Auxiliary features for Continuous Space Language Model

Walid Aransa

LIUM, University of Maine

Talk at LIMSI, Paris, 21 June 2016

Presentation outline

1. Auxiliary features for continuous space language model
2. Experiments results and analysis
3. WMT'16 Multimodal Task 1 and auxiliary features
4. Conclusion and prospects

Presentation outline

1. **Auxiliary features for continuous space language model**
2. WMT'16 Multimodal Task 1 and auxiliary features
3. Experiments results and analysis
4. Conclusion and prospects

CSLM

What is CSLM?

- A multi-layer neural network model which learns the words projection and the probabilities jointly [Bengio et al., 2003].

Mainly used to re-score SMT n-best list [Schwenk et al., 2006].

CSLM advantages

- Better estimation of the probability of non-observed n-gram.
- Directly can estimate the probability of long context.

CSLM

What is CSLM?

- A multi-layer neural network model which learns the words projection and the probabilities jointly [Bengio et al., 2003].

Mainly used to re-score SMT n-best list [Schwenk et al., 2006].

CSLM advantages

- Better estimation of the probability of non-observed n-gram.
- Directly can estimate the probability of long context.

CSLM

What is CSLM?

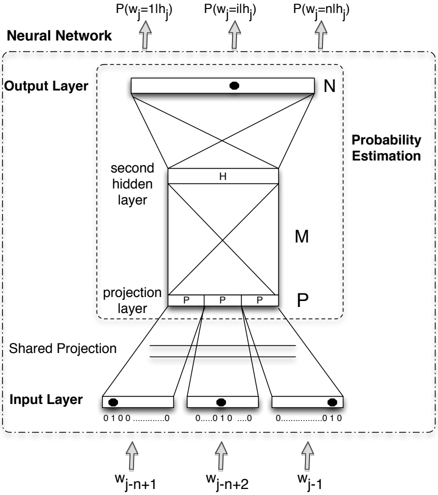
- A multi-layer neural network model which learns the words projection and the probabilities jointly [Bengio et al., 2003].

Mainly used to re-score SMT n-best list [Schwenk et al., 2006].

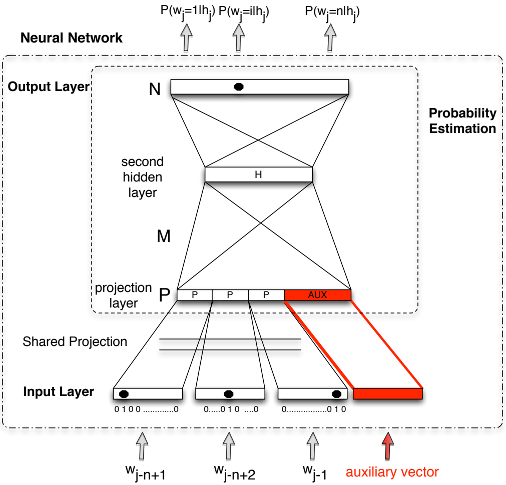
CSLM advantages

- Better estimation of the probability of non-observed n-gram.
- Directly can estimate the probability of long context.

The architecture of CSLM



The modified architecture of the CSLM



Some related works

- Interpolates two LMs, cache LM trained on the last N words: Kuhn and De Mori [1990]
- Integrating semantic knowledge in 2nd LM using LSA and clustering techniques: Bellegarda [2000] and Coccaro and Jurafsky [1998]
- LM that takes advantage of the topic in a sentence or article: Iyer and Ostendorf [1999] and Khudanpur and Wu [2000]
- Topic-conditioned RNNLM: Mikolov and Zweig [2012]

Auxiliary features types

1. Line or corpus characteristics

The first type provides additional information on the current line except the context representation.

$\Rightarrow$ Motivated by MT quality estimation literature.

- Line length.
- Text genre to train genre-conditioned CSLM.

2. Context features

This auxiliary type aims at providing a larger and different context.

Auxiliary features types

1. Line or corpus characteristics

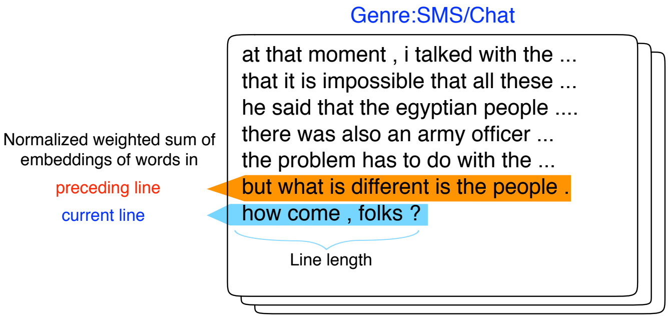
The first type provides additional information on the current line except the context representation.
$\Rightarrow$ Motivated by MT quality estimation literature.

- Line length.
- Text genre to train genre-conditioned CSLM.

2. Context features

This auxiliary type aims at providing a larger and different context.

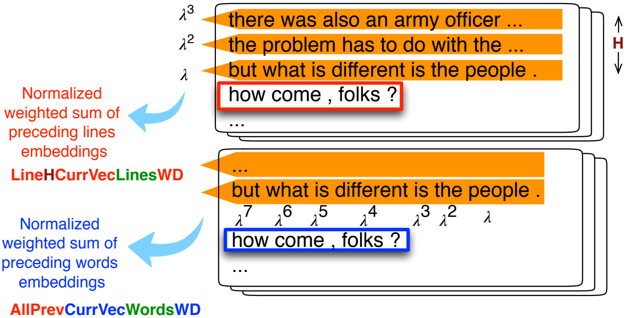
Auxiliary features



Current line embeddings:

$$\hat{\alpha}_l = \frac{\sum_{w \in l} e_w}{|\sum_{w \in l} e_w|} \quad (1)$$

Auxiliary features



Example: **LineHCurrVecLinesWD**:

$$\hat{\eta}_{l,h} = \frac{\sum_{i=l-h}^{l} \hat{\alpha}_i \lambda^{l-i}}{\left|\sum_{i=l-h}^{l} \hat{\alpha}_i \lambda^{l-i}\right|} \tag{2}$$

Presentation outline

1. Auxiliary features for continuous space language model
2. **Experiments results and analysis**
3. WMT'16 Multimodal Task 1 and auxiliary features
4. Conclusion and prospects

Evaluation on Penn Treebank (PPL)

Evaluated using one auxiliary feature: *PrevLineVec*

System	Auxiliary layer	1st layer lr scale	DevSet PPL	TestSet PPL
Baseline1	N/A	1	133.19	127.66
Baseline2	N/A	2	130.48	125.28
CSLM1	Copy	1	128.26	123.45
CSLM2		2	124.80	120.32
CSLM3	Seq. of two tanh	1	127.15	121.93
CSLM4		2	124.22	118.57
CSLM5		3	**122.98**	**118.08**
Mikolov ICLR'15 LSTM 100 hidden	-	-	**120**	**115**

- Using better meta-configuration or topology proved to increase the accumulated gains up to 10.21 (8.3%) PPL on dev and 9.58 (7.5%) PPL on test.

Evaluation on Penn Treebank (PPL)

Evaluated using one auxiliary feature: *PrevLineVec*

System	Auxiliary layer	1st layer lr scale	DevSet PPL	TestSet PPL
Baseline1	N/A	1	133.19	127.66
Baseline2	N/A	2	130.48	125.28
CSLM1	Copy	1	128.26	123.45
CSLM2		2	124.80	120.32
CSLM3	Seq. of two tanh	1	127.15	121.93
CSLM4		2	124.22	118.57
CSLM5		3	**122.98**	**118.08**
Mikolov ICLR'15 LSTM 100 hidden	-	-	**120**	**115**

- Our results are similar to the best results obtained by Mikolov ICLR'15.

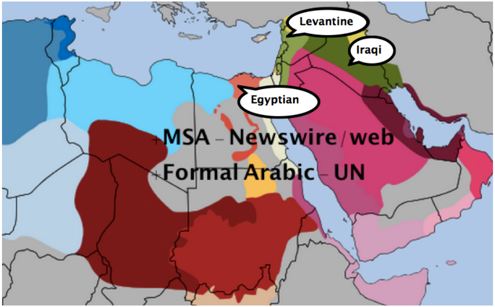
DARPA Broad Operational Language Translation (BOLT)

Objective

Enable communication with non-English-speakers and identify important information in foreign-language sources.

- Information retrieval, automatic speech recognition for foreign-language and two-way speech-to-speech translation.
- Several genres: **SMS/chat** and **informal conversation.**
- Languages: **Egyptian dialect** and **Chinese**
- Three phases (2012-2014) with three official NIST evaluations.
- Two teams: Delphi and Astral.
- LIUM & other universities were part of Delphi, leaded by IBM.

BOLT data resources and genres



Bilingual corpora (Arabic/English)	Monolingual corpora
Formal MSA (UN), GALE (newswire/web) Egyptian dialect (Conversational telephone speech (**CTS**), Discussion forum (**DF**) and **SMS/chat**), Iraqi & Levantine dialects	English (gigaword, DF) Egyptian dialect (DF)

Baseline System

- Standard phrase-based SMT with Moses toolkit, alignment using GIZA++.
- Standard 14 features optimized using MERT.
- 4-gram LM and Kneser-Ney smoothing using SRILM toolkit.

Evaluation metric:

- NIST official evaluation uses human-targeted TER (HTER).
- For system development, (TER - BLEU)/2 is used => TB2

STOA methods integrated into LIUM's BOLT system

- Arabic segmentation (using IBM, MADA [Habash and Rambow, 2005] and MADA Arz [Habash et al., 2013])
- Domain adaptation
    - Monolingual data selection [Moore and Lewis, 2010].
    - Bilingual data selection [Axelrod et al., 2011].
    - Translation model domain adaptation [Sennrich, 2012].
    - Multi-domain translation model [Sennrich et al., 2013]
    - Lightly supervised training [Schwenk, 2008b].
- Operation sequence model [Durrani et al., 2011].
- CSLM rescoring [Schwenk et al 2014].

SMT experiments

SMT System

BOLT Phase 3 SMS/Chat system presented in BOLT section.

CSLM model training corpora:

type	data set	Arabic tokens	English tokens	genre
train	gale	4.28m	5.01	MSA
	bolt	1.70m	2.05m	DF
	smschat	648k	845k	SMS/CHAT
	Total	6.63m	7.9m	-
tune	smschat tune	19.7k	25.6k	SMS/CHAT
test	smschat dev	19.4k	24.6k	SMS/CHAT

Results of re-scoring n-best list (BLEU)

- Summary of best scores per each auxiliary feature (explained in next slides)

System	Tune	Test
Baseline	27.35	25.72
LineLen	28.65	26.14
GenreVec	**28.90**	**26.32**
CurrLineVec	28.29	26.09
PrevLineVec	28.67	**26.33**
LineHCurrVecLinesWD, λ=0.95, h=50	**28.92**	26.26
AllPrevCurrVecWordsWD, λ=0.75	28.52	25.86
AllPrevVecWordsWD, λ=0.95	**28.77**	**26.82**
AllPrevVecLinesWD, λ=0.98	28.63	**26.52**

Results of re-scoring n-best list (BLEU) - analysis

System	**Auxiliary input**		**Tune**	**Test**
	dim.	**layer**		
Baseline	-	-	27.35	25.72
LineLen	1/200	Proj. 200x320	28.65	26.14
GenreVec	5/-	Copy 5x5	**28.90**	**26.32**

- Observed a good improvement of **LineLen**, but **GenreVec** gives relatively better gain on both tune and test.

Results of re-scoring n-best list (BLEU) - analysis 2

Q. How can we explain the score gain for **GenreVec**?.

The improvement factors:

1. Better training of auxiliary-conditioned CSLM.
2. Using discriminative auxiliary feature.

GenreVec gain is because of the first one.

Observation

Good non-discriminative auxiliary features can be useful for CSLM re-scoring.

Results of re-scoring n-best list (BLEU) - analysis 3

System	Auxiliary input		Tune	Test
	dim.	layer		
Baseline	-	-	27.35	25.72
CurrLineVec	320/-	Seq. of two tanh 320x320	28.29	26.09
PrevLineVec			28.67	**26.33**

- **PrevLineVec** has better context information compared to **CurrLineVec**.
- Current line auxiliary features generally have lower BLEU.

Results of re-scoring n-best list (BLEU) - analysis 4

PrevLineVec has good BLEU scores 28.67, 26.33 on tune and test respectively.

Assumption

two or more preceding lines may be more useful (possibly weighted).

Verification is needed for this assumption using *AllPrevVecLinesWD* CSLM, which uses auxiliary feature that does not contain the **current** line.

Results of re-scoring n-best list (BLEU) - analysis 5

System	λ	Tune	Test
SMT baselessine	-	27.35	25.72
CurrLineVec	-	28.29	26.09
PrevLineVec	-	28.67	26.33
AllPrevVecLinesWD	0.85	28.06	25.52
	0.95	28.59	26.42
	0.98	**28.63**	**26.52**
AllPrevVecWordsWD	0.75	28.37	26.36
	0.85	28.74	26.49
	0.95	**28.77**	**26.82**

- This confirms that more weighted preceding lines are more useful and provide better context information.

Results of re-scoring n-best list (BLEU) - analysis 6

System	H	Tune	Test
SMT baseline	-	27.35	25.72
CurrLineVec	-	28.29	26.09
PrevLineVec	-	28.67	**26.33**
LineHCurrVecLinesWD	10	28.70	26.21
	30	28.28	**26.26**
	50	**28.92**	**26.26**

- But even with H=50, the scores are not better than just **one** preceding line **PrevLineVec** on test.

Conclusion

Using current line in the auxiliary feature gives inconsistent results.

Results of re-scoring n-best list (BLEU) - analysis 7

System	Tune	Test
Baseline	27.35	25.72
LineLen	28.65	26.14
GenreVec	**28.90**	**26.32**
CurrLineVec	28.29	26.09
PrevLineVec	28.67	**26.33**
LineHCurrVecLinesWD, λ=0.95, h=50	**28.92**	26.26
AllPrevVecWordsWD, λ=0.95	**28.77**	**26.82**
AllPrevVecLinesWD, λ=0.98	28.63	**26.52**

Conclusion

- Improvement using weighted sum of preceding words embeddings: 1.42 BLEU on tune (5%) and 1.1 on test (4%).

Reference

Walid Aransa, Holger Schwenk, and Loic Barrault. 2015. Improving continuous space language models using auxiliary features. In Proceedings of the 12th International Workshop on Spoken Language Translation, pages 151-158, Da Nang, Vietnam, December.

Presentation outline

1. Auxiliary features for continuous space language model
2. Experiments results and analysis
3. **WMT'16 Multimodal Task 1 and auxiliary features**
4. Conclusion and prospects

WMT 2016 Multimodal Tasks

Task1: Multimodal Machine Translation

- This task consists in translating an English sentence that describes an image into German, given the English sentence itself and the image that it describes.

Task2: Multimodal Image Caption Generation

- The objective of Task 2 is to produce German descriptions of images given the image itself and one or more English descriptions as input.

