First Insight into the Processing of the Language Consulting Center Data

Zbyněk Zajíc[0000−0002−4153−6560], Lucie Zajícová[0000−0002−0021−5365], Josef V. Psutka, Petr Salajka[0000−0003−4238−8077], Jaromír Novotný, Aleš Pražák, and Luděk Müller[0000−0002−6581−6348]

University of West Bohemia, Faculty of Applied Sciences,
NTIS - New Technologies for the Information Society
and Department of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
{zzajic, lskorkov, psutka_j, salajka,
fallout7, aprazak, muller, ircing}@ntis.zcu.cz

Abstract. In this paper, we describe the initial stages of the project "Access to a Linguistically Structured Database of Enquiries from the Language Consulting Center". This project is attempting to provide an improved access to the large archives of mainly telephone conversations collected continuously by the Institute of the Czech Language. The main goal is to open up the unique Czech data acquired from the queries to the Language Consulting Center and to build the semi-automatic system that will facilitate searching and categorizing of these queries. For this purpose, the Automatic Speech Recognizer (ASR) and the language processing methods are being designed. The vocabulary used in such queries contains many unusual words unlike the common speech (e.g. linguistic terms). In order to train the ASR system, it is necessary to manually transcribe a large amount of speech data, identify the appropriate vocabulary, and obtain relevant text for language modeling purposes. In this paper, the proposed telephone system for recording the new data and the baseline speech recognition on these data is described. The first experiments with the topic detection on these data aimed at discovering what can be found in them and also how to preprocess them is also described.

Keywords: automatic speech recognition, topic detection, telephony system, language consulting

1 Introduction

The Language Consulting Center (LCC) of the Czech Language Institute of the Academy of Sciences of the Czech Republic provides a unique language consultancy service in the matters of the Czech language. The counselors of the LCC are answering questions regarding the Czech language problems on a telephone line open to public calls. The data, gathered from these language queries are unique in several aspects. The Language Consulting Center deals with completely new language material so it is the only source of advice for new language problems. It also records peripheral matters that will never be explained in dictionaries and grammar books as these are focused on the core of the language system.

The main goal of the project “Access to a Linguistically Structured Database of Enquiries from the Language Consulting Center” is to publish these unique data acquired from the queries from the LCC and to deal with them in a new, user-friendly language software database. In order to make these data more accessible a semi-automatic system for processing the queries and recording them in the database is being created. The aim of publishing these data is not only to provide the practical help for users solving language problems but also to create a tool that would preserve those language data as a national attribute. These data are not preserved in any other linguistic source as these sources have a different focus (see section 2 for a detailed description).

The final system is designed to be a flexible tool that would work long after the project is finished. A semi-automatic system for processing the queries and recording them in the database is being created. The system will facilitate searching and categorizing the queries by the language counselors and database users. For this purpose, the Automatic Speech Recognizer (ASR) and the language processing methods (like topic detection, keyword spotting, etc.) are being designed to describe the speech data to allow their better accessibility. From the description of the nature of the data can be seen a clear challenge for the ASR and the language processing methods. The queries often contain a new language material (new expressions, foreign words, etc.) for language counselors themselves and the vocabulary also contains many unusual words compared to the normal speech (e.g. linguistic terms). From this, it appears that the query dictionary contains words that are not used in the common Czech language, they either fall into the domain of linguistics, or may be completely new to the language itself.

The goal of our system is to help the language counselors with the description of the queries, the system will provide the recognized text and the suggestions for topics etc., but the final decision is on the counselor himself. Furthermore, the automatically recognized text can be used for searching in the database and creating statistics about the content of it. In this initial phase of the project the ASR and the topic detection methods are developed to help the language counselors and other users to work with these data.

The telephone calls from the LCC are considered to be the primary source for the database and also for our training process, but for the start of the project, the Czech Language Institute also has some stored old email communication with user queries which can be used for the first experiments. Before this project, the LCC has been recording data only on the analog telephone line (8kHz, μ-law resolution) stored only in mono (counselor and user mixed in one channel). These data are very insufficient for the automatic recognition and subsequent categorization because of their bad quality and the difficulty in separating the question and the answer properly. For this purpose, the new recording system was applied to store the queries called to LCC with better quality (8kHz, 16bit resolution) and with separated channels. The biggest improvements are the significantly higher signal-to-noise ratio and separated speakers in different channels. Because of the difficulty to describe the challenging language material in queries fully automatically, the recording system allows the user to specify the content of speech query during the recording to give more precise information for further semi-automatic processing. The proposed telephone recording system is described in section 2.1 and the results of the ASR are shown in section 3.

