"SINAI at CLEF 2005: The evolution of the CLEF2003 system."

Fernando Martínez-Santiago
Miguel Ángel García-Cumbreras

University of Jaén

Content

- CLEF 2003. Experiment Framework
- CLEF 2005. Experiment Framework
  - Machine Translators and Alignment
  - Aligned and non-aligned scores
  - IR systems and Pseudo-Relevance Feedback
- Results and comparative
- Conclusions and Future Work

CLEF 2003

Experiment Framework

- ***Pre-processing of monolingual collections***:
  stopword lists and stemming algorithms.
- ***Decompounding algorithm***
  for Dutch, Finnish, German and Swedish.
- ***IR System***: *Zprise IR system, using the OKAPI probabilistic model.*
- ***Machine Dictionary Readable (MDR)***
  FinnPlace for Finnish / Babylon for the rest.
- *With and without Pseudo-relevance feedback (**PRF**):*
  Robertson-Croft's approach (no more than 15 search keywords, extracted from the 10-best ranked documents).
- ***Fusion Algorithms***:
  2-Step RSV and raw mixed 2-Step RSV

CLEF 2005

Experiment Framework - Resume

- ***Pre-processing of monolingual collections***:
  stopword lists and stemming algorithms.
- ***Decompounding algorithm***
  for Dutch, Finnish, German and Swedish.
- ***IR Systems***:
  - *Document Retrieval: Zprise IR system*
  - *Passage Retrieval: IRn*
  - *Several lists of relevant documents available from Multi-8 Merging-only task (DataFusion, OKAPI and Prosit)*
- ***Machine Translators (MT) vs. MDR.***
- ***Alignment Algorithm.***
- *With and without Pseudo-relevance feedback (**PRF**): Robertson-Croft's approach (10-15 search keywords, extracted from the 10-best ranked documents).*
- ***Fusion Algorithms***:
  raw mixed 2-Step RSV with and without machine learning

CLEF 2005
Experiment Framework – MTs and Alignment

- *Since 2-step RSV requires to group together the DF for each concept (term and its translations), and MTs translate better the whole phrase →* ***Alignment Algorithm*** *[SINAI at CLEF 2004: Usign Machine Translation Resources with 2-step RSV Merging Algorithm]*

- *Percent of aligned non-empty words (CLEF 2005, Title+Desc)*

Language	Translation Resource	Alignment percent
Dutch	Prompt (MT)	85.4%
Finnish	FinnPlace (MDR)	100%
French	Reverso (MT)	85.6%
German	Prompt (MT)	82.9%
Italian	FreeTrans (MT)	83.8%
Spanish	Reverso (MT)	81.5%
Swedish	Babylon (MDR)	100%

CLEF 2005

Experiment Framework – Aligned and non-aligned scores.

- *We use two subqueries → Q_1: aligned terms / Q_2: non-aligned terms*
- *For each subquery we obtain an RSV score.*
  Several ways to combine both scores:
  1. *Raw mixed 2-step RSV. Combining the RSV value of the aligned words and not aligned words with the formula:*
     0.6 * <RSV_aligned_doc> + 0.4 * <RSV_not_aligned>
  2. *Mixed 2-step RSV by using Logistic Regression. It applies the formula:*
     $e^{(alpha * <RSV\_aligned\_doc> + beta * <RSV\_not\_aligned>)}$
  3. *Mixed 2-step RSV by using Logistic Regression and local score. The last one also uses Logistic Regression but include a new component, the ranking of the doc. It applies the formula:*
     $e^{(alpha * <RSV\_aligned\_doc> + beta * <RSV\_not\_aligned> + gamma * <ranking\_doc>)}$
  4. *Mixed 2-step RSV by using Bayesian Logistic Regression and local score. similar to the previous approach, but it is based on bayesian logistic regression instead of logistic regression.*
- *2, 3 y 4 use Machine Learning Methods.*
- *CLEF 2005 → 60 queries: 20 for training / 40 for evaluation*

CLEF 2005
Experiment Framework – IR Systems and PRF

- ***3 Sources of Relevant Lists:***
  - *Documents Retrieval IR System: Zprise using the OKAPI probabilistic model.*
  - *Passage Retrieval: IR-n [F. Llopis, R. Muñoz, R. M. Terol and E. Noguera. IR-n r2: Usign Normalized Passages. CLEF 2004. University of Alicante] modified to return lists of relevant documents (that contain relevant passages).*
  - *Several lists of relevant documents, available from Multi-8 Merging-only task (DataFusion, Okapi and Prosit, thanks to Jacques Savoy)*

- ***PRF:***
  - *Some experiments based on Zprise.*
  - *Robertson-Croft's approach in the first step. 10-15 search keywords expanded, extracted from the 10-best ranked documents.*
  - *Second step does not make use of automatic query expansion techniques such as relevance feedback (RF) or pseudo-relevance feedback (PRF) applied to monolingual queries.*

CLEF 2005

Comparative Results

- ***CLEF 2005 Vs. CLEF 2003***

Year	Main Features	AvgP
2003	Case Base 2003 (OKAPI Zprise IR, **MDR**, 2-Step RSV)	24.18
2005	Case Base 2005 (OKAPI **Zprise IR**, no PRF, MT, raw mixed 2-Step RSV)	28.78
2005	Case Base 2005 + **PRF**	29.01
2005	Case Base 2005 (**IRn IR**, no PRF, MT, raw mixed 2-Step RSV)	28.81
2005	OKAPI Zprise IR, no PRF, MT, Mixed 2-Step RSV by using **Logistic Regresion** and local score	29.19
2005	OKAPI Zprise IR, **PRF**, MT, Mixed 2-Step RSV by using **Logistic Regresion** and local score	**29.57**

↑ **19.02%**

↑ **22.29%**

CLEF 2005

Multi-8 Merging-only Results

- ***The relevant documents are available for the task from Neuchatel Bilingual Runs from CLEF 2003.***

Year	Main Features	AvgP
2003	Case Base 2003 (OKAPI Zprise IR, **MDR**, 2-Step RSV)	24.18
2005	OKAPI Zprise IR, **PRF**, MT, Mixed 2-Step RSV by using **Logistic Regression** and local score	29.57
2005	OKAPI Zprise IR, no PRF, MT, raw mixed 2-Step RSV **PROSIT Documents**	28.40
2005	OKAPI Zprise IR, no PRF, MT, raw mixed 2-Step RSV **OKAPI Documents**	28.87
2005	OKAPI Zprise IR, **PRF**, MT, Mixed 2-Step RSV by using **Logistic Regression** and local score **DataFusion Documents**	**30.37**

↑ **22.29%**

↑ **25.60%**

CLEF 2005

Conclusions and Future Work

- ***Conclusions:***
  - *Our 2003 CLIR system has been improved using Machine Translators instead of Machine Dictionaries Readable, around a **20%** in terms of Average Precision.*
  - *This year we have tested more IR systems, based on Document and Passage Retrieval. Last one gets a bit better results.*
  - *Machine Learning Methods not produce an important improvement, because monolingual CLEF collections are quite comparable.*

- ***Future Work:***
  - *To improve the translation and the Alignment Algorithm.*
  - *To test and to improve the IR system for the second phase.*

Thanks

“SINAI at CLEF 2005: The evolution of the CLEF2003 system.”