Documentation & Data for *Prediction of Learning Curves in Machine Translation*

Prasanth Kolachina* **Nicola Cancedda†** **Marc Dymetman†** **Sriram Venkatapathy†**
∗ LTRC, IIIT-Hyderabad, Hyderabad, India
† Xerox Research Centre Europe, 6 chemin de Maupertuis, 38240 Meylan, France

Abstract

Parallel data in the domain of interest is the key resource when training a statistical machine translation (SMT) system for a specific purpose. Since ad-hoc manual translation can represent a significant investment in time and money, a prior assesment of the amount of training data required to achieve a satisfactory accuracy level can be very useful. In this work, we show how to *predict* what the learning curve would look like *if* we were to manually translate increasing amounts of data.

As a part of this work, we trained a multitude of instances of the same phrase-based translation system for 30 distinct combinations of language-pair and domain, each with fourteen distinct training sets of increasing size and tested these instances on multiple in-domain datasets, generating 96 learning curves. We provide the BLEU measurements for all 96 learning curves and other datasets obtained from our experiments.

1 Contribution Summary

This document summarizes the settings used in our experiments and data created as a byproduct of our work, for the purposes of replicating and building on our work.

The structure of this document is as follows: Section 2 lists the corpora and the external software used in our work. Section 3 shows the settings used in creating the translation models and the various preprocessing steps required. In Section 4, we provide the necessary details for using the datasets provided here.

2 Corpora and External Dependencies

We use the following corpora resources in our experiments to train the translation models. All of the resources are available online and free-of-cost.

1. Europarl parallel corpus (v6) for European languages (Koehn, 2005)
2. News Commentary corpus (v6) [1]
3. EMEA medical domain (v2) parallel corpus (Tiedemann, 2009)
4. KFTT parallel corpus (v1.0) for Japanese-English language pair (Neubig, 2011)

The following external toolkits/packages are used in our experiments.

1. Moses: Statistical Machine Translation (SMT) Toolkit (Koehn et al., 2007)
2. SRI Language Modeling Toolkit (Stolcke, 2002)
3. Part-of-Speech taggers: Stanford Part-of-Speech Tagger (Toutanova et al., 2003), SVMTool (Giménez and Màrquez, 2004), MElt (Denis and Sagot, 2009), RFTagger (Schmid and Laws, 2008), KyTea (Neubig et al., 2011)
4. Python packages: SciPy (Jones et al., 2001), Matplotlib (Hunter, 2007), Scikits-learn (Pedregosa et al., 2011)

3 Moses Translation systems

In this section, we provide the steps followed to create the translation models. During the course of our experiments, we trained a large number of translation models for 30 distinct combinations of language-pair and domain, each with distinct training sets of increasing size and tested these models on multiple in-domain test sets. We use the phrase-based SMT model available in Moses (Koehn et al., 2007) to train the translation models for all configurations.

*This research was carried out during at internship at Xerox Research Centre Europe.

[1] Both available from `http://statmt.org/wmt11/`

3.1 Baseline systems

The translation models trained follow the settings used to create the baseline systems in the WMT shared task (Callison-Burch et al., 2011). Models trained on Europarl corpus and News Commentary corpus use seperate in-domain monolingual data of the target language to train the language model[2]. In the case of EMEA and KFTT models, we use the entire target side of the parallel corpus due to lack of in-domain monolingual data in the target language. The language model used is the same for all translation models trained on different sizes of the corpus.

3.2 Corpora Preprocessing

The preprocessing stage consists of the three steps described in the baseline system description: *i*) tokenization *ii*) cleaning *iii*) lowercasing . The built-in tokenizer available in Moses is used to tokenize corpora in English and other European languages prior to building the translation models. For the Japanese data, we use KyTea segmentation toolkit to tokenize the raw text.

We also run part-of-speech taggers on the tokenized text in order to extract features described in the paper for the prediction models. Table 1 shows the list of the pos taggers used and their respective settings.

Language	Tagger
Czech	RFTagger
Danish	Stanford (trained on Danish treebank)
English	Stanford (left3words-distsim-wsj-0-18.tagger)
French	MElt
German	Stanford (german-fast.tagger)
Japanese	KyTea
Spanish	SVMTool (-S LRL -T 0)

Table 1: Settings used for various POS taggers

4 Datasets

4.1 BLEU scores

The BLEU scores obtained on different configurations for all sizes are provided in `expt-data/curve-fitting/points/bleu` directory. There are 30 files in this directory, each corresponding to a distinct combination of language pair and domain. Each file in this directory contains the following fields:

1. Number of sentences in the parallel corpus (column 1).
2. Number of tokens in the source language portion of the parallel corpus (column 2).
3. BLEU scores on multiple test sets for each configuration (column 3-).

The scorer used to compute the BLEU scores is provided as a part of Moses toolkit (Koehn et al., 2007).

4.2 Curve Fitting

The BLEU scores are used to fit the different family of curves described on Section 3 of the paper. The implementation of the Levenberg-Marquardt algorithm used to compute the best parameter values is available in the SciPy package (Jones et al., 2001). Table 2 shows the intial values of the parameters for curve fitting.

Model	Initial parameter values
Exp_3	$a = 0, b = 0, c = 1$
Exp_4	$a = 0, b = 0, c = 1, \alpha = 1$
ExpP_3	$b = 1, c = 1, \alpha = 0$
Pow_3	$a = 0, c = 1, \alpha = 1$
Pow_4	$a = 0, b = 1, c = 1, \alpha = 1$
ILog_2	$a = 0, c = 1$

Table 2: Intial parameter values for curve fitting

The BLEU scores at the three anchor points (10K, 75K, 500K) given by the Pow_3 curve for all the 96 learning curves are given in `expt-data/curve-fitting/evaluation/pow-values.txt` file. The optimal parameter values given by the Pow_3 curve family used to compute these values can be obtained by solving the curve family for the values at the three anchor points.

