Domain Adaptation in Machine Translation

Cristina España-Bonet, **Georg Heigold**

Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

15th March, 2017

This project has received funding from the
European Union's Horizon 2020 research and innovation programme
under grant agreement no. 645452.

Problem and First Solutions, SMT

Observation: Systems trained on small in-domain data perform better than systems trained on large out-of-domain data or its combination.

A waste of data?

1. Select a subset of in-domain sentences from the out-of-domain corpus $\Rightarrow$ LM, TM

   Ex: Cross-entropy difference between in-domain and general data

Data Selection

Selection of a subset of in-domain sentences according to:

- Perplexity with respect to an in-domain LM

  Ref: Yasuda et al. (2008)

- Cross-entropy difference between in-domain and general LMs

  Ref: Moore and Lewis (2010); Duh et al. (2013)

- TF-IDF similarity between a sentence and the collection of in-domain sentences

  Ref: Lü et al. (2007)

Problem and First Solutions, SMT

Observation: Systems trained on small in-domain data perform better than systems trained on large out-of-domain data or its combination.

A waste of data?

2. Interpolate in-domain and out-of-domain models
   $\Rightarrow$ LM, TM

   Ex: Fit weights of the models on an in-domain development set

Model Combination

Use information of both in-domain and out-of-domain models by:

- Combining LMs and/or TMs in a log-linear model
  Ref: Koehn and Schroeder (2007)
- Phrase table augmentation
  Ref: Bisazza et al. (2011)
- Interpolation of TMs (LMs) even at sentence level
  Ref: Finch and Sumita (2008); Sennrich (2011)

Data Acquisition

Obtaining additional parallel data from monolingual/comparable corpora:

- Translation of monolingual (comparable) in-domain corpora

  Ref: Schwenk (2008); Lambert et al. (2011); Sennrich et al. (2016)

- Extraction of parallel sentences from comparable corpora

  Ref: Abdul Rauf and Schwenk (2011); Skadina et al. (2012); Barrón-Cedeño et al. (2015)

Data Acquisition II

Extraction of parallel sentences from comparable corpora

	GNOME	WP_{CS}
Europarl	18.15	27.99
WP_{CS}	22.41	64.65
$WP_{CS+SP+SC}$	20.63	64.47

BLEU score for En2Es systems on IT domain

Data Acquisition II

Extraction of parallel sentences from comparable corpora

	GNOME	WP_{CS}
Europarl	18.15	27.99
WP_{CS}	22.41	64.65
$WP_{CS+SP+SC}$	20.63	64.47
EP+WP_{CS}	22.37	66.22
EP+$WP_{CS+SP+SC}$	21.43	65.67

BLEU score for En2Es systems on IT domain

Domain Adaptation in NMT

Transfer learning with in-domain corpus

What worked best:

- Ensemble with general and adapted models

An easy addition:

- + generated parallel corpora (either from mono translated or comparable)

DFKI's Domain Adaptation for NMT

Transfer Learning: train an NMT system on general data and then do transfer learning on a small amount (e.g., 20k sentence pairs) of in-domain data.

EN2DE System	Dev	Test
(char, word)	33.99	33.47
(char, word)+adaptation	40.66	40.12
(char, word)+adaptation+aug. vocab.	41.45	41.41
(char, BPE)	35.37	34.65
(char, BPE)+adaptation	44.09	44.44
(BPE, BPE)	23.63	22.74
(BPE, BPE)+adaptation	41.65	41.64

DFKI's Domain Adaptation for NMT

Transfer Learning: train an NMT system on general data and then do transfer learning on a small amount (e.g., 20k sentence pairs) of in-domain data.

EN2DE System	Dev	Test
(char, word)	33.99	33.47
(char, word)+adaptation	40.66	40.12
(char, word)+adaptation+aug. vocab.	41.45	41.41
(char, BPE)	35.37	34.65
(char, BPE)+adaptation	44.09	44.44
(BPE, BPE)	23.63	22.74
(BPE, BPE)+adaptation	41.65	41.64

Questions?

QT21 QUALITY TRANSLATION

Thank you for your attention

Cristina España-Bonet

cristinae@dfki.de

References

References are not exhaustive

References I

Abdul Rauf, S. and Schwenk, H. (2011). Parallel sentence generation from comparable corpora for improved smt. *Machine Translation*, 25(4):341–375.

Barrón-Cedeño, A., España-Bonet, C., Boldoba, J., and Màrquez, L. (2015). A Factory of Comparable Corpora from Wikipedia. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 3–13.

Bisazza, A., Ruiz, N., and Federico, M. (2011). Fill-up versus interpolation methods for phrase-based smt adaptation. In *IWSLT*, pages 136–143.

Duh, K., Neubig, G., Sudoh, K., and Tsukada, H. (2013). Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683, Sofia, Bulgaria. Association for Computational Linguistics.

References II

Finch, A. and Sumita, E. (2008). Dynamic model interpolation for statistical machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 208–215, Columbus, Ohio. Association for Computational Linguistics.

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic. Association for Computational Linguistics.

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.

References III

Lü, Y., Huang, J., and Liu, Q. (2007). Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 343–350.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Schwenk, H. (2008). Investigations on large-scale lightly-supervised training for statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 182–189.

References IV

Sennrich, R. (2011). Combining multi-engine machine translation and online learning through dynamic phrase tables. In Forcada, M. L., Depraetere, H., and Vandeghinste, V., editors, *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 89–96.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Skadina, I., Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., Vasiljevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M. L., and Pinnis, M. (2012). Collecting and using comparable corpora for statistical machine translation. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and*

References V

Evaluation (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA).

Yasuda, K., Zhang, R., Yamamoto, H., and Sumita, E. (2008). Method of selecting training data to build a compact and efficient translation model. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*.

DFKI's Domain Adaptation for NMT II

Source: char, Target: