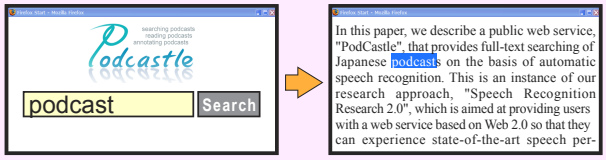
PodCastle

Goal

❑ Full-text retrieval of speech data
- Podcasts (audio blogs)
- Individual audio files
- Video clips *(YouTube, Ustream.tv, and Nico Nico Douga)*



❑ ASR (automatic speech recognition) for text transcription
- Difficult to achieve high accuracy
- Diversity of topics, vocabularies, and speaking styles

PodCastle

http://podcastle.jp

❑ Speech retrieval web service based on ASR and crowdsourcing
- Collect and amplify voluntary contributions by anonymous users

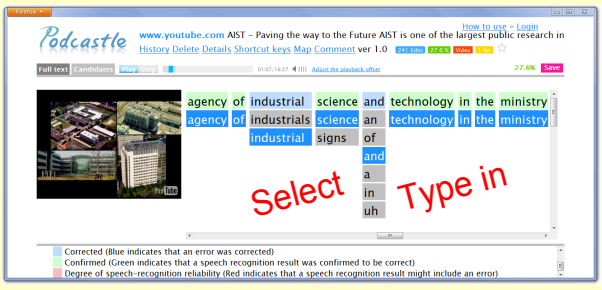
❑ Automatic learning from the web
- Automatically collect new words/phrases, their pronunciation, and usage examples
  News articles (Yahoo! news) and web dictionaries
- Add new words to ASR dictionary (0.24M words)

❑ Users can find and correct ASR errors
- Original efficient error correction interface
  [Ogata & Goto, Interspeech 2005]
- Improve retrieval performances by correct indices
- Improve recognition performances by automatic learning (adaptation/training)

Collaborative Training of ASR

❑ Incorporate error corrections (wisdom of crowds) into ASR training
- Collaborative training of acoustic models (AMs)
  Podcast-dependent acoustic model trained using transcripts corrected by users
  [Ogata & Goto, Interspeech 2007, 2009, SSCS 2009]
  Relative error reduction of 21-33%
  [Ogata & Goto, Interspeech 2009]
- Collaborative training of language models (LMs)
  [Ogata & Goto, Interspeech 2012: this paper]

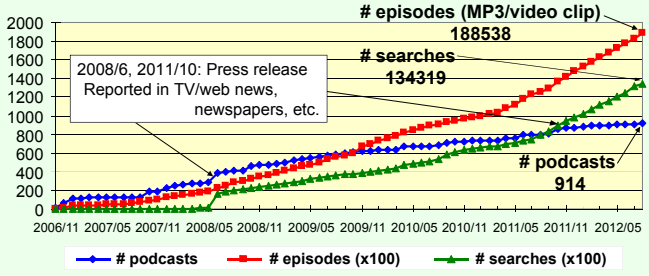
❑ Generate content-specific LM training data
- Overcome difficulties in preparing task-specific language models in advance
  A common approach of building a task-specific corpus in advance is impractical because it will be too costly and time consuming to cover the diversity
- Use the benefits of our web service PodCastle

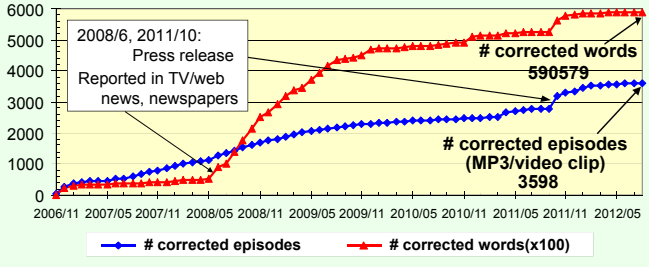


Candidate list is generated by using a confusion network
- Transcripts voluntarily corrected by end users are used for language-model training

Usage Analysis of PodCastle

❑ How widely used? (as of Jan 1st, 2012)
- 877 Japanese speech programs
  Podcasts and YouTube channels
- Consist of 147,280 audio files in total
- 3,279 audio files were partially corrected
- 580,765 corrected words (errors)
  177.1 corrected words per corrected file on average
  51.8% were corrected by the candidate selection
  48.2% were corrected by the text typing
- There are users who voluntarily cooperate in the correction
  Speech data recorded by famous artists and TV personalities tend to receive many corrections





