CLEF-2006 CL-SR at Maryland: English and Czech

Jianqiang Wang
Department of Library and Information Studies
State University of New York at Buffalo, Buffalo, NY 14260
jw254@buffalo.edu

Douglas W. Oard
College of Information Studies and Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20740
oard@glue.umd.edu

Abstract

The University of Maryland participated in the English and Czech tasks. For English, one monolingual run using only fields based on fully automatic transcription (the required condition) and one (otherwise identical) cross-language run using French queries were officially scored. Three contrastive runs in which manually generated metadata fields in the English collection were indexed were also officially scored to explore the applicability of recently developed "meaning matching" approaches to cross-language retrieval of manually indexed interviews. Statistical translation models trained on European Parliament proceedings were found to be poorly matched to this task, yielding 38% and 44% of monolingual mean average precision for indexing based on automatic transcription and manually generated metadata, respectively. Weighted use of alternative translations yielded an apparent (but not statistically significant) 7% improvement over one-best translation when bi-directional meaning matching techniques were employed. Results for Czech were not informative in this first year of that task, perhaps because no accommodations were made for the unique characteristics of Czech morphology.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Measurement, Performance, Experimentation

Keywords

Speech Retrieval, Cross-Language Information Retrieval, Statistical Translation

1 Introduction

Previous experiments have shown that limitations in Automatic Speech Recognition (ASR) accuracy were an important factor degrading the retrieval effectiveness for spontaneous conversational

speech [2, 4]. In this year's CLEF CL-SR track, ASR text with lower word error rate (hence, better recognition accuracy) was provided for the same set of segmented English interviews used in last year's CL-SR track. Therefore, one of our goals was to determine the degree to which improved ASR accuracy could measurably improve retrieval effectiveness.

This year's track also introduced a new task, searching unsegmented Czech interviews. Unlike more traditional retrieval tasks, the objective in this case was to find points in the interviews that mark the beginning of relevant segments (i.e., points at which a searcher might wish to begin replay). A new evaluation metric, Generalized Average Precision (GAP), was defined to evaluate system performance on this task. The GAP measure takes into account the distance between each system-suggested start time in a ranked list and the closest start time found by an assessor—the greater the time between the two points, the lower the contribution of that match to a system score. A more detailed description of GAP and details of how it is computed can be found in the track overview paper. In this study, we were interested in evaluating retrieval based on overlapping passages, a retrieval techniques that has proven to be useful in more traditional document retrieval settings, while hopefully also gaining some insight into the suitability of the new evaluation metric.

2 Techniques

In this section, we briefly describe the techniques that we used in our study.

2.1 Combining Evidence

Both test collections provide several types of data that are associated with the information to be retrieved, so we tried different pre-indexing combinations of that data. For the English test collection, we combined the ASR text generated with the 2004 system and the ASR text generated with the 2006 system, and we compared the result with that obtained by indexing each alone. Our expectation was that the two ASR engines could produce different errors, and that combining their results might therefore yield better retrieval effectiveness. Thesaurus terms generated automatically using a kNN classifier offer some measure of vocabulary expansion, so we added them to the ASR text combination as well.

The English test collection also contains a rich set of metadata that was produced by human indexers. Specifically, a set of (on average, five) thesaurus terms were manually assigned, and a three-sentence summary was written for each segment. In addition, the first names of persons that were mentioned in each segment are available, even if the name itself was not stated. We combined all three types of human-generated metadata to create an index for a contrastive condition.

The Czech document collection contains, in addition to ASR text, manually assigned English thesaurus terms, automatic translations of those terms into Czech, English thesaurus terms that were generated automatically using a kNN classifier that was trained on English segments, and automatic translations of those thesaurus terms into Czech. We tried two ways of combining this data. Czech translations of the automatically generated thesaurus terms were combined with Czech ASR text to produce the required automatic run. We also combined all available keywords (automatic and manual, English and Czech) with the ASR text, hoping that some proper names in English might match proper names in Czech.

2.2 Pruning Bidirectional Translations for Cross-language Retrieval

In our previous study, we found that using several translation alternatives generated from bidirectional statistical translation could significantly outperform techniques that utilize only the most probable translation, but that using all possible translation alternatives was harmful because statistical techniques generate a large number of very unlikely translations [5]. The usual process for statistically deriving translation models is asymmetric, so we produce an initial bidirectional model by multiplying the translation probabilities between source words and target words from models trained separately in both directions and then renormalizing. This has the effect of driving

run name	umd.auto	umd.auto.fr0.9	umd.manu	umd.manu.fr0.9	umd.manu.fr0
CL-SR?	monolingual	CL-SR	monolingual	CL-SR	CL-SR
doc fields	ASR 2004A ASR 2006A autokey04A1A2	ASR 2004A ASR 2006A autokey04A1A2	NAME manualkey SUMMARY	NAME manualkey SUMMARY	NAME manualkey SUMMARY
MAP	0.0543	0.0209	0.235	0.1026	0.0956

