Abstract. This paper describes implementation and evaluation of an Estonian large vocabulary continuous speech recognition system prototype for the radiology domain. We used a 44 million word corpus of radiology reports to build a word trigram language model. We recorded a test set of dictated radiology reports using ten radiologists. Using speaker independent speech recognition, we achieved a 9.8% word error rate. Recognition worked in around 0.5 real-time. One of the prominent sources of errors were mistakes in writing compound words.

Keywords. speech recognition, radiology, applications

Introduction

Radiology has historically been one of the pioneer domains in large vocabulary continuous speech recognition (LVCSR) among several languages. Radiologists’ eyes and hands are busy during the preparation of a radiological report, creating thus a suitable condition for speech based input as an alternative to keyboard based text entry. In many hospitals, radiologists dictate the reports which are then converted to text by human speech transcribers. Speech recognition systems have the potential to replace human transcribers and enable faster and less expensive report delivery. For example, [1] describes a case study where the use of speech recognition decreased the mean report turnaround time by almost 50%. In radiology, a typical active vocabulary is much smaller than in general communication and the sentences have a more well-defined structure, following certain patterns. This enables to create statistical language models for the radiology domain that are accurate and have a good coverage, given enough training data [2]. Over the past several years, speech recognition for radiology has greatly improved and some vendors claim up to 99% accuracy in word error rate (WER) [3]. However, most vendors provide systems only for the biggest languages (although Nuance’s SpeechMagic supports 25 languages [4]) and there have been no known attempts of building a speech recognition system for radiology for the Estonian language.

We have found one report on building a radiology-focused speech recognition system for medium or little resourced languages. In [5], a radiology dictation system for Turkish is developed, with an emphasis on improving modeling of pronunciation variation. A very low word error rate of 4.4% was reported, however, the vocabulary of the system was only 1562.
In this paper, we describe our implementation of the Estonian speech recognition system for the radiology domain. In the first section, we describe the training data and different aspects of building the system. In the second section, the process of collecting test data is explained and the experimental results are reported. Some error analysis is performed. The paper ends with a discussion and gives some plans for future work.

1. LVCSR system

1.1. Language model

For training a language model (LM), our industry partner made a large anonymized corpus of digital radiology reports available to us. The corpus contains over 1.6 million distinct reports, and has 44 million word tokens before normalization. We randomly selected 600 reports for development and testing and used the rest for training. The corpus contains over 480 thousand distinct words (including all different numbers).

We created a word trigram LM from the corpus. In previous LVCSR systems for Estonian, sub-word units have been used instead of words as basic units of the LM, since the inflectional and compounding nature of the language makes it difficult to achieve good vocabulary coverage with words [6]. However, our initial investigations revealed that in the radiology domain, the vocabulary is much more compact and words can be used as basic units.

The corpus was normalized using a multi-step procedure:

1. A large set of hand-made rules (implemented using regular expressions) was applied that expanded and/or normalized different types of abbreviations often used by radiologists.
2. A morphological analyzer [7] was used for determining the part-of-speech (POS) properties of all words.
3. Numbers were expanded into words, using the surrounding POS information for determining the correct inflections for expansion.
4. The resulting corpus was used for producing two corpora: one including verbalized punctuations and another without punctuations.
5. A candidate vocabulary of all words that occur at least five times was generated.
6. Pronunciations for all words were generated using simple Estonian pronunciation rules [6]. A list of exceptions was used to acquire pronunciations for abbreviations not expanded in the first step.
7. Words which had pronunciations that were now composed only of consonants were removed. Such words are mostly spelling errors and abbreviations whose pronunciations has not been defined in the exception dictionary. However, during recognition, such mis-modeled words are often mis-inserted instead of fillers.
8. The LM vocabulary was composed of the most frequent 50,000 words from the remaining candidate vocabulary.
9. Two trigram LMs – one with verbalized punctuations and another without punctuations – were built, using interpolated modified Kneser-Ney discounting [8]. The two LMs were interpolated into one final model. We model optional verbalized punctuations in our LM because some recordings in our test set contain them.
The LM perplexity against the development texts without verbalized punctuation is 37.9. The perplexity against development texts with verbalized punctuations is 31.6. The out-of-vocabulary (OOV) rate is 2.6%. However, the majority of the OOV words are compound words that are not present as compounds in the LMs but whose compound constituents are present. There are also many spelling errors that contribute to the relatively high OOV rate.

1.2. Acoustic models

We used off-the-shelf acoustic models trained on various wideband Estonian speech corpora: the BABEL speech database [9] (phonetically balanced dictated speech from 60 different speakers, 9h), transcriptions of Estonian broadcast news (mostly read news speech from around 10 different speakers, 7.5h) and transcriptions of live talk programs from three Estonian radio stations (42 different speakers, 10h). The latter material consists of two or three hosts and guests half-spontaneously discussing current affairs or certain topics, and includes passages of interactive dialogue and long stretches of monologue-like speech. Speech signal is sampled at 16 kHz and digitized using 16 bits. From a window of 25 ms, taken after every 10 ms, 13 MFC coefficients are calculated, together with their first and second order derivatives. A maximum likelihood linear transform, estimated from training data, is applied to the features, with an output dimensionality of 29. The data is used to train triphone HMMs for 25 phonemes, silence and nine different noise and filler types. The models comprise 2000 tied states, each state is modeled with eight Gaussians.

