Hierarchical Incremental Adaptation for Statistical Machine Translation

Joern Wuebker
Lilt Inc.
joern@lilt.com

Spence Green
Lilt Inc.
spence@lilt.com

John DeNero
Lilt Inc.
john@lilt.com

Abstract

We present an incremental adaptation approach for statistical machine translation that maintains a flexible hierarchical domain structure within a single consistent model. Both weights and rules are updated incrementally on a stream of post-edits. Our multi-level domain hierarchy allows the system to adapt simultaneously towards local context at different levels of granularity, including genres and individual documents. Our experiments show consistent improvements in translation quality from all components of our approach.

1 Introduction

Suggestions from a machine translation system can increase the speed and quality of professional human translators (Guerberof, 2009; Plitt and Masselot, 2010; Green et al., 2013a, *inter alia*). However, querying a single fixed model for all different documents fails to incorporate contextual information that can potentially improve suggestion quality. We describe a model architecture that adapts simultaneously to multiple genres and individual documents, so that translation suggestions are informed by two levels of contextual information.

Our primary technical contribution is a hierarchical adaptation technique for a post-editing scenario with incremental adaptation, in which users request translations of sentences in corpus order and provide corrected translations of each sentence back to the system (Ortiz-Martínez et al., 2010). Our learning approach resembles Hierarchical Bayesian Domain Adaptation (Finkel and Manning, 2009), but updates both the model weights and translation rules in real time based on these corrected translations (Mathur et al., 2013; Denkowski et al., 2014). Our adapted system can provide on-demand translations for any genre and document to which it has ever been exposed, using weights and rules for domains associated with each translation request.

Our weight adaptation is performed using a hierarchical extension to fast and adaptive online training (Green et al., 2013b), a technique based on AdaGrad (Duchi et al., 2011) and forward-backward splitting (Duchi and Singer, 2009) that can accurately set weights for both dense and sparse features (Green et al., 2014b). Rather than adjusting all weights based on each example, our extension adjusts offsets to a fixed baseline system. In this way, the system can adapt to multiple genres while preventing cross-genre contamination.

In large-scale experiments, we adapt a multi-genre baseline system to patents, lectures, and news articles. Our experiments show that sparse models, hierarchical updates, and rule adaptation all contribute consistent improvements. We observe quality gains in all genres, validating our hypothesis that document and genre context are important additional inputs to a machine translation system used for post-editing.

2 Background

The log-linear appoach to statistical machine translation models the predictive translation distribution $p(e|f; w)$ directly in log-linear form (Och and Ney, 2004):

$$p(e|f; w) = \sum_{\substack{r: \\ src(r)=f \\ tgt(r)=e}} \frac{1}{Z(f)} \exp\left[w^\top \phi(r; c)\right] \quad (1)$$

where $f \in \mathcal{F}$ is a string in the set of all source language strings $\mathcal{F}$, $e \in \mathcal{E}$ is a string in the set of all target language strings $\mathcal{E}$, r is a phrasal derivation with source and target projections $src(r)$ and $tgt(r)$, $w \in \mathbb{R}^d$ is the vector of model parameters, $\phi(\cdot) \in \mathbb{R}^d$ is a feature map computed using corpus c, and $Z(f)$ is an appropriate normalizing constant. During search, the maximum approximation is applied rather than summing over the derivations r.

Model. We extend a phrase-based system for which $\phi(r; c)$ includes 16 dense features:

- Two phrasal channel models and two lexical channel models (Koehn et al., 2003), the (log) count of the rule in the training corpus c, and an indicator for singleton rules in c.
- Six orientation models that score ordering configurations in r by their frequency in c (Koehn et al., 2007).
- A linear distortion penalty that promotes monotonic translation.
- An n-gram language model score, $p(e)$, which scores the target language projection of r using statistics from a monolingual corpus.
- Fixed-value phrase and word penalties.

The elements of $\phi(r; c)$ may also include sparse features that have non-zero values for only a subset of rules, but typically do not depend on c (Liang et al., 2006). In this paper, we use four types of sparse features: rule indicators, discriminative lexicalized reordering indicators, rule shape indicators and alignment features (Green et al., 2014b).

The model parameters w are chosen to maximize translation quality on a tuning set.

Adaptation. Domain adaptation for machine translation has improved quality using a variety of approaches, including data selection (Ceauşfu et al., 2011), regularized online learning (Simianer et al., 2012; Green et al., 2013b), and input classification (Xu et al., 2007; Banerjee et al., 2010; Wang et al., 2012) and has also been investigated for multi-domain tasks (Sennrich et al., 2013; Cui et al., 2013; Simianer and Riezler, 2013). Even without domain labels at either training or test time, multi-task learning can boost translation quality in a batch setting (Duh et al., 2010; Song et al., 2011).

Post-editing with incremental adaptation describes a particular mixed-initiative setting (Ortiz-Martínez et al., 2010; Hardt and Elming, 2010). For each f in a corpus, the machine generates a hypothesis e, then a human provides a corrected translation e^* to the machine. Observing e^* can affect both the model weights w and corpus c used for rule extraction and dense feature estimation.[1] To translate the ith sentence f_i, the system uses weights w_{i-1} and corpus c_{i-1}. The new corpus c_i results from adding (f_i, e_i^*) to c_{i-1}. For incremental adaptation, speed is essential, and so w_i is typically computed with a single online update from w_{i-1} using (f_i, e_i^*) as the tuning example.

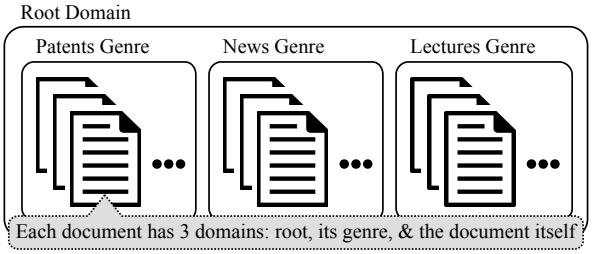

Figure 1: The weights used to translate a document in the patent genre include three domains.

To alleviate the need for human intervention in the experiment cycle, *simulated post-editing* (Hardt and Elming, 2010; Denkowski et al., 2014) replaces each e^* with a reference that is not a corrected variant of e. Thus, a standard test corpus can be used as an adaptation corpus. Prior work on online learning from post-edits has demonstrated the benefit of adjusting only c (Ortiz-Martínez et al., 2010; Hardt and Elming, 2010) and further benefit from adjusting both c and w (Mathur et al., 2013; Denkowski et al., 2014). Incremental adaptation of both c and the weights w for sparse features is reported to yield large quality gains by Wäschle et al. (2013).[2]

3 Hierarchical Incremental Adaptation

Our hierachical approach to incremental adaptation uses document and genre information to adapt appropriately to multiple contexts. We assume that each sentence f_i has a known set D_i of domains, which identify the genre and individual document origin of the sentence. This set could be extended to include topics, individual translators, etc.

Figure 1 shows the domains that we apply in experiments. All sentences in the baseline training corpus, the tuning corpus, and the adaptation corpus share a ROOT domain.

[1] For the purpose of our description, the corpus c is equivalent to the set of phrases and their scores in the rule table. We prefer this notation because it is consistent with our stream-based rule table, where the models are computed on-the-fly from the indexed training corpus c.

[2] Language model adaptation also has a rich literature, but it is beyond the scope of this paper.

Our adaptation is conceptually similar to hierarchical Bayesian domain adaptation (Finkel and Manning, 2009), but both weights and feature values depend on D_i, and we use L_1 regularization.