Table 1: Mean average precision (MAP) for official runs, TD queries.

the modeled probabilities for translation mappings that are not well supported in both directions to relatively low values. Synonymy knowledge (in this case, from round trip translation using cascaded unidirectional mappings) is then used to aggregate probability mass that would otherwise be somewhat eclectically divided across translation pairings that actually share similar meanings. To prune the resulting translations, document-language synonym sets that share the meaning of a query word are then arranged in decreasing order of modeled translation probability and a cumulative probability threshold is used to truncate that list. In our earlier study, we found that a cumulative probability threshold of 0.9 typically results in near-peak cross-language retrieval effectiveness. Therefore, in this study, we tried two conditions: one-best translation (a threshold of zero) and multiple alternatives with a threshold of 0.9.

We used the same bilingual corpus that we used in last year's experiment. Specifically, to produce a statistical translation table from French to English, we used the freely available GIZA++ toolkit [3][1] to train translation models with the Europarl parallel corpus [1]. Europarl contains 677,913 automatically aligned sentence pairs in English and French from the European Parliament. We stripped accents from every character and filtered out implausible sentence alignments by eliminating sentence pairs that had a token ratio either smaller than 0.2 or larger than 5; that resulted in 672,247 sentence pairs that were actually used. We started with 10 IBM Model 1 iterations, followed by 5 Hidden Markov Model (HMM) iterations, and ending with 5 IBM Model 4 iterations. The result is a a three-column table that specifies, for each French-English word pair, the normalized translation probability of the English word given the French word. Starting from the same aligned sentence pairs we used the same process to produce a second translation table from French to English.

Due to the time and resource constraints, we were not able to try a similar technique for Czech this year. All of our processing for Czech was, therefore, monolingual.

3 English Experiment Results

The required run for the CLEF-2006 CL-SR track called for use of the *title and description fields* as a basis for formulating queries. We therefore used all words from those fields as the query (a condition we call "TD") for our five official submissions. Stopwords in each query (as well as in each segment) were automatically removed (after translation) by InQuery, which is the retrieval engine that we used for all of our experiments. Stemming of the queries (after translation) and segments was performed automatically by InQuery using kstem. Statistical significance is reported for $p < 0.05$ by a Wilcoxon signed rank test for paired samples.

3.1 Officially Scored Runs

Table 1 shows the experiment conditions and the Mean Average Precision (MAP) for the five official runs that we submitted. Not surprisingly, every run based on manually created metadata statistically significantly outperformed every run based on automatic data (ASR text and automatic keywords).

[1] http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

run name	umd.auto	umd.asr04a	umd.asr06a	umd.asr06b
MAP	0.0543	0.0514	0.0517	0.0514

Table 2: Mean average precision (MAP) for 4 monolingual automatic runs, TD queries.

As Table 1 shows, our one officially scored automatic CLIR run, umd.auto.fr0.9 (with a threshold of 0.9) achieved only 38% of the MAP of the corresponding monolingual MAP, a statistically significant difference. We suspect that domain mismatch between the corpus used for training statistical translation model and the document collection might be a contributing factor, but further investigation of this hypothesis is clearly needed. A locally scored one-best contrastive condition (not shown) yielded about the same results.

Not surprisingly, combining first names, segment summaries, and manually assigned thesaurus terms produced the best retrieval effectiveness as measured by MAP. Among the three runs with manually created metadata, umd.manu.fr0.9 and umd.manu.fr0 are a pair of comparative cross-language runs: umd.manu.fr0.9 had a threshold of 0.9 (which usually led to the selection of multiple translations), while umd.manu.fr0 used only the most probable French translation for each English query word. These settings yielded 44% and 41% of the corresponding monolingual MAP, respectively. The 7% apparent relative increase in MAP with a threshold of 0.9 (compared with one-best translation) was not found to be statistically significant.

3.2 Additional Locally Scored Runs

In additional to the officially scored monolingual run umd.auto that used four automatically generated fields (ASRTEXT2004A, ASRTEXT2006A, AUTOKEYWORD2004A1, and AUTOKEYWORD2004A2), we scored three additional runs based on automatically generated data locally: umd.asr04a, umd.asr06a, and umd.asr06b. These runs used only the ASRTEXT2004A, ASRTEXT2006A, or ASRTEXT2006B fields, respectively (ASRTEXT2006A is empty for some segments; in ASRTEXT2004B those segments are filled in with reportedly less accurate data from ASRTEXT2004A). Table 2 shows the MAP for each of these runs and (again) for umd.auto. Combining the four automatically generated fields yielded a slight apparent improvement in MAP over any run that used only one of those four fields, although the differences are not statistically significant. Interestingly, there is no noticeable difference among umd.asr04a, umd.asr06a and umd.asr06b even when average precision is compared on a topic-by-topic basis, despite the fact that the ASR text produced by the 2006 system is reported to have a markedly lower word error rate than that produced by the 2004 system.

