Dublin City University at CLEF 2006: Cross-Language Speech Retrieval (CL-SR) Experiments

Gareth J. F. Jones, Ke Zhang and Adenike M. Lam-Adesina

Centre for Digital Video Processing & School of Computing

Dublin City University, Dublin 9, Ireland

Overview

- Motivation
- Retrieval of Multi-Field Documents
  - Multi-Field Term Weighting
  - Multi-Field Relevance Feedback
- Results using Multi-Field Retrieval Methods
- Results using Summary-Based Pseudo Relevance Feedback
- Conclusions and Further Work

Motivation

- The CL-SR documents comprise a number of different fields derived from different sources.
- The different fields have different utility for Speech Retrieval.
- Question arises: can Speech Retrieval be improved by appropriate combination of the fields?

Very interesting recent paper on this topic:

S. E. Robertson, H. Zaragoza and M. Taylor, *Simple BM25 Extension to Multiple Weighted Fields*, Proceedings of the 13th ACM CIKM, pp42-49, 2004.

Retrieval of Multi-Field Documents

Each “document” segment has associated with it the following fields:

- ASR2003A, ASR2004A, ASR2006A, ASR2006B: automatic transcription of the spoken content.
- AKW1, AKW2: two assigned sets of keywords generated automatically.
- MKW: one assigned set of manually generated keywords.
- SUMMARY: a short three sentence manually written summary of each document.
- NAME: manually determined list of the names of all the individuals appearing in the interview.

For some documents the NAME field is empty.

Retrieval of Multi-Field Documents

Two approaches are typically taken to retrieval of documents with multiple fields:

- Pretend they are not multi-field documents!

  Simply merge the multiple fields, effectively losing the document field structure, and then perform standard information retrieval.

- Index the fields separately, and then retrieve separate ranked lists for each field.

  Then combine the lists, typically in a linear sum, to arrive at a final document score.

  This can lead to problems for term weighting in functions with nonlinear treatment of $tf(i,j)$ such as BM25.

Retrieval of Multi-Field Documents

- Most modern term weighting functions include a nonlinear $tf(i,j)$ component.
- This is desirable since the information gained on observing a term once is greater than that on each subsequent occasion.

Consider the standard BM25 weighting function,

$$cw(i,j) = \frac{tf(i,j) \times (k_1 + 1)}{k_1((1-b) + b \times ndl(j)) + tf(i,j)} cfw(i)$$

Retrieval of Multi-Field Documents

The basic $tf(i,j)$ function for a document of average length with maximal length normalisation is,

$$f(tf(i,j)) = \frac{tf(i,j)}{k_1 + tf(i,j)}$$

- In BM25 the term frequency saturates after a few occurrences.
- The rate at which the saturation point is reached is controlled by the k_1 factor.

Retrieval of Multi-Field Documents

Simple linear summation of scores across multiple fields breaks the nonlinear $tf(i,j)$ relation.

For example, for a query term in a document with Title $tf(i,j[1]) = 1$ and Body $tf(i,j[2]) = 2$.

For a standard unstructured document these will be combined to give an overall $tf(i,j) = 3$ in a single BM25 combined weight for this term i in document j.

This works fine - but we've lost any information from the document structure.

Retrieval of Multi-Field Documents

Define a field weight $v[f]$.

Suppose we wish to weight the Title $v[f] = 2$ and the Body $v[f] = 1$.

This should boost the weight of this term somewhat overall in the matching score of the document.

The linear combination of scores would give the number high score for this term shown as ScoreComb.

This is equivalent to an effective $tf(i,j)$ contribution of:
$2 \times f(BM25_{Title}(tf(i,j[1]) = 1) + f(BM25_{Body}(tf(i,j[2]) = 2).$

Thus a document matching a single query term over several fields could score much higher than a document matching several terms in one field only.

Multi-Field Term Weighting

Proposed method: Robertson et al. 2004.

Modify standard ranking functions to exploit multiple weighted fields, while satisfying the following requirements:

- **preserve term frequency non-linearity** which has been shown repeatedly to improve retrieval performance.
- **give a simple interpretation** to collection statistics and to document length incorporating field weights.
- **revert to the unstructured case** when field weights are set to 1.

Multi-Field Term Weighting

Combine the term frequencies of the different fields by forming a linear combination weighted by the corresponding field weights:

$$\mathbf{j}' = \sum_{f=1}^{K} v[f].j[f]$$

where $j[f]$ is the field f of document j.

