Query Recovery of Short User Queries: On Query Expansion with Stopwords

Johannes Leveling and Gareth J. F. Jones
School of Computing, CNGL
Dublin City University
Dublin, Ireland
{jleveling, gjones}@computing.dcu.ie

ABSTRACT

User queries to search engines are observed to predominantly contain inflected content words but lack stopwords and capitalization. Thus, they often resemble natural language queries after case folding and stopword removal. Query recovery aims to generate a linguistically well-formed query from a given user query as input to provide natural language processing tasks and cross-language information retrieval (CLIR). The evaluation of query translation shows that translation scores (NIST and BLEU) decrease after case folding, stopword removal, and stemming. A baseline method for query recovery reconstructs capitalization and stopwords, which considerably increases translation scores and significantly increases mean average precision for a standard CLIR task.

Categories and Subject Descriptors:
H.3.3 [INFORMATION STORAGE AND RETRIEVAL] Information Search and Retrieval—*Query formulation, Search process*

General Terms: Experimentation, Performance, Measurement

Keywords: Query Reformulation, Query Expansion, CLIR

1. INTRODUCTION

Query processing for experimental information retrieval (IR) systems typically involves transforming the *original user query* (OQ) by successively applying *case folding* (CF), *stopword removal* (SR), and *stemming* (ST). However, real user queries to search engines usually consist of 2-3 words [5, 6] and seldom take the form of full sentences or questions [4]. Thus, they already resemble results from query preprocessing (as shown in Table 1) in that they typically lack capitalization and stopwords, but still contain full word forms. This paper proposes query recovery (QR), a method which seeks to restore a fully capitalized query with syntactic structure from its input. For example, the query *embargo iraq* (topic C046) is transformed into *The embargo against Iraq*, which results in a better query translation for CLIR.

Reconstructing punctuation and capitalization has been applied to automatic speech recognition and machine translation (MT) [1, 2], but focuses on processing full text instead of short queries. Query modification for IR has been concerned with query expansion by adding content terms to the query (see, for example [7]). In contrast, QR aims to expand a query by adding stopwords and capitalization.

Copyright is held by the author/owner(s).
SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

2. CORPORA AND QUERY ANALYSIS

Our experiments and analyses are performed on the following corpora and data sets: The Excite query log (ENEx) of user queries, as distributed in the Pig query log analysis tool[1]; the 1M sentence English (EN1M) and 3M sentence German corpus from the Leipzig Corpora Collection[2]; the English and German Wikipedia article names (ENWi)[3]; and the titles of 160 English and German topics which have been used in ad-hoc retrieval experiments at CLEF from 2003-2006 (see, for example [3]).

Results of an analysis of this data are shown in Table 1, confirming that the average length of user queries (in column ENEx) is 2-3 words. In addition, the following observations can be made: user queries rarely contain stopwords, punctuation symbols (e.g. *"!"*, *"?"*), or numeric terms; special characters (e.g. quotation marks or *"-"*) often indicate queries with special syntax, e.g. a phrase search or exclusion of terms. Topic titles contain capitalization in expected places, e.g. at the beginning of sentences. Thus, the proportion of capitalized words is actually much higher in comparison with corpora containing full sentences. Compared to the original and processed CLEF topic titles (OQ-ST), user queries are most similar to topic titles after CF and SR with respect to their average length in tokens, the number of lowercase words, stopwords, and stems (base forms). Users often enter full word forms as query terms (52.9% stems, 47.1% non-stems for ENEx), assuming that the search engine will handle morphological variation or exact matching of query terms. In contrast, topic titles after stemming mostly consist of stems only (94.9% stems).

3. BASELINE QUERY RECOVERY

A user query can be represented as a sequence of content words W_i. Between any two tokens W_i and W_{i+1}, a stopword sequence S_i occurs (a special case is the empty sequence). Thus, a query is a sequence of content words and stopword sequences (S_0, W_1, S_1, ..., S_{n-1}, W_n, S_n).

The baseline QR method uses the 1M sentence English and 3M sentence German corpora and the Wikipedia article names as training data. The method consists of replacing

[1] `http://hadoop.apache.org/pig/`
[2] `http://corpora.uni-leipzig.de/`
[3] `http://dumps.wikimedia.org/`

Table 1: Analysis of English corpora and topics.

	Corpus			CLEF topic titles			
	ENEx	EN1M	ENWi	OQ	CF	SR	ST
entries	0.94M	1M	5.24M	160	160	160	160
tokens	2.45M	25.1M	16.8M	577	577	458	458
avg. length	2.6	25.1	3.2	3.6	3.6	2.9	2.9
uppercase [%]	0.7	13.8	66.6	45.8	0.00	0.0	0.0
lowercase [%]	81.8	70.6	17.7	50.6	96.4	99.2	99.2
numeric [%]	4.9	2.1	2.4	0.5	0.5	0.8	0.8
punct. [%]	6.8	11.2	5.3	2.3	2.3	0.0	0.0
special [%]	5.8	2.3	7.9	0.8	0.8	0.0	0.0
stopwords [%]	7.8	49.0	11.7	18.3	18.3	0.0	1.3
non-stopw. [%]	92.2	51.0	88.3	81.7	81.7	100.0	98.7
stem [%]	52.9	28.5	8.3	13.6	47.7	47.7	94.9
non-stem [%]	47.1	71.5	91.7	86.4	52.3	52.3	5.1

Table 2: NIST/BLEU scores for CLEF topics.