The final goal is to classify the query into the linguistic topics and store it in the linguistics database. The query will be stored with automatically recognized speech transcript and semi-automatically assigned linguistic categories. The definition of these actual topics is one of the goals of this project (the part of the partner of the project, the Czech Language Institute). The first experiments with the topic detection methods on these data can be seen in section 4.

2 Data

The previous calls to the LCC were stored as a low quality recordings and are not accessible to the public, so there was no way to search in a previously answered questions. The newly designed system would allow the users to search in the previous questions and look for the answers in them. The database would also serve as a collective historic memory of the evolution of the Czech language and grammar rules.

2.1 Telephony System

High-quality records are essential for automatic speech recognition (ASR). Before this project started, the LCC used an analog phone line of a low voice quality. Moreover, both channels (incoming and outgoing) were mixed. After the first experiments on the existing low-quality mono data, we decided to look for another approach to improve the quality of the newly collected data. After examining possible options, we decided to design a new system based on the Voice over Internet Protocol (VOIP) and Asterisk [2] as an open source telephony platform.

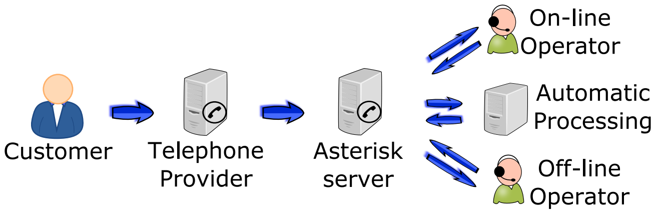

Fig. 1. Telephony System

The Figure 1 shows the architecture of the telephony system. “On-line Operator” responds to the customer queries in the time of the LCC open hours. He also uses the web application to write down some notes about the responded queries, and he assigns the call to one or more “Off-line Operators” (linguistic experts) for post-processing according to their expertise. At this time, the “Off-line Operators” divide the call into queries and write down their classification based on the linguistic topic tree. These data are used as labels for the automatic topic detection. This post processing is really time consuming for the “Off-line Operators”, so we suppose, in future, the process will reverse and the calls will be automatically spit into the queries concerning a single linguistic question and assigned the automatic topic at the time of the call and the “Off-line Operators” will only check the correctness or fine-tune the data.

3 Baseline Automatic Speech Recognition System

This section describes the baseline Automatic Speech Recognition system designed to describe the speech data to allow their better accessibility in the final online database. The data represent a clear challenge for the ASR system as they often contain a new language material and the vocabulary also consists of many unusual words compared to the normal speech.

3.1 Front-End and Acoustic Modeling

For the stereo data the standard Kaldi “nnet” recipe [14] (with RBM pre-training for every layer) was used for hybrid DNN/HMM model, with approx. 5000 states, DNN 6 layers with 2048 neurons each, learning rate starts at 0.008 and using early stopping criteria. The model was trained on 500 hours of spontaneous telephone speech, all converted into the quality of target data. As the feature descriptor, we used a DNN introduced in the paper [20]. Short-time feature vectors are computed from an absolute spectrum by the means of a small NN (which has 256 neurons in an only one hidden layer and 16 neurons in an output layer), mean and variance normalization and all relevant delta and delta-delta coefficient are then computed. Finally, a linear transformation (trained simultaneously with the small NN) is used to splice (21-vectors-long) the time window of all the previous features into the resultant vectors (with 40 features).

The model was trained on various 500 hours of spontaneous telephone speech corpora ([3], [13] and other unpublished sources), all converted into target quality.

3.2 Language Modeling

For the better aim of the language model, we trained a domain LM [11] with the dictionary size 174k as a standard trigram language model with Kneser-Ney smoothing with maximal entropy criterion. This model was trained on the available transcribed data from LCC of the language counselor (0.5 million tokens) and the client (0.5 million tokens) and from the email communication (counselor 3.8 million tokens and client 3.4 million tokens). For the resulting LM model we have mixed these four available data sets with weights 0.70, 0.16, 0.10 and 0.04 for counselor LM and with weights 0.62, 0.23, 0.03 and 0.12 for client LM. As you can see the main influences were on the data from transcription of the language counselor calls in both cases.

