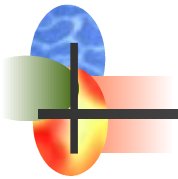

REINA at CLEF 2006 Robust Task: Local Query Expansion Using Term Windows for Robust Retrieval

Angel Zazo
Carlos G. Figuerola
José Luis A. Berrocal

REINA Research Group
University of Salamanca
http://reina.usal.es

REINA participation at CLEF

- CLEF 2000
- CLEF 2001:
  - Monolingual Experiments
- CLEF 2002:
  - Monolingual Experiments
- CLEF 2003
- CLEF 2004:
  - iCLEF

- CLEF 2005:
  - iCLEF
  - WebCLEF
- CLEF 2006:
  - iCLEF
  - WebCLEF
  - Ad-hoc: Robust task

Objective

- Robust task: ad-hoc task that explores methods for stable retrieval by focusing on poorly performing topics
- Subtask:
  - Monolingual: EN, ES, DE, FR, IT, NL
  - Bilingual: IT→ES, FR→NL, EN→GE
  - Multilingual: All six languages
- Robustness: all topics obtain minimum effective levels

MAP ↔ GMAP (TREC 2004)

Experiments

- Subtask:
  - Monolingual: EN, ES, DE, FR, IT, NL
  - Bilingual: IT→ES, FR→NL, EN→GE
  - Multilingual: ES→[EN ES FR IT]

- Previous approaches (TREC):
  - Use of external corpus (Web) to perform query expansion
  - Tokenization techniques or weighting schemes
  - New retrieval models or (re-)ranking functions
  - Local query expansion

Experiments. Our approach

- Monolingual retrieval
  - We have some experience on query expansion:
    - Testing some **local query expansion** techniques with training topics and select the best one:
      - Blind/Pseudo relevance feedback
      - Co-occurrence based thesaurus
      - Co-occurrence based thesaurus built with term windows
  - Focussing on topics:
    - "OK" if average precision > MAP
    - "bad" if average precision is only > MAP/2
    - "hard" if average precision < MAP/2

Experiments. Our approach

- Bilingual retrieval: IT→ES
  - Merging some MT translations (is another expansion process)
  - Doing monolingual retrieval

- Multilingual retrieval: ES→[EN ES FR IT]
  - The same as bilingual retrieval for target languages
  - List fusion

Query expansion depends on ...

- How we obtain relations between terms ?
  - relations with all query terms, not with only one separately !!
- What is the importance of added terms (weight) ?
- Sort queries vs. long queries
  - Long queries: adding more terms no improves performance
- **Local expansion**: the **first retrieval** is fundamental
  - A good information retrieval system is better than a good expansion technique

Monolingual retrieval

- Our “good” information retrieval system:
  - Simple vector space model: weighting schema
  - No plugins for word sense disambiguation
  - No linguistic techniques

- Experiments:
  - In mind:
    - Best **MAP** and **GMAP**
    - **Hard** topics
    - Sort and long queries:
      - **td**: `title` + `description` topic fields (mandatory)
      - **t**: only the `title` field

Monolingual retrieval: weighting schema

- Vector space model: weighting schema?
  - No stop-word removing
  - No stemming
  - A lot of tests to obtain the best performance (training topics):
    - **dnu-ntc** (u: pivoted document normalization)
      - pivot: average document length
      - slope=0.1 for all EN, ES, FR, IT !!
- Predicting topic difficulty (as in TREC 2004, 2005):
  - IDF, TF, combinations...
    - **No heuristic was found** 

Stop word removing

- Not easy to build the set of stop words
  - terms with DF > 50% | 25% | 15%

- Performance:
  - For **all topics**:
    - no difference in MAP and GMAP (all languages)
  - For **hard topics**:
    - EN, IT: slight improvement
    - ES, FR: no improvement
  - We decide to remove words if DF>25%

Remove stop words !

