SCIENCESPRINGDAY

FCT FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE NOVA DE LISBOA

DEPARTAMENTO DE INFORMÁTICA

Supervised Phrase-Based SMT

Human Language Technologies Group - HLT

CENTER FOR INFORMATICS AND INFORMATION TECHNOLOGIES

José Aires

(Ph.D. Student)

Research in Machine Translation using Aligned Parallel Corpora, Suffix Arrays and Phrase-Based Bi-Lingual Lexica

Objectives

- Simple combination of phrases to produce translations of complete texts.
- Including verified lexica to provide additional quality.
- Integration of translation patterns to support efficient order change rules and generic elements, like numbers.
- Use of a small number of feature models, focusing on translation and language models, but also attributing some relevance to the number of phrases.
- Submit a patent for the work.

Methodology

- Efficient use of Suffix Arrays for both phrase translations extraction and language model.
- Phrase-based aligned parallel corpora used for high quality phrase translations extraction.
- Simple phrase-based language model supported by monolingual corpora.
- Inclusion of verified lexica.
- Efficient integration and use of translation patterns.
- Simple combination of translation and language models, alternative to the log-linear models, which allow a countless number of features with no significant improvement.

Obtained Results

- On the side, Table1 shows the BLEU scores obtained by this system, translating four different files in several language pairs, also using our lexicon.
- The remaining two tables measure the difference between our scores and the scores obtained by Moses for the same files: Table 2 without our lexicon, and Table 3 including it. Negative differences indicate Moses has better results while the positive ones favor this system.

file	file 1	file 2	file 3	file 4	Avg
enpt	77,9	95,7	75,7	79,4	82,2
pten	78,7	95,8	80,7	79	83,6
espt	75,3	90,5	72,3	49,4	71,9
ptes	74,1	90,7	70,6	47,6	70,8
enes	80,8	95,3	74,5	32,3	70,7
esen	80,1	95,8	72,9	34,9	70,9
deen	80,9	79,4	72,9	22,2	63,9
ende	78,9	78,7	67,6	15,2	60,1
dept	71,8	75	63,1	22	58
ptde	58,2	69,7	50,2	12,5	47,7

Table 1

file	file 1	file 2	file 3	file 4	Avg
enpt	19,3	14,2	19,9	44,3	24,4
pten	21,7	14,2	25,6	46,4	27
espt	11,4	6,9	9,6	0,4	7,1
ptes	10	8,6	8,1	1,2	7
enes	20,2	12,7	18,5	-2,2	12,3
esen	20,5	11,6	17,2	1,9	12,8
deen	30,1	3,4	25,6	-6,3	13,2
ende	31,4	5,3	27,5	-4,1	15
dept	23,3	4,4	19,3	-4,3	10,7
ptde	17	-1,1	14,9	-5	6,5

Table 2

file	file 1	file 2	file 3	file 4	Avg
enpt	17,8	14,1	15,7	30,9	19,6
pten	19,6	15,4	20,9	28,3	21,1
espt	10,6	6,6	8,7	-2,8	5,8
ptes	10,1	8,5	8,1	-2	6,2
enes	20,7	12,1	18,2	-4,8	11,6
esen	20,7	11,6	17,4	-1,4	12,1
deen	30,3	3,1	26	-9,4	12,5
ende	32,9	4,8	26,3	-6,5	14,4
dept	23	3,1	18	-8,9	8,8
ptde	18,5	-2,7	14,1	-7	5,7

Table 3

Funding:

FCT Fundação para a Ciência e a Tecnologia

SFRH/BD/48839/2008
PTDC/EIAEIA/114521/2009