From the earlier example, combining the term frequencies and field weights $2 + 2 \times 1 = 4$ gives a slight boost to the weight of the term in each field.

Multi-Field Relevance Feedback

Standard okapi (BM25) relevance feedback selects top ranked expansion terms using the Robertson offer weight $ow(i)$.

$$ow(i) = r(i) \times rw(i)$$

where,

$$rw(i) = \log \frac{(r(i) + 0.5)(N - n(i) - R + r(i) + 0.5)}{(n(i) - r(i) + 0.5)(R - r(i) + 0.5)}$$

This is ordinarily applied to non-structured documents, and is unaffected by the multi-field weighted $tf(i, j)$ combination.

Multi-Field Relevance Feedback

Proposed multi-field method:

- Merge fields as described above and perform baseline retrieval for the tpoic.
- Calculate ow separately for each field of the original document, but where document rank is determined using the merged document.
- Ranked ow list for each field is then normalised, and then summed to form a single merged ow list from the expansion terms are selected. Fields optionally be weighted in the summation stage.

The objective is to favour selection of expansion terms which are ranked highly by multiple fields, rather than those which may obtain a high ow value based on their association with a minority of the fields.

Results using Multi-Field Retrieval Methods

Development runs using CLEF 2005 CL-SR topics.

System parameters for submitted automatic only field experiments:
$k_1 = 5.2$ and $b = 0.2$, and document fields were weighted as follows:

- ASR2006B $\times$ 2;
- Autokeyword1 $\times$ 1;
- Autokeyword2 $\times$ 1.

For pseudo-relevance feedback (PRF) 30 expansion terms are added, with original terms upweighted by 3.0.

Submitted Weighted Auto-Only Field Combination

TD	Recall	MAP	P5	P10	P30
Baseline	1290	0.071	0.163	0.163	0.149
New PRF*	1361	0.073	0.152	0.142	0.146

Monolingual English with parameters trained on CLEF 2005 data.

TD	Recall	MAP	P5	P10	P30
Baseline	1070	0.047	0.119	0.113	0.106
New PRF*	1097	0.047	0.106	0.094	0.102

French-English bilingual with parameters trained on CLEF 2005 data.

Data Fusion Auto-Only Field

TD	Recall	MAP	P5	P10	P30
Baseline	1299	0.080	0.212	0.197	0.166
Std PRF	1291	0.071	0.194	0.164	0.137

Simple list summation, monolingual English (CLEF 2005 training).

TD	Recall	MAP	P5	P10	P30
Baseline	1250	0.078	0.255	0.218	0.169
Std PRF	1294	0.075	0.224	0.197	0.154

Weighted list summation, monolingual English (CLEF 2005 training).

Auto-Only Field Combination

TD	Recall	MAP	P5	P10	P30
Baseline	1399	0.077	0.182	0.164	0.160
Std PRF	1346	0.072	0.170	0.182	0.157
New PRF	1390	0.078	0.170	0.164	0.163

Unweighted combination, monolingual English (CLEF 2005 training).

TD	Recall	MAP	P5	P10	P30
Baseline	1290	0.071	0.163	0.163	0.149
Std PRF	1292	0.064	0.164	0.158	0.151
New PRF*	1361	0.073	0.152	0.142	0.146

Weighted combination, monolingual English (CLEF 2005 training).

Results using Multi-Field Retrieval Methods

To explore the stability of the system settings.

Development runs using CLEF 2006 CL-SR topics.

System parameters modified as follows: $k_1 = 40.0$ and $b = 0.3$, and document fields were weighted as before:

ASR2006B $\times$ 2;

Autokeyword1 $\times$ 1;

Autokeyword2 $\times$ 1.

For pseudo-relevance feedback (PRF) 40 expansion terms are added, with original terms upweighted by 3.0.

Optimised Weighted Auto-Only Field Combination

TD	Recall	MAP	P5	P10	P30
Baseline	1335	0.080	0.224	0.215	0.169
New PRF	1379	0.094	0.188	0.206	0.184

Monolingual English with only auto document fields (CLEF 2006 training).

TD	Recall	MAP	P5	P10	P30
Baseline	1110	0.050	0.121	0.127	0.123
New PRF	1167	0.055	0.127	0.142	0.124

French-English bilingual with only auto document fields (CLEF 2006 training).

