SCIENCE*SPRING*DAY

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

DEPARTAMENTO DE INFORMÁTICA

Compression in Machine Translation

MULTIMODAL SYSTEMS

CENTER FOR INFORMATICS AND INFORMATION TECHNOLOGIES

Jorge Costa

(PhD Student)

Advised by Gabriel P. Lopes and Luís Russo.

Research focus on Algorithms and Compressed Data Structures

Objectives

Machine Translation (MT) tasks are typically space demanding. Tasks like parallel text alignment, extraction of translation pairs or concordancer applications, demand querying over the texts, which can have sizes of 1 (or more) Gigabyte per language. The space consuming nature of text indexes like suffix trees or suffix arrays (at least 4 times the text size), makes it difficult to index the texts in main memory, which slows down the applications.

Our aim is to solve this space consumption problem, by developing a framework, or several frameworks, based on data compression, for supporting the mentioned MT tasks in main memory, without losing efficiency on query time response.

Methodology

For development a compressed framework, we need several layers:

- **Compressed Text Indexing:** Word-Based Compressed Data Structures.
  - Byte-Oriented Wavelet Tree [1] – text words are codified in codewords and organized in a wavelet tree-like fashion.
  - Word-Based Compressed Suffix Array [2] – words replaced by integer ids (Sid) and added to an integer compressed suffix array (iCSA).
- **Alignment Layer:** represent the links between aligned segments [3], independently of the granularity, which can be at sentence or word level. This representation works as well for bilingual lexicon entries [4].
- **Document Layer:** represent the several documents indexed and their home directories, distinguishing them in the compressed index and enabling the filtering of the occurrences to certain documents.

Expected Results

- Space consumption lower than the size of the texts indexed.
- Logarithmic query response, as the known text indexes, where a potential slowdown due to compression is compensated by avoiding secondary memory.
- Ability to index several texts at the same time in one single index, reducing the number of loading operations from disk to main memory.
- Ability to index and query not only aligned parallel corpora, but also bilingual lexicons and monolingual text collections for capturing language models.
- Support gapped alignments (Hierarchical Phrase Based Translation) and gapped bilingual lexicon entries, with variables ($) between string literals as in Figure [5].

Funding:

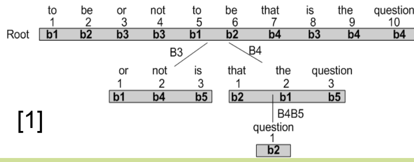
FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

IStRioN Project

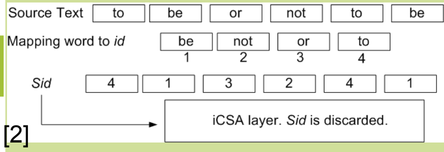
PEst-OE/EEI/UI0527/2011

TEXT: to be or not to be that is the question

WORD	FREQ	CODE
to	2	b1
be	2	b2
or	1	b3b1
not	1	b3b4
that	1	b4b2
is	1	b3b5
the	1	b4b1
question	1	b4b5b2



[1]



Source Text: to | be | or | not | to | be

Mapping word to *id*: be 1 | not 2 | or 3 | to 4

Sid: 4 | 1 | 3 | 2 | 4 | 1

iCSA layer. *Sid* is discarded.

[2]

English	Segments	Portuguese
the	1	a
Slovak Republic	2	República Eslovaca
acceded	3	aderiu
	4	em 1 de Maio de 2004
to the	5	às
European Communities	6	Comunidades Europeias
and	7	e
to the	8	à
European Union	9	União Europeia
established	10	instituída
by	11	por
the	12	o
Treaty on European Union	13	Tratado da União Europeia
on 1 May 2004	14	

[3]

ENGLISH	are translated by	PORTUGUESE
established		instituída
European		Europeia
European		Europeu
European Union		União Europeia
Slovak Republic		República Eslovaca
to the		às

[4]

ENGLISH	PORTUGUESE
has already provided	já disponibilizou
has only provided	apenas disponibilizou
has not provided	não disponibilizou
has not only provided	não só disponibilizou
↓ REPLACED BY	↓
has $ provided	$ disponibilizou

[5]