LIUM @ IWSLT'15 evaluation campaign

Mercedes García-Martínez, ***Loïc Barrault***, *Anthony Rousseau*,
Paul Deléglise, *Yannick Estève*

German ASR task and English to French SLT task

Overview

- Data
- Description of the German ASR system
- Description of the English-French SLT system
  - Phrase-based SMT system
  - Neural MT system
- Results
- Discussion

German ASR task

First participation of LIUM for that language

Data selection for acoustic models

Sources of speech:

- Euronews ASR 2013 Dataset as primary source
- in-house sources
- extracted TEDx Talks

Corpus	Duration	Segments	Words
Euronews	62.5h	20 187	506 019
In-house	23.9h	6 196	232 716
TEDx	38.0h	42 633	312 142
Total	124.4	69 016	1 050 877

Characteristics of the acoustic data used in the LIUM ASR system acoustic models.

Data selection for language models

Sources:

- all of publicly available data from WMT15
- collection of TEDx Talks closed-captions

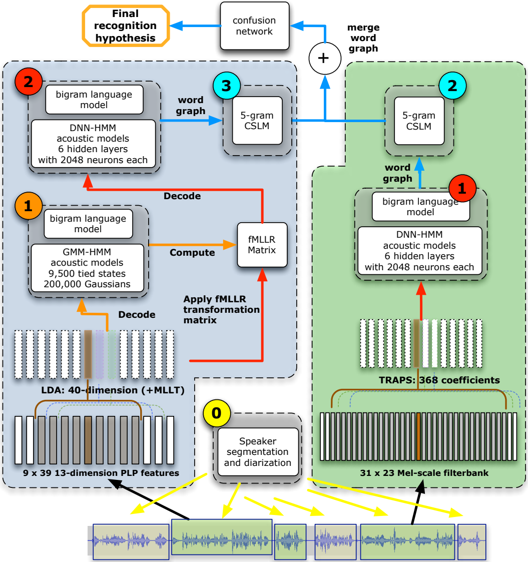
Data selection:

- data selection tool XenC [Rousseau, 2013]
- cross-entropy difference [Moore & Lewis, 2010, Axelrod, 2011]

Data selection for language models

Corpus	Original # of words	Selected # of words	% of Orig.
IWSLT14	2.85M	2.85M	100.00
Common Crawl	48.04M	4.24M	8.82
Europarl	47.40M	3.20M	6.74
News Crawl	1 409.62M	130.60M	9.26
News-Comm.	5.06M	0.62M	12.25
Total (w/o IWSLT14)	1 510.12M	138.66M	9.18

Architecture



Architecture of the LIUM ASR systems

- Two separate systems
- Based on Kaldi open-source speech recognition toolkit

Two-pass systems:

- first pass
  - decode with 2-gram LM and DNNs
  - generate word-lattice
- second pass
  - word-lattice rescoring with 3-gram, 4-gram back-off LMs and 5-gram CSLM
  - apply an accelerated version of the consensus algorithm to the confusion networks from rescored graphs

Acoustic modeling

GMM-HMM acoustic models:

- 13 PLP + 1st & 2nd derivatives : 39 features per frame
- left & right 4-frames context (9 frames in total)
- $39 * 9 = 351$ features projected to 40 dimensions by LDA and MLLT
- speaker adaptive training with fMLLR
- models trained on the full 124 hours, with 9 500 tied triphones and 325 000 states

Acoustic modeling

System 1 DNN (TRAP system):

- Input is 368 TRAP coefficients
  - computed on a sliding window of 31 frames
- Frames are from the output of 23 Mel-scale filterbanks
- 6 hidden layers with 2048 units, softmax layer is 4 627 outputs

System 2 DNN (fMLLR system):

- Input is 440 LDA parameters on a sliding window of 11 frames
- discriminative criterion is sMBR
- 6 hidden layers with 2048 units, softmax layer is 7 827 outputs

Each DNN is trained using GPUs and the CUDA toolkit.

Language modeling

- Rely on two toolkits:
  - SRILM language modeling toolkit
  - CSLM toolkit
- Vocabulary is 131 425 entries
- Separate sets of LMs are trained for each system
- 2G, 3G and 4G models:
  - trained individually from each source
  - modified KN discounting, no cut-offs
  - then linearly interpolated
- 5G CSLM, also with modified KN and no cut-offs

Word-lattice merging

- Same audio segmentation for both systems, using LIUMSpkDiarization toolkit
- Final output by merging word-lattices from both systems
- Standard word-lattices with word, temporal information, acoustic & linguistic scores

Process:

- Compute *a posteriori* probabilities for each lattice
- Weight the probs by $1/n$, where n is the number of lattices
- Replace scores with these probabilities for each edge
- Merge start and end nodes from lattices into a single lattice
- Process the merged lattice with an optimized version of the consensus network confusion algorithm

Results

Results on development corpus (% WER):

- fMLLR system: 17.6
- TRAP system: 16.8

  $\rightarrow$ Fusion: 15.1

Official results for the LIUM German ASR system (% WER):

- Before adjudication: 17.8
- After adjudication: 17.6

English French SLT task

Original plan

Plan: combining Phrase-Based and Neural MT systems

- System complementarity?
- engine, model, etc.

