Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200615

Creation of Language Resources for the Development of a Medical Speech Recognition System for Latvian

Roberts DARĢIS [1], Normunds GRŪZĪTIS, Ilze AUZIŅA and Kaspars STEPANOVS
Institute of Mathematics and Computer Science, University of Latvia, Latvia
Riga East University Hospital, Latvia

Abstract. This paper describes an ongoing work on the creation of Latvian language resources for the medical domain focusing on digital imaging to develop a medical speech recognition system for Latvian. The language resources include a pronunciation lexicon, a text corpus for language modelling, and an orthographically transcribed speech corpus for the (i) adaptation of the acoustic model, (ii) evaluation of the speech recognition accuracy, (iii) development and testing of rewrite rules for automatic text conversion to the spoken form and back to the written form. This work is part of a larger industry-driven research project which aims at the development of specific Latvian speech recognition systems for the medical domain.

Keywords. Speech recognition, language resources, medical domain, Latvian language

1. Introduction

This paper describes the creation of domain-specific language resources required for the development of a medical speech recognition system. The language resources of interest are: an anonymised text corpus of medical reports, namely digital imaging reports and epicrisis reports (excerpts from an archive) for language modelling; a pronunciation lexicon of medical terms, abbreviations and named entities for their recognition and consistent transcription; an anonymised and orthographically transcribed speech corpus for adapting the acoustic model, evaluating the adapted speech recognition systems, developing and testing automatic pre-editing and post-editing rules to rewrite the existing final reports to their spoken form (as if dictated) and dictations to the expected written form.

Since the creation of a relatively large[2] general-domain speech corpus for Latvian [1], various automatic speech recognition (ASR) systems of industrial applicability have been developed for Latvian [2], [3]. ASR systems trained on general-purpose

[1] Corresponding Author: Roberts Darģis, Artificial Intelligence Laboratory, Institute of Mathematics and Computer Science, University of Latvia, Raiņa bulv. 29, Riga, LV-1459, Latvia; E-mail: roberts.dargis@lumii.lv.

[2] 100 hours of annotated and orthographically transcribed balanced audio data, of which 4 hours are also phonetically transcribed.

speech and text corpora, however, are not applicable[3] for the very specific language of medical reports. Domain-adapted language model and pronunciation lexicon (both derived from a text corpus of written medical reports) give the most significant boost in ASR accuracy, while a domain-adapted acoustic model (derived from a speech corpus of dictated medical reports) makes a considerable impact as well [4].

The work presented in this paper is a part of an ongoing collaborative project between a language technology research group and the largest hospital in Latvia on Latvian ASR for medical applications. Although modern medical technology is widely used in Latvia, particularly for imaging diagnostics, medical reports are still produced completely manually. The largest healthcare institutions in Latvia maintain or outsource services of transcriptionist centres to produce medical reports. However, the number of diagnostic examinations is constantly growing, and clinicians and patients must wait up to several days for the reports. Moreover, transcriptionist services are expensive, and regional healthcare institutions cannot afford them. Inspired by the successful implementation of Estonian ASR for radiology [4], [5], our goal is to create the essential language resource and technology components for Latvian ASR in the medical domain, particularly in radiology, and showcase their usage by developing and validating an automated dictation platform for radiology reporting. The platform will be customised for two usage scenarios. First, we expect that for subdomains that typically produce simpler and more fluent dictations (e.g. X-ray and ultrasonography), radiologists will use ASR in a self-service manner. Second, we expect that a transcriptionist centre might still be preferred by radiologists in the case of more complicated dictations (e.g. computed tomography and magnetic resonance); however, a centralised transcriptionist centre would become more productive since part of the workload would be moved to the self-service scenario, and draft reports (via ASR) would be available for the complex dictations.

2. Text Corpus

The data source for the creation of the text corpus is an archive of medical reports produced over the last decade in a large multi-profile hospital.

The corpus creation involves several steps. The first step is the extraction of plain-text paragraphs from the description part and the conclusion part of the actual reports stored in the archive. The hospital's transcriptionist centre produces, so far manually, Microsoft Word documents from audio files dictated and submitted by radiologists, clinicians and other doctors.

Over the time, there are many different templates used for the documents, and multiple nested tables are used to create layout for the templates. The documents also contain various text formatting elements, such as underlines, bold, and italic. In the text extraction step, it is important to preserve the text segmentation, since language modelling relies on correct text division into sentences. The documents contain multiple short-text segments without any punctuation marks, such as table column names and field names. If all the extracted text segments would be concatenated as is, such short-text segments would be incorrectly added to other sentences. An opposite issue would occur if larger text paragraphs would be divided into smaller segments due to mistakenly recognised text formatting elements as paragraph separators.

[3] The word error rate (WER) is way too high for efficient usage of non-adapted ASR systems.

Multiple third-party plain-text extraction tools were tested. Unfortunately, all of them had the same issue: either too long or too short text segments are extracted. Therefore a custom plain-text extraction tool was developed using the underlying XML structure of a Word document. Each paragraph is represented by a *p* tag, and each text segment inside the paragraph is encoded by a *t* tag. This method correctly extracts text segments from table cells as well, since text in each cell is encoded in its own paragraph.

The second step is the anonymisation of the extracted texts to avoid any sensitive personal data (name, surname, ID number, etc.) to be included in the text corpus and, thus, in the language model. Although the description and conclusions parts of medical reports should not contain any personal data, it is still possible that some irrelevant text segments containing personal data are extracted by mistake.

The available amount of archived text is more than necessary for language modelling, therefore, text anonymisation was prioritised over text preservation. To reduce the risk of personal information leaking into text corpus, anonymisation is done at the paragraph level. If a paragraph contains any tokens recognised as potential personal data, the whole paragraph is excluded from the corpus. The remaining paragraphs are split into sentences, since language modelling is done at the sentence level.

The first two steps are executed in the hospital's IT infrastructure, so that no sensitive data is handed over to the research partner. Anonymised plain-text were extracted from 100k reports covering 8 years of reporting. The number of reports produced each year is steadily increasing, reaching 15k in 2018.

The third step is text normalisation w.r.t. the tokenization and correction of typical typos. If a sentence contains words recognised as typos, they are automatically corrected according to the annotations made during the lexicon development (Section 3).

The fourth step is automatic expansion (verbalization) of numbers, abbreviations, symbols and other such tokens. In the result, we effectively acquire a parallel corpus of collapsed (original) and expanded texts. We follow the successful approach used by Alumäe et al. [5]. A small but representative part of the corpus is expanded manually by domain experts. Additional parallel texts are acquired through the speech corpus creation (Section 4). Based on these subsets, context-sensitive rewrite rules are defined for automatic expansion of the rest of the text corpus. Additionally, random punctuation marks are verbalised as commands for structural formatting. The language models are acquired from the verbalised version of the text corpus.

3. Pronunciation Lexicon

Development of the lexicon is the next step after text extraction (Section 2). The lexicon contains a list of words and their pronunciation that the speech recognition system should recognise.

The lexicon is derived from the text corpus. First, the texts are tokenized, and all the unique words are extracted. Their pronunciation is generated semi-automatically.

There are four types of words w.r.t. the pronunciation generation:

- words in Latvian, which are pronounced exactly as written (the majority of words);
- words containing typos, which should not be included in the lexicon as they are;
- abbreviations and symbols that need to be expanded to generate pronunciation;

- words in a language other than Latvian (e.g. drug names and Latin terms), which are specifically pronounced in the Latvian context, i.e., phonetic transcription based on Latvian phonemes has to be provided.

Radiologists tend to pronounce some words incorrectly because it is easier, faster or just more convenient. Additional pronunciation variants are extracted from the speech corpus (Section 4) to better reflect pronunciation variation in the actual dictations.

Words that are part of personal data (name, surname, address) are semi-automatically separated and used as a supplementary filter in text anonymisation (Section 2).

In total, there are 1.8M unique tokens in the extracted text corpus, of which 1.1M are tokens that contain at least one letter and are, thus, considered for inclusion in the lexicon.

Manual categorisation of all words into the above mentioned four groups would be an inefficient and time consuming task.

