TCD-DCU at TEL@CLEF 2009: Document Expansion, Query Translation and Language Modeling

Johannes Leveling[1], Dong Zhou[2], Gareth F. Jones[1], and Vincent Wade[2]

[1] Centre for Next Generation Localisation
School of Computing
Dublin City University
Dublin 9, Ireland
{johannes.leveling, gareth.jones}@computing.dcu.ie

[2] Centre for Next Generation Localisation
Computer Science Department
Trinity College Dublin
Dublin, Ireland
{dong.zhou, vincent.wade}@cs.tcd.ie

Abstract

For the multilingual ad-hoc document retrieval track (TEL@CLEF) at at the Cross-Language Retrieval Forum (CLEF) Trinity College Dublin and Dublin City University participated in collaboration. Our retrieval experiments focus on i) investigating document expansion using an entry vocabulary module, ii) translating queries with Google translate and a statistical MT system, and iii) investigating language modeling as a retrieval method. The major results are that the document expansion approach did not increase MAP; topic translation using the statistical MT system resulted in about 70% of the mean average precision (MAP) achieved when using Google translate for topic translation, and language modeling performs equally or better in comparison with BM25. The bilingual retrieval French and German to English experiments obtained 89% and 90% of the best MAP for monolingual English.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods; Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation; Search process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Experimentation, Measurement, Performance

Keywords

Information Retrieval, Document Expansion, Machine Translation, Language Modeling

1 Introduction

The TEL (The European Library) task at CLEF is concerned with ad-hoc information retrieval (IR) [1]. The TEL document subcollections in English, German, and French consist of about 1 million bibliographic records. The data is provided by the archives of the British Library (English), the Austrian National Library (German), and Bibliothèque nationale de France (French) of The European Library. TEL documents follow the Dublin Core metadata standard and contain multiple fields including title, contributors, language, and subject terms.

Our IR experiments for the ad-hoc task at CLEF 2009 aim at investigating several aspects of retrieval: 1. employing and evaluating EVM [8] for document expansion (DE) to obtain longer documents for the TEL collection (see [4] for a comparison of query and document expansion), 2. applying a statistical MT system [7] for topic translation and comparing it to Google translate, and 3. comparing language modeling (LM) [5] as a retrieval method to Okapi BM25 [10].

2 Retrieval Experiments

2.1 Topic Processing

The Lemur toolkit[1] was employed to index and retrieve documents. Two different retrieval models were employed: BM25 [10] with default parameters ($b = 1.2$, $k1 = 2.0$, $k3 = 7$) and language modeling with Jelinek-Mercer smoothing [5, 6]. The text of different fields was extracted and processed to produce a single flat index:

- all: all fields
- set1: dc:title, dc:description, dcterms:alternative, and dc:subject.
- set2: dc:language, dc:identifier, dc:rights, dc:type, dc:creator, dc:publisher, dc:date, dc:relation, dc:contributor, dcterms:issued, dcterms:extent, mods:location
- set3: dc:language, dc:identifier, dc:rights, dc:type, dc:creator, dc:publisher, dc:date, dc:contributor, mods:location
- set4: dc:language, dc:identifier, dc:rights, dc:type, dc:creator, dc:publisher, dc:date, dc:contributor, dcterms:spatial, dcterms:isPartOf, dcterms:edition, dcterms:issued, dcterms:available, mods:location

All other document fields were discarded. Prior to indexing the documents, their contents were preprocessed with the Snowball stemmer[2] for the corresponding language and stopwords were removed.

For most runs, pseudo-relevance feedback was applied for query expansion (QE): the top ten ranked documents and 30 terms were used for BM25 and the top five documents and 20 added terms for LM. A variant of query expansion using information an external resource was also explored (QE2) for bilingual retrieval. The top 10 results for the query in the source language were extracted and translated with Google translate. Highly co-occurring terms were extracted for query expansion [2], using the mutual information to calculate co-occurrence and select the highest score for target translation.

For the bilingual retrieval experiments, topics were translated using either Google translate (GT) [3] or a statistical machine translation system (MT) [7].

