Machine Translation

09: Monolingual Data

Rico Sennrich

University of Edinburgh

Refresher

why monolingual data?

language models are an important component in statistical machine translation

- monolingual data is far more abundant than parallel data
- phrase-based SMT models suffer from independence assumption; LMs can mitigate this
- monolingual data may better match target domain

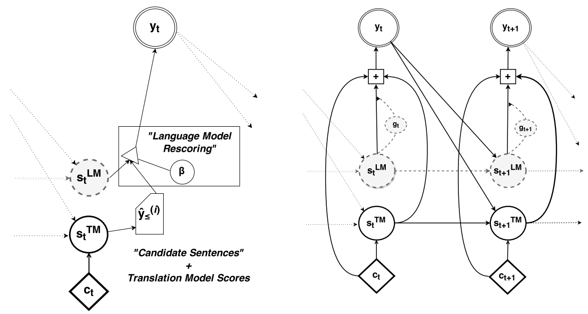
MT – 2018 – 09

Language Models in NMT

[Gülçehre et al., 2015]

shallow fusion: rescore beam with language model (≈ ensembling)

deep fusion: extra, LM-specific hidden layer



(a) Shallow Fusion (Sec. 4.1) (b) Deep Fusion (Sec. 4.2)

	De-En Dev	De-En Test	Cs-En Dev	Cs-En Test
NMT Baseline	25.51	23.61	21.47	21.89
Shallow Fusion	25.53	23.69	21.95	22.18
Deep Fusion	**25.88**	**24.00**	**22.49**	**22.36**

MT – 2018 – 09

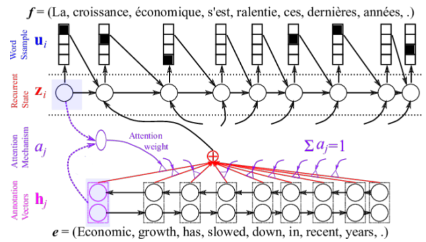
Monolingual Data in NMT

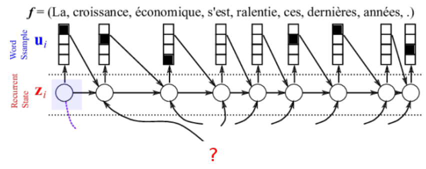
NMT is a conditional language model

$$p(u_i) = f(z_i, u_{i-1}, c_i)$$

Problem

for monolingual training instances, source context c_i is missing

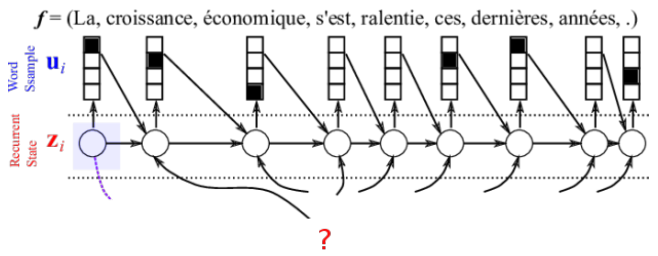




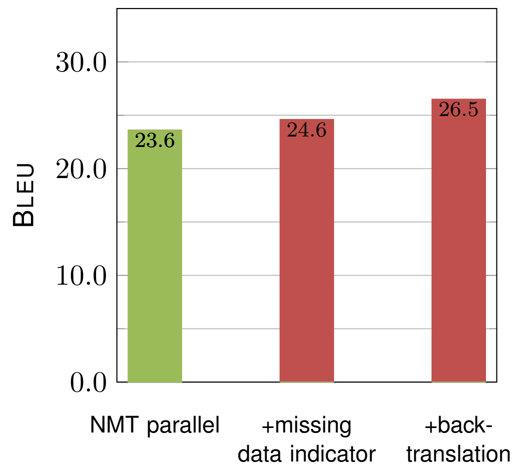
Monolingual Training Instances

solutions: missing data imputation for c_i

- missing data indicator: $\vec{0}$
  $\rightarrow$ works, but danger of catastrophic forgetting
- impute c_i with neural network
  $\rightarrow$ we do this indirectly by back-translating the target sentence



Evaluation: English→German



Back-Translation: Comparison to Phrase-based SMT

back-translated parallel data

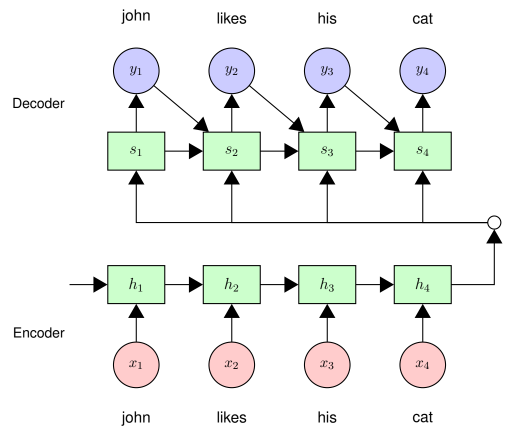
- back-translation has been proposed for phrase-based SMT [Schwenk, 2008, Bertoldi and Federico, 2009, Lambert et al., 2011]
- PBSMT already has LM
  → main rationale: phrase-table domain adaptation
- rationale in NMT: train end-to-end model on monolingual data

system	BLEU	
	WMT (in-domain)	IWSLT (out-of-domain)
PBSMT gain	+0.7	+0.1
NMT gain	+2.9	+1.2

Table: Gains on English→German from adding back-translated News Crawl data.

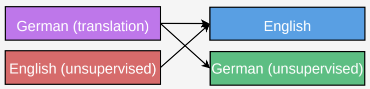
Autoencoders

general principle: train network that encodes input, and learns to reconstruct input from encoded representation
→ unsupervised representation learning



Autoencoders in Neural Machine Translation

- autoencoders are used via multi-task learning:
  shared models, multiple task-specific objectives



[Luong et al., 2016]

- does idea still work if we use attention mechanism?
  (far less of a representation bottleneck)
- apparently, yes (for low-resource language pairs):
  [Currey et al., 2017]
- analysis: BPE-based system gets better at copying unknown names:

source	Les **Dissonances** a aparut pe scena muzicala în 2004 ...
reference	Les **Dissonances** appeared on the music scene in 2004 ...
baseline	Les **Dissonville** appeared on the music scene in 2004 ...
+ copied	Les **Dissonances** appeared on the music scene in 2004 ...

Dual Learning [He et al., 2016]

dual-learning game

- closed loop of two translation systems
- translate sentence from language A into language B and back
- loss functions:
  - is sentence in language B natural?
    → loss is negative log-probability under (static) LM
  - is second translation similar to original?
    → loss is standard cross-entropy, with original as reference
- use reinforcement learning to update weights
- we can also start with sentence in language B

Parameter Pre-Training

[Ramachandran et al., 2017]

- core idea: pre-train encoder and decoder on language modelling task
- models are fine-tuned with translation objective, along with continued use of LM objective (with shared parameters)

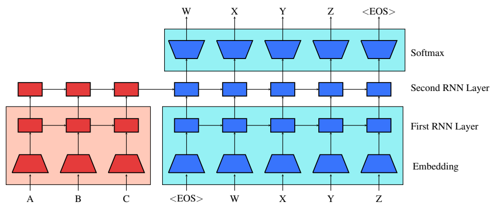


Figure 1: Pretrained sequence to sequence model. The red parameters are the encoder and the blue parameters are the decoder. All parameters in a shaded box are pretrained, either from the source side (light red) or target side (light blue) language model. Otherwise, they are randomly initialized.

MT – 2018 – 09

1. Language Models in NMT
2. Training End-to-End NMT Model with Monolingual Data
3. "Unsupervised" MT from Monolingual Data

Bilingual Lexicon Induction

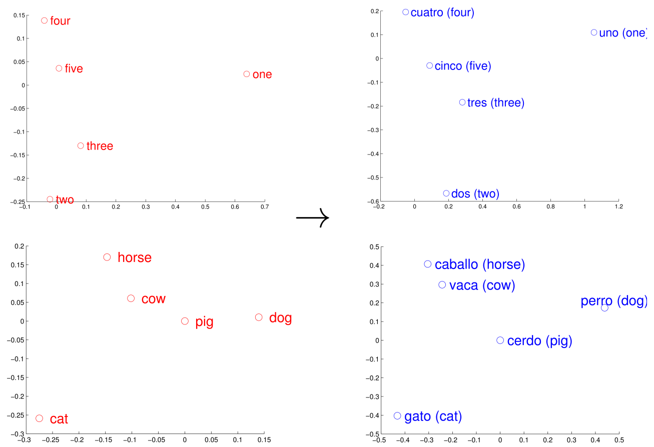
learn lexical correspondences from monolingual data

correspondences are based on various types of similarity:

- contextual similarity
- temporal similarity
- orthographic similarity
- frequency similarity

today we look at distributional word representations (contextual similarity)

Embedding Space Similarities Across Languages



[Mikolov et al., 2013]

Learning to Map Between Vector Spaces

supervised mapping [Mikolov et al., 2013]

- we can learn linear transformation between embedding spaces with small dictionary.
- given linear transformation matrix W, and two vector representations x_i, y_i in source and target language
- training objective (optimized with SGD):

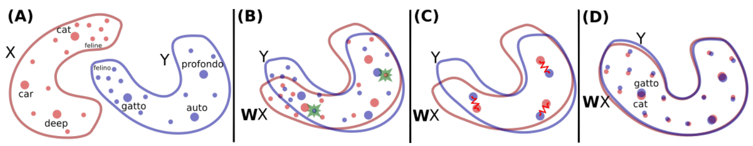
$$\underset{W}{\operatorname{argmin}} \sum_{i=1}^{n} ||Wx_i - y_i||^2$$

- training requires small seed lexicon of (x, y) pairs
- after mapping, induce bilingual lexicon via nearest neighbor search

Learning to Map Between Vector Spaces

unsupervised mapping [Miceli Barone, 2016, Conneau et al., 2017]

- adversarial training:
  - co-train classifier (adversary) that predicts whether embedding represents source or target language word
  - objective of linear, orthogonal transformation:
    fool classifier by making embeddings as similar as possible



[Conneau et al., 2017]

Learning to Map Between Vector Spaces

warning

these are recent research results – open questions remain

- under what conditions will this method succeed / fail?
- method was tested with typologically relatively similar languages
- method was tested with similar monolingual data (same domains and genres)

Improving Word Order

[Lample et al., 2017]

- joint training of both translation directions
- use translation model to back-translate monolingual data
- learn encoder-decoder to reconstruct original sentence from noisy translation
- iterate several times
- use various other tricks and objectives to improve learning
  - pre-trained embeddings
  - denoising autencoder as additional objective
  - shared encoder / decoder parameters in both directions
  - adversarial objective

	BLEU	
system	en-fr	en-de
supervised	28.0	21.3
word-by-word [Conneau et al., 2017]	6.3	7.1
[Lample et al., 2017]	15.1	9.6

Conclusion

- there are various ways to learn from monolingual data
  - combination with language model
  - pre-training and parameter sharing
  - creating synthetic training data
- methods are especially useful when:
  - parallel data is sparse
  - monolingual data is highly relevant (in-domain)
- hot research topic: learning to translate without parallel data

Bibliography I

Bertoldi, N. and Federico, M. (2009).
Domain adaptation for statistical machine translation with monolingual resources.
In Proceedings of the Fourth Workshop on Statistical Machine Translation StatMT 09. Association for Computational Linguistics.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017).
Word Translation Without Parallel Data.
CoRR, abs/1710.04087.

Currey, A., Miceli Barone, A. V., and Heafield, K. (2017).
Copied Monolingual Data Improves Low-Resource Neural Machine Translation.
In Proceedings of the Second Conference on Machine Translation, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

Gülçehre, c., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H., Bougares, F., Schwenk, H., and Bengio, Y. (2015).
On Using Monolingual Corpora in Neural Machine Translation.
CoRR, abs/1503.03535.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T., and Ma, W.-Y. (2016).
Dual Learning for Machine Translation.
In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, Advances in Neural Information Processing Systems 29, pages 820–828. Curran Associates, Inc.

Lambert, P., Schwenk, H., Servan, C., and Abdul-Rauf, S. (2011).
Investigations on Translation Model Adaptation Using Monolingual Data.
In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 284–293, Edinburgh, Scotland. Association for Computational Linguistics.

Bibliography II

Lample, G., Denoyer, L., and Ranzato, M. (2017).
Unsupervised Machine Translation Using Monolingual Corpora Only.
CoRR, abs/1711.00043.

Luong, M., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2016).
Multi-task Sequence to Sequence Learning.
In ICLR 2016.

Miceli Barone, A. V. (2016).
Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders.
In Proceedings of the 1st Workshop on Representation Learning for NLP, pages 121–126, Berlin, Germany. Association for Computational Linguistics.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013).
Exploiting Similarities among Languages for Machine Translation.
CoRR, abs/1309.4168.

Ramachandran, P., Liu, P., and Le, Q. (2017).
Unsupervised Pretraining for Sequence to Sequence Learning.
In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 383–391, Copenhagen, Denmark. Association for Computational Linguistics.

Schwenk, H. (2008).
Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation.
In International Workshop on Spoken Language Translation, pages 182–189.