 - Experiment with the contents of the anchor list
 - Use other texts, other language pairs

The data

- 16 English-Portuguese “texts”
 - 15 English different texts (14 authors)
 - 6 translations into Brazilian Portuguese
 - 10 translations into European Portuguese
- Final reviewed alignment available

Example of I/O of the TCA

English original (extract from Doris Lessing's <i>The Good Terrorist</i>)	Portuguese translation by Bernardette Pinto Leite
<p><s>She faced him, undefiant but confident, and said, "I wonder if they will accept us?"</s> <s>And, as she had known he would, he said, "It is a question of whether we will accept them."</s></p> <p><s>She had withstood the test on her, that bony pain, and he let her wrist go and went on to the door.</s></p></p>	<p><p><s>Ela encarou-o, sem desafio mas confiante, e perguntou:</s></p> <p><s>&mdash; Achas que nos aceitam?</s> </p> <p><s>&mdash; E, conforme sabia que Jasper responderia, este retorquiu:</s></p> <p><s>&mdash; É tudo uma questão de nós os aceitarmos a eles.</s></p> <p><s>Alice resistira ao teste sobre a sua pessoa, à dor óssea, e ele largou-lhe o pulso e dirigiu-se para a porta.</s></p></p>

<s id=DL2.1.s15 corresp='DL2TP.1.s15 DL2TP.1.s16'>She faced him, undefiant but confident, and said, "I wonder if they will accept us?"</s> <s id=DL2.1.s16 corresp='DL2TP.1.s17 DL2TP.1.s18'>And, as she had known he would, he said, "It is a question of whether we will accept them."</s></p> <s id=DL2.1.s16 corresp='DL2TP.1.s17 DL2TP.1.s18'>And, as she had known he would, he said, "It is a question of whether we will accept them."</s></p> <p id=DL2.1.p3> <s id=DL2.1.s17 corresp=DL2TP.1.s19>She had withstood the test on her, that bony pain, and he let her wrist go and went on to the door.</s>

<s id=DL2TP.1.s15 corresp=DL2.1.s15>Ela encarou-o, sem desafio mas confiante, e perguntou:</s> </p> <p id=DL2TP.1.p6> <s id=DL2TP.1.s16 corresp=DL2.1.s15> — Achas que nos aceitam?</s></p> <p id=DL2TP.1.p7> <s id=DL2TP.1.s17 corresp=DL2.1.s16> — E, conforme sabia que Jasper responderia, este retorquiu:</s></p> <p id=DL2TP.1.p8> <s id=DL2TP.1.s18 corresp=DL2.1.s16>— É tudo uma questão de nós os aceitarmos a eles.</s></p> <p id=DL2TP.1.p9> <s id=DL2TP.1.s19 corresp=DL2.1.s17>Alice resistira ao teste sobre a sua pessoa, à dor óssea, e ele largou-lhe o pulso e dirigiu-se para a porta.</s>

Anchor word list

Original (882)	Modified (1,022)
is, 's / é, está	's / ((é) (está)) is / ((é) (está))
became, becom* / torn*, volt*, fic*	became / ((torn.*) (volt.*) (fic.*)) becom.* / ((torn.*) (volt.*) (fic.*))
English* / ingl*	English.* / ingl.*
has, have, 've / tenho, tens, tem, temos, têm	has / ((tenho) (tens) (tem) (temos) (têm)) have / ((tenho) (tens) (tem) (temos) (têm)) 've / ((tenho) (tens) (tem) (temos) (têm))
7*, seven / 7*, sete	7.* : ((7.*) (sete)) seven : ((7.*) (sete))

First and second tasks

Estimate 1. the revision and correction involved for a given E-P text set, and 2. the contribution of the anchor word list

- Run the program
- Count the matrices to review
- Count the changes relative to the final result
 - a. With anchor list
 - b. Without

Third task

3. Investigate in detail the contents of the anchor list

- For each anchor pair, look at
 - the number of occurrences in the E. Texts
 - the number of occurrences in the P. Texts
 - the number of joint occurrences
 - the number of s-units (E. and P.) involved
- For each s-unit, look at
 - the number of relevant anchor pairs

Qualitative remarks

- Many independent indicators make for a robust performance
- Cross-categorial match
 - nation.*? / ((país.*?)|(naç.*?)|(naciona.*?))
 - imagin.*? / imagin.*?
- Success not always linguistically motivated
 - not expected
 - wrong match; right alignment

Unexpected hits

- **buil.* / constru.***
 - *Once the job is finished the **builders** are killed*
 - *Terminada a obra, os **construtores** serão mortos.*
- **earli.* / cedo:**
 - *She worked harder than anybody else, got up **earlier**, came to bed long after the others*
 - *Ela trabalhava mais que qualquer um, acordava **cedo**, era a que ia dormir mais tarde.*
- **crim.* / crim.*:**
 - *... arrived in Sicily from somewhere far away, perhaps the **Crimea**, and within a few days ...*
 - *... chegou à Sicília vinda de um lugar distante, talvez da **Criméia**, e em poucos dias ...*

Hits by chance

- look.* / olh.*
 - *which makes my face **look** pallid and ill, with circles under the eyes.*
 - *que dá à minha cara um aspecto pálido e doentio, com **olheiras**.*
- be / ((ser)|(estar))
 - *one will always **be** a stranger*
 - *um **ser** ('being', n.) humano sempre será um estranho*
- couple.* / par.*
 - *After a **couple** of years to settle down,...*
 - *Depois de alguns anos **para** ('for') se estabelecer, ...*
- little / pequen.*
 - *was a little group of children, the eldest girl wheeling a pushchair with two **smaller** children,*
 - *ia um grupo de crianças, a mais velha a empurrar um carrinho, com duas crianças mais **pequenas**, uma de cada lado*

Some anchor pairs

File	&mdash	&mdash	both
AB	57	247	16
AH	64	454	25
AT	53	343	46

File	I	eu	both
AB	238	105	87
AH	171	61	40
AT	73	32	19

File	been	sido, estado	both	A\B
AB	62	16	16	46
AH	45	9	1	44
AT	33	10	7	26
All	632	216	134	498

estado 'state'

simple MQP

File	could	podia.* etc.	both
AB	29	132	12
AH	19	120	10
AT	48	216	28

podia 'might'

Concluding remarks

- Evaluation is a time and resource consuming task
- The strengths of an NLP system may not depend crucially on the quality of the linguistic information
- A language-dependent system may not be that language-dependent after all
- Decrease of performance by adapting to a new language pair is hard to measure
- Such a detailed study pinpoints interesting areas of textual differences between the two languages, and thus areas for contrastive research

Results, 1st task (for Abr1)

- Total number of matrices: 120
- Matrices to check: 78 (65%), corresponding to 740 s-units
- Number of corrections: 33
- Percentage of corrected s-units: 4.4% of those inspected, 2.9% of total s-units

Results, 2nd task (for ABR1)

- Differences between the final version and the raw version with anchor list: **33** (ranges from 4 to 68)
- Differences between the final version and the raw version without anchor list: **139** (ranges from 11 to 139)
- Differences between raw versions with and without anchor list: **126** (ranges from 9 to 126)

Note: For 6 out of 16 texts it was not possible to produce a result without anchor list

Results, 3rd task (for ABR1)

- English matches: 695 out of 1022 (68%), resulting in 7,194 English hits (pointing to 1,014 different target s_units)
- Portuguese matches: 780 out of 1022 (76%), amounting to 50,226 Portuguese hits
- Pairs with hits on both languages: 563 (55%), providing **3,585** successes, relevant to X different sentence pairs