Therefore, several existing dictionaries were used to categorise most of the words automatically. First, words that are included in the largest open dictionary of Latvian Tezaurs.lv [6] were automatically marked as standard words. In the next iteration, an open-source Latvian NLP pipeline [7] (namely, a morphological tagger and a named entity recogniser) was used to recognise words that are possibly part of personal data. Words that were not classified automatically need to be manually classified and transcribed. Words that occur at least 1,000 times in the text corpus (almost 14k words)[4], were selected for the manual review process in the first iteration.

The lexicon and the frequency information is also used as a filter in the development of the language model. Rarely or incorrectly used words and words that possibly constitute personal data are excluded from the language model.

4. Speech Corpus

A domain-specific orthographically transcribed speech corpus is a key component in the adaptation of the acoustic model as well as for the evaluation of the specialised ASR systems.

Adaptation of the acoustic model generally helps to reduce the word error rate (WER). It makes even a bigger impact in ASR settings where the potential user pool is limited, as it is in the medical domain, and the ASR system can be adapted to the speakers. In the past year, there have been 71 unique speakers.

Speech corpus also allows to evaluate how well all the ASR components work together: how well the acoustic model predicts the environment, how well the pronunciation modelled in the lexicon matches the actual pronunciation, and how well the language model predicts the text.

Actual dictation records are used for the creation of the speech corpus. Records are selected from the hospital's transcriptionist centre's archive of recordings. Low quality records that have been dictated over a telephone, or contain significant background noise or parallel speech, or are dictated by non-native speakers with sever accent and many pronunciation errors are omitted.

[4] An empirical threshold based on the word frequency distribution in the extracted text corpus.

```
1 Izmeklējums CT         vēdera dobumam -
2 izmeklējums CT [cē tē] vēdera dobumam {dash}
3 izmeklējums CT         vēdera dobumam domuzīme
4 Izmeklējums CT         vēdera dobumam -

1 Natīvs izmeklējums ar p/o      kontrastētu kuņģa-zarnu traktu
2 natīvs izmeklējums ar perorāli kontrastētu kuņģa zarnu traktu
3 natīvs izmeklējums ar perorāli kontrastētu kuņģa zarnu traktu
4 Natīvs izmeklējums ar p/o      kontrastētu kuņģa-zarnu traktu

1 izmeklējums pēc i/v         ievadīta Ultravist 300
2 izmeklējums pēc intravenozi ievadīta Ultravist trīssimt [trīssimti]
3 izmeklējums pēc intravenozi ievadīta Ultravist trīssimt
4 izmeklējums pēc i/v         ievadīta Ultravist 300

1 Pancreas            difūza tauku involūcija , veidojumus        neredzu
2 pancreas [pankreas] difūza tauku involūcija   veidojumus nesas* neredzu
3 pancreas            difūza tauku involūcija   veidojumus        neredzu
4 Pancreas            difūza tauku involūcija   veidojumus        neredzu

1 kreisā niere mazāka apjomā ,        plānāku parenhīmu .
2 kreisā niere mazāka apjomā {comma}  plānāku parenhīmu {full-stop}
3 kreisā niere mazāka apjomā komats   plānāku parenhīmu punkts
4 kreisā niere mazāka apjomā ,        plānāku parenhīmu .

1 Kreisās       nieres konkrements 0,39              cm         / Ø
2        labās nieres konkrements nulle trīs deviņi centimetri   diametrā
3        labās nieres konkrements nulle trīs deviņi centimetri   diametrā
4        Labās nieres konkrements 0,39              cm         / Ø
```

Figure 1. Sample excerpts from the aligned text and speech corpora. Lines: 1 – a text span from an archived written report; 2 – an orthographically transcribed and annotated segment from the corresponding dictation record; 3 – the corresponding text span from the derived text corpus for language modelling; 4 – the expected final output of the ASR system

We are aiming at a 30 hour orthographically transcribed corpus, part of which will be used for the evaluation purposes. The transcriptions are bootstrapped by aligning the corresponding expanded text segments of the extracted text corpus (Section 2). The automatic alignment is, in general, partial, and it is manually post-edited by experts. Simple annotation guidelines are used, e.g. on how to annotate pronunciation of specific terms, based on the experience gained in the creation of the general-purpose Latvian speech corpus [1]. Additionally, text formatting commands are annotated, based on the previous work on designing a general-purpose dictation corpus of Latvian [8].