Optimised Data Fusion Auto-Only Field

TD	Recall	MAP	P5	P10	P30
Baseline	1294	0.081	0.218	0.212	0.169
Std PRF	1282	0.076	0.176	0.170	0.169

Simple list summation, monolingual English (CLEF 2006 training).

TD	Recall	MAP	P5	P10	P30
Baseline	1249	0.080	0.0242	0.230	0.170
Std PRF	1288	0.080	0.194	0.212	0.173

Weighted list summation, monolingual English (CLEF 2006 training).

Optimised Auto-Only Field Combination

TD	Recall	MAP	P5	P10	P30
Baseline	1348	0.084	0.206	0.215	0.176
Std PRF	1368	0.086	0.121	0.200	0.188
New PRF	1410	0.090	0.200	0.206	0.175

Unweighted combination, monolingual English (CLEF 2006 training).

TD	Recall	MAP	P5	P10	P30
Baseline	1335	0.080	0.224	0.215	0.169
Std PRF	1326	0.076	0.230	0.212	0.171
New PRF	1379	0.094	0.188	0.206	0.184

Weighted combination, monolingual English (CLEF 2006 training).

Results using Multi-Field Retrieval Methods

Development runs using CLEF 2005 CL-SR topics.

System parameters for submitted automatic only field experiments: $k_1 = 6.2$ and $b = 0.4$, and document fields were weighted as follows:

- Name field $\times$ 1;
- Manualkeyword field $\times$ 10;
- Summary field $\times$ 10;
- ASR2006B $\times$ 2;
- Autokeyword1 $\times$ 1;
- Autokeyword2 $\times$ 1.

For pseudo-relevance feedback (PRF) 20 expansion terms are added, with original terms upweighted by 3.0.

Submitted Weighted All-Field Combination Results

TD	Recall	MAP	P5	P10	P30
Baseline:	1844	0.223	0.366	0.293	0.255
New PRF*	1864	0.202	0.321	0.288	0.252

Monolingual English with all document fields (CLEF 2005 training).

TD	Recall	MAP	P5	P10	P30
Baseline	1491	0.158	0.306	0.256	0.204
New PRF*	1567	0.160	0.291	0.252	0.199

French-English bilingual with all document fields (CLEF 2005 training).

Results using Multi-Field Retrieval Methods

Development runs using CLEF 2006 CL-SR topics.

System parameters modified as follows: $k_1 = 10.5$ and $b = 0.35$, and document fields were weighted as follows:

Name field $\times$ 1;

Manualkeyword field $\times$ 5; *

Summary field $\times$ 5; *

ASR2006B $\times$ 1; *

Autokeyword1 $\times$ 1;

Autokeyword2 $\times$ 1,

For pseudo-relevance feedback (PRF) 40 expansion terms are added, with original terms upweighted by 33.0.

Optimised Weighted All-Field Combination Results

TD	Recall	MAP	P5	P10	P30
Baseline:	1908	0.234	0.364	0.342	0.303
New PRF	1929	0.243	0.364	0.370	0.305

Monolingual English with all document fields (CLEF 2006 training).

TD	Recall	MAP	P5	P10	P30
Baseline	1560	0.172	0.315	0.267	0.231
New PRF	1601	0.173	0.315	0.267	0.225

French-English bilingual with all document fields (CLEF 2006 training).

Summary-Based Pseudo Relevance Feedback

- Unweighted document field combination.
- Okapi BM25 term weighting with document summaries used for PRF.

  Basically the same system used by DCU for CLEF 2005.

  Details in the CLEF 2005 and CLEF 2006 papers.

Summary-Based Pseudo Relevance Feedback

TDN	Recall	MAP	P10	P30
Baseline	1832	0.246	0.391	0.321
PRF*	1895	0.277	0.439	0.357

Results for monolingual English with all document fields.

Results using Summary-Based Pseudo Relevance Feedback

TDN	Recall	MAP	P10	P30
Baseline	633	0.029	0.069	0.068
PRF	993	0.047	0.118	0.107

Results for monolingual English with only auto document fields.

TD	Recall	MAP	P10	P30
Baseline	627	0.025	0.069	0.061
PRF	900	0.039	0.091	0.089

Results for monolingual English with only auto document fields.

Conclusions and Further Work

- Exploitation of document field weighting can be useful in term weighting and PRF for the CLEF CL-SR English task.
- Further investigation is required to make setting of system parameters more stable.
- It would be interesting to combine summary-based feedback methods with the field weighting methods.
- Perhaps we should try it for the Czech task!