Weight Updates. Model tuning and adaptation are performed with AdaGrad, an online subgradient method with an *adaptive learning rate* that comes with good theoretical guarantees. AdaGrad makes the following update:

$$w_t = w_{t-1} - \eta \Sigma_t^{1/2} \nabla \ell_t(w_{t-1}) \quad (2)$$

$$\Sigma_t^{-1} = \Sigma_{t-1}^{-1} + \nabla \ell_t(w_{t-1}) \nabla \ell_t(w_{t-1})^\top$$

$$= \sum_{i=1}^{t} \nabla \ell_i(w_{i-1}) \nabla \ell_i(w_{i-1})^\top \quad (3)$$

The loss function ℓ reflects the pairwise ordering between hypotheses. For feature selection, we apply an L_1 penalty via forward-backward splitting (Duchi and Singer, 2009). η is the initial learning rate. See (Green et al., 2013b) for details.

Our adaptation schema is an extension of *frustratingly easy domain adaptation* (FEDA) (Daumé III, 2007) to multiple domains with different regularization parameters, similar to (Finkel and Manning, 2009). Each feature value is replicated for each domain. Let $\mathcal{D}$ denote the set of all domains present in the adaptation set. Given an original feature vector $\phi(r; c)$ for derivation r of sentence f_i with $D_i \subseteq \mathcal{D}$, the replicated feature vector includes $|\mathcal{D}|$ copies of $\phi(r; c)$, one for each $d \in |\mathcal{D}|$, such that

$$\phi_d(r; c) = \begin{cases} \phi(r; c), & d \in D_i \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

The weights of this replicated feature space are initialized using the weights w tuned for the baseline $\phi(r; c)$.

$$w_d = \begin{cases} w, & d \text{ is ROOT} \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

In this way, the ROOT domain corresponds to the unadapted baseline weights, denoted as Θ_* in (Finkel and Manning, 2009). The idea is that we simultaneously maintain a generic set of weights that applies to all domains as well as their domain-specific "offsets", describing how a domain differs from the generic case. Model updates during adaptation are performed according to the same procedure as tuning updates, but now in the replicated space.

Different from (Finkel and Manning, 2009), this generalized FEDA model does not restrict the domains to be strictly hierarchically structured. We could, for example, include a domain for each translator that crossed different genres. However, all of our experimental evaluations maintain a hierarchical domain structure, leaving more general setups to future work.

Rules and Feature Values. A derivation r of sentence f_i has features that are computed from the combination of the baseline training corpus c_0 and a genre-specific corpus that includes all sentence pairs from the tuning corpus as well as from the adaptation corpus (f_j, e_j^*) with $j < i$ sharing f_i's genre. We refer to this combined corpus as c_i. The tuning corpus is the same that is used for parameter tuning in the baseline system. The adaptation corpus is our test set. Note that in our evaluation, each sentence is translated before it is used for adaptation, so that there is no contamination of results.

In order to extend the model efficiently within a streaming data environment, we make use of a suffix-array implementation for our phrase table (Levenberg et al., 2010).

Rather than combining corpus counts across these different sources, separate rules extracted from the baseline corpus and the genre-specific corpus exist independently in the derivation space, and features of each are computed only with one corpus. In this configuration, a large amount of out-of-domain evidence from the baseline model will not dampen the feature value adaptation effects of adding new sentence pairs from the adaptation corpus. The genre-specific phrases are distinguished by an additional binary provenance feature.

In order to extract features from the genre-specific corpus, a word-level alignment must be computed for each (f_j, e_j^*). We force decode using the adapted translation model for f_j. In order to avoid decoding failures, we insert high-cost single-word translation rules that allow any word in f_j to align to any word in e_j^*.

Sparse Features. Applying a large number of sparse features would compromise responsiveness of our translation system and is thus a poor fit for real-time adaptive computer-assisted translation. However, features that can be learned on a single document are limited in number and can be discarded after the document has been processed. Therefore, document-level sparse features are a powerful means to fit our model to local context with a comparatively small impact on efficiency.

4 Experiments

We performed two sets of German→English experiments; Table 1 contains the results for both. Our first set of experiments was performed on the *PatTR* corpus (Wäschle and Riezler, 2012). We divided the corpus into training and development data by date and selected 2.4M parallel segments dated before 2000 from the "claims" section as bilingual training data, taking equal parts from each of the eight patent types A–H as classified by the Cooperative Patent Classification (CPC). From each type we further drew separate test sets and a single tune set, selecting documents with at least 10 segments and a maximum of 150 source words per segment, with around 2,100 sentences per test set and 400 sentences per type for the tune set. The "claims" section of this corpus is highly repetitive, which makes it ideal for observing the effects of incremental adaptation techniques.

