$B_1A_2D_3$ LUC@WMT 2016: a Bilingual$_1$ Document$_3$ Alignment$_2$ Platform Based on Lucene

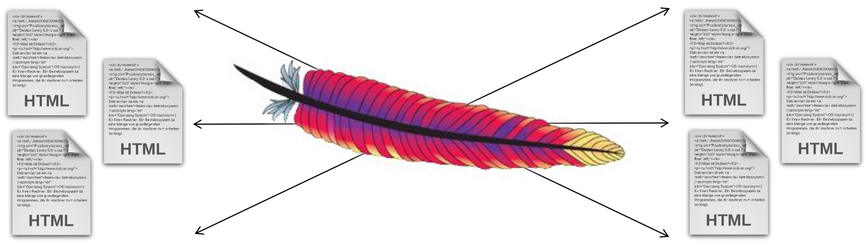
Laurent Jakubina and Philippe Langlais

RALI – Université de Montréal, Québec, Canada
`rali.iro.umontreal.ca`

Abstract

- Motivation: plain cross-lingual IR system using

Apache Lucene framework
(https://lucene.apache.org/**)**



- We developed several variants, including one which only uses URLs, which offers (surprisingly) high performance.
- We submitted a variant which combines text and URL and which **achieves 92% on the development set** provided.

Architecture

2 Indexes

- One per language.
- 1 **`Lucene`** document = 1 web page.
- 3 **`Lucene` in-house tokenized fields** on:
  - ***`Text`*** (as provided)
  - ***`Url`***
  - `Size` (of the text, in # tokens)

Own Tokenizer: blanks and special characters ► Splits ***`Url`***
http://creationwiki.org/Earth
► *url*_tok: ***earth creationwiki / org http . :***

Queries

- **Bag-of-Word Queries** (*MoreLikeThis Query)*
  - on each field.
- Several meta-parameters:
  - *Term Frequency*
  - *Min/Max Doc Frequency*
  - *Min/Max Word Length*
  - *Nb. terms in query (size)*
  - *Stop-word List*
  - *(boost factor per query term)*
- **Monolingual** `(mono)`
  - capture named entities, numbers, URLs, etc.
- **Bilingual** `(bili)`, thanks to a lexicon
  - Keep untranslated terms (`K`, `-K`).
  - Nb. of translations to take (`all`, `first`).
- **Queries on Multiple Fields** `(both)`
  - queries on ***`both`*** fields (***`text`*** and ***`url`***) in one pass.
- **Length-based Filter (**`+size`**)**
  - preprocessor (Gaussian model) that filters out implausible target documents.

	http://creationwiki.org/Earth **Earth - CreationWiki, the encyclopedia of creation science [...] 23.439281 0.409rad 26.044grad Physical characteristics Mass 5.9736 * 1024 kg [...] taking 23 hours, 56 minutes, and 4.091 seconds to line up relative to the stars (Sidereal day), and 24 hours plus or minus 20 seconds to line up relative to the sun [...] is closer to the sun at some times of the year than others; the Earth moves faster [...] Kepler's laws of planetary motion [...] Saturnine - Uranian - Neptunian [...]**
`both` `bili` **QUERY**	text_tok: hours neptunian 1024 penchant théorie tennessee absorbant prononcée prénommé 2008 métrique 397 équateur … url_tok: déblai masse tanière terre earth creationwiki / org http . : …

Post-treatment

- Each query returns independently a ranked list of documents (thus a target document can be associated with more than one source document).
- We tested 3 **strategies for selecting exactly one candidate per source document**:
  - **`hungarian`**: the *Hungarian* Algorithm.
  - **`b-greedy`**: *batch* greedy (keeps the best association of the 1st source doc, removes the target document, continue).
  - **`o-greedy`**: *online* greedy (keeps the best association, removes the pair of documents, loops).

Experiments

Protocol

- Distributed dataset :
  - Indexes built on **lett.train (49)** webcrawls.
  - 1624 source URLs as test set.
- In-house lexicon (108k entries).
- Accuracy at rank 1,5 and 100 (TOP@i).