3.3 Experiments with ASR

In the results of the proposed system on the stereo data (shown in Table 1) can be seen the higher correctness of ASR for the language counselor. There is a limited amount of

role	Corr [%]	Acc [%]
language counselor	85.03	82.41
client of LCC	82.52	78.62

Table 1. Correctness [%] (Corr) and Accuracy [%] (Acc) for the ASR system on stereo data separately evaluated for the language counselor and the client.

transcribed data from LCC, so the test was made only on a small number of sentences. From the results can be seen that the language counselors speech has better accuracy of transcription. It is mainly caused by the professionality of the speaker (his confidence about the topic, the consistent vocabulary and less colloquial speech) compared to the client of LCC. Still, the accuracy of this ASR can be enhanced by improving the AM and with the use of LM trained on more transcribed data of LCC which are now of limited amount. Also, the knowledge that the specific language counselor is speaking (from the finite set of speakers) can be used to improve the accuracy of ASR, e.g. by adapting the AM on that speaker or by choosing the speaker dependent AM for each particular speaker in the counselor part of the conversation. Those are the aims of the project in the next years.

4 Topic Detection

As stated before, one of the goals of this project is to develop a semi-supervised topic detection system, which would help the counselors from the LCC to faster categorize the newly recorded query into the database. The idea is, that the topic detection system would try to identify the topic of the query as soon as the call ends and the recording is processed with the ASR system, so the counselor would be provided with this information when he starts to fill in the metadata of the query call. The topics would correspond to the linguistic categories defined by the Czech Language Institute, that is organized in a rich hierarchical structure - linguistic topic tree with hundreds of leaves as linguistics categories. In this state of the project, our goal is to find the essence of the particular query and categorized it into the higher level of the topic tree (find the meta-category). According this classification, the suitable language counselor will be chosen with an appropriate expertise to precise decision about the topic in the query.

4.1 Unsupervised Topic Detection

For the first experiments, the LCC has supplied us with the stored older queries in the text only form, consisting of 2126 letters with only the counselors answer (without the original query) and 70718 email communications, mostly containing both the query and the answer. All text parts were lemmatized since lemmatization was shown to improve the effectiveness of natural language processing methods in highly inflected languages (as is the Czech language) [4][15][5]. For the lemmatization, an automatically trained lemmatizer described in [6][7] was used. The comparison of the use of the original and the lemmatized text is also shown in section 4.1.

We have tried three approaches, the first one was the classic K-means clustering algorithm [10] on the texts preprocessed with the TF-IDF weighting and Latent Semantic Analysis (LSA) [8]. The top terms from each cluster were printed out to find the contents of the cluster. The other two approaches are the Latent Dirichlet Allocation (LDA) [1] method applied on the raw TF count vectors and the Non-negative Matrix Factorization (NMF) [12] method applied on the TF-IDF weights vectors, similarly, as in the K-means method, the top topics were printed out. Since we do not have the "ground truth" topic annotations for our data, the experiments were aimed mostly on discovering some properties of the data, which we can use in our future system development. For this experiment only the counselors answers were used. For all algorithms, we have tried the setting of the number of clusters (topics) from 2 to 50, subjectively the best number of clusters seems to be around 20 clusters, where the clusters are general enough. The results of all approaches were quite similar, the top words or topics found in them are almost identical. In the lack of annotated data for proper objective evaluation, we can only state that the results are very promising and the most common topics distinctively emerged. The most common queries (most distinctive clusters in the setting of around 20 clusters, the rest was either incomplete questions, greetings, or infrequent questions merged together) are about (this was also confirmed by the LCC counselors): How to form feminine surnames from masculine ones; How the write the capital letters in street names with a preposition in it; If it is correct to write Romany or Gypsy; What case to use when addressing somebody; How to write capital letters in the geographical terms (actually split into two clusters about towns and generally); How is the official form of degree abbreviations; How to properly write punctuation when using different subordinate clauses; How to decline the adjectives connected with a noun; What is the meaning of some word; And also a quite big cluster of answers, that the LCC provides only the linguistic consulting, not the legal consulting.