To train the language and translation model we additionally leveraged all available bilingual and monolingual data provided for the *EMNLP 2015 Tenth Workshop on Machine Translation*[3]. The total size of the bitext used for rule extraction and feature estimation was 6.4M sentence pairs. We trained a standard 5-gram language model with modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998) using the *KenLM* toolkit (Heafield et al., 2013) on 4 billion running words. The bitext was word-aligned with *mgiza* (Och and Ney, 2003), and we used the *phrasal* decoder (Green et al., 2014a) with standard German-English settings for experimentation.

Our second set of experiments was performed on a mixed-genre corpus containing lectures, patents, and news articles. The standard dev and test sets of the IWSLT 2014 shared task[4] were used for the *lecture* genre. Each document corresponded to an entire lecture. For the *news* genre, we used *newstest2012* for tuning, *newstest2013* for meta-parameter optimization, and *newstest2014* for testing. The tune set for the *patent* genre is identical to the first set of experiments, while the test set consists of the first 300 sentence pairs of each of the patent type specific test sets of the previous experiment. The documents in the news and patent genres contain around 20 segments on average.

Our evaluation proceeded in multiple stages. We first trained a set of background weights on the concatenated tune sets (*baseline*). Keeping these weights fixed, we performed an additional tuning run to estimate genre-level weights (+ *genre weights*).[5] In the patent-only setup, we used patent CPC type as genre. Next, we trained a genre-specific translation model for each genre by first feeding the tune set and then the test set into our incremental adaptation learning method as a continuous stream of simulated post edits (+ *genre TM*). After each sentence, we performed an update on the genre-specific weights. In separate experiments, we also included document-level weights as an additional domain (+ *doc. weights*) and included sparse features at the document level (+ *sparse features*).[6]

	PatTR	**heterogeneous data**		
	avg	lecture	news	patent
repetition rate	*27.80*	*5.46*	*3.13*	*27.42*
baseline	48.89	25.82	24.92	48.97
+ genre weights	49.05	26.64	25.12	49.39
+ genre TM	53.25	27.67	25.66	53.22
+ doc. weights	53.56	27.98	25.71	53.40
+ sparse features	54.53	28.09	25.89	54.30

Table 1: Results in uncased BLEU [%]. Each component is added on top of the previous line. All results in line + *genre TM* and below are statistically significant improvements over the baseline with 95% confidence. We also report the repetition rate of the test corpora as propsed by Bertoldi et al. (2013).

Table 1 demonstrates that each component of this approach offered consistent incremental quality gains, but with varying magnitudes. For the patent experiments we report the average over our eight test sets (A-H) due to lack of space, but total improvement varied from +4.92 to +6.46 BLEU. In the mixed-genre experiments, BLEU increased by +2.27 on *lectures*, +0.97 on *news*, and +5.33 on *patents*. On all tasks, we observed statistically significant improvements over the baseline (95% confidence level) in the + *genre TM*, + *doc. weights* and + *sparse features* experiments using bootstrap resampling (Koehn, 2004).

These results demonstrate the efficacy of hierarchical incremental adaptation, although we would like to stress that the patent data was selected specifically for its high level of repetitiveness, and the

[3] http://www.statmt.org/wmt15/

[4] http://workshop2014.iwslt.org/

[5] Learning rates and regularization weights for this step were selected on *newstest2013*.

[6] Learning rates and regularization weights for each genre were selected on the genre-specific tune sets.

