SCIENCESPRINGDAY

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

DEPARTAMENTO DE INFORMÁTICA

Translation Alignment and Extraction

MULTIMODAL SYSTEMS / HLT Team

CENTER FOR INFORMATICS AND INFORMATION TECHNOLOGIES

Luís Gomes

(PhD Student)

2007 – Licenciatura em Engenharia Informática (FCT-UNL)

2009 – Mestrado em Engenharia Informática

Objectives

My work is focused on the alignment and extraction of word and multi-word translations from parallel texts (texts that are translations of each other). Aligned texts and bilingual lexica (large collections of word and multi-word translations) are crucial for various translation-related tasks, in particular for Phrase-Based Statistical Machine Translation (PBSMT) and Computer-Assisted Translation (CAT).
The HLT team is commited to create a Machine Translation system capable of producing high-quality translations that require very little post-editing, as well as developing CAT tools for helping human translators in the revision of generated translations.

Methodology

To extract word and multi-word translations we first align large collections of parallel texts using our manually verified bilingual lexica, and then we collect co-occurrence statistics for neighbouring misaligned segments. Combining these statistics, and exploiting the fact that pairs of expressions that are translations tend to co-occur more systematically than non-translations, we extract pairs of expressions that are more likely to be translations. Afterwards, a team of linguists manually validates extracted pairs, augmenting our bilingual lexica, which in turn allows a more accurate alignment of the base collection of parallel documents.
By iterating over alignment, extraction and validation we continuously improve the quality of the alignment and the coverage of our lexica.

Expected Results

Today we have lexica for more than a dozen language pairs, including European languages, Hindi, Arabic and Chinese. The current sizes (in terms of pairs of expressions) of our largest lexica are:

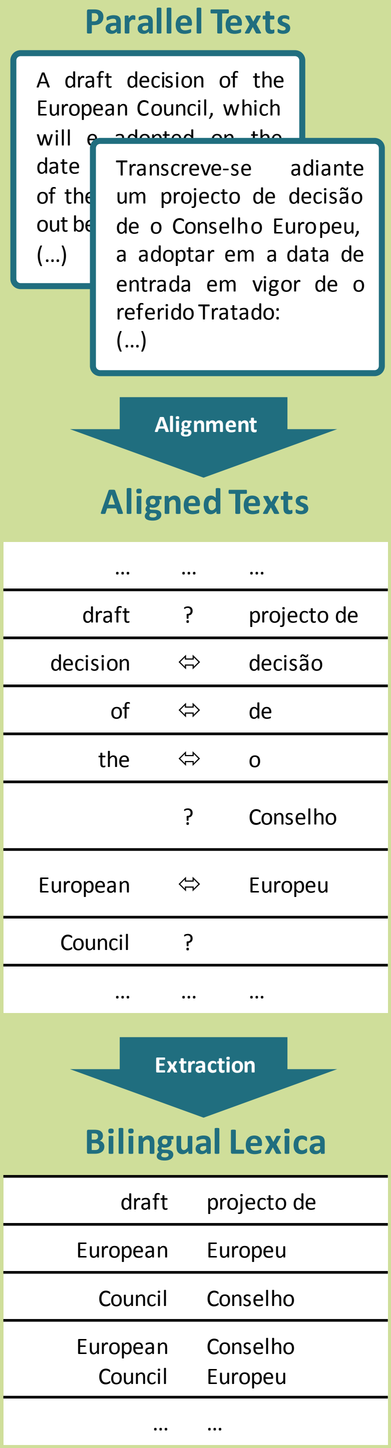
EN-PT	FR-PT	EN-FR	DE-PT	EN-ES	ES-PT	DE-EN	DE-ES	EN-ZH
748.360	287.603	253.488	237.295	213.935	162.903	131.611	71.784	13.474

In the near future we expect that our English-Portuguese lexicon (on which we have been working for longer than other language pairs) reaches one million entries.

Funding:

FCT Fundação para a Ciência e a Tecnologia

PEst-OE/EEI/UI0527/2011 SFRH/BD/65059/2009 PTDC/EIA-EIA/114521/2009



Parallel Texts

A draft decision of the European Council, which will e adopted on the date of the out be (…)

Transcreve-se adiante um projecto de decisão de o Conselho Europeu, a adoptar em a data de entrada em vigor de o referido Tratado:
(…)

Aligned Texts

…	…	…
draft	?	projecto de
decision	⇔	decisão
of	⇔	de
the	⇔	o
	?	Conselho
European	⇔	Europeu
Council	?	
…	…	…

Bilingual Lexica

draft	projecto de
European	Europeu
Council	Conselho
European Council	Conselho Europeu
…	…