Lemmatization The results presented in the previous section were achieved on the lemmatized data since from our previous experience with the processing of the Czech data the lemmatization tends to improve the results. Our next experiment was aimed at confirming or disproving this assumption. We have repeated the previous experiment also on the non-lemmatized data and it can be confirmed, that the results are better with the use of the lemmatization to preprocess the text data. With the non-lemmatized data the formed clusters seems to be less meaningful and also the description with the top words/topics is less readable. The lemmatized results contain only descriptive words and the topic of the cluster can be easily derived from them. On the other hand, with the non-lemmatized data, the top words contain nondescriptive and also common words so the actual topic of the cluster is hard to derive (only by looking at the contained data).

Query or Answer (or Both)? The next set of experiments was aimed at finding out if it would be better to use only the answers (like in the first experiments) or the queries, or both, since the new stereo recording (described in section 2.1) would allow us to distinguish between them. For these experiments, we have used the set of emails and we have adopted all methods described in previous section 4.1. We found out that on this bigger set, the K-means algorithm seems to be overperformed by the NMF and

LDA, which performed quite alike. The most common topics found in the previous experiments were confirmed. We have found out that the best option is to use only the counselor answer since it tended to form the most compact clusters/topics, using only the queries seems to form the least meaningful clusters. We have looked into the data for the reason of this effect and we have found out, that the counselors tend to form the answers in a more general way than the query is.

4.2 Supervised Topic Detection

In the second experiment, we focus on the query from the telephone calls. Each call contains generally more than one query. Therefore, we cut each call into parts with one question about a particular topic. This division of calls was made manually by LCC counselors, the automatic division will be solved later in this project using methods for spoken language understanding [18].

Our dataset consist of manually transcribed 607 parts of historical mono phone calls and automatically transcribed (by our ASR system) 3128 parts of actual stereo phone call, all divided into 20 categories by their topic. This 20 categories were manually assigned by counselors from LCC and corresponds with higher level of the linguistic topic tree (for example ”semantics” or ”lexicology”). The division of phone call into categories is not uniform, some categories contain only a few parts. The setting of this experiment (mainly the number of categories, the using of lemmatization and experiments only on answer) is based on previous findings (see section 4.1).

In preprocessing stage, all uppercase characters were lower-cased and all digits were replaced by one universal symbol, lemmatization using MorphoDiTa [17] tool [1] and stop word removal was applied to all data.

Our results (see Table 4.2) were gained using simple supervised classificaton algorithm Linear Support Vector Machine (SVM). This algorithm uses a different inputs: *TF-IDF* weights with dimension $D = 5000$, *doc2vec* features also with dimension $D = 5000$ created by Gensim package [16], both of these vectors after LSA dimensionality reduction ($D = 200$) and their combination (by concatenation of these two vectors) after LSA dimensionality reduction ($D = 200 + 200$). The Accuracy measure is applied on different parts of our dataset and represents the percentage of correctly classified parts of transcribed phone calls (i.e., show what percentage of the parts is assigned to the correct topic). 10-fold cross-validation was used to get the results.

The amount of manually transcribed data is considerably smaller in comparison with the ASR transcriptions but contains less mistakes in annotation, therefore the results on these different data are comparable.

The experiments shows the best results using all data (manual and ASR transcription, query and answer) for training the classifier. The main reason for this is the limited amount of training data, so the premise about the superiority of answers data from the previous experiment in 4.1 has not been confirmed.

The *TF-IDF* approach exceeded the *doc2vec* in this task, as in work [19, 9]. The special character of the data in this project (e.g. linguistic terms) is not precisely represented by generally trained feature extractor based on *doc2vec* from Gensim package.

[1] `ufal.morphodita` at `https://pypi.python.org/pypi/ufal.morphodita`

Data	*Accuracy* of methods [%] Linear SVM method with features				
	TF-IDF	*TF-IDF (LSA)*	*doc2vec*	*doc2vec (LSA)*	*TF-IDF (LSA) + doc2vec (LSA)*
manual transcription	76.58	75.20	69.87	66.45	76.84
ASR transcription	76.56	70.33	69.28	66.93	73.44
ASR transcription - answer only	74.56	65.23	63.78	61.54	69.05
all	77.92	71.14	70.86	68.36	75.19

Table 2. Accuracy [%] of topic detection on transcribed telephone data.

Also the limited length of a query (400 word on average) tends to prefer the *TF-IDF* approach before *doc2vec* approach.

After the dimensionality reduction of the input vector by LSA the result decrease minimally. Then, the combination of these two methods (after LSA) brings almost the same accuracy as the *TF-IDF* approach with full dimension.