Final submissions

- Primary system:
  - 1000-best list generated by Phrase based MT system
  - rescored by CSLM and NMT
- Contrastive systems: baseline and individual systems rescored (for the sake of comparison)

Data

Preprocessing

- *ASR-ization* of the English portion of the available bitexts
  - rewrite numbers in letters, lowercase and remove punctuation
- No change on the French (target) side

Dev and test corpora

- *liumdev15*: dev2010 + tst2010 + tst2013
- *liumtst15*(internal) : tst2011 + tst2012

Data

Data selection

- Based on Moore & Lewis, ACL'10 and Axelrod, EMNLP'11

  $\rightarrow$ select a small subset containing most relevant data based on cross-entropy difference

  $\rightarrow$ speed-up training considerably (translation and language model)

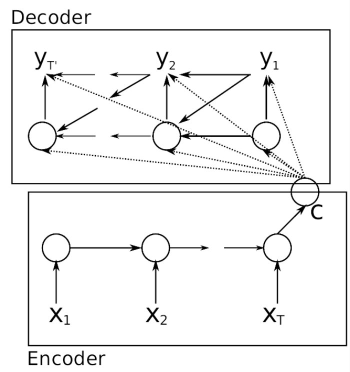
  $\Rightarrow$ keep around 33% of the data

Neural MT system

Model Details

- Given source sequence $\mathbf{X} = (x_1, \ldots, x_T)$ and target sequence $\mathbf{Y} = (y_1, \ldots, y_{T'})$,
- Model $p(\mathbf{Y}|\mathbf{X})$ directly with two RNN's
- **c** is a representation of source sentence (Cho et al., 2014)
- Train to maximize *log* $p(\mathbf{Y}|\mathbf{X})$ (end-to-end)

Encoder-Decoder Architecture



Baseline Neural MT system with Alignment

Model Details

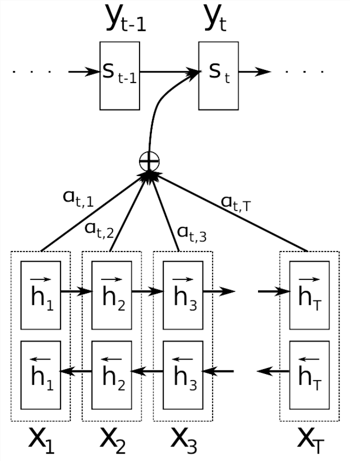
Bi-directional RNN for encoder,

- Get annotation vector h_i, where $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$

For each time step t in decoder,

- Compute a relevance score $a_{t,i}$ for each annotation h_i
- Use the weighted sum of the annotations as a context c_t
- Train end-to-end again with SGD (Bahdanau et al., 2015)

Alignment Module



Neural network machine translation system results

Corpus	*Beam size*		
	10	100	1000
liumtst15	36.79	36.1	35.24
liumdev15	31.62	30.95	30.12

- The larger the beam size, the lower the results

  $\rightarrow$ problematic behaviour
- Impact of beam size:
  - Partial hypothesis with low score is not early pruned anymore
  - In the end: gets high score, BUT this is actually a worse translation (regarding BLEU)

  $\rightarrow$ sharp NN output distributions

  $\rightarrow$ BLEU differs from internal score (correlation?)
- Deeper analysis needed

Phrase-based SMT system

Architecture

- PBSMT based on Moses
- standard 14 feature functions
- + Operation Sequence Model (5 feats.)
- 1000-best list rescoring with a large context CSLM
  - $\Rightarrow \sim 1$ BLEU point improvement

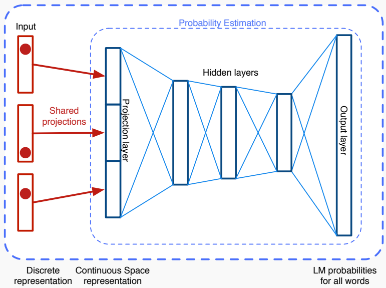
Continuous Space Language Model

Architecture

- Feed-forward NN
- Output : softmax
- Trained with SGD to minimize cross-entropy
- PPL reduction ∼38%

different configurations

CSLM



Name	Order	Proj. size	#hidd. x size	PPL
BO LM	4	-	-	67.85
CSLM11	11	512	3 x 1024	41.98
CSLM19	19	320	3 x 1024	41.38

Results

Name	*liumdev15*	*liumtst15*	*test2015* Case	
	%BLEU	%BLEU	%BLEU	%TER
NMT	31.62	36.79	14.88	84.69
Moses	31.81	37.35	16.95	80.61
Moses+CSLM11	32.81	38.36	17.54	80.04
Moses+CSLM19	32.70	38.28	17.56	80.07
Moses+CSLM11+NMT	33.81	39.61	18.51	79.06
Moses+CSLM19+NMT	33.82	39.65	18.53	78.96

- Same improvement with two different CSLMs
- around +1 BLEU point by rescoring with NMT
- Absolute scores lower than previous years

  $\rightarrow$ impact of text segmentation : - 5 to 6 BLEU point (compared to last year)

Conclusion

What did not worked (as expected)

- NMT system still provides lower results compared to PBSMT
- Rescoring NMT with CSLM

  → Tentative explanation

  - Search space not as furnished as for PBSMT
    → cf. problem with beam-size

What worked

- Rescoring PBSMT with CSLM → +1 BLEU (as expected)
- Rescoring PBSMT with NMT → +1 BLEU on top of CSLM
  → not expected

→ NMT is good for rescoring while getting low scores alone

→ we can do better with it! (needs a better search-space)

References I

- Anthony Rousseau, XenC: An Open-Source Tool for Data Selection in Natural Language Processing, The Prague Bulletin of Mathematical Linguistics, p73–82, vol. 100, 2013.
- Axelrod, A. and He, X. and Gao, J. Domain Adaptation via Pseudo In-Domain Data Selection. EMNLP, 2011.
- Moore, R. C. and Lewis, W., Intelligent selection of language model training data, ACL, 2010
- D. Bahdanau, K Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate, ICLR 2015.
- K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, EMNLP, 2014.

References II

- C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, On using monolingual corpora in neural machine translation, arXiv:1503.03535