Language-dependent and Language-independent Approaches to Document Retrieval

Jaap Kamps and Maarten de Rijke

University of Amsterdam

Cross-Language Evaluation Forum 2003

Trondheim, Norway, August 21, 2003

Cross-Language Evaluation Forum (CLEF)

- CLEF greatly helped our IR research:

  2000 only postsubmission experiments
  2001 monolingual
  2002 mono-, bilingual, and domain-specific
  2003 mono-, bi-, multilingual, domain-specific and QA

- Thank you to all the CLEF organizers!

Overview

- Language-dependent Approaches
  - ⋆ Stemming
  - ⋆ Compound splitting
- Language-independent Approaches
  - ⋆ n-Gramming
- Combining Approaches
  - ⋆ Monolingual
  - ⋆ Multilingual
- Domain-Specific Information Retrieval

Experimental Set-up

- FlexIR system:
  - ⋆ Default settings: Lnu.ltc ($slope = 0.2$) with Rocchio feedback.
- Baseline: a vanilla word-based run.

Baseline: word-based run.									
	NL	EN	FI	FR	DE	IT	RU	ES	SV
Wrd	0.4800	0.4483	0.3175	0.4313	0.3785	0.4631	0.2551	0.4405	0.3485

- Conclusions
  - ⋆ Baseline is fairly high performing run for most languages.

Stemming Algorithm

- Martin Porter's Snowball project.
  - ⋆ Snowball is string-processing language.
  - ⋆ Porter-style stemming algorithms for all 9 CLEF languages.
- Example: `Information Retrieval` $\leadsto$ `inform retriev`
- In English: mixed evidence on the effectiveness.

Stemming Results

Snowball stemming algorithms.									
	NL	EN	FI	FR	DE	IT	RU	ES	SV
Wrd	0.4800	0.4483	0.3175	0.4313	0.3785	0.4631	0.2551	0.4405	0.3485
Stm	0.4652	0.4273	0.3998	0.4511	0.4504	0.4726	0.2536	0.4678	0.3707
%Ch	-3.1	-4.7	+25.9	+4.6	+19.0	+2.1	-0.6	+6.2	+6.4
Signif.	-	-	$\star$	-	$\star\star\star$	-	-	$\star$	-

- Conclusions
  - $\star$ Mixed results:
    - $*$ decrease for NL, EN, RU;
    - $*$ increase for FI, FR, DE, IT, ES, SV.
  - $\star$ Improvements for FI, DE, ES are significant.

Compound Splitting

- We decompound the compound-rich languages
  - ⋆ Dutch, German, Finnish, and Swedish.
- The potential compound-parts are simply the words occurring in the corpus.
- We also use their associated collection frequencies.
  - ⋆ We only regard compound parts that have a higher collection frequency than the compound itself.
- We add the compound parts but retain also the compound:

  Dutch *boekenkast* $\rightsquigarrow$ *boekenkast boek kast*

Compound Splitting Results

Compound splitting and stemming algorithms.									
	NL	EN	FI	FR	DE	IT	RU	ES	SV
Wrd	0.4800	0.4483	0.3175	0.4313	0.3785	0.4631	0.2551	0.4405	0.3485
Stm	0.4652	0.4273	0.3998	0.4511	0.4504	0.4726	0.2536	0.4678	0.3707
C.S	0.4984	–	0.4453	–	0.4840	–	–	–	0.3957
%Ch	+3.8	-4.7	+40.3	+4.6	+27.9	+2.1	-0.6	+6.2	+13.5
Signif.	-	-	⋆⋆⋆	-	⋆⋆⋆	-	-	⋆	-

- Conclusions
  - ⋆ Positive results overall:
    - ∗ NL improves; and
    - ∗ FI, DE, SV improve further.
  - ⋆ The (Split+)Stemmed runs improve for all languages, except for EN and the low-performing RU.

n-Gramming

- Zero-knowledge language independent runs were generated using character n-grams.
- We used with $n = 4$ for all languages; n-grams were not allowed to cross word boundaries.
- We add the n-grams but retain also the full words:

  `Information Retrieval` $\rightsquigarrow$ `information info nfor form orma rmat mati atio tion retrieval retr etri trie riev ieva eval`

n-Gramming Results

4-Gramming.									
	NL	EN	FI	FR	DE	IT	RU	ES	SV
Wrd	0.4800	0.4483	0.3175	0.4313	0.3785	0.4631	0.2551	0.4405	0.3485
4-Gr	0.4996	0.4119	0.4905	0.4616	0.5005	0.4227	0.3030	0.4733	0.4187
%Ch	+4.1	-8.1	+54.5	+7.0	+32.2	-8.7	+18.8	+7.4	+20.1
Signif.	-	*(!)	***	-	***	-	*	*	*

- Conclusions
  - ⋆ Decrease in performance for EN, IT; the rest improves.
  - ⋆ The improvement is significant for FI, DE, RU, ES, SV.
  - ⋆ The decrease is significant for EN.