Figure 1 illustrates the orthographic transcription and annotation of the speech corpus (see Line 2 in each of the six samples). For each token, pronunciation is given in square brackets next to the token if its pronunciation differs from the written form.[5] For instance, *CT* (‘computed tomography’) is pronounced as *cē tē* in the 1st sample. Pronunciation is given also in cases where a word that should be pronounced as it is written, is

[5] In Latvian, words are mostly pronounced as they are written.

pronounced incorrectly, as it often occurs with numbers: e.g., *trīssimt* ('three hundred') is pronounced as *trīssimti* in the 3rd sample. Pronunciation annotations are used for automatic extension of the pronunciation dictionary. Text formatting instructions (radiologist-to-transcriptionist) are segmented within braces as in the 1st and 5th samples (in Figure 1, instructions are translated in English for clarity). Numbers are expanded to their spoken form as in the 3rd and 6th samples. In each sample, Line 1 represents the corresponding text span in the archived report. Observing the differences between Line 1 and Line 2 in the aligned speech and text corpora, rewrite rules are specified to automatically expand the text corpus of archived reports into a derived text corpus (Line 3) for language modelling. Rewrite rules typically deal with abbreviations, symbols and numbers (see the 2nd, 3rd and 6th examples). Line 4 represents the expected final output of the ASR system, where transcriptions are converted from the spoken form into the standard form of written reports by using reverse rewrite rules.

Anonymisation of specific segments in audio files is a lot more challenging task than text anonymisation. It would involve transcripts of inaccurate ASR, therefore, it could not be guaranteed that the anonymisation is done accurately. Since personal data should be dictated only at the beginning of a report, we are using a more simple and reliable approach instead: given the anonymised text corpus (Section 2), we extract the main body of the report (cropping off any metadata before and after the main text) and align it with the ASR result to find the segment of the given audio file, which corresponds to the main body of the text; the rest of the audio file is trimmed.

5. Conclusion

The language resources described in this paper are the most crucial part in the adaptation of a speech recognition system for the medical domain or any other highly specialised domain. The expected end result is dual: a self-service platform for instant radiology reporting, and a semi-automated platform for a more productive work at the transcriptionist centre. The next steps are further development of the text rewriting system for automatic verbalisation and deverbalisation, adaption of the ASR system for the medical domain, and the platform development, followed by evaluation and user studies. High accuracy ASR accompanied by productive user interfaces must be achieved for the overall system to be accepted and used in practice.

Acknowledgements

This work has received financial support from the European Regional Development Fund under the grant agreement No. 1.1.1.1/18/A/153 (*Latvian Speech Recognition and Synthesis for Medical Applications*).

References

[1] Pinnis M, Auzina I, Goba K. Designing the Latvian speech recognition corpus. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC). Reykjavik, Iceland; 2014. p. 1547–1553.

[2] Salimbajevs A, Strigins J. Latvian speech-to-text transcription service. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH). Dresden, Germany; 2015. p. 722–723.

[3] Znotins A, Polis K, Dargis R. Media monitoring system for Latvian radio and TV broadcasts. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH). Dresden, Germany; 2015. p. 732–733.

[4] Paats A, Alumäe T, Meister E, Fridolin I. Retrospective Analysis of Clinical Performance of an Estonian Speech Recognition System for Radiology: Effects of Different Acoustic and Language Models. Journal of Digital Imaging. 2018;31(5):615–621.

[5] Alumäe T, Paats A, Fridolin I, Meister E. Implementation of a Radiology Speech Recognition System for Estonian Using Open Source Software. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH); 2017. p. 2168–2172.

[6] Spektors A, Auzina I, Dargis R, Gruzitis N, Paikens P, Pretkalnina L, et al. Tezaurs.lv: the largest open lexical database for Latvian. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC). Portoroz, Slovenia; 2016. p. 2568–2571.

[7] Znotins A, Cirule E. NLP-PIPE: Latvian NLP Tool Pipeline. In: Human Language Technologies - The Baltic Perspective. vol. 307 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2018. p. 183–189.

[8] Pinnis M, Salimbajevs A, Auzina I. Designing a speech corpus for the development and evaluation of dictation systems in Latvian. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC). Portoroz, Slovenia; 2016. p. 775–780.