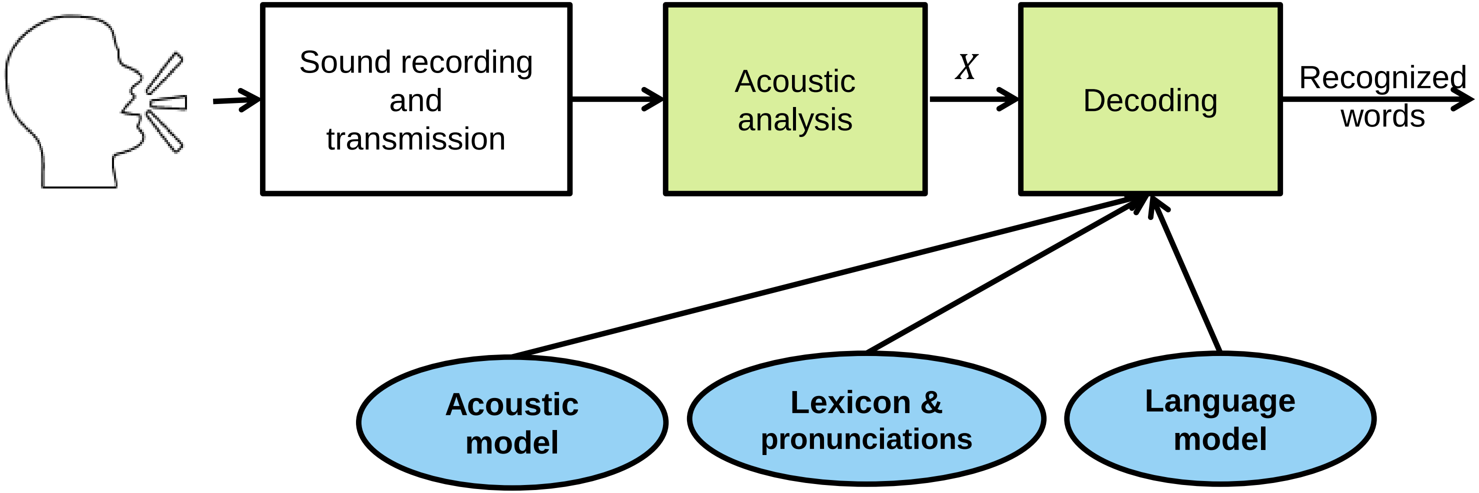
Language Models for Speech Recognition

Denis Jouvet
Inria, CNRS, Université de Lorraine
June 2022

Automatic speech recognition



Represents how sounds are pronounced (for one or several speakers, in a given environment)

List of the words known by the system, with their pronunciation variants

Possible sequences of words

About manually transcribed data

Best speech recognition performance is achieved with **models trained** on **in domain data** using **manually transcribed data**

However, in early development stages of a new application

- Only limited amounts of in domain data is available
- And manual transcription may not be available (*expensive, time consuming*)

Hence, investigation of **training / adaptation using uncertain transcriptions** (obtained through automatic speech recognition) on a limited amount of in-domain data

Note: this concerns both acoustic and language models

About privacy

Speech data contains a lot of personal information

- Identity of speakers can be recovered from speech signal
- Linguistic content can refer to personal data, such as person names, locations, telephone numbers, …

Hence the interest of sharing anonymized data, where

- Speech signal is transformed to sound as pronounced by another speaker (e.g., voice conversion)
- Lexical content is modified to remove personal information

But this impacts on language model training

Overview

- Training models using uncertain transcriptions (i.e., from automatic speech recognition)
- Impact of anonymization process on training language models

Based on results from the COMPRISE project

- EU Horizon 2020 Research and Innovation Program
- COMPRISE: cost-effective, multilingual, privacy-driven voice-enabled services [Dec. 2018 – Nov. 2021]
- https://www.compriseh2020.eu/

Training models using uncertain transcriptions (from automatic speech recognition)

Semi-supervised training of acoustic models

1/ Train initial acoustic models from labeled speech data

- i.e. speech data with associated correct transcriptions
- Whether from a different domain (*domain mismatch*)
  or from same domain (*matched-domain*)