2.2 Document Expansion using EVM

The TEL collection contains documents of ranging from very short documents, because the presence or absence of a field with bibliographic information leads to documents with varying length. Furthermore, some fields contain information not in natural language (i.e. alphanumeric codes or classifications). The

[1] http://www.lemurproject.org/
[2] http://snowball.tartarus.org/
[3] http://translate.google.com/

Table 1: Bias in long vs. short and classified vs. unclassified documents.

documents	rel		rel_ret		MAP	GMAP	P@10
all	2097	(100%)	2533	(100%)	0.3474	0.1875	0.4900
long	1160	(55%)	1396	(55%)	0.2048	0.0874	0.2400
short	937	(45%)	1137	(45%)	0.2096	0.0793	0.2500
with DDC	1582	(75%)	1815	(72%)	0.3326	0.1460	0.4060
without DDC	515	(25%)	718	(28%)	0.0909	0.0230	0.0840

main idea behind some of our experiments was to apply a document expansion method to obtain longer documents.

In the English documents, there are more than 585,000 fields containing a valid Dewey Decimal Code (DDC), and about 50% of all documents contain a corresponding field. The percentage of LCC in the English collection is considerably lower and DDC and LCC for the German and French collection are not present or occur only sparsely, so the experiments were focused on the DDC classification and the English document collection.

Before conducting the official experiments, we performed a test experiment with the English TEL documents. We divided the document collection into short (less than 80 characters) and long documents and into document with a DDC and without. Results for the test run based on CLEF 2008 data for the different sets of documents are shown in Table 1. For short and long documents, retrieval performance is very similar, but less short documents were assessed as relevant. In contrast, documents with DDC make up a large portion of relevant documents (72%), while about half of all English documents are associated with a DDC. Possible explanations might be that in previous experiments, the DDC has been treated as a separate index term which could be used in relevance feedback or that longer documents provide more context for relevance assessment. The relative and absolute performance for documents without DDC classification is lower. As a result of this analysis, we tried to expand documents via an automatic DDC classification to create documents with a more evenly distributed length.

The DDC is a hierarchical library classification. The classification system defines ten main classes, 100 divisions, and 1000 sections, each denoted by digits. For example, the DDC 627 represents the main class *"technology"*, division *"engineering and applied operations"*, section *"hydraulic engineering"*.

The main idea for document expansion was to train a classifier on documents containing a DDC and apply it to obtain classification codes for all other documents. All classification codes are then replaced with their natural language description, which is preprocessed and added to the index. The natural language descriptions are available in English only and originate from the OCLC web site[4]. The natural language description for these codes was compiled into a machine-readable format using the sources from OCLC. The resulting description contained all 1110 entries for the DDC of which 933 were actually used in the document collection. The documents were modified as follows: documents with a DDC are expanded by appending the natural language description of the DDC to their content; documents without a DDC are first classified using an EVM and then processed as described above.

Entry Vocabulary Modules (EVM, [8]) have been successfully employed to map uncontrolled vocabulary (free text) to a controlled vocabulary or classification for query expansion [9, 3]. EVM determine a ranking of most likely classifications. The top-ranked classification is used for document expansion. The EVM used for our experiments was trained on all documents with a DDC assigned to them. As the EVM returns a ranking of classification, only the top ranked DDC was considered and its description used to expand the documents.

[4] `http://www.oclc.org/dewey/`

Table 2: Results for monolingual and bilingual IR experiments for the ad-hoc task.

Run ID	source	target	description	rel_ret	MAP	GMAP	P@10
TCDDCU_EN1F	EN	EN	BM25, set1, QE	2075	0.3640	0.1926	0.5080
TCDDCU_EN2F	EN	EN	BM25, set1, QE, DE	1990	0.3426	0.1869	0.4980
TCDDCU_EN3	EN	EN	LM, set2, QE	2059	0.3696	0.2414	0.5060
TCDDCU_EN4	EN	EN	LM, all, QE	2122	0.3688	0.2675	0.5200
TCDDCU_FR1	FR	FR	BM25, set1	999	0.1783	0.0982	0.3340
TCDDCU_FR1F	FR	FR	BM25, set1, QE	1020	0.1831	0.0919	0.3420
TCDDCU_FR3	FR	FR	LM, set3, QE	1114	0.1758	0.0434	0.2327
TCDDCU_FR4	FR	FR	LM, all, QE	1135	0.1749	0.0417	0.2224
TCDDCU_DE1	DE	DE	BM25, set1	945	0.2329	0.1221	0.3540
TCDDCU_DE1F	DE	DE	BM25, set1, QE	1052	0.2561	0.1137	0.3580
TCDDCU_DE3	DE	DE	LM, set4, QE	1097	0.2686	0.1291	0.3840
TCDDCU_DE4	DE	DE	LM, all, QE	1063	0.2439	0.1258	0.3460
TCDDCU_DEEN1	DE	EN	LM, GT, set2, QE	1966	0.3333	0.1981	0.4420
TCDDCU_DEEN3	DE	EN	LM, GT+QE, set2, QE2	1895	0.2947	0.1351	0.3900
TCDDCU_FREN1F	FR	EN	BM25, GT, set1, QE	1827	0.3323	0.1761	0.4820
TCDDCU_FREN2	FR	EN	BM25, MT set1,	1523	0.2072	0.0533	0.3800
TCDDCU_FREN2F	FR	EN	BM25, MT, set1, QE	1681	0.2551	0.0497	0.3920

