Estonian Large Vocabulary Speech Recognition System for Radiology

Tanel Alumäe, Einar Meister

Institute of Cybernetics
Tallinn University of Technology, Estonia

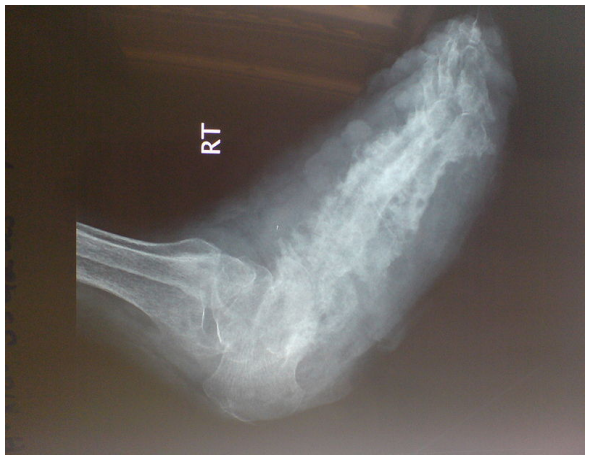
October 8, 2010

Radiology

Radiology

Radiology is the branch or specialty of medicine that deals with the study and application of imaging technology (such as X-ray, ultrasound, computer tomography) and radiation to diagnosing and treating disease.

Radiologist views and interprets a radiology image and creates a report that describes the findings.

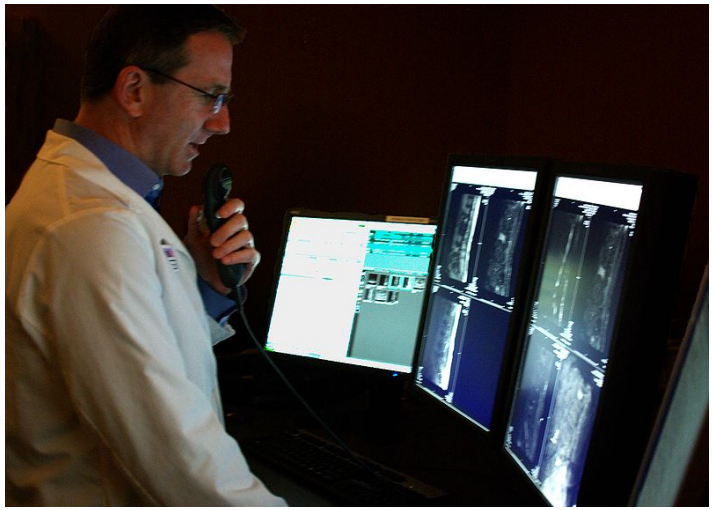


Achilles tendon has a uniform structure, tears are not detected. Left, there is a small tissue swelling of the tendon side.

Motivation

- Radiologists' eyes and hands are busy during the preparation of a radiological report
- 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
- In radiology, a typical active vocabulary is much smaller than in general communication and the sentences usually have a well-defined structure



Acoustic models

We used various wideband Estonian speech databases for training acoustic models:

- BABEL speech database (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)
- transcriptions of live talk programs from three Estonian radio stations (42 different speakers, 10h)

Models:

- MFCC features
- 25 phonemes, silence, nine fillers
- triphone models
- 2000 tied states, 8 Gaussian per state
- CMU Sphinx

Language model

For training a language model:

- 1.6 million distinct reports, with 44 million word tokens
- Normalization:
  - expanded and/or normalized different abbreviations using hand-made rules
  - used a morphological analyzer for determining the part-of-speech (POS) properties of all words
  - expanded numbers
- the resulting corpus was used for producing two corpora: one including verbalized punctuations and another without punctuations.
- vocabulary was composed of the most frequent 50 000 words
- Two trigram LMs – one with verbalized punctuations and another without punctuations – were built. The two LMs were interpolated into one final model. Perplexity 35, OOV 2.6%

Evaluation

- Recorded a small test corpus of radiology reports
- Dictated by 10 radiologists, 26 minutes per speaker on average
- Actual reports from our test set
- Recordings were manually trascribed

Results

- Used CMU Sphinx 3.7, running in 0.5 x real-time
- one-pass speaker independant system

Speaker	WER
AL	7.3
AR	8.5
AS	8.5
ER	10.3
JH	13.3
JK	9.2
SU	10.7
VE	8.7
VS	11.9
Average	9.8

Error analysis

Errors:

- 17% of errors 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)
- 17% due to spelling errors
- 11% normalization mistmatches (e.g., *C kuus* ’C six’ vs. *C6*)

Thus, only around 55% of the errors were “real” recognition errors

Future work

- We have now more wideband speech data (55 hours in total)
- Adaptation:
  - Acoustic model: adapt to the voice of a speaker
  - Language model: adapt to the typical report content of a speaker (e.g., one radiologists might be specialized on MRI images)
- Perform Wizard of Oz style experiments where radiologists produce reports spontaneously for previously unseen images
- Post-processing: consistent normalization of read numbers, dates, abbreviations, and proper structuring of the generated reports
- Integrate the system into the radiology information system (RIS)

Thanks!