5 Conclusions

The objective of this paper is to verify the feasibility of the goals of this project and to introduce the first results made on the available data. Looking at the first results of our ASR system (with the stereo data, domain LM) we can say that our system is designed correctly and the results are promising.

The topic-oriented text together with a large portion of manual transcripts could improve the performance of the recognition task and also enable the appropriate topic identification. The experiments with topic detection introduced a reasonably effective approach of the textual transcribed phone calls query. The manually transcribed data shows a slightly better results, nevertheless, the ASR transcription proved the practicability of our approach and with increasing the number of data the accuracy will rise.

The presented results with the processing of the available data show the first step to fulfill the goal of the project: to publish these unique data acquired from the queries from the LCC and to create an on-line linguistic database.

Acknowledgements. This research was supported by the Ministry of Culture Czech Republic, project No.DG16P02B009. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (Mar 2003)
2. Bryant, R., Madsen, L., Meggelen, J.V.: Asterisk: The Definitive Guide: The Future of Telephony Is Now. O'Reilly Media, 4th edn. (2013)

3. Černocký Jan, Pollák Petr, H.V.: Czech speechdat(e) database. ELRA-S0077, ELRA (2000)
4. Ircing, P., Müller, L.: Benefit of Proper Language Processing for Czech Speech Retrieval in the CL-SR Task at CLEF 2006. In: Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the CLEF. pp. 759–765. LNCS, Alicante, Spain (2007)
5. Ircing, P., Psutka, J., Vavruška, J.: What Can and Cannot Be Found in Czech Spontaneous Speech Using Document-Oriented IR Methods – UWB at CLEF 2007 CL-SR Track, pp. 712–718. Springer-Verlag, Berlin, Heidelberg (2008).
6. Kanis, J., Müller, L.: Automatic Lemmatizer Construction with Focus on OOV Words Lemmatization. In: TSD 2005, LNCS, vol. 3658, p. 742. Springer, Heidelberg (2005)
7. Kanis, J., Skorkovská, L.: Comparison of Different Lemmatization Approaches through the Means of Information Retrieval Performance. In: TSD 2010, LNCS, vol. 6231, pp. 93–100. Springer, Heidelberg (2010)
8. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review **104**, 211–240 (1997)
9. Lilleberg, J., Zhu, Y., Zhang, Y.: Support vector machines and Word2vec for text classification with semantic features. In: IEEE Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC). pp. 136–140. Beijing (2015).
10. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. pp. 281–297. University of California Press, Berkeley, Calif. (1967)
11. Maergner, P., Waibel, A., Lane, I.: Unsupervised vocabulary selection for real-time speech recognition of lectures. In: ICASSP. pp. 4417–4420. Kyoto (2012)
12. Paatero, P., Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics **5**(2), 111–126 (1994). https://doi.org/10.1002/env.3170050203
13. Pollák Petr, Černocký Jan, H.V.: Telephone speech data collection for czech. ELRA-S0094, ELRA (1999)
14. Povey, D.: nnet architecture (2017), `https://github.com/kaldi-asr/kaldi/tree/master/egs/wsj/s5/steps/nnet`
15. Psutka, J., Jan, Š., Psutka, J.V., Van, J., Lubo, Š., Ircing, P.: System for fast lexical and phonetic spoken term detection in a Czech cultural heritage archive. EURASIP Journal on Audio, Speech, and Music Processing pp. 1–11 (2011). https://doi.org/10.1186/1687-4722-2011-10
16. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50 (2010),
17. Straková, J., Straka, M., Hajič, J.: Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 13–18 (2014)
18. Švec, J., Šmídl, L., Ircing, P.: Hierarchical discriminative model for spoken language understanding. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 8322–8326. IEEE, Vancouver (2013).
19. Wang, Y., Zhou, Z., Jin, S., Liu, D., Lu, M.: Comparisons and Selections of Features and Classifiers for Short Text Classification. In: International Conference on Artificial Intelligence Applications and Technologies (AIAAT). vol. 261, pp. 1–7. IEEE, Hawaii (2017).
20. Zelinka, J., Vaněk, J., Müller, L.: Neural-Network-Based Spectrum Processing for Speech Recognition and Speaker Verification. In: Statistical Language and Speech Processing. vol. 9449, pp. 288–299. Budapest (2015).