Combining Approaches

- We combine both types of approaches.
- A weighted combination was produced as follows.
  - ⋆ rerank the similarity values in $[0, 1]$.
  - ⋆ use a linear interpolation factor for the relative weight of a run.
- For equal weights, this is the combSUM function of Fox and Shaw.
- The interpolation factors were obtained from experiments on the CLEF 2002 data sets.

Combination Results

Combination of Language-dependent and Language-independent Approaches.									
	NL	EN	FI	FR	DE	IT	RU	ES	SV
Wrd	0.4800	0.4483	0.3175	0.4313	0.3785	0.4631	0.2551	0.4405	0.3485
Sp+St	0.4984	0.4273	0.4453	0.4511	0.4840	0.4726	0.2536	0.4678	0.3957
4-Gr	0.4996	0.4119	0.4905	0.4616	0.5005	0.4227	0.3030	0.4733	0.4187
Comb	0.5072	0.4575	0.5236	0.4888	0.5091	0.4781	0.2988	0.4841	0.4371
%Ch	+5.7	+2.1	+64.9	+13.3	+34.5	+3.2	+17.1	+9.9	+25.4
Signif.	-	-	$\star\star\star$	$\star\star$	$\star\star\star$	-	$\star$	$\star\star\star$	$\star$

- Conclusions
  - $\star$ All languages improve over the baseline, even EN!
  - $\star$ The improvement for FI, FR, DE, RU, ES, SV is significant.

Multilingual Base-Runs

- We used the English topic set for all the multilingual runs.
  - ⋆ WorldLingo machine translation for NL, FR, DE, IT, ES.
  - ⋆ Babylon online dictionary for SV.
  - ⋆ PROMT-Reverso machine translation for RU.
- We created bilingual base-runs with the same settings as for the monolingual task.

Multilingual Base-Run Results

Bilingual runs using EN topic set.									
	NL	EN	FI	FR	DE	IT	RU	ES	SV
Wrd	0.3554	–	–	0.3547	0.3378	0.3810	0.1379	0.3246	0.1187
Sp+St	0.4043	–	–	0.3567	0.3968	0.3860	0.2270	0.3588	0.1898
4-Gr	0.3690	–	–	0.3762	0.4228	0.3801	0.1983	0.3775	0.2371
Comb	0.3971	–	–	0.3951	0.4479	0.3927	0.2195	0.3888	0.2478
Mono	0.4800	0.4483	0.3175	0.4313	0.3785	0.4631	0.2551	0.4405	0.3485

- Conclusions
  - ⋆ Improvements for bilingual runs mimick the monolingual runs.
  - ⋆ Exception is RU: Stemming even outperforms 4-gramming.
  - ⋆ The bilingual runs approach the monolingual baseline score, and exceed it for DE.

Multilingual Run

- Three sets of experiments
  - ⋆ Small multilingual task (EN,FR,DE,ES).
  - ⋆ Large multilingual task (Small + NL,IT; no FI/SV).
  - ⋆ Large multilingual task (Small + NL,IT,SV; no FI).
- Three unweighted combinations per experiment.

  For each language:
  - ⋆ a 4-grammed run.
  - ⋆ a weighted combination 4-grammed/Split+Stemmed run.
  - ⋆ both a 4-grammed run and a Split+Stemmed run.

Multilingual Results

Multilingual combinations.			
	Multi-4	Multi-8	
		No Fl/SV	No Fl
4-Gram	0.2953	0.2417	0.2467
Combined 4-Gram/(Split+)Stem	**0.3341**	**0.2797**	0.2871
4-Gram and (Split+)Stem	0.3292	0.2755	**0.2884**

- Conclusions
  - ⋆ For the small task, the weighted combinations outperform the unweighted combination of all base-runs.
  - ⋆ For seven languages, the unweighted combination of all base-runs scores best.

Domain Specific Information Retrieval

- Experimented with re-ranking initially retrieved documents.
- Key idea is to exploit the human assigned keywords.
  - ⋆ Use LSI-style approach to derive vectors for all the keywords.
  - ⋆ Determine a vector for a document as the mean of its keywords.
  - ⋆ Determine a vector for a topic as the weighted mean of the top 10 documents.
- Rerank the initially retrieved documents based on
  - ⋆ the similarity score and
  - ⋆ the distance between document and topic vectors.

GIRT Results

Domain Specific results.				
	Base-Run	Rerank	% Change	Signif.
Words	0.2360	0.2863	+21.31%	***
Stemmed	0.2832	0.3361	+18.68%	***
4-Gram	0.3449	**0.3993**	+15.77%	***

- Conclusions
  - ⋆ All three runs improve significantly.
  - ⋆ The gain is additional to blind feedback.

CLEF 2003 Results

- Document retrieval in non-English is different.
  - ⋆ Stemming and compound splitting helps.
  - ⋆ Character n-gramming helps.
- Combination generally exploits the best of both worlds.
- Approaches also beneficial for multilingual retrieval.
  - ⋆ Importance of weighting decreases with the number of languages.
- Reranking strategy improves domain-specific retrieval.

End!