UCAM Biomedical translation at WMT19: Transfer learning multi-domain ensembles

Danielle Saunders[†] and **Felix Stahlberg**[†] and **Bill Byrne**[‡†]

[†]Department of Engineering, University of Cambridge, UK

[‡]SDL Research, Cambridge, UK

Abstract

The 2019 WMT Biomedical translation task involved translating Medline abstracts. We approached this using transfer learning to obtain a series of strong neural models on distinct domains, and combining them into multi-domain ensembles. We further experiment with an adaptive language-model ensemble weighting scheme. Our submission achieved the best submitted results on both directions of English-Spanish.

1 Introduction

Neural Machine Translation (NMT) in the biomedical domain presents challenges in addition to general domain translation. Firstly, available corpora are relatively small, exacerbating the effect of noisy or poorly aligned training data. Secondly, individual sentences within a biomedical document may use specialist vocabulary from small domains like health or statistics, or may contain generic language. Training to convergence on a single biomedical dataset may therefore not correspond to good performance on arbitrary biomedical test data.

Transfer learning is an approach in which a model is trained using knowledge from an existing model (Khan et al., 2018). Transfer learning typically involves initial training on a large, general domain corpus, followed by fine-tuning on the domain of interest. We apply transfer learning iteratively on datasets from different domains, obtaining strong models that cover two domains for both directions of the English-German language pair, and three domains for both directions of English-Spanish.

The domain of individual documents in the 2019 Medline test dataset is unknown, and may vary sentence-to-sentence. Evenly-weighted ensembles of models from different domains can give good results in this case (Freitag and Al-Onaizan, 2016). However, we suggest a better approach would take into account the likely domain, or domains, of each test sentence. We therefore investigate applying Bayesian Interpolation for language-model based multi-domain ensemble weighting.

1.1 Iterative transfer learning

Transfer learning has been used to adapt models both across domains, e.g. news to biomedical domain adaptation, and within one domain, e.g. WMT14 biomedical data to WMT18 biomedical data (Khan et al., 2018). For en2de and de2en we have only one distinct in-domain dataset, and so we use standard transfer learning from a general domain news model.

For es2en and en2es, we use the domain-labelled Scielo dataset to provide two distinct domains, health and biological sciences ('bio'), in addition to the complete biomedical dataset (Neves et al., 2016). We therefore experiment with iterative transfer learning, in which a model trained with transfer learning is then trained further on the original domain.

NMT transfer learning for domain adaptation involves using the performance of a model on some general domain A to improve performance on some other domain B: $A \rightarrow B$. However, if the two domains are sufficiently related, we suggest that task B could equally be used for transfer learning A: $B \rightarrow A$. The stronger general model A could then be used to achieve even better performance on other tasks: $B \rightarrow A \rightarrow B$, $B \rightarrow A \rightarrow C$, and so on.

1.2 Adaptive decoding

Previous work on transfer learning typically aims to find a single model that performs well on a known domain of interest (Khan et al., 2018).

The biomedical translation task offers a scenario in which the test domain is unknown, since individual Medline documents can have very different styles and topics. Our approach is to decode such test data with an ensemble of distinct domains.

For intuitive ensemble weights, we use sequence-to-sequence Bayesian Interpolation (BI) as described in Saunders et al. (2019), which also contains a more in-depth derivation and discusses possible hyperparameter configurations. We consider models $p_k(\mathbf{y}|\mathbf{x})$ trained on K domains, used for $T = K$ domain decoding tasks. We assume throughout that $p(t) = \frac{1}{T}$, i.e. that tasks are equally likely absent any other information. Weights $\lambda_{k,t}$ define a task-conditional ensemble. At step i, where $h_i = \mathbf{y}_{1:i-1}$ is decoding history:

$$p(y_i|h_i, \mathbf{x}) = \sum_{k=1}^{K} p_k(y_i|h_i, \mathbf{x}) \sum_{t=1}^{T} p(t|h_i, \mathbf{x})\lambda_{k,t} \tag{1}$$

This is an adaptively weighted ensemble where, for each source sentence $\mathbf{x}$ and output hypothesis $\mathbf{y}$, we re-estimate $p(t|h_i, \mathbf{x})$ at each step:

$$p(t|h_i, \mathbf{x}) = \frac{p(h_i|t, \mathbf{x})p(t|\mathbf{x})}{\sum_{t'=1}^{T} p(h_i|t', \mathbf{x})p(t'|\mathbf{x})} \tag{2}$$

$p(h_i|t, \mathbf{x})$ is found from the last score of each model:

$$p(h_i|t, \mathbf{x}) = p(y_{i-1}|h_{i-1}, t, \mathbf{x}) \tag{3}$$
$$= \sum_{k} p_k(y_{i-1}|h_{i-1}, t, \mathbf{x})\lambda_{k,t}$$

We use G_t, an n-gram language model trained on source training sentences from task t, to estimate initial task posterior $p(t|\mathbf{x})$:

$$\frac{p(\mathbf{x}|t)p(t)}{\sum_{t'=1}^{T} p(\mathbf{x}|t')p(t')} = \frac{G_t(\mathbf{x})^{\alpha}}{\sum_{t'=1}^{T} G_{t'}(\mathbf{x})^{\alpha}} \tag{4}$$

Here α is a smoothing parameter. If $\overline{G}_{k,t} = \sum_{\mathbf{x} \in \text{Test}_t} G_k(\mathbf{x})$, we take:

$$\lambda_{k,t} = \frac{\overline{G}_{k,t}^{\alpha}}{\sum_{k'} \overline{G}_{k',t}^{\alpha}} \tag{5}$$

Figure 1 demonstrates this adaptive decoding when weighting a biomedical and a general (news) domain model to produce a biomedical sentence. The model weights are even until biomedical-specific vocabulary is produced, at which point the in-domain model dominates.
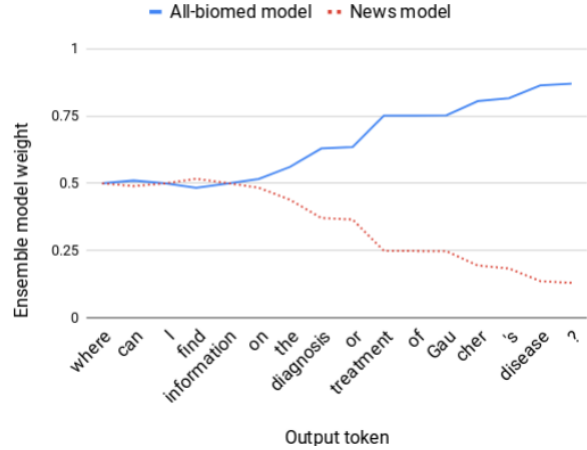

Figure 1: Adaptively adjusting model weights during decoding with Bayesian Interpolation

1.3 Related work

Transfer learning has been applied to NMT in many forms. Luong and Manning (2015) use transfer learning to adapt a general model to in-domain data. Zoph et al. (2016) use multilingual transfer learning to improve NMT for low-resource languages. Chu et al. (2017) introduce mixed fine-tuning, which carries out transfer learning to a new domain combined with some original domain data. Kobus et al. (2017) train a single model on multiple domains using domain tags. Khan et al. (2018) sequentially adapt across multiple biomedical domains to obtain one single-domain model.

At inference time, Freitag and Al-Onaizan (2016) use uniform ensembles of general and fine-tuned models. Our Bayesian Interpolation experiments extend previous work by Allauzen and Riley (2011) on Bayesian Interpolation for language model combination.