3.3 Detailed Failure Analysis

To investigate the factors that could have had a major influence on the effectiveness of runs with automatic data, we conducted a topic-by-topic comparison of average precision. To facilitate that analysis, we produced two additional runs with queries that were formed using words from the title field only, with one run searching the ASRTEXT2006B field only (AUTO) and the other the three metadata fields (MANU). We focused our analysis on those topics that have a MAP of 0.2 or larger in the MANU run for consistency with the failure analysis framework that we have applied previously. With this constraint applied, 15 topics remain for analysis. Figure 1 shows the topic-by-topic comparison of average precision. As we saw in 2006, the difference in average precision between the two conditions is quite large for most topics.

Using title-only queries allowed us to look at the contribution of each query word in more detail. Specifically, we looked at the number of segments in which query word appears (a statistic normally referred to as "document frequency"). As Table 3 shows, there are some marked differences in the prevalence of query term occurrences between automatically and manually generated fields. In last year's study, poor relative retrieval effectiveness with automatically-generated data was most

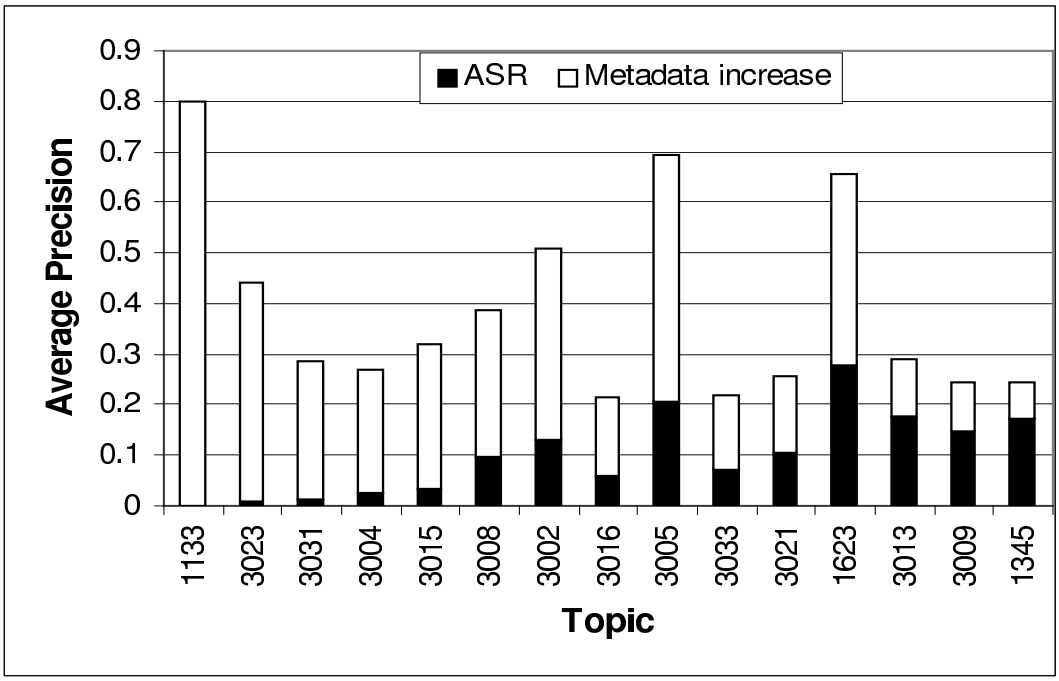

Figure 1: Query-by-query comparison of retrieval effectiveness (average precision) between ASR text and metadata, 15 title queries with average precision of metadata equal to or higher than 0.2. in increasing order of (ASR MAP) / (metadata MAP).

often associated with a failure to recognizing at least one important query word (often a person or location name). This year, by contrast, we see only two obvious cases that fit that pattern ("varian" and "dp"). This may be because proper names are less common in this year's topic titles. Instead, the pattern that we see is that query words actually appears quite often in ASR segments, and in some cases perhaps too often. Said another way, our problem last year was recall; this year our problem seems to be precision. We'll need to actually read some of the ASR text to see if this supposition is supported, of course. But it is intriguing to note that the failure pattern this year seems to be very different from last year's.

4 Czech Experiment Results

The University of Maryland was also one of three teams to participate in the first year of the Czech evaluation. Limited time prevented us from implementing any language processing techniques that were specific to Czech, so all of our runs were based on string matching without any morphological normalization. Automatic segmentation into overlapping passages was provided by the organizers , and we used that segmentation unchanged. Our sole research question for Czech was, therefore, whether the new start-time evaluation metric yielded results that could be used to compare variant systems.