Stemming

- Stemmers:
  - EN: Porter stemmer
  - ES, FR, IT: stemmers from http://www.unine.ch/info/clef/ (thanks to Jacques Savoy)

- Performance
  - For **all topics**:
    - Improvement for all collections
  - For **hard topics**:
    - EN: no change
    - ES: little improvement
    - FR, IT: important improvement

Do stemming !

Blind Relevance Feedback

- Rocchio formula ($\gamma=0$):

$$\vec{q'} = \alpha\vec{q} + \frac{\beta}{n_r}\sum_{i\in rel}^{n_r}\vec{d_i} - \frac{\gamma}{n_{nr}}\sum_{j\in norel}^{n_{nr}}\vec{d_j}$$

  - A lot of tests with:
    - first 5, 10, 15 and 20 retrieved documents
    - probe α, β values
    - how many terms in expanded query?
  - Best results with:
    - Docs = 5 or 10
    - $\alpha=1$, $\beta\sim2.5$ (rare)
    - ~ 50 terms

Blind Relevance Feedback

- Performance (5 docs, 50 terms):
  - For **all topics**:

		α=1, β=1		α=1, β=2.5	
		MAP	GMAP	MAP	GMAP
EN	td	-1.36	-3.99	-1.15	-7.51
	t	+2.33	+0.32	+0.72	-11.46
ES	td	+4.17	+3.29	+6.73	+4.51
	t	+5.62	+5.89	+5.89	+5.95
FR	td	+0.56	+1.40	+0.56	+11.17
	t	+3.02	+7.66	+4.49	+12.75
IT	td	+3.01	+2.54	+5.20	+3.89
	t	+3.42	+3.77	+7.88	+7.54

  - For **hard topics**:
    - EN: worsening (td, t)
    - ES, FR, IT: td: no changes
      t: little improvement

Really do BRF ?

Co-occurrence based thesaurus

- Term co-occurrence in first retrieved documents
  - Association function:

$$\text{Tanimoto}(t_i, t_k) = \frac{n_{ik}}{n_i + n_k - n_{ik}}$$

$$\text{Coseno}(t_i, t_k) = \frac{n_{ik}}{\sqrt{n_i \cdot n_k}}$$

$$\text{Dice}(t_i, t_k) = \frac{2 \cdot n_{ik}}{n_i + n_k}$$

  - Relation measurement:

$$\text{rel}(q, t_e) = \vec{q}^T * \vec{t_e} = (\sum_{t_i \in q} q_i \cdot \vec{t_i})^T * \vec{t_e} =$$

$$= \sum_{t_i \in q} q_i \cdot (\vec{t_i}^T * \vec{t_e}) = \sum_{t_i \in q} q_i \cdot \text{ASS}(t_i, t_e)$$

  - Weight of added terms:

$$q_e = \frac{\text{rel}(q, t_e)}{\sum_{t_i \in q} q_i}$$

Co-occurrence based thesaurus

- Co-occurrence unit: complete document
- A lot of tests with:
  - First 2, 5, 10, 25, 50 and 100 retrieved documents
  - How many terms to add to original query?

- Results: Performance worsening in all cases:
  - Problems with weighting schema for topics: **ntc**

    c = cosine normalization

    Added terms obtain much importance than original terms

- Repeat tests with **ntn** schema for topics →

Co-occurrence based thesaurus - ntn

- Performance
  - For **all topics**:
    (5-10 docs, 10-20 added terms)

		MAP	GMAP
EN	td	-1,56	-7,47
	t	+2,15	-24,76
ES	td	+8,61	+7,49
	t	+10,74	+6,75
FR	td	+1,88	+14,07
	t	+8,20	+6,50
IT	td	+4,89	+3,25
	t	+8,47	+0,04

  - For **hard topics**:
    - EN: worsening (td, t)
    - ES: td: little improvement
      t: no changes
    - FR: td: important improvement
      t: no changes
    - IT: little deterioration (td, t)

Apply this technique ?
- Only for ES and FR.

Co-occurrence based thesaurus built with term windows

- Co-occurrence unit: term windows. Terms close to query terms are more important to expansion
  - Distance: 0 (adjacent), 1 (one term between two ones), ...
- Example: Topic 141: (title: "*Letter Bomb for Kiesbauer*"). Distance = 1

VIENNA: Letter-bomb attacks yesterday injured two women in the Austrian town of Linz and another woman in Munich, Germany. However they missed one target -- 29-year-old Arabella Kiesbauer, a television talk show hostess of mixed race and outspoken human rights campaigner. Police linked the bombs to anti-foreigner attacks in Austria by suspected right-wing extremists in the past 18 months. Letter bombs strike women.

Co-occurrence based thesaurus built with term windows

- Performance
  - For **all topics**:

		MAP	GMAP	Settings
EN	td	+4,14	+2,74	Local expansion: 10 docs Distance = 1 Terms added to original query: 10
	t	+3,64	+3,18	
ES	td	+8,46	+6,31	Local expansion: 10 docs Distance = 2 Terms added to original query: 30
	t	+9,51	+6,08	
FR	td	+0,19	+13,13	Local expansion: 5 docs Distance = 1 Terms added to original query: 30
	t	+6,30	+18,99	
IT	td	+2,29	+0,09	Local expansion: 10 docs Distance = 1 Terms added to original query: 30
	t	+8,00	+5,07	

  - For **hard topics**:
    - EN: td: medium improvement
t: slight improvement
    - ES: td, t: slight improvement
    - FR: td, t: important improvement
    - IT: td: no changes
t: little deterioration

The best result for hard topics !!

Bilingual retrieval: IT→ES

- Previous step:
  - MT translating topics:
    - Power Translation Pro 7.0
    - Worldlingo (http://www.worldlingo.com)
  - Merging translations (is another expansion process)

- Doing monolingual retrieval
  - Applying query expansion:
    - Co-occurrence based thesaurus built with term windows

Multilingual retrieval: ES→[EN ES FR IT]

- Steps:
  - Translating ES topics:
    - ES→EN: Power Translation Pro 7.0, Systran, Reverso
    - ES→FR: Systrans, Reverso
    - ES→IT: Power Translation Pro 7.0, Wordlingo
  - Merging translations
  - Doing monolingual retrieval
    - Applying the same query expansion
    - Monolingual ES
  - Doing data fusion: MAX-MIN
    - Coefficients: ES: **1.02**
      EN, FR, IT: **1.00**

Official Results

- Monolingual retrieval
  - EN: td: MAP > mean; MAP > Q3
    t: MAP > mean; Q1 < MAP < Q2
  - ES: td: MAP > mean; MAP = Q2
    t: worst run in competition (the only run with **t** field)
  - FR: td: MAP > mean; MAP ~ Q3
    t: MAP > mean; MAP ~ Q3
  - IT: td: MAP > mean; MAP > Q3
    t: MAP > mean; MAP ~ Q1

Official Results

- Bilingual retrieval
  - IT→ES: td: **best run** in competition !!
    t: MAP < mean; MAP ~ Q1

- Multilingual retrieval:
  - ES→[EN ES FR IT]: very bad results

Conclusions

- Monolingual retrieval
  - We use a simple vector space model IR system
  - We look for a good document-query weighting schema
  - Local query expansion using co-occurrence based thesauri built with term windows in an effective and relative simple expansion technique
- Bilingual retrieval:
  - Collecting terms from some translations + query expansion obtains performance improvement
- Multilingual retrieval:
  - The problem is the data fusion procedure

REINA at CLEF 2006 Robust Task: Local Query Expansion Using Term Windows for Robust Retrieval

Thanks very much !!

Angel Zazo
Carlos G. Figuerola
José Luis A. Berrocal

REINA Research Group
University of Salamanca
http://reina.usal.es