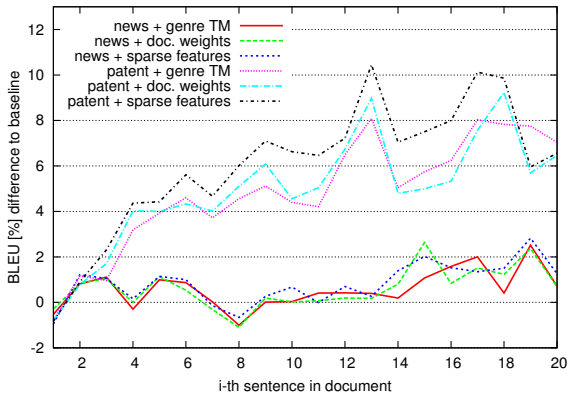

Figure 2: BLEU difference between *baseline + genre weights* and our incremental adaptation approach, computed on a single segment from each document according to their order, i.e. the first segment from each document, then the second segment from each document, etc.

large improvement in this genre would only be expected to arise in similarly structured domains. This property is quantified by the *repetition rate* measure (RR) (Bertoldi et al., 2013) reported in Table 1, which confirms the finding by Cettolo et al. (2014) that RR correlates with the effectiveness of adaptation.

Analysis. Figure 2 shows BLEU score differences to the *baseline + genre weights* system for different subsets of the news and patent test sets. Each point is computed by document slicing, i.e. on a single segment from each document. The rightmost data point is the BLEU score we obtain by evaluating on the 20th segment of each document, grouped into a pseudo-corpus. Note that this group does not correspond to any number in Table 1, which reports BLEU on the entire test sets. Thus, we evaluate on all sentences that have learned from exactly $(i-1)$ segments of the same document, with $i = 1, \ldots, 19$. Although the graph is naturally very noisy (each score is computed on roughly 150 segments), we can clearly see that incremental adaptation learns on the document level: on average, the improvement over the baseline increases when proceeding further into the document.

Decoding speed. In our real-time computer-assisted translation scenario, a certain translation speed is required to allow for responsive user interaction. Table 2 reports the speed in words per second on the lecture data. Adding a genre-specific translation model results in a speed reduction by a factor of 12.6 due to the additional (forced) decoding run and weight updates. Sparse features slows the system down further by a factor of 2.4. However, the largest part of the computation time incurs only when the user has finalized collaborative translation of one sentence and is busy reading the next source sentence. Further, the speed/quality tradeoff can be adjusted with pruning parameters.

	words / sec
baseline	177.6
+ genre weights	58.5
+ genre TM	14.1
+ doc. weights	9.8
+ sparse features	5.8

Table 2: Decoding speed on the lecture data.

5 Conclusion

We have presented an incremental learning approach for MT that maintains a flexible hierarchical domain structure within a single consistent model. In our experiments, we define a three-level hierarchy with a global root domain as well as genre- and document-level domains. Further, we perform incremental adaptation by training a genre-specific translation model on the stream of incoming post-edits and adding document-level sparse features that do not significantly compromise efficiency. Our results show consistent contributions from each level of adaptation across multiple genres.

References

Pratyush Banerjee, Jinhua Du, Baoli Li, Sudip Kr. Naskar, Andy Way, and Josef van Genabith. 2010. Combining multi-domain statistical machine translation models using automatic classifiers. In *Proceedings of the Association for Machine Translation in the Americas*, Denver, Colorado, October.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *Proceedings of the XIV Machine Translation Summit*, pages 36–42, Nice, France, September.

Alexandru Ceauşfu, John Tinsley, Jian Zhang, and Andy Way. 2011. Experiments on domain adaptation for patent machine translation in the PLuTO project. In Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste, editors, *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 21–28, Leuven, Belgium.

Mauro Cettolo, Nicola Bertoldi, and Marcello Federico. 2014. The repetition rate of text as a predictor of the effectiveness of machine translation adaptation. In *Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 166–179, Vancouver, Canada, October.

Stanley F. Chen and Joshuo Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, MA, August.

Lei Cui, Xilun Chen, Dongdong Zhang, Shujie Liu, Mu Li, and Ming Zhou. 2013. Multi-domain adaptation for SMT using multi-task learning. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1055–1065, Seattle, Washington, USA, October.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June.

Michael Denkowski, Chris Dyer, and Alon Lavie. 2014. Learning from post-editing: Online model adaptation for statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 395–404, Gothenburg, Sweden, April.

John Duchi and Yoram Singer. 2009. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, December.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, July.

Kevin Duh, Katsuhito Sudoh, Hajime Tsukada, Hideki Isozaki, and Masaaki Nagata. 2010. N-best reranking by multitask learning. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 375–383, Uppsala, Sweden, July.

Jenny Rose Finkel and Christopher D. Manning. 2009. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610, Boulder, Colorado.

Spence Green, Sida Wang, Daniel Cer, and Christopher D. Manning. 2013a. The efficacy of human post-editing for language translation. In *ACM CHI Conference on Human Factors in Computing Systems*, Paris, France, April.

Spence Green, Sida Wang, Daniel Cer, and Christopher D. Manning. 2013b. Fast and adaptive online training of feature-rich translation models. In *51st Annual Meeting of the Association for Computational Linguistics*, pages 311–321, Sofia, Bulgaria, August.

Spence Green, Daniel Cer, , and Christopher D. Manning. 2014a. Phrasal: A Toolkit for New Directions in Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 114–121, Baltimore, Maryland USA, June.

Spence Green, Daniel Cer, and Christopher D. Manning. 2014b. An empirical comparison of features and tuning for phrase-based machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 466–476, Baltimore, Maryland, USA, June.

A. Guerberof. 2009. Productivity and quality in the post-editing of outputs from translation memories and machine translation. *International Journal of Localization*, 7(1):11–21.

Daniel Hardt and Jakob Elming. 2010. Incremental re-training for post-editing SMT. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado, October.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

Reinerd Kneser and Hermann Ney. 1995. Improved backing-off for M-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, May.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. of the Human Language Technology Conf. (HLT-NAACL)*, pages 127–133, Edmonton, Canada, May/June.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantine, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. pages 177–180, Prague, Czech Republic, June.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 388–395, Barcelona, Spain, July.

Abby Levenberg, Chris Callison-Burch, and Miles Osborne. 2010. Stream-based translation models for statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the*

North American Chapter of the Association for Computational Linguistics, pages 394–402, Los Angeles, California, June.

Percy Liang, Alexandre Buchard-Côté, Dan Klein, and Ben Taskar. 2006. An End-to-End Discriminative Approach to Machine Translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 761–768, Sydney, Australia.

Prashant Mathur, Cettolo Mauro, and Marcello Federico. 2013. Online learning approaches in computer assisted translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 301–308, Sofia, Bulgaria, August.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

F. J. Och and H. Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics,*, 30(4):417–450.

Daniel Ortiz-Martínez, Ismael García-Varea, and Francisco Casacuberta. 2010. Online learning for interactive statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 546–554, Los Angeles, California, June.

M. Plitt and F. Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Rico Sennrich, Holger Schwenk, and Walid Aransa. 2013. A multi-domain translation model framework for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 832–840, Sofia, Bulgaria, August.

Patrick Simianer and Stefan Riezler. 2013. Multi-task learning for improved discriminative training in SMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 292–300, Sofia, Bulgaria, August.

Patrick Simianer, Stefan Riezler, and Chris Dyer. 2012. Joint Feature Selection in Distributed Stochastic Learning for Large-Scale Discriminative Training in SMT. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 11–21, Jeju Island, Korea, July.

Linfeng Song, Haitao Mi, Yajuan Lü, and Qun Liu. 2011. Bagging-based system combination for domain adaption. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, pages 293–299, Xiamen, China.

Wei Wang, Klaus Macherey, Wolfgang Macherey, Franz Och, and Peng Xu. 2012. Improved domain adaptation for statistical machine translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, California.

Katharina Wäschle and Stefan Riezler. 2012. Analyzing Parallelism and Domain Similarities in the MAREC Patent Corpus. *Multidisciplinary Information Retrieval*, pages 12–27.

Katharina Wäschle, Patrick Simianer, Nicola Bertoldi, Stefan Riezler, and Marcello Federico. 2013. Generative and Discriminative Methods for Online Adaptation in SMT. In *Proceedings of Machine Translation