EVALITA 2011

Evaluation of NLP and Speech Tools for Italian

EVALITA 2011

Automatic Speech Recognition
Large Vocabulary Transcription

M. Matassoni, F. Brugnara, R. Gretter

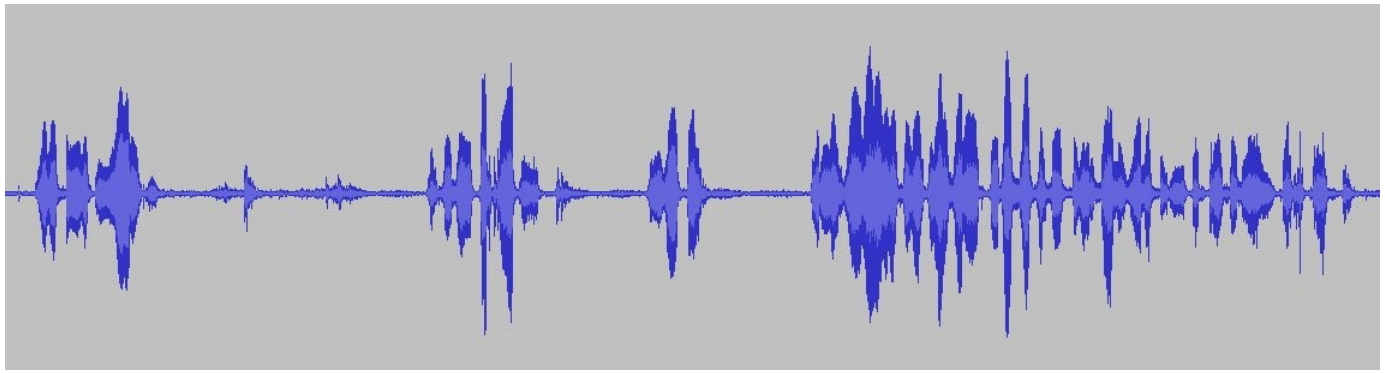
FONDAZIONE
BRUNO KESSLER

Outline

- Introduction
  - motivation
- Evaluation and results
  - datasets and metrics
  - participants and results
  - discussion
- Conclusion

Introduction

- Automatic Speech Recognition (ASR) toward increasingly complex models
- General goal: improve accuracy in different acoustic conditions and with larger vocabularies
- Transcription: interesting ASR application



Task features

Selected features:

- large vocabulary;
- large number of speakers;
- controlled recording conditions;
- spontaneous speech but with limited colloquial or dialectal expressions;
- availability of data for acoustic and language model training;
- “free” distribution of data

Parliament speeches

Italian Parliament speeches satisfy these requirements:

- audio and minutes of all the sessions publicly available;
- additional effort to manually annotate a (small) portion of the corpus already made by FBK;
- opportunity of defining two realistic subtasks, with different goals and different prospective application scenarios

Task description

Large Vocabulary Transcription task:

- Transcribe recordings of sessions of the Italian Parliament

Two sub-tasks:

- **transcription subtask**: an automatic transcription of the session by exploiting only the corresponding audio signal
- **constrained transcription subtask**: the accompanying minutes are provided, and participant can exploit them to produce more accurate results

Training modality

Two training modality:

- **Closed**: only distributed data are allowed for training and tuning the system
- **Open**: the participant can use any type of data for system training, declaring and describing the proposed setup.

Dataset for training

The training set consists in:

- ~30h of parliament audio sessions along with corresponding automatic transcriptions;
- 5-years (1 legislature) minutes of parliament sessions, for a total of about 32 millions running words;
- 74K-word lexicon covering acoustic training data and most of language model data.

```
75.645 75.725 è
75.725 76.405 legittimo
76.405 77.265 ammantare
77.265 77.455 @e
77.455 77.965 @bh
77.965 78.185 @ne
78.185 78.425 un
78.425 78.875 basso
78.875 79.635 concetto
79.635 79.785 @u
79.785 80.015 @n
80.015 80.145 di
80.145 80.535 nobili
80.535 80.965 parole
80.965 81.915 @n
```

Development/test sets

The development set contains:

- about 1 hour parliament audio session
- the minutes of the session
- the reference transcription (manually annotation)

The test set contains:

- about 1 hour parliament audio session
- minutes

Metrics

Evaluation based on word accuracy, computed as Minimum Edit Distance (Levenshtein) between the recognizer output and the reference annotation.

The resulting Word Error Rate computes:

- substitutions
- deletions
- insertions

Scores: (#C #S #D #I) 25 2 1 0
REF: in concreto significa voler imporre l' applicazione degli accordi collettivi stipulati anche nei confronti di quelle regioni che fossero CONTRARIE a detti accordi o A PARTE di essi
HYP: in concreto significa voler imporre l' applicazione degli accordi collettivi stipulati anche nei confronti di quelle regioni che fossero CONTRARI a detti accordi o * PARTI di essi

Sclite as evaluation tool:

- developed by NIST
- provided in the distributed package

Participants

- 5 prospective participants
- 3 submissions
- 1 withdrawal

- Vocapia Research, Orsay, France
- Fondazione Bruno Kessler, Trento, Italy

Results

Transcription task

Closed	System	WER (%)
	FBK	8.4
Open		
	Vocapia (run 1)	6.4
	Vocapia (run 2)	5.4

Vocapia (run1): single-pass real-time (RT) system,
Vocapia (run2): two-pass system that includes AM adaptation and word-lattice rescoring (~5RT)
FBK: two-pass system that includes acoustic normalization (~3RT)

Constrained Transcription task

Closed	System	WER (%)
	FBK	7.2

After evaluation, FBK reported a mistake in the 2 runs so current results are better

Discussion

Difficult comparison but:

- provided data, although limited, sufficient for a performing system
- acoustic feature: combination of PLP and MLP-based probabilistic features by a bottleneck architecture seems more effective than standard MFCC (+HLDA)
- acoustic model: gender-models with ~200K Gaussians vs. gender-independent model with ~40K Gaussians
- language model: continuous space Neural Network LM vs. 4-gram LM

Conclusions

- although modular task, few participants (no students)
- already raised issue: encourage wider participation of companies

- in terms of results, mature technology
- still requires to properly design different components (AM, LM, lexicon) and strategies (2-pass decoding)

Thanks!

Any questions?