SINAI at CLEF 2004: Using Machine Translation resources with mixed 2-step RSV merging algorithm

University of Jaén - SPAIN

Fernando Martínez Santiago
Miguel Ángel García Cumbreras
Manuel Carlos Díaz Galiano
Alfonso Ureña López

SINAI at CLEF 2004

- Multilingual CLEF task
- Only queries are translated: merging algorithms are required
- Main interest: testing Machine Translation(MT) with mixed 2-step RSV merging algorithm.

- 2-step RSV calculus requires that given a word of the original query, its translation to the rest of languages must be known.
- MT translates the whole of the phrase better than word for word, we can't use MT with 2-step RSV merging algorithm directly.
- We propose an straightforward and effective algorithm to align the original query and its translation at term level.

Content

- How 2-step RSV merging algorithm works?
- An algorithm to align parallel text at term level based on MT
- Experimentation framework
- Results
- Global pseudo-relevance feedback
- Conclusions a future work

2-step RSV method

Conflict of Interest in Italy?

Conflicto intereses Italia

Conflitto interessi Italia

Conflits intérêt Italie

Spanish IR

Italian IR

French IR

#1 (conflicto, conflitto, conflits)
#2 (intereses, interessi, intérêt)
#3 (Italia, Italia, Italie)

Spanish, Italian, French retrieved documents

STEP 1

STEP 2

Indexing *concepts* (#1,#2,#3)

***Concept* IR**

Multilingual list of Ranked documents

Spanish		Italian		French		*Concept*	
Term	*df*	Term	*df*	Term	*df*	Id.	*df*
conflicto	1000	conflitto	1500	conflits	1250	#1	1000+1500 +1250=3750
intereses	5000	interessi	6000	Intérêt	4000	#2	5000+6000 +4000=15000
Italia	3500	Italia	6000	Italie	4500	#3	3500+6000 +4500=14000

An algorithm to align at term level a phrase and its translations by using machine translation resources

P_{en} → Pesticides in baby food

STEP 1

Machine Translation

P_{sp} → Pesticidas en alimento para niños

STEP 2

Unigrams (P_{en}) = {Pesticides, baby, food}
Bigrams (P_{en}) = {Pesticides baby, baby food}

STEP 3

U (P_{sp}) = {Pesticidas,alimento,bebés}
B (P_{sp}) = {Pesticidas bebés, alimento para niños}

...

...

STEP 4

Non Stopwords Stemmed

P_{sp} → Pesticida alimento niño
U (P_{sp}) = {Pesticida,alimento,bebé}
B (P_{sp}) = {Pesticida bebé, alimento niño}

STEP 5

To Align Unigrams

Pesticida **alimento** niño, {Pesticida, **alimento**, bebé}

Pesticide baby food, { Pesticide , Food , baby}

...

...

STEP 6

To Align Bigrams

Pesticida **alimento niño**, {Pesticida bebé, **alimento niño**}

Pesticide baby food, {Pesticide baby, **baby food** }

An algorithm to align at term level a phrase and its translations by using machine translation resources

Spanish	German	French	Italian
91%	87%	86%	88%

Percentage of aligned non-empty words (CLEF2001+CLEF2002+CLEF2003 query set, Title+Description fields, Babelfish machine Translation)

Finnish	French	Russian
100%	85%	80%

Percentage of aligned non-empty words (CLEF2004 query set, Title+Description fields, MT for French and Russian. MDR for Finnish)

Mixed 2-step RSV:Queries partially aligned

- Raw [illegible] 2-step RSV:

$$RSV_i [illegible] V_i^{align} + (1-\alpha) \cdot RSV_i^{nonalign}$$

Conflicto de intereses en Italia

Aligned terms

Dinamic 2-step RSV index

RSV$^{\text{aligned terms}}$

573

SP123

- Log[illegible]egression:

$$Prob[D_i [illegible] ank_i, rsv_i^{align}, rsv_i^{nonalign}] = \frac{e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i^{align}+\beta_3 \cdot rsv_i^{nonalign}}}{[illegible] \ln(rank_i)+\beta_2 \cdot rsv_i^{align}+\beta_3 \cdot rsv_i^{nonalign}}$$

Non- aligned terms

RSV$^{\text{not aligned terms}}$

39

Monolingual index (Spanish)

Experimentation framework – language dependent features

	English	Finnish	French	Russian
Preprocessing	stop words removed and stemming			
Additional preprocessing		compounds words to simple words		Cyrillic→ASCII
			Query aligment at word level algorithm based on MT	
Translation approach		FinnPlace MDR	Reverso MT	Prompt MT

Experimentation framework – language independent features

- ZPrise IR engine
- OKAPI probabilistic model (fixed at b = 0.75 and k1 = 1.2 for every language and for the 2-step RSV index)
- Neither blind feedback nor query expansion (no improvement except of French)