2/ Automatically annotate (transcribe) speech data from the new domain and use that data for fine tuning the acoustic model
several approaches are possible:

- Use directly the recognized words (provided by the speech recognition system)
- Or, apply an “error detection” module based on a deep neural network
  and replace words tagged as “error” by “unknown word”

Speech data

Two corpora are considered

- Librispeech (LS): English **read speech**, clean condition, 100 hours
- Verbmobil (VM): **conversational speech** corpus, English speech

Experiments

- Initial model
  - ***Domain mismatch***: trained on Librispeech (→ 100 hours)
  - ***Matched-domain***: trained on a subset of Verbmobil corpus (→ 5 hours)
- Semi-supervised training
  - Using another subset of Verbmobil corpus (→ 20 hours)
- Evaluation
  - Word error rates computed on a Verbmobil test set (→ 3 hours)

Evaluation of initial acoustic models

Word error rates computed on a Verbmobil test set (⬄ 3 hours)

	Init: Librispeech 100 h (*domain mismatch*)		Init: Seed Verbmobil 5 h (*matched domain*)	
Initial model	LS 100 h	41.0 %	VM 5 h	**37.8 %**

Better performance when training on matched domain data (even if lower amount of data)

Evaluation of semi-supervised training

Adaptation using a subset of Verbmobil corpus (20 hours, non-transcribed)

	Init: Librispeech 100 h (*domain mismatch*)		**Init: Seed Verbmobil 5 h (*matched domain*)**	
Initial model	LS 100 h	41.0 %	VM 5 h	37.8 %
Semi-supervised training using speech recognition hypotheses		38.0 %		33.7 %
Semi-supervised training using speech recognition output **& error detection**		37.6 %		**32.0 %**

Better performance with error detection module (thus ignoring a few words)

Comparison with oracle performance

	Init: Librispeech 100 h (*domain mismatch*)		Init: Seed Verbmobil 5 h (*matched domain*)	
Initial model	LS 100 h	41.0 %	VM 5 h	37.8 %
Semi-supervised training using speech recognition hypotheses		38.0 %		33.7 %
Semi-supervised training using speech recognition output **& error detection**		37.6 %		**32.0 %**
Oracle (i.e., using correct transcriptions of adaptation data)	LS 100 h + VM 20 h	30.2 %	VM 5 h + VM 20 h	**26.4 %**

Language models

N-gram based language models

- Used in speech recognition since many years
- Provide the probability of a word knowing the *N-1* previous words
- A good compromise is $N = 3$

Neural network based language models

- More recent approaches
- Leads to better performance, especially when large amounts of training data are available

Training language models

Conventional training

- Relies on text data
  - From written texts
  - Manual transcription of speech data

Training from speech recognition hypotheses

- Problem of speech recognition errors (some words are incorrect)
  This is taken into account in the training process, by considering alternate hypotheses

Language models

Word error rates computed on a Verbmobil test set

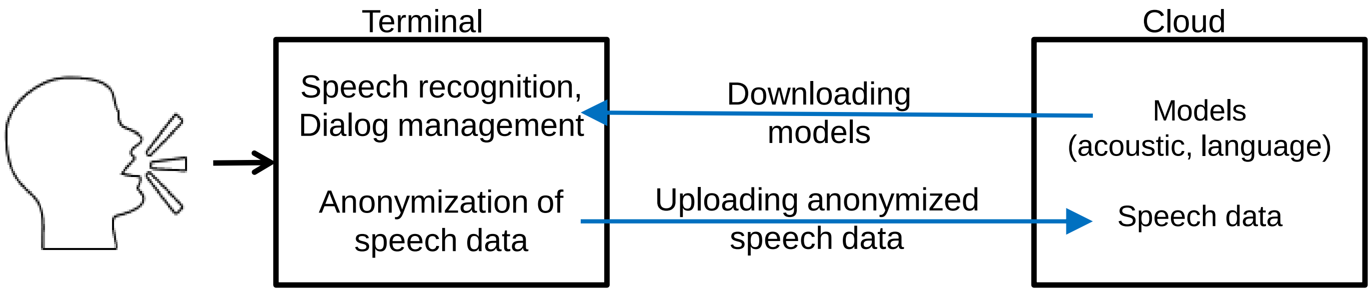
	Verbmobil English test data
3-gram trained on Verbmobil training labeled data (5 h)	39.7 %
Neural network LM trained on Verbmobil labeled data (5 h) & Verbmobil **unlabeled** data (20 h)	**36.1 %**
Oracle :: Neural network LM trained on Verbmobil labeled data (5 h) & Verbmobil labeled data (20 h)	32.9 %

