LIMSI English-French Speech Translation System

Natalia Segal **Hélène Bonneau-Maynard** *Quoc Khanh Do*
Alexandre Allauzen *Jean-Luc Gauvain* *Lori Lamel* *François Yvon*

LIMSI-CNRS and Université Paris-Sud

December 4th, 2014

Motivation

LIMSI Spoken Language Processing group

- ASR team
- SMT team
- Joint projects: Quaero, U-STAR, RAPMAT

IWSLT 2014 LIMSI participation

- A nice opportunity to continue the collaboration
- Towards a tighter integration of both processes

Main contributions / outline

- Adapting LIMSI ASR system to TED talks transcription
- Adapting MT system to ASR
  - Punctuation and number normalization
  - Adaptation to ASR transcriptions
  - Application of SOUL NN models

Speech recognizer overview

- Adaptation of the LIMSI ASR system for broadcast data
- Adaptation concerns acoustic and language models, and pronunciation dictionary.
- Audio partitioning to separate speech/nonspeech and assign speaker labels to segment clusters
- Two pass decoding with lattice generation and consensus decoding

Acoustic Models

- Acoustic features:
  - 12-dimensional PLP features (cep, Δ, $\Delta\Delta$)
  - + 3-dimensional F0 features (pitch, Δ, $\Delta\Delta$)
  - 39 dimensional probabilistic features produced by a Multi-Layer Perceptron from raw TRAP-DCT features
  - cepstral normalization on a segment-cluster basis
  - 81-dimensional feature vector (MLP+PLP+F0)
- Gender-independent, tied-state, left-to-right 3-state HMMs with Gaussian mixture observation densities
- Word position-dependent states tied using decision tree
- Speaker-adaptive (SAT) and Maximum Mutual Information (MMIE) trained

ASR Language Models

- N-gram language models obtained by interpolating TED LM with existing 78k LM from the BN system
- LM texts
  - IWSLT14 TED LM transcriptions (3.2M words)
  - Various texts (LDC, web downloads) all predating December 31, 2010
- Resulting vocabulary size: 95k words

ASR Results

- First pass decoding with modified Quaero 2011 system for English broadcast data, replacing LM and pronunciation dictionary
- Second decoding pass with same interpolated language model TED-specific acoustic models, trained only on 180 hours of transcribed TED talks predating December 31, 2010.

ASR Results

- First pass decoding with modified Quaero 2011 system for English broadcast data, replacing LM and pronunciation dictionary
- Second decoding pass with same interpolated language model TED-specific acoustic models, trained only on 180 hours of transcribed TED talks predating December 31, 2010.

dataset	WER (del., ins.)
dev2010	15.0 (4.0, 3.5)
tst2010	12.7 (3.3, 2.7)

Case-insensitive recognition results on the 2010 dev and tst data, scored using sclite

MT: N-code, *n*-gram based approach

The starting assumption [Casacuberta and Vidal, 2004, Mariño et al., 2006]

training the translation model given
a fixed segmentation and reordering.

MT: N-code, *n*-gram based approach

The starting assumption [Casacuberta and Vidal, 2004, Mariño et al., 2006]

training the translation model given
a fixed segmentation and reordering.

Break up the translation process [Crego and Mariño, 2006]

1. Source re-ordering
2. Monotonic decoding

MT: N-code, *n*-gram based approach

The starting assumption [Casacuberta and Vidal, 2004, Mariño et al., 2006]

training the translation model given
a fixed segmentation and reordering.

Break up the translation process [Crego and Mariño, 2006]

1. Source re-ordering
2. Monotonic decoding

The translation model is a *n*-gram model of tuples
(*i.e* phrase pairs):

$$P(\mathbf{s},\mathbf{t}) = \prod_{i=1}^{L} P(u_i|u_{i-1},...,u_{i-n+1})$$

See `http://ncode.limsi.fr/`

MT baseline

Pre-processing

- Cleaning (comments, speaker names, etc.)
- Tokenization using MT-specific in-house tool
- Word alignments using MGIZA++
- POS tagging using TreeTagger

Target Language Model

- Log-linear interpolation: TED LM and WMT LM

Narrowing the gap between ASR and MT

Normalization of numbers

- Spelled-out numbers in ASR output are converted to digits
- Digital numbers in MT train are converted to text and back to digits (for better consistency)

Narrowing the gap between ASR and MT

Normalization of numbers

- Spelled-out numbers in ASR output are converted to digits
- Digital numbers in MT train are converted to text and back to digits (for better consistency)

Results

training corpora	normalization	BLEU (tst2010) manual	BLEU (tst2010) auto
TED	no norm	33.2	20.5
	norm	33.0	**21.0**

Manual punctuation in manual transcriptions, no punctuation in ASR transcriptions

- 17% WER and no punctuation for ASR results in -13 BLEU points
- Small drop of performance after normalization for manual transcriptions

Punctuation

Punctuation has to be produced in the translations [Peitz et al., 2011].

- Implicit (retrain bilingual MT system): no punctuation in MT train source, punctuation in MT train target
- Explicit (bilingual MT system unchanged): automatic punctuation of ASR output
  - Via monolingual monotonic MT systems (TED, News-Commentary and Europarl)
    - ALL (all the punctuation symbols)
    - 6-MAIN (only simple unpaired punctuation symbols)

Punctuation

Punctuation has to be produced in the translations [Peitz et al., 2011].