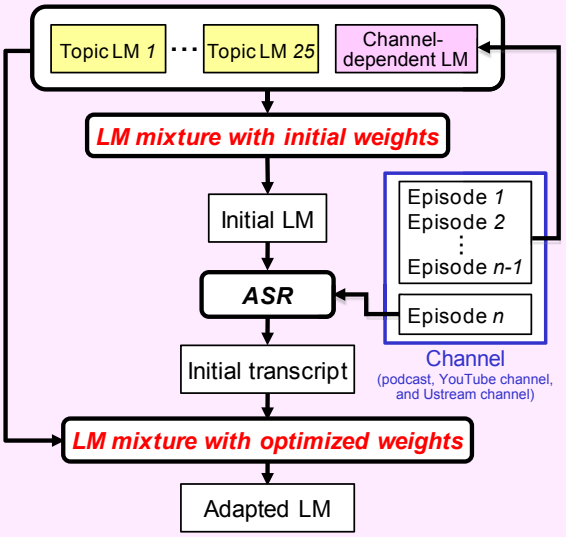
Collaborative Training of Language Models on the Basis of Wisdom of Crowds

Jun Ogata and Masataka Goto (AIST, Japan)

Dynamic Language Modeling System

❑ Dynamically mix two kinds of LMs for each input episode (audio file)

- Use corrected ASR transcripts on PodCastle
- Iterative adaptation system for LVCSR



❑ Topic LMs (background LMs)

- Use web news text from news aggregation site
  Text articles (Yahoo! news) published during 40 months
- Each covers a different topic
  Train 25 topic LMs (6 main topics with 25 sub-topics)

❑ Channel-dependent LM

- Use ASR transcripts corrected by end users
- Generate LM on the fly by using all the other episodes within the same channel/program
  Assume that episodes belonging to each channel have a linguistic similarity in terms of topics, vocabularies, and frequent phrases

❑ Dynamic mixture of topic LMs

- Model-level mixture scheme
  Combine n-gram probabilities of 26 component LMs (25 topic LMs with channel-dependent LM)

$$p_{mix}(w \mid h) = \sum_i \lambda_i p_i(w \mid h)$$

Mixture weights λ_i are computed by minimizing the perplexity on initial ASR transcript via EM algorithm

Experiments

❑ Speech data (partially corrected on PodCastle)

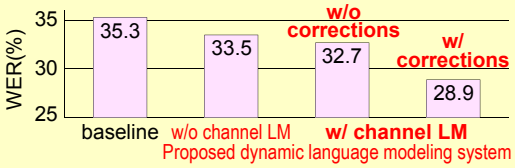
- Test data: 45 episodes from 8 Japanese podcasts
- Training data for channel-dependent LM

ID	Domain	Main topic	Test data #episodes (#words)	Training data #episodes (#words)
A	news	*multi*	4 (11,170)	94 (259,840)
B	news	*multi*	4 (4,937)	23 (24,908)
C	lecture	politics	20 (13,876)	79 (65,984)
D	lecture	business	5 (10,763)	35 (125,432)
E	quiz show	general knowledge	2 (1,910)	30 (39,098)
F	lecture	health & care	2 (3,292)	71 (193,569)
G	chitchat	music	4 (25,239)	56 (283,414)
H	chitchat	show business	4 (9,251)	60 (127,840)

❑ ASR system [Ogata, Goto & Eto, Interspeech 2007]

- Multi-pass decoding with unsupervised MLLR
  GMM-based audio segmentation (speech/music/others)
  Confusion network generation
- Baseline acoustic model:
  Tied-state triphone (4513 states, 16 Gaussians/state)
  Trained with 600 h of presentation speech data (CSJ)
- Language model: *3-gram (vocabulary size: 286,345)*

❑ WER results (average)



Corrected transcripts are effective in language modeling

- Comparison of weight optimization (WER%)

	unsupervised	supervised
w/o channel LM	33.5	33.4
w/ channel LM	28.9	28.6

unsupervised: w/ recognition hypothesis
supervised: w/ manual transcriptions

History

❑ http://podcastle.jp since 2006

- 2006/01 Started the project
- 2006/12 Released to the public
  The world's first speech retrieval using crowdsourcing
- 2008/06 Press release *(media coverage)*
- 2011/10 Press release *(media coverage)*
- 2011/10 Launch the English version
  Powered by ASR of CSTR, Univ. of Edinburgh
  in collaboration with CSTR and AIST

The Centre for Speech Technology Research, The University of Edinburgh

Summary

❑ Dynamic language modeling for spoken content on the web

- Use of topic LMs based on web news
- Use of channel-dependent LM generated by user transcripts (wisdom of crowds)
- Model-level mixture with optimal weights
- Plan to deploy the proposed system on the PodCastle web service to improve ASR results

Video clip of PodCastle:
http://staff.aist.go.jp/m.goto/PodCastle/

2012/09/13 Interspeech 2012 poster