Results

- ***`url`*** Queries: Impressive results for such a simple approach.
  - Tokenizes the URL, translates its words, uses the filter.

	Query Type	TOP@1	TOP@5	TOP@100
best-***`url`***	`bili+size`	**80.1**	**88.6**	**95.6**

- ***`text`*** Queries: Decent results without translation.
  - Better with translation.
    - One translation per word, keeping unknown words.
  - **Default *`Lucene`* configuration is useless.**
    - Importance of *MoreLikeThis* parameters.
    - Our **tokenizer** really improves performances.

	Query Type	TOP@1	TOP@5	TOP@100
	`mono`	**64.9**	**87.2**	**96.8**
best-***`text`***	`bili+size`	**83.3**	**96.2**	**98.2**
(default)	`mono`	**6.4**	**15.8**	**49.5**
(def+tok)	`mono`	**35.4**	**57.0**	**83.9**

- ***`both`*** Queries (best-***`url`*** **+** best-***`text`***): Impressive combination.

	Query Type	TOP@1	TOP@5	TOP@100
RALI	`bili+size`	**88.6**	**97.6**	**98.3**

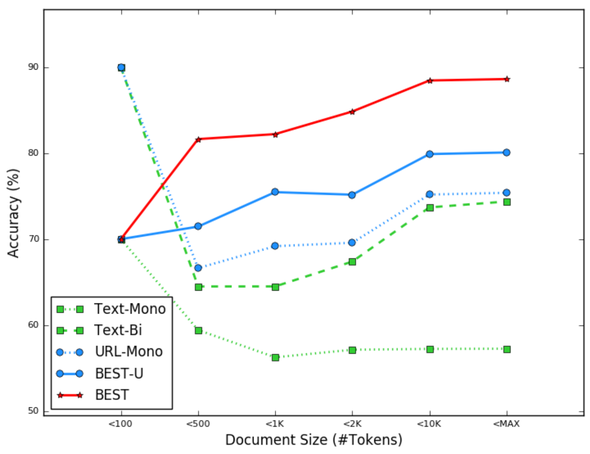
- Length-based filter (`+size`): Positive impact.
  - Improves performances (2 to 40 points).
  - Significant speedup (2 to 10 times).

- Post-treatments: Applying one improves TOP@1.
  - Does not matter much which algorithm is used.
  - Final submission: o-greedy because the others could not handle the size of the data set.

	TOP@1
`w/o` **(RALI)**	**88.6**
`o-greedy`	**91.6**
`b-greedy`	**92.1**
`hungarian`	**92.1**

Analysis

Sensitivity to Source Document Length



- RED (RALI submission) improves as source documents get larger.
- GREEN (mono and bili ***`text`***-based IR): benefit of embedding translation increases with doc size.
- BLUE (***`url`***-based IR): unclear why performance increases with doc size.

Error Analysis

- Often, French pages contain English material.
- Many source documents have several URLs associated.
- RALI failed at identifying the reference document in 1.7% of the cases:
  - (almost) no text.
  - Reference errors (see the paper for examples).
  - Incompatibility of our bilingual lexicon with specific documents.

	Almost no text inside
src	http://rehazenter.lu/en/medical/explorations_fonctionnelles/explorations_posture/laboratoire_de_biomecanique
trg	http://rehazenter.lu/fr/medical/explorations_fonctionnelles/explorations_posture/laboratoire_de_biomecanique
src	http://www.dakar.com/2009/DAK/RIDERS/us/equipage/57.html
trg	http://www.dakar.com/2009/DAK/RIDERS/fr/equipage/57.html
	Reference problem
src	http://www.nauticnews.com/en/2009/06/23/burger-_boat-_company-_launches-_151-_03-_fantail-_motor-_yacht-_sycara-_iv
trg	http://www.nauticnews.com/2009/07/13/ishares-_cup-_2009-_a-_bord-_dholmatro

Acknowledgments

This work has been founded by the *Fonds de Recherche du Québec en Nature et Technologies* (FRQNT).