4.3 Feature Extraction

The features obtained for all the 96 learning curves are given in `expt-data/feature-extraction` directory. All the files in the directory are in xml format and contain the feature values extracted from the corpora used to infer the learning curves. Table 3 shows the description of different features present in these files.

[2]Available from `http://statmt.org/wmt11`

FeatureName	Description
test-src-segcount	Number of segments in the test set
test-src-tokcount	Number of tokens in the test set
test-src-avgseglen	Average length of segment in the test set
test-src-avgtoklen	Average length of token in the test set
train-src-avgseglen	Average length of segment in the source monolingual corpus
train-src-avgtoklen	Average length of tokens in the source monolingual corpus
train-src-ttr_ord1	Type-token ratio of order 1 in the source monolingual corpus
train-src-ttr_ord2	Type-token ratio of order 2 in the source monolingual corpus
train-src-ttr_ord3	Type-token ratio of order 3 in the source monolingual corpus
train-src-ttr_ord4	Type-token ratio of order 4 in the source monolingual corpus
test-src-ttr_ord1	Type-token ratio of order 1 in the source test set
test-src-ttr_ord2	Type-token ratio of order 2 in the source test set
test-src-ttr_ord3	Type-token ratio of order 3 in the source test set
test-src-ttr_ord4	Type-token ratio of order 4 in the source test set
train-tgt-ttr_ord1	Type-token ratio of order 1 in the target monolingual corpus
train-tgt-ttr_ord2	Type-token ratio of order 2 in the target monolingual corpus
train-tgt-ttr_ord3	Type-token ratio of order 3 in the target monolingual corpus
train-tgt-ttr_ord4	Type-token ratio of order 4 in the target monolingual corpus
train+test-src-perplex2	Perplexity on test set using bigram language model trained on source monolingual corpus
train+test-src-perplex3	Perplexity on test set using trigram language model trained on source monolingual corpus
train+test-src-perplex4	Perplexity on test set using 4-gram language model trained on source monolingual corpus
train+test-src-perplex5	Perplexity on test set using 5-gram language model trained on source monolingual corpus
test-src+ref-seglenratio	Ratio of sentences in test and reference datasets
train-src+tgt-ttrratio_ord1	Ratio of type-token ratios from source and target monolingual corpora (order = 1)
train-src+tgt-ttrratio_ord2	Ratio of type-token ratios from source and target monolingual corpora (order = 2)
train-src+tgt-ttrratio_ord3	Ratio of type-token ratios from source and target monolingual corpora (order = 3)
train-src+tgt-ttrratio_ord4	Ratio of type-token ratios from source and target monolingual corpora (order = 4)
test-src+ref-ttrratio_ord1	Ratio of type-token ratios from test and reference datasets (order = 1)
test-src+ref-ttrratio_ord2	Ratio of type-token ratios from test and reference datasets (order = 2)
test-src+ref-ttrratio_ord3	Ratio of type-token ratios from test and reference datasets (order = 3)
test-src+ref-ttrratio_ord4	Ratio of type-token ratios from test and reference datasets (order = 4)
pos-src+tgt-crossentropy_ord2	Cross-entropy between distributions of part-of-speech tags in source and target monolingual corpus (order = 2)
pos-src+tgt-crossentropy_ord4	Cross-entropy between distributions of part-of-speech tags in source and target monolingual corpus (order = 4)
pos-src+tgt-crossentropy_ord6	Cross-entropy between distributions of part-of-speech tags in source and target monolingual corpus (order = 6)
pos-src+tgt-crossentropy_ord8	Cross-entropy between distributions of part-of-speech tags in source and target monolingual corpus (order = 8)

Table 3: Descriptions of different features extracted for inferring learning curves

4.4 Inference

The implementations for the Lasso and the Ridge linear regression models are available in (Pedregosa et al., 2011). The predictions at the three anchor points from all the three models (Baseline, Lasso and Ridge) can be found in `expt-data/inference` directory. Each prediction file has the following fields:

1. Name of the learning curve (column 1).
2. Actual value at 10K (from the gold curve), Predicted value at 10K and (Actual-Predicted) value (column 2, 3, 4).
3. Actual value at 75K (from the gold curve), Predicted value at 75K and (Actual-Predicted) value (column 5, 6, 7).
4. Actual value at 500K (from the gold curve), Predicted value at 500K and (Actual-Predicted) value (column 8, 9, 10).

Associated with each model, is also given the evaluation of the predictions.

References

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July. Association for Computational Linguistics.

Pascal Denis and Benoît Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Pacific Asia Conference on Language, Information and Computation*, Hong Kong, China.

Jesús Giménez and Lluís Màrquez. 2004. Svmtool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of the Fourth conference on International Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

John D. Hunter. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science and Engineering*, 9(3):90–95, May.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. 2001–. SciPy: Open source scientific tools for Python.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand, September.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 529–533, Portland, Oregon, USA, June. Association for Computational Linguistics.

Graham Neubig. 2011. The Kyoto Free Translation Task. http://www.phontron.com/kftt.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Helmut Schmid and Florian Laws. 2008. Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 777–784, Manchester, UK, August. Coling 2008 Organizing Committee.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver, USA, September.

Jörg Tiedemann. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of Human Language Technologies: The 2003 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259, Edmonton, Canada, May. Association for Computational Linguistics.