WMT 2016 Multimodal Tasks

Task1: Multimodal Machine Translation

- This task consists in translating an English sentence that describes an image into German, given the English sentence itself and the image that it describes.

Task2: Multimodal Image Caption Generation

- The objective of Task 2 is to produce German descriptions of images given the image itself and one or more English descriptions as input.

Auxiliary Features Types

- **VGG19-FC7 image features:** The image features provided by the organizers which are extracted from the FC7 layer (relu7) of the VGG-19 network [Simonyan and Zisserman, 2014]. This allows us to train a multimodal CSLM that uses additional context learned from the image features.
- **Source side sentence representation vectors:** We used the method described in [Le and Mikolov, 2014] to compute continuous space representation vector for each source (i.e. English) sentence. The idea behind this is to condition our target language model on the source side as additional context.

Training Data for WMT'16 Multimodal Task 1

Side	Vocabulary	Words
English	10211	377K
German	15820	369K

Results

System Description	Validation Set METEOR (norm)	Validation Set BLEU	Test Set METEOR (norm)	Test Set BLEU
Phrase-based Baseline (BL)	53.71 (58.43)	35.61	52.83 (57.37)	33.45
BL+IMG	53.57(58.31)	35.47	52.72(57.29)	33.65
BL+EN2V	53.63(58.34)	35.35	52.95(57.49)	33.68
BL+NMT	54.02(58.73)	36.01	52.83 (57.35)	33.70
BL+RNN	53.78 (58.42)	35.75	53.14 (57.74)	34.27
BL+3Features	54.29 (58.99)	36.52	53.19 (57.76)	34.31
BL+4Features	54.40 (59.08)	36.63	53.18 (57.76)	34.28

Table: BLEU and METEOR scores on detokenized outputs of baseline and submitted Task 1 systems. The METEOR scores in parenthesis are computed with `-norm` parameter.

Presentation outline

1. Auxiliary features for continuous space language model
2. Experiments results and analysis
3. WMT'16 Multimodal Task 1 and auxiliary features
4. **Conclusion and prospects**

Conclusions

- Introduced a novel method to improve the continuous space language model using auxiliary features.
- Used different auxiliary features and presented results analysis for each feature including source-side sentence representation vectors and image features.
- Using the current line embeddings in the calculation of the auxiliary feature vector gives inconsistent results.
- Weighted sum of the individual word embeddings is more stable and outperforms the line level weighted sum of embeddings.

Conclusions

- Introduced a novel method to improve the continuous space language model using auxiliary features.
- Used different auxiliary features and presented results analysis for each feature including source-side sentence representation vectors and image features.
- Using the current line embeddings in the calculation of the auxiliary feature vector gives inconsistent results.
- Weighted sum of the individual word embeddings is more stable and outperforms the line level weighted sum of embeddings.

Conclusions

- Introduced a novel method to improve the continuous space language model using auxiliary features.
- Used different auxiliary features and presented results analysis for each feature including source-side sentence representation vectors and image features.
- Using the current line embeddings in the calculation of the auxiliary feature vector gives inconsistent results.
- Weighted sum of the individual word embeddings is more stable and outperforms the line level weighted sum of embeddings.

Conclusions

- Introduced a novel method to improve the continuous space language model using auxiliary features.
- Used different auxiliary features and presented results analysis for each feature including source-side sentence representation vectors and image features.
- Using the current line embeddings in the calculation of the auxiliary feature vector gives inconsistent results.
- Weighted sum of the individual word embeddings is more stable and outperforms the line level weighted sum of embeddings.

Prospects

- Study the use of additional auxiliary features extracted from the source language side of the bitext.
- Use topics instead of genres as auxiliary feature and assign the topic ID dynamically by using automatic clustering algorithm.
- Use auxiliary features with recurrent neural network language models (RNNLMs) architecture.
- Study the impact of using auxiliary features in neural network machine translation (NMT).

Thanks

Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.