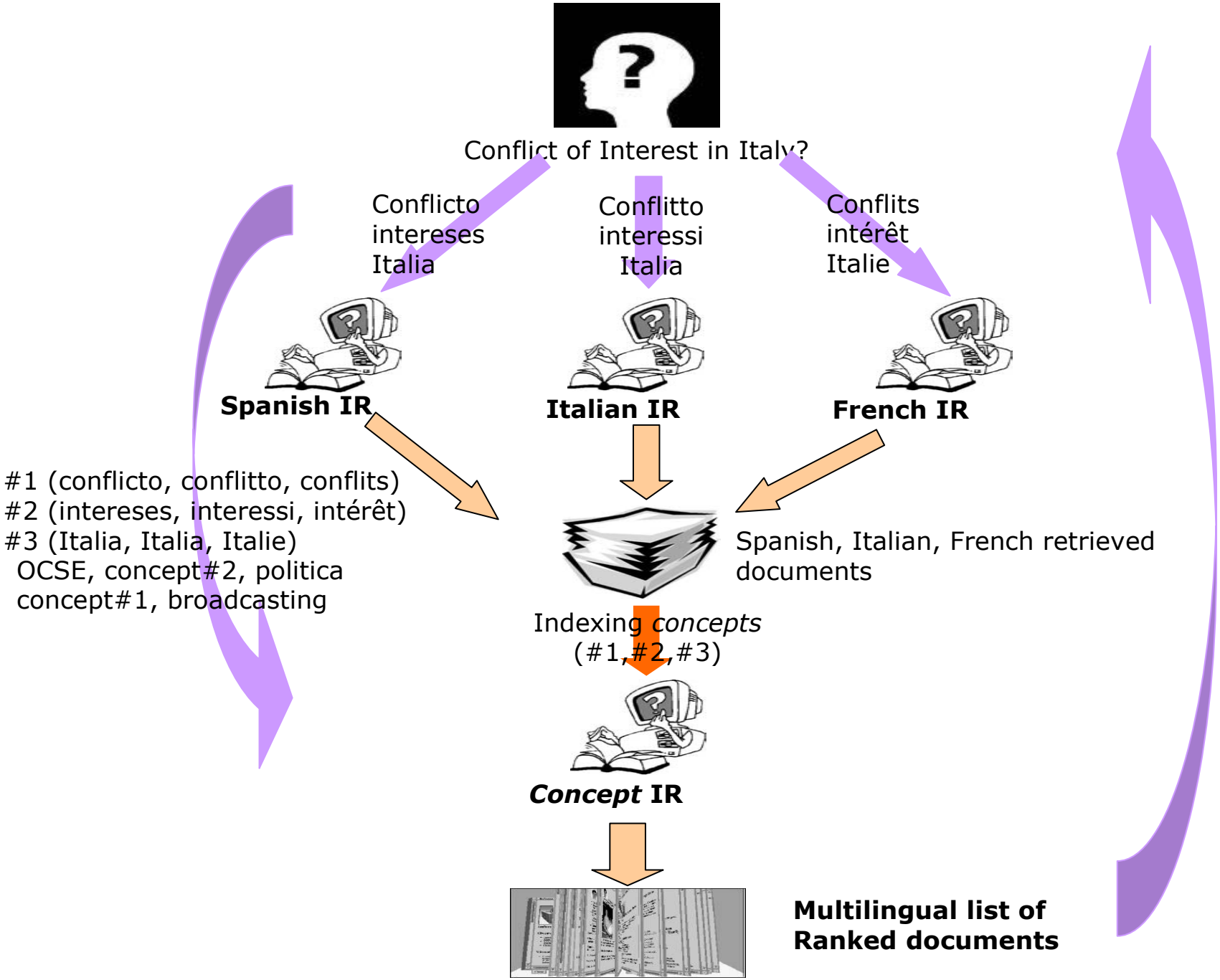
Results

Merging strategy	Experiment	AvgPrec
Round robin	unofficial	0.220
Raw scoring	unofficial	0.280
Formula 2 (logistic regression)	UJAMLRL	0.277
Formula 1 (raw mixed 2-step RSV)	**UJAMLRSV2**	**0.334**
Formula 3 (logistic regression and 2-step RSV)	UJAMLRL2P	0.333
Formula 4 (logistic regression and 2-step RSV)	UJAMLRL3P	0.301

Global pseudo-relevance feedback

- The idea:
  - Since 2-step RSV creates a only index for all collections and it returns a only multingual list of documents, why don't apply PRF with such index?

- The implementation:
  1. Merge the document rankings using 2-step RSV.
  2. Apply blind relevance feedback to the top-N documents ranked into the multilingual list of documents.
  3. Add the top-N more meaningful terms to the query.
  4. Expand the concept query with the selected terms.
  5. Apply again 2-step RSV over the ranked lists of documents, but by using the expanded query instead of the original query.


Conflict of Interest in Italy?
Conflicto intereses Italia
Conflitto interessi Italia
Conflits intérêt Italie
Spanish IR
Italian IR
French IR
#1 (conflicto, conflitto, conflits)
#2 (intereses, interessi, intérêt)
#3 (Italia, Italia, Italie)
OCSE, concept#2, politica
concept#1, broadcasting
Spanish, Italian, French retrieved documents
Indexing *concepts* (#1,#2,#3)
Concept IR
Multilingual list of Ranked documents

Global pseudo-relevance feedback

Merging strategy	AvgPrec	
	without global BRF	with global BRF
Formula 1 (raw mixed 2-step RSV)	0.334	0.331
Formula 3 (logistic regression and 2-step RSV)	0.333	0.332
Formula 4 (logistic regression and 2-step RSV)+global BRF	0.301	0.309

- But global PRF doesn't work:
  - Usually, blind relevance feedback is poorly suited to CLEF document collections.
  - We use the expanded query to apply 2-step RSV re-weighting the documents retrieved for each language, but the list of retrieved documents does not change (it only changes the score of such documents).

Conclusions and future work

- 2-step RSV merging algorithm works well with Machine Translation!
- We propose a new word-level aligment algorithm based on MT
- In the future:
  - Partially aligned queries → integration of two scores for documents must be improved by using other algorithms (normalized versions of raw mixed 2-step RSV, SVM and neural and bayesian networks...)
  - Global PRF idea must be more investigated