2 Experimental setup

2.1 Data

We report on two language pairs: Spanish-English (es-en) and English-German (en-de). Table 1 lists the data used to train our biomedical domain evaluation systems. For en2de and de2en

---

[1] https://ufal.mff.cuni.cz/ufal_medical_corpus
[2] Dušek et al. (2017)
[3] Neves et al. (2016)
[4] https://github.com/biomedical-translation-corpora/medline (Yepes et al., 2017)
[5] http://www.himl.eu/test-sets

	Domain	Training datasets	Sentence pairs	Dev datasets	Sentence pairs
es-en	All-biomed	UFAL Medical[1]	639K	Khresmoi[2]	1.5K
		Scielo[3]	713K		
		Medline titles[4]	288K		
		Medline training abstracts	83K		
		Total (pre) / post-filtering	(1723K) / **1291K**		
	Health	Scielo health only	587K	Scielo 2016 health	5K
		Total post-filtering	**558K**		
	Bio	Scielo bio only	126K	Scielo 2016 bio	4K
		Total post-filtering	**122K**		
en-de	All-biomed	UFAL Medical	2958K	Khresmoi	1.5K
		Medline training abstracts	33K	Cochrane[5]	467
		Total (pre) / post-filtering	(2991K) / **2156K**		

Table 1: Biomedical training and validation data used in the evaluation task. For both language pairs identical data was used in both directions.

we additionally reuse strong general domain models trained on the WMT19 news data, including filtered Paracrawl. Details of data preparation and filtering for these models are discussed in Stahlberg et al. (2019).

For each language pair we use the same training data in both directions, and use a 32K-merge source-target BPE vocabulary (Sennrich et al., 2016) trained on the 'base' domain training data (news for en-de, Scielo health for es-en)

For the biomedical data, we preprocess the data using Moses tokenization, punctuation normalization and truecasing. We then use a series of simple heuristics to filter the parallel datasets:

- Detected language filtering using the Python langdetect package[6]. In addition to mislabelled sentences, this step removes many sentences which are very short or have a high proportion of punctuation or HTML tags.

- Remove sentences containing more than 120 tokens or fewer than 3.

- Remove duplicate sentence pairs

- Remove sentences where the ratio of source to target tokens is less than 1:3.5 or more than 3.5:1

- Remove pairs where more than 30% of either sentence is the same token.

2.2 Model hyperparameters and training

We use the Tensor2Tensor implementation of the Transformer model with the transformer_big setup for all NMT models (Vaswani et al., 2018). By default this model size limits batch size of

2K due to memory constraints. We delay gradient updates by a factor of 8, letting us effectively use a 16K batch size (Saunders et al., 2018). We train each domain model until it fails to improve on the domain validation set in 3 consecutive checkpoints, and perform checkpoint averaging over the final 10 checkpoints to obtain the final model (Junczys-Dowmunt et al., 2016).

At inference time we decode with beam size 4 using SGNMT (Stahlberg et al., 2017). For BI we use 2-gram KENLM models (Heafield, 2011) trained on the source training data for each domain. For validation results we report cased BLEU scores with SacreBLEU (Post, 2018)[7]; test results use case-insensitive BLEU.

2.3 Results

Our first experiments involve iterative transfer learning in es2en and en2es to obtain models on three separate domains for the remaining evaluation. We use health, a relatively clean and small dataset, as the initial domain to train from scratch. Once converged, we use this to initialise training on the larger, noiser all-biomed corpus. When the all-biomed model has converged, we use it to initialise training on the health data and bio data for stronger models on those domains. Figure 2 shows the training progression for the health and all-biomed models, as well as the standard transfer learning case where we train on all-biomed from scratch.

Table 2 gives single model validation scores for es2en and en2es models with standard and iterative transfer learning. We find that the all-biomed domain gains 1-2 BLEU points from transfer learning. Moreover, the health domain gains on