We submitted three officially scored runs for Czech: umd.all used all available fields (both automatically and manually generated, in both English and Czech), umd.asr used only the Czech ASR text, and umd.akey.asr used both ASR text and the automatic Czech translations of the automatically generated thesaurus terms. Table 4 shows the resulting mean Generalized Average Precision (mGAP) values. About all that we can conclude form these results is that they do not serve as a result for making meaningful comparisons. We have identified four possible causes that merit investigation: (1) our retrieval system may indeed be performing poorly, (2) the

Topic	Word	ASR#	Metadata#	Topic	Word	ASR#	Metadata#
1133	varian	0	4	3005	death	287	1013
	fry	4	4		marches	1327	736
3023	attitudes	60	242	3033	immigration	153	325
	germans	4235	1254		palestine	188	90
3031	activities	305	447	3021	survivors	724	169
	dp	0	192		contributions	50	6
	camps	3062	2452		developing	121	26
3004	deportation	277	568		israel	641	221
	assembly	59	9	1623	jewish	4083	1780
	points	1144	16		partisans	469	315
3015	mass	107	302		poland	1586	3694
	shootings	580	82	3013	yeshiva	101	23
3008	liberation	594	609		poland	1586	3694
	experience	811	658	3009	jewish	4083	1780
3002	survivors	724	169		children	2551	426
	impact	35	107		schools	2812	448
	children	2551	426	1345	bombing	593	71
	grandchildren	219	90		birkenau	167	677
3016	forced	654	1366		buchenwald	139	144
	labor	399	828				
	making	2383	80				
	bricks	197	12				

Table 3: Query word statistics in the document collection. ASR#: the number of ASR segments that the query word appears; Metadata#: the number of segments in at least one of the three metadata fields of which the query word appears. Statistics were obtained after both the queries and the document collection were stemmed.

run name	umd.all	umd.asr	umd.akey.asr
mGAP	0.0003	0.0005	0.0004

Table 4: Mean generalized average precision (mGAP) for the three official runs, TD queries.

evaluation metric may have unanticipated weaknesses, (3) the scripts for computing the metric may be producing erroneous values, or (4) the relevance assessments may contain errors. There is some reason to suspect that the problem may lie with our system design, which contains two known weaknesses. Most obviously, Czech is a highly inflected language in which some degree of morphological normalization is more important than it would be, for example, for English. Good morphological analysis tools are available for Czech, so this problem should be easily overcome. The second known problem is more subtle: overlapping segmentation can yield redundant highly ranked segments, but the mGAP scoring process penalized redundancy. Failing to prune redundant segments prior to submission likely resulted in a systematic reduction in our scores. It is not clear whether these two explanations together suffice to explain the very low reported scores this year, but the test collection that is now available from this year's Czech task is exactly the resource that we need to answer that question.

5 Conclusion

Earlier experiments with searching broadcast news yielded excellent results, leading to a plausible conclusion that searching speech was a solved problem. As with all such claims, that is both true and false. Techniques for searching broadcast news are now largely well understood, but searching spontaneous conversational speech based solely on automatically generated transcripts remains a very challenging task. We continue to be surprised by some aspects of our English results, and we are only beginning to understand what happens when we look beyond manually segmented English to unsegmented Czech. Among the things that we don't yet understand are the degree to which the differences we observed this year in English are due to differences in the topics or differences in the ASR, how best to affordable analyze retrieval failures when the blame points more to precision than to recall, whether our unexpectedly poor CLIR results were caused solely by a domain mismatch or principally by some other factor, and whether our Czech test collection and retrieval evaluation metric are properly constructed. In each case, we have so far looked at only one small part of the opportunity space, and much more remains to be done. We look forward to this year's CLEF workshop, where we will have much to discuss!

References

[1] Philipp Koehn. Europarl: A multilingual corpus for evaluation of machine translation. unpublished draft. 2002.

[2] Douglas W. Oard, Dagobert Soergel, David Doermann, Xiaoli Huang, G. Craig Murray, Jianqiang Wang, Bhuvana Ramabhadran, Martin Franz, Samuel Gustman, James Mayfield, Liliya Kharevych, and Stephanie Strassel. Building an information retrieval test collection for spontaneous conversational speech. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–38, 2004.

[3] F. J. Och and H. Ney. Improved statistical alignment models. In *ACL'00*, pages 440–447, Hongkong, China, October 2000.

[4] Jianqiang Wang and Douglas W. Oard. Clef-2005 cl-sr at maryland: Document and query expansion using side collections and thesauri. In *Proceedings of the CLEF 2005 Workshop*, 2005.

[5] Jianqiang Wang and Douglas W. Oard. Combining bidirectional translation and synonymy for cross-language information retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 202–29, 2006.