3 Results

Results for the ad-hoc IR experiments are shown in Table 2. Some experiments achieved a performance among the top five participants at the TEL track at CLEF 2009, i.e. run TCDDCU_DEEN1 was 4th in bilingual English (0.3333 MAP), run TCDDCU_DE3 was 4th in monolingual German (0.2686 MAP), and run TCDDCU_EN3 was 5th in monolingual English (0.3696 MAP).

In all cases, runs with blind relevance feedback to expand queries yield a higher MAP compared to the corresponding runs without blind feedback. The query expansion variant based on external information from web pages found by Google web search did not show the expected results as it degraded the performance (TCDDCU_DEEN3 vs. TCDDCU_DEEN1).

Obviously using only a subset of the document fields yields a slightly higher precision (e.g. TCDDCU_DE3 vs. TCDDCU_DE4).

BM25 and language modeling perform similar for the retrieval experiments in all languages. Because of small differences in the experimental setup (e.g. the fields indexed), some additional experiments will have to be conducted before testing for significant differences.

For the bilingual runs with target language English, 89.9% and 90.1% of the MAP for the best monolingual English runs was achieved for French and German, respectively. Using the MaTrEx system for topic translation achieves a MAP of 70.1% in comparison to topic translation by Google translate (TCDDCU_FREN2 vs. TCDDCU_FREN1).

4 Future Work

Future work will include an analysis of the accuracy of the DDC classification based on a manually extracted and annotated sample of the English document collection. Also, blind relevance feedback using external resources will be further investigated.

References

[1] Eneko Agirre, Giorgio M. Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters. CLEF 2008: Ad hoc track overview. In *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark*, 2008.

[2] Lisa Ballesteros and Bruce W. Croft. Resolving ambiguity for cross-language retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71, Melbourne, Australia, 1998. ACM. address = New York, USA.

[3] Benjamin Berghaus, Thomas Mandl, Christa Womser-Hacker, and Michael Kluck. An entry vocabulary module for a political science test collection. In Witold Abramowicz and Dieter Fensel, editors, *Business Information Systems, 11th International Conference, BIS 2008*, volume 7 of *Lecture Notes in Business Information Processing*, pages 1–11. Springer, Berlin, 2008.

[4] Bodo Billerbeck and Justin Zobel. Document expansion versus query expansion for ad-hoc retrieval. In Andrew Turpin and Ross Wilkinson, editors, *Proceedings of the Tenth Australasian Document Computing Symposium*, pages 34–41, Sydney, Australia, December 2005.

[5] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[6] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *34th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 310–318, Santa Cruz, USA, 1996. Morgan Kaufmann/ACL.

[7] Jinhua Du, Yifan He, Sergio Penkale, and Andy Way. MaTrEx: the DCU MT system for WMT 2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 95–99, Athens, Greece, 2009.

[8] Fredric C. Gey, Michael Buckland, Aitao Chen, and Ray R. Larson. Entry vocabulary – a technology to enhance digital search. In *Proceedings of the First International Conference on Human Language Technology*, San Diego, USA, March 2001.

[9] Vivien Petras. GIRT and the use of subject metadata for retrieval. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, volume 3491 of *LNCS*, pages 298–309. Springer, Berlin, 2005.

[10] Stephen E. Robertson, Steve Walker, Susan Jones, and Micheline Hancock-Beaulieu. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC 1994)*, Gaithersburg, USA, 1994.

Acknowledgments

Thanks to Andy Way's group at DCU for providing topic translations with the MaTrEx statistical MT system.
This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project.