---

[6]https://pypi.org/project/langdetect/

[7]SacreBLEU signature: BLEU+case.mixed
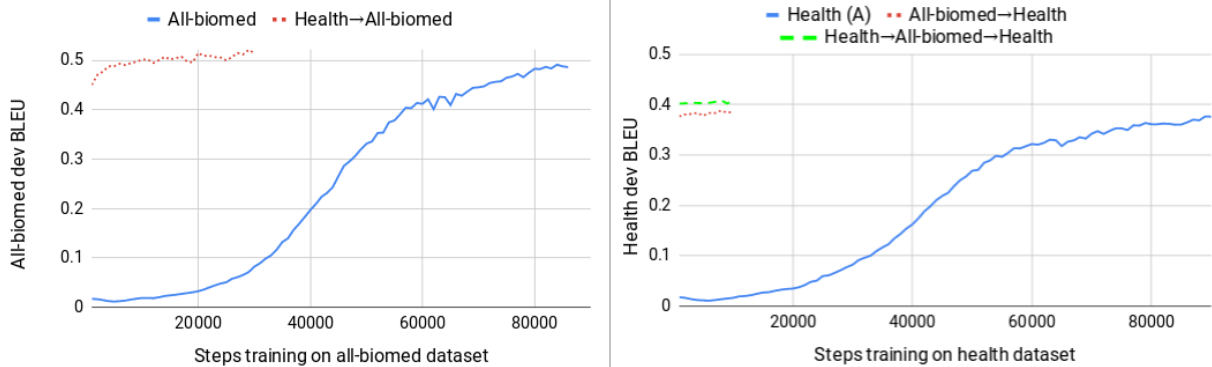+numrefs.1+smooth.exp+tok.13a+version.1.3.2

Figure 2: Transfer learning for es2en domains. Left: standard transfer learning improves performance from a smaller (health) to a larger (all-biomed) domain. Right: returning to the original domain after transfer learning provides further gains on health.

Transfer learning schedule	es2en			en2es		
	Khresmoi	**Health**	**Bio**	**Khresmoi**	**Health**	**Bio**
Health	45.1	35.7	34.0	41.2	34.7	36.1
All-biomed	49.8	35.4	35.7	43.4	33.9	37.5
All-biomed → Health	48.9	36.4	35.9	43.0	35.2	38.0
All-biomed → Bio	48.0	34.6	37.2	43.2	34.1	40.5
Health → All-biomed	**52.1**	36.7	37.0	44.2	35.0	39.0
Health → All-biomed → Health	51.1	**37.0**	37.2	44.0	**36.3**	39.5
Health → All-biomed → Bio	50.6	36.0	**38.0**	**45.2**	35.3	**41.3**

Table 2: Validation BLEU for English-Spanish models with transfer learning. We use the final three models in our submission.

all domains from iterative transfer learning relative to training from scratch and relative to standard transfer learning(All-biomed → Health), despite being trained twice to convergence on health.

We submitted three runs to the WMT19 biomedical task for each language pair: the best single all-biomed model, a uniform ensemble of models on two en-de and three es-en domains, and an ensemble with Bayesian Interpolation. Tables 3 and 4 give validation and test scores.

We find that a uniform multi-domain ensemble performs well, giving 0.5-1.2 BLEU improvement on the test set over strong single models. We see small gains from using BI with ensembles on most validation sets, but only on en2es test.

Following test result release, we noted that, in general, we could predict BI ($\alpha = 0.5$) performance by comparing the uniform ensemble with the oracle model performing best on each validation domain. For en2es uniform ensembling underperforms the health and bio oracle models on their validation sets, and the uniform ensemble slightly underperforms BI on the test data. For en2de, by contrast, uniform ensembling is consistently better than oracles on the dev sets, and outperforms BI on the test data. For de2en and es2en,

uniform ensembling performs similarly to the oracles, and performs similarly to BI.

From this, we hypothesise that BI ($\alpha = 0.5$) has a tendency to converge to a single model. This is effective when single models perform well (en2es) but ineffective if the uniform ensemble is predictably better than any single model (en2de). Consequently in Table 5 we experiment with BI ($\alpha = 0.1$). In this case BI matches or out-performs the uniform ensemble. Notably, for en2es, where BI ($\alpha = 0.5$) performed well, taking $\alpha = 0.1$ does not harm performance.