Processing	EN	DE→EN	DE	EN→DE
OQ	9.45/0.97	5.80/0.37	9.66/1.00	5.30/0.39
CF	4.65/0.22	4.26/0.21	2.47/0.07	5.07/0.37
SR	3.34/0.10	2.98/0.10	0.49/0.00	4.00/0.14
ST	1.15/0.00	2.09/0.06	0.20/0.00	1.67/0.00
QR	6.79/0.32	4.84/0.20	5.72/0.24	4.31/0.21

lowercase words W_i in the input with the most frequent capitalized variant found in the training corpus and inserting the most frequent stopword sequence S_i occurring between two words W_i and W_{i+1}. If W_i is unknown, its initial character is capitalized, according to the observation by [1] that most out-of-vocabulary words are proper nouns (which are capitalized in English and German). If W_i or W_{i+1} is unknown, the empty stopword sequence is selected for S_i.

4. QUERY TRANSLATION EXPERIMENTS

The effect of QR for NLP is investigated by evaluating the baseline method for query translation, which is a typical task for CLIR. Translation experiments and CLIR experiments are based on the CLEF topic titles (C041-C200), which are capitalized, contain stopwords and full word forms. For comparison with real user queries, the original topics are preprocessed by applying case folding (CF), stopword removal (SR), and stemming (ST). Table 2 shows NIST and BLEU scores for CLEF topics after query processing and translation by the Google translate web service[4]. The original parallel English (EN) and German (DE) topics and variants with corrected orthography were used as reference translations. Query translation scores decrease after each processing step, i.e. translating queries lacking stopwords, case information, or full word forms adversely affects MT quality. As expected, QR for short queries reverses the effects of CF and SR and increases the the quality of translations. For monolingual QR, the English (German) queries achieve 71.8% (59.2%) of the score for the original queries. For a translation of queries after QR to English (German), the baseline QR yields 83.4% (81.3%) of the NIST score for translating the original query. The QR scores are considerably higher than scores for processed queries.

[4]`http://translate.google.com/`

Table 3: MAP for CLIR on 160 CLEF topics.

	EN	DE→EN	DE	EN→DE
SR	0.241	0.221	0.350	0.335
QR	0.241	0.235 (+6.5%)*	0.350	0.353 (+5.2%)*
OQ	0.241	0.237 (+7.5%)	0.350	0.342 (+1.9%)

5. CLIR EXPERIMENTS

Results for IR experiments on the English and German CLEF ad hoc document collections are shown in Table 3. Significance using the Wilcoxon test with $p < 5\%$ is indicated by '*'. For comparison with an upper baseline, performance for the unprocessed original queries (OQ) is also shown. MAP does not change at all for monolingual IR experiments, i.e. applying processing steps to the query twice (e.g. stemming) will not affect IR results. For bilingual IR, a slight but significant increase in MAP is observed for query translation after QR (more specifically: after CF, SR, and QR have been applied) compared to query translation after SR (SR and CF). Stemming was not included as a preprocessing step because users do not typically enter stems in queries. For DE→EN, QR achieves almost the same MAP compared to using OQ, which demonstrates the usefulness of QR for CLIR. For EN→DE, MAP is even slightly higher, due to hyphenated compounds in the German translation of recovered topics, i.e. compound splitting.

6. CONCLUSIONS AND FUTURE WORK

The major findings are: User queries to search engines lack capitalization and stopwords, and are most similar to topic titles after CF and SR (e.g. in average length). Restoring capitalization and adding stopwords to user queries benefits MT and CLIR which was shown by calculating translation scores for various processing stages and after QR. Translation scores for the baseline QR are considerably higher than for preprocessed queries. The proposed baseline QR method serves as a proof of concept for different approaches at QR (e.g. with language models). Future work will include recovery of questions for question answering.

Acknowledgments

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/CE/I1142.

7. REFERENCES

[1] E. Brown and A. Coden. Capitalization recovery for text. In *Information Retrieval Techniques for Speech Applications*, volume 2273 of *LNCS*, pages 11–22. Springer, 2002.

[2] A. Gravano, M. Jansche, and M. Bacchiani. Restoring punctuation and capitalization in transcribed speech. In *ICASSP 2009*, pages 4741–4744. IEEE, 2009.

[3] G. M. D. Nunzio, N. Ferro, T. Mandl, and C. Peters. CLEF 2006: Ad hoc track overview. In *Evaluation of Multilingual and Multi-modal Information Retrieval, CLEF 2006*, volume 4730 of *LNCS*. Springer, 2007.

[4] S. Ozmutlu, H. Ozmutlu, and A. Spink. Are people asking questions of general web search engines? *Online Information Review*, 27(6):396–406, 2003.

[5] A. Spink, D. Wolfram, M. Jansen, and T. Saracevic. Searching the web: The public and their queries. *JASIST*, 52(3), 2001.

[6] J. Teevan, E. Adar, R. Jones, and M. Potts. History repeats itself: Repeat queries in Yahoo's query logs. In *SIGIR 2006*, pages 703–704. ACM, 2006.

[7] J. Xu and W. Croft. Query expansion using local and global document analysis. In *SIGIR '96*, pages 4–11. ACM, 1996.