2. Evaluation

2.1. Data acquisition

For performing speech recognition experiments, we recorded a small speech corpus of radiology reports, using 10 subjects (5 male and 5 female). The recordings were performed using a close-talking microphone. The subjects dictated different reports in the test set of our radiology text corpus. As all subjects were professional radiologists with variable degree of experience (from 1 to 15 years) in writing radiology reports, they had no difficulties in reading specific medical terminology and in interpretation of different abbreviations. Nine subjects were native Estonian speakers, one subject was a non-native speaker with a slight perceived foreign accent (her speech samples were not used in testing). The speakers did not receive any special training before the recordings. Thus, some radiologists chose to dictate the reports with verbalized punctuations, while the majority did not include them. The total length of the recordings was 4 hours and 23 minutes, with 26 minutes per speaker on average.

The recordings were then manually transcribed, using the report texts given to the speakers as templates. The total number of running words in the transcripts is 19,486. The number of unique words is 4317.
Table 1. Word error rates for different speakers and the average.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>7.3</td>
</tr>
<tr>
<td>AR</td>
<td>8.5</td>
</tr>
<tr>
<td>AS</td>
<td>8.5</td>
</tr>
<tr>
<td>ER</td>
<td>10.3</td>
</tr>
<tr>
<td>JH</td>
<td>13.3</td>
</tr>
<tr>
<td>JK</td>
<td>9.2</td>
</tr>
<tr>
<td>SU</td>
<td>10.7</td>
</tr>
<tr>
<td>VE</td>
<td>8.7</td>
</tr>
<tr>
<td>VS</td>
<td>11.9</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>9.8</strong></td>
</tr>
</tbody>
</table>

2.2. Recognition experiments

The recognition experiments were performed using the CMU Sphinx 3.7 open source recognizer\(^1\). The recognizer was configured to use relatively narrow search beam and ran in 0.5 × real time on a machine with Intel Xeon X5550 processor (2.66 GHz, 8 MB cache, 667 MHz DDR3 memory).

The WER results per speaker and on average are given in Table 1.

We analyzed the recognition errors and found that around 17% of them are “word compounding” errors – a compound word is recognized as a non-compound, or vice versa (i.e., the only error is in the space between compound constituents). Such errors have a high impact on the WER since they are counted as one substitution error and one deletion error, or one substitution error and one insertion error. However, such errors have probably the lowest impact on the perceived quality of the recognized text. Often the fact, whether a word is a real compound word or not is arguable even for humans. Other prominent sources of errors were spelling errors in the reference transcripts (17% of all errors) and normalization mismatches, i.e., situations where reference and hypothesis represent the same term using different levels of expansion or normalization (e.g., \(C\ kuus\) ‘C six’ vs. \(C6\), 11%). Thus, only around 55% of the errors were “real” recognition errors.

3. Discussion and future work

The paper described a pilot study of an Estonian LVCSR system for radiology. Using off-the-shelf acoustic models, and a word trigram language model built from a large corpus of proprietary radiology reports, a word error rate of 9.8% was achieved, using one-pass speaker independent recognition. Brief error analysis suggests that almost half of the errors can be contributed to different normalization and compound word representation problems.

The system can be improved in many ways. First, we have gathered additional transcribed wideband speech data for training acoustic models since doing the experiments reported here (over 13 hours of conference speech recordings, additional 10 hours of broadcast conversations from radio). Second, our system did not use adaptation of any

\(^1\)http://cmusphinx.org
kind, while in a such system, both acoustic model adaptation towards specific speakers, as well as language model adaptation towards certain parameters of the radiological study will probably improve the accuracy of the system.

The speech recognition experiments were conducted using speech from written radiology reports read out aloud. Such setting might have skewed the results in a positive direction, since when dictating spontaneously new reports, the concentration of speech disfluencies at lexical, syntactic, and acoustic-prosodic levels is probably much higher. In order to measure the WER more realistically, we should perform Wizard of Oz style experiments where radiologists produce reports for previously unseen images. However, to gain an objective insight of the system potential, the subjects should also receive some training on dictating the reports using a speech recognition system.

This study concentrated only on the speech recognition aspects of voice-automated transcription of radiology reports. There are many post-processing steps, such as consistent normalization of read numbers, dates, abbreviations, and proper structuring of the generated reports, that are perhaps even more important for the report availability and time efficiency of the voice-automated reporting process. Also, the best benefit from speech recognition can be obtained if it is fully integrated into a radiology information system (RIS) [10]. We are planning to continue the cooperation with our industry partner on such aspects of the system.

Acknowledgments

This research was partly funded by the target-financed theme No. 0322709s06 of the Estonian Ministry of Education and Research and by the National Programme for Estonian Language Technology, and by Cybernetica Ltd’s project CyberDiagnostika supported by Enterprise Estonia.

References