Free English and Czech telephone speech corpus

shared under the CC-BY-SA 3.0 license

Matěj Korvas, Ondřej Plátek,
Ondřej Dušek, Lukáš Žilka, Filip Jurčíček

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague

May 30th, 2014
LREC, Reykjavík, Iceland

Introduction

The Vystadial 2013 telephone speech corpus

- Two corpora of transcribed telephone speech, English and Czech
- Under a free license
- Distributed with scripts for ASR training

Introduction

The Vystadial 2013 telephone speech corpus

- Two corpora of transcribed telephone speech, English and Czech
- Under a free license
- Distributed with scripts for ASR training

Outline

1. Acquiring the data using crowdsourcing
2. ASR training scripts
3. Evaluation

Motivation

ASR for a spoken dialogue system?

- Commercial (Nuance & others) – costly, restrictive license
- Cloud-based (Google, Nuance) – costly or unclear licensing
- Custom ASR model – data needed
  - Available for English
  - Restrictive license and/or costly for non-LDC members

Motivation

ASR for a spoken dialogue system?

- Commercial (Nuance & others) – costly, restrictive license
- Cloud-based (Google, Nuance) – costly or unclear licensing
- Custom ASR model – data needed
  - Available for English
  - Restrictive license and/or costly for non-LDC members

The Vystadial 2013 Speech corpus

- English and Czech, telephone speech
- CC-BY-SA 3.0 license: for research and commercial use
- Training scripts for HTK and Kaldi ASR toolkits

English Data

Collection

- Using crowdsourcing via Amazon Mechanical Turk
- Most speakers: American English
- Interaction with a spoken dialogue system – restaurant information domain

English Data

Collection

- Using crowdsourcing via Amazon Mechanical Turk
- Most speakers: American English
- Interaction with a spoken dialogue system – restaurant information domain

Transcription

- Also using Amazon Mechanical Turk
- Quality checks, restricted to experienced workers
- Orthographic, with non-speech events
  - `__NOISE__`, `__LAUGH__`

Data Collection – Czech

Collection

- Using crowdsourcing, free Czech phone numbers (AMT unavailable)
  - Call-a-friend
  - Repeat-after-me
  - Spoken dialogue system – public transport information
- License agreement at the beginning of the call

Data Collection – Czech

Collection

- Using crowdsourcing, free Czech phone numbers (AMT unavailable)
  - Call-a-friend
  - Repeat-after-me
  - Spoken dialogue system – public transport information
- License agreement at the beginning of the call

Transcription

- Similar to English
- Hired transcribers
- Anonymization (personal information excluded)

Data

Size

- English: 41 hours, 47k sentences (178k words)
- Czech: 15 hours, 22k sentences (126k words)
- + 2k sents dev, 2k sents test in both languages (ca. 1.5 hr each)

Data

Size

- English: 41 hours, 47k sentences (178k words)
- Czech: 15 hours, 22k sentences (126k words)
- + 2k sents dev, 2k sents test in both languages (ca. 1.5 hr each)

Characteristics

- Different sources (no problem for a general acoustic model)
  - English: narrow domain
  - Czech: general domain (multiple domains)
- 16kHz mono WAV files (`X.wav`)
  + matching plain text files with transcription (`X.wav.trn`)

ASR Acoustic Modelling Scripts

- Scripts to create acoustic models for ASR
- Coding recordings into MFCCs + Δ + ΔΔ features
- For both languages, for HTK and Kaldi

ASR Acoustic Modelling Scripts

- Scripts to create acoustic models for ASR
- Coding recordings into MFCCs + Δ + ΔΔ features
- For both languages, for HTK and Kaldi
- Easily applicable to other data sets (and other languages):
  - Just need `X.wav` + `X.wav.trn`
- Language-specific parts:
  - List of phones in the language
  - Orthography-to-phonetics mapping (dictionary and/or rules)
  - “Phonetic questions” – to group similar triphones (HTK only)

HTK vs. Kaldi

HTK

- Hidden Markov models, Gaussian mixtures
- EM training: uniform $\rightarrow$ monophone $\rightarrow$ triphone model
- Triphones clustered using phonetic questions

HTK vs. Kaldi

HTK

- Hidden Markov models, Gaussian mixtures
- EM training: uniform $\rightarrow$ monophone $\rightarrow$ triphone model
- Triphones clustered using phonetic questions

Kaldi

- Finite state transducers
- Generative models parallel to HTK (but Viterbi training)

HTK vs. Kaldi

HTK

- Hidden Markov models, Gaussian mixtures
- EM training: uniform $\rightarrow$ monophone $\rightarrow$ triphone model
- Triphones clustered using phonetic questions

Kaldi

- Finite state transducers
- Generative models parallel to HTK (but Viterbi training)
- Discriminative models:
  - Multiple methods and feature transformations available
  - Our models: non-speaker-adaptive
  - BMMI training (with unigram LM), LDA + MLLT transformations

Evaluation

- Generative with similar complexity + discriminative for Kaldi
- 0-gram and bigram LMs (testing acoustic models & real use)
- Czech: bigger dictionary & higher perplexity than English

Evaluation

- Generative with similar complexity + discriminative for Kaldi
- 0-gram and bigram LMs (testing acoustic models & real use)
- Czech: bigger dictionary & higher perplexity than English

Word Error Rate

	kit	method	0-gram	bigram
Czech	HTK	tri $\Delta + \Delta\Delta$	64.5	60.4
	Kaldi	tri $\Delta + \Delta\Delta$	69.3	53.8
		tri LDA + MLLT	65.4	51.2
		tri LDA + MLLT / BMMI	–	48.0
English	HTK	tri $\Delta + \Delta\Delta$	50.0	17.5
	Kaldi	tri $\Delta + \Delta\Delta$	41.1	17.5
		tri LDA + MLLT	37.3	17.2
		tri LDA + MLLT / BMMI	–	12.0

Thank you for your attention

Links

- The corpora (CC-BY-SA 3.0 + Apache 2.0):
  http://bit.ly/free-phone-corp

Thank you for your attention

Links

- The corpora (CC-BY-SA 3.0 + Apache 2.0):
  `http://bit.ly/free-phone-corp`
- Online lattice decoding for Kaldi:

  Plátek & Jurčíček: *Free on-line speech recogniser based on Kaldi ASR toolkit producing word posterior lattices.* To appear at SIGDIAL in June.

Thank you for your attention

Links

- The corpora (CC-BY-SA 3.0 + Apache 2.0):
  `http://bit.ly/free-phone-corp`
- Online lattice decoding for Kaldi:

  Plátek & Jurčíček: *Free on-line speech recogniser based on Kaldi ASR toolkit producing word posterior lattices.* To appear at SIGDIAL in June.
- Our spoken dialogue systems framework (Apache 2.0):
  `https://github.com/UFAL-DSG/alex`

Thank you for your attention

Links

- The corpora (CC-BY-SA 3.0 + Apache 2.0):
  http://bit.ly/free-phone-corp
- Online lattice decoding for Kaldi:

  Plátek & Jurčíček: *Free on-line speech recogniser based on Kaldi ASR toolkit producing word posterior lattices.* To appear at SIGDIAL in June.
- Our spoken dialogue systems framework (Apache 2.0):
  https://github.com/UFAL-DSG/alex

Contact us

Ondřej Dušek
Institute of Formal and Applied Linguistics
Charles University in Prague
odusek@ufal.mff.cuni.cz