- Implicit (retrain bilingual MT system): no punctuation in MT train source, punctuation in MT train target
- Explicit (bilingual MT system unchanged): automatic punctuation of ASR output
  - Via monolingual monotonic MT systems (TED, News-Commentary and Europarl)
    - ALL (all the punctuation symbols)
    - 6-MAIN (only simple unpaired punctuation symbols)

training corpora	punct test	BLEU (tst2010 auto)
TED (implicit punct)	none	**24.4**
TED (manual punct)	none	21.0
	ALL	24.0
	6-MAIN	**24.4**

- On manual transcription: no punctuation in source results in -3 BLEU points

Adaptation of the MT system to ASR transcriptions

Consider automatic transcription as a source of variability

- include the automatic transcriptions of the source part of the parallel corpus in the training process
- for both SMT training and development corpus
- TED auto corpus: transcriptions by ASR baseline (unpunctuated)

Adaptation of the MT system to ASR transcriptions

Consider automatic transcription as a source of variability

- include the automatic transcriptions of the source part of the parallel corpus in the training process
- for both SMT training and development corpus
- TED auto corpus: transcriptions by ASR baseline (unpunctuated)
- Different configurations:

training corpora	BLEU (tst2010, no punct)	
	manual	auto
TED man only	29.9	24.4
TED auto only	28.8	24.2
TED man+auto (2 tables)	29.5	24.6
TED man+auto (1 table)	29.3	**24.8**

Adaptation of MT systems to ASR transcriptions

Examples of MT improvement

- Repeated words

manual source	*and it just disturbed me so much .*
ASR source	*and* **it it** *just to scare me so much .*
trans. without adaptation	*et ça , ça ne m' effraie beaucoup .*
trans. with adaptation	*et il m' effraie beaucoup .*

Adaptation of MT systems to ASR transcriptions

Examples of MT improvement

- Replacement of phonetically close words

manual source	*those who were still around in school*
ASR source	*those who were still around* **and** *school*
trans. without adaptation	*ceux qui étaient encore* **et** *l' école*
trans. with adaptation	*ceux qui étaient encore* **dans** *l' école*

manual source	*what does that have to do with the placebo effect .*
ASR source	*was* **that** *have to do with the placebo effect .*
trans. without adaptation	*que* **nous devons faire** *avec l' effet placebo .*
trans. with adaptation	*qu'***est -ce que cela a à voir** *avec l' effet placebo .*

Final MT system configuration and ASR quality impact

- Corpora: TED man+auto (concatenated), Gigaword (filtered)
- Test: ASR baseline (WER=17%) vs ASR adapted (WER=12.7%)

training corpora	punctuation	BLEU (test2010 auto)	
		ASR (WER=17%)	ASR (WER=12.7%)
TED man+auto (1 table)	no punct	24.8	-
+ GIGA	no punct	25.0	-
	punct main	25.5	**27.7**

Continuous Space Translation Models

Continuous space *n*-gram models

n-gram distributions can be estimated using neural network models [Bengio et al., 2003, Schwenk, 2007], for translation models:

- SOUL models can efficiently deal with large vocabularies [Le et al., 2011]
- The bilingual extension the SOUL model is used [Le et al., 2012]

Word factored translation models

- Translation distributions (*n*-gram) can be decomposed at the word level in different ways
- By considering the source and target parts of tuples

$\Rightarrow$ 4 bilingual *n*-gram distributions of words

For more details see the presentation of Quoc Khanh Do

Experimental results

Systems	**dev**	**test**
Before SOUL	23.7	27.7
Adding all 4 SOUL TMs		
+ TMs TED manual	24.1	27.9
+ TMs TED auto	24.2	28.0
+ TMs mixing TED-GIGA	24.4	27.9
Adding all 4 SOUL TMs and SOUL target LM		
+ TMs TED manual + LM	24.3	27.9
+ TMs TED auto + LM	24.3	27.6
+ TMs mixing TED-GIGA + LM	**24.4**	**28.3**

Conclusion and future work

Primary submission

- ASR adaptation: -4.3%WER $\rightarrow$ +2 BLEU for MT
- MT adaptation to ASR
  - Punctuation and number normalization (+4 BLEU)
  - Adaptation by training MT models on ASR transcriptions (+0.5 BLEU)
  - Rescoring with SOUL NN models (0.5 BLEU), to be analyzed

Conclusion and future work

Primary submission

- ASR adaptation: -4.3%WER $\rightarrow$ +2 BLEU for MT
- MT adaptation to ASR
  - Punctuation and number normalization (+4 BLEU)
  - Adaptation by training MT models on ASR transcriptions (+0.5 BLEU)
  - Rescoring with SOUL NN models (0.5 BLEU), to be analyzed

Next step

- Automatic segmentation
- Adapt the ASR system to translation


Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003).
A neural probabilistic language model.
Journal of Machine Learning Research, 3:1137–1155.


Casacuberta, F. and Vidal, E. (2004).
Machine translation with inferred stochastic finite-state transducers.
Computational Linguistics, 30(3):205–225.


Crego, J. M. and Mariño, J. B. (2006).
Improving statistical MT by coupling reordering and decoding.
Machine Translation, 20(3):199–215.


Le, H.-S., Allauzen, A., and Yvon, F. (2012).
Continuous space translation models with neural networks.
In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 39–48, Montréal, Canada.


Le, H.-S., Oparin, I., Allauzen, A., Gauvain, J.-L., and Yvon, F. (2011).
Structured output layer neural network language model.
In *Proceedings of ICASSP*, pages 5524–5527.

Mariño, J. B., Banchs, R. E., Crego, J. M., de Gispert, A., Lambert, P., Fonollosa, J. A., and Costa-Jussà, M. R. (2006).
N-gram-based machine translation.
Computational Linguistics, 32(4):527–549.

Peitz, S., Freitag, M., Mauser, A., and Ney, H. (2011).
Modeling punctuation prediction as machine translation.
In *International Workshop on Spoken Language Translation (IWSLT 2011)*, pages 238–245.

Schwenk, H. (2007).
Continuous space language models.
Computer Speech and Language, 21(3):492–518.