

First Experiments Searching Spontaneous Czech Speech

Pavel Ircing
Department of Cybernetics
University of West Bohemia
Plzen, Czech Republic
ircing@kky.zcu.cz

Douglas W. Oard
College of Information
Studies/UMIACS
University of Maryland
College Park, Maryland
oard@glue.umd.edu

Jan Hoidekr
Department of Cybernetics
University of West Bohemia
Plzen, Czech Republic
hoidekr@kky.zcu.cz

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing

General Terms

Experimentation

Keywords

Speech retrieval; Spontaneous speech

1. INTRODUCTION

This paper reports on experiments with the first available Czech IR test collection. The collection consists of a continuous stream from automatic transcription of spontaneous speech (see [3] for details) and the task of the IR system is to identify appropriate replay points where the discussion about the queried topic starts. The collection thus lacks clearly defined document boundaries. Moreover, the accuracy of the transcription is limited (around 35% word error rate), mostly due to the nature of the speech—interviews with Holocaust survivors, which are sometimes emotional, accented, and exhibiting age-related speech impediments. This collection therefore offers an excellent opportunity to explore both effects present in Czech (e.g., morphology) and effects that result from processing spontaneous speech. It was also used in the CL-SR track at the CLEF 2006 evaluation campaign (<http://www.clef-campaign.org/>).

2. METHODS

Retrieval from a speech stream with unknown topic boundaries is an interesting challenge, but that is not our principal focus in these experiments. We therefore transformed the collection into artificially defined set of “documents” by removing all recognized pauses between words and then sliding a 3-minute window over the transcripts with a 1-minute step size. This resulted in a collection of 11,377 overlapping passages, each containing an average of 390 recognized words (denoted as the **asr** field) and a set of automatically produced Czech translations (using techniques described in [2]) for 20 automatically assigned thesaurus keywords (using techniques described in [4]) (the **ak** field). Each

	word	stem	lemma
asr	0.0256	0.0494	0.0506
ak	0.0018	0.0022	0.0023
asr.ak	0.0241	0.0447	0.0467

Table 1: Mean GAP, long queries.

field was indexed separately, and a unified index (**asr.ak**) was also constructed.

Twenty-nine topics were initially created in English in the usual TREC-style format (<title>, <desc> and <narr> fields), translated into Czech by a native speaker, and then checked for natural expression by a second native speaker. We performed monolingual experiments with “long” queries constructed by concatenating the words from all three topic fields.

A morphological analyser was used to obtain the information about the lemma (linguistic root form), stem (approximation to that root form using truncation alone) and part-of-speech for each Czech word [1]. Three variants of the collection were indexed, one with only words, one with only lemmas and one with only stems. Part-of-speech tags were used as a basis for stopword removal—as we could not find any decent stoplist for Czech, we simply removed all words that were tagged as preposition, conjunction, particle or interjection. In each case, identical processing was done for the queries. We used Lemur to implement a simple *tf.idf* model with blind feedback (using Lemur’s standard parameters). Length normalization was not performed because the collection preprocessing resulted in documents with nearly identical lengths.

3. EVALUATION

Relevance assessors identified appropriate start times by interactively searching using manually assigned English thesaurus terms and the same automatically transcribed content, ultimately confirming their decisions by listening to the audio when the automatically produced transcripts were not sufficiently accurate to make a definitive judgment. Table 1 reports the mean Generalized Average Precision (mGAP), which is computed in a manner similar to mean average precision (for details see [3]).

Indexing the **ak** field, alone or in combination with **asr**, proved not to be helpful (although the apparent reduction when indexed together is not statistically significant ($p > 0.05$)). Manual examination of a few **ak** fields indeed indicates a low density of terms that appear as if they match

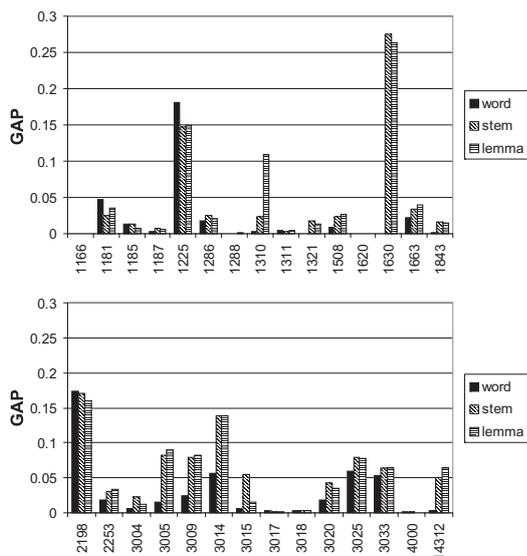


Figure 1: GAP by topic, asr field, long queries.

the content of the passage, but additional analysis will be needed before we can ascribe blame between the transcription, classification and translation stages in the cascade that produced those keyword assignments. We therefore focus on results obtained using the `asr` field alone for the remainder of our analysis.