3 Conclusions

Our WMT19 Biomedical submission covers the English-German and English-Spanish language pairs, achieving the best submitted results on both directions of English-Spanish. We use transfer learning iteratively to train single models which perform well on related but distinct domains, and show further gains from multi-domain ensembles. We explore Bayesian Interpolation for multi-domain ensemble weighting, and find that a strongly smoothed case gives small gains over uniform ensembles.

	es2en				en2es			
	Khresmoi	**Health**	**Bio**	**Test**	**Khresmoi**	**Health**	**Bio**	**Test**
Health → All-biomed	52.1	36.7	37.0	42.4	44.2	35.0	39.0	44.9
Health → All-biomed → Health	51.1	37.0	37.2	-	44.0	36.3	39.5	-
Health → All-biomed → Bio	50.6	36.0	38.0	-	45.2	35.3	41.3	-
Uniform ensemble	**52.2**	36.9	37.9	**43.0**	**45.1**	35.6	40.2	45.4
BI ensemble (α=0.5)	52.1	**37.0**	**38.1**	42.9	44.5	**35.7**	**41.2**	**45.6**

Table 3: Validation and test BLEU for models used in English-Spanish language pair submissions.

	de2en			en2de		
	Khresmoi	**Cochrane**	**Test**	**Khresmoi**	**Cochrane**	**Test**
News	43.8	46.8	-	30.4	40.7	-
News → All-biomed	44.5	47.6	27.4	31.1	39.5	26.5
Uniform ensemble	45.3	48.4	**28.6**	**32.6**	42.9	**27.2**
BI ensemble (α=0.5)	**45.4**	**48.8**	28.5	32.4	**43.1**	26.4

Table 4: Validation and test BLEU for models used in English-German language pair submissions.

	es2en	en2es	de2en	en2de
Uniform	**43.2**	45.3	28.3	25.9
BI (α=0.5)	43.0	**45.5**	28.2	25.2
BI (α=0.1)	**43.2**	**45.5**	**28.5**	**26.0**

Table 5: Comparing uniform ensembles and BI with varying smoothing factor on the WMT19 test data. Small deviations from official test scores on submitted runs are due to tokenization differences. $\alpha = 0.5$ was chosen for submission based on results on available development data.

Acknowledgments

References

Cyril Allauzen and Michael Riley. 2011. Bayesian Language Model Interpolation for Mobile Speech Input. In *Proceedings of the Twelfth Annual Conference of the International Speech Communication Association*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391.

Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Jindřich Libovický, Pavel Pecina, Aleš Tamchyna, and Zdeňka Urešová. 2017. Khresmoi summary translation test data 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for Neural Machine Translation. *CoRR*, abs/1612.06897.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. 2016. The AMU-UEDIN submission to the WMT16 news translation task: Attention-based NMT models as feature functions in phrase-based SMT. In *Proceedings of the First Conference on Machine Translation*, pages 319–325, Berlin, Germany. Association for Computational Linguistics.

Abdul Khan, Subhadarshi Panda, Jia Xu, and Lampros Flokas. 2018. Hunter nmt system for wmt18 biomedical translation task: Transfer learning in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 655–661.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford Neural Machine Translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.

Mariana L Neves, Antonio Jimeno-Yepes, and Aurélie Névéol. 2016. The ScieLO Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. In *LREC*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. *CoRR*, abs/1804.08771.

---

[8] http://www.hpc.cam.ac.uk

Danielle Saunders, Felix Stahlberg, Adrià de Gispert, and Bill Byrne. 2018. Multi-representation ensembles and delayed sgd updates improve syntax-based nmt. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 319–325.

Danielle Saunders, Felix Stahlberg, Adrià de Gispert, and Bill Byrne. 2019. Domain adaptive inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1715–1725.

Felix Stahlberg, Eva Hasler, Danielle Saunders, and Bill Byrne. 2017. SGNMT–A Flexible NMT Decoding Platform for Quick Prototyping of New Models and Search Strategies. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 25–30.

Felix Stahlberg, Danielle Saunders, Adrià de Gispert, and Bill Byrne. 2019. CUED@WMT19:EWC&LMs. In *Proceedings of the Fourth Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for Neural Machine Translation. *CoRR*, abs/1803.07416.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondrej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, et al. 2017. Findings of the wmt 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.