Impact of anonymization process on training language models

Automatic speech recognition and privacy

Speech recognition as local processing on terminal

Sharing of anonymized data (for training models) in cloud



Privacy

Removing of personal information, such as person names, organizations, locations, …

Original	Hi, Mister Miller, the Lufthansa flight from Frankfurt Airport to Rome is leaving at six pm.

Privacy

Removing of personal information, such as person names, organizations, locations, …

Original	Hi, Mister Miller, the Lufthansa flight from Frankfurt Airport to Rome is leaving at six pm.

Privacy

Removing of personal information, such as person names, organizations, locations, …

Original	Hi, Mister Miller, the Lufthansa flight from Frankfurt Airport to Rome is leaving at six pm.
Typed place holder	Hi, Mister PER, the ORG flight from LOC to LOC is leaving at TIME.
Entity replacement	Hi, Mister John, the Bosch flight from New York to Berlin is leaving at twelve pm.

Language models

Three models are considered

- N-gram word-based
  - Provide probability of a word knowing previous words
- N-gram class-based
  - Similar to previous one, but considers a few classes: person-names, organizations, locations, …
  - And takes into account the probability of the words inside classes
- Neural network based

Language models

Evaluation on original data: performance expressed as word error rate

			Training on original data
3-gram word-based			28.8 %
3-gram class-based			29.3 %
Neural network LM (LSTM-based)			**27.6 %**

- Training on original data ➔ Neural-network based model is the best

Language models

Evaluation on original data: performance expressed as word error rate

	Training on anonymized data		Training on original data
3-gram word-based	32.3 %		28.8 %
3-gram class-based	**30.2 %**		29.3 %
Neural network LM (LSTM-based)	30.5 %		**27.6 %**

- Training on original data ➔ Neural-network based model is the best
- Training on anonymized data ➔ Best results obtained with the class-based model

Adaptation using a small amount of original data

Evaluation on original data: performance expressed as word error rate

	Training on anonymized data	**+ adaptation on small amount of original data**	Training on original data
3-gram word-based	32.3 %	31.2 %	28.8 %
3-gram class-based	**30.2 %**	**29.8 %**	29.3 %
Neural network LM (LSTM-based)	30.5 %	29.9 %	**27.6 %**

- With adaptation on small amount of original (i.e. NON-anonymized data)
  ➔ Best results still obtained with the class-based model

Conclusion

Conclusion

Training / adapting for targeted application

- Better to have a limited amount of transcribed data from the targeted application (*in-domain data*) rather than just training on a larger generic dataset (*domain mismatch*)
- Semi-supervised training improves the performance of the speech recognition models, and benefits from the use of a module to detect “speech recognition errors” that are later ignored in training

Dealing with privacy transformed (anonymized) data

- Anonymization modifies named entities (such as person names, locations, telephone numbers, …). This impacts the estimation of the language models
- Adaptation of language models using a limited amount of original (non anonymized) data improves the performance

References

- Imran Sheikh, Emmanuel Vincent, Irina Illina.
  On semi-supervised LF-MMI training of acoustic models with limited data.
  INTERSPEECH 2020, October 2020, Shanghai, China. hal-02907924

- Imran Sheikh, Emmanuel Vincent, Irina Illina.
  Transformer versus LSTM language models trained on uncertain ASR hypotheses in limited data scenarios.
  LREC 2022 - 13th Language Resources and Evaluation Conference, June 2022, Marseille, France. hal-03362828v2

- Mehmet Ali Tugtekin Turan, Dietrich Klakow, Emmanuel Vincent, Denis Jouvet.
  Adapting language models when training on privacy-transformed data.
  LREC 2022 - 13th Language Resources and Evaluation Conference, Jun 2022, Marseille, France. hal-03189354v2