It is apparent that some form of linguistic preprocessing is indeed crucial for Czech. Both lemmatization and stemming boosted the performance almost by a factor of two in comparison with the `word` runs, and a Wilcoxon signed-rank test shows that difference to be statistically significant ($p < 0.005$). The slight apparent advantage of the `lemma` run over the `stem` run is not statistically significant ($p > 0.05$). As Figure 1 shows, substantial variation in GAP is evident across topics. The four topics with the highest GAP values (1225, 1630, 2198, 3014) each contain highly discriminative terms that were correctly transcribed. Topic 1630 exhibits an enormous difference between word matching and matching either stems or lemmas, a vivid reminder of how the recall-enhancing effect of linguistic analysis can dominate averaged measures (a similar effect is also apparent for topic 1310). While a few cases of adverse effects from linguistic analysis are visible (most notably with topics 1225 and 1181), these effects are generally relatively small. The occasional differences between stems and lemmas suggests that combining evidence from both might help in some cases.

Unsuccessful topics generally either asked about abstract concepts without using many discriminative terms (e.g., topic 1288: “strengthening faith during the Holocaust”), or the discriminative terms for the topic happened to be missing from the collection. For example, topic 3018 contained a single discriminative term that was simply spelled differently in the ASR lexicon (and consequently in the transcripts). Manually conforming the spelling in the topic to that found in the lexicon would have increased the GAP for that topic (with `lemma`) from 0.0026 to 0.1175.

Interestingly, it turned out that every term that we (manually) judged to be highly discriminating in our analysis of

successful and unsuccessful topics was a named entity (NE). This prompted us to perform a more systematic analysis of the vocabulary coverage for the NEs present in all 29 topics. If we leave out the NEs that are widespread in the collection and thus useless for IR (Jew, Holocaust, Hitler, etc.), there are 42 NEs in the topic set; only 13 of them are present in the ASR lexicon, only 11 of those 13 actually appeared anywhere in the transcripts, and only 5 of those 11 substantially contributed to successful IR (or, if we manually conform the spelling in topic 3018, 6 of 12). The overall “query rare named entity error rate” for this collection is therefore $(42-5)/42=88\%$, more than double the overall word error rate. Rare NEs are quite naturally not well represented in the materials from which ASR systems are trained; integrating phone-lattice term detection with large-vocabulary recognition offers one promising research direction. Inconsistent spelling is probably the more easily rectified problem; annotators of ASR training materials are typically not domain experts, and in some cases valid alternate transliterations (e.g., from Yiddish roots) result in disagreement even among experts. One useful approach would be to adjust the topics to conform to the ASR lexicon, thus simulating a similar process an interactive searcher could perform if notified that one of their query terms is outside the known vocabulary.

4. NEXT STEPS

In addition to the ideas above for dealing with rare terms, another obvious next step would be to optimize our system design to better reflect the task characteristics that motivated the design of the mean GAP measure. We have shown that passage retrieval can indeed sometimes get us in the right neighborhood, but overlapping passages may not be the best way of identifying optimal replay start times. Another question that we need to explore is whether some other retrieval model might be more effective. Finally, extending our work to include on the far larger CLEF 2007 Czech news test collection will allow us to enrich our comparison between lemmas and stems for Czech indexing.

5. ACKNOWLEDGMENTS

This work was supported in part by projects MSMT LC536, GACR 1ET101470416 and NSF IIS-0122466.

6. REFERENCES

- [1] J. Hajič. *Disambiguation of Rich Inflection. (Computational Morphology of Czech)*. Karolinum, Prague, 2004.
- [2] C. Murray *et al.* Leveraging Reusability: Cost-effective Lexical Acquisition for Large-scale Ontology Translation. In *Proceeding of ACL 2006*, pages 945–952, Sydney, Australia, 2006.
- [3] D. Oard *et al.* Overview of the CLEF-2006 Cross-Language Speech Retrieval Track. In *CLEF 2006 - revised selected papers - Springer LNCS*, 2007.
- [4] S. Olsson, D. Oard, and J. Hajič. Cross-Language Text Classification. In *Proceedings of SIGIR 2005*, pages 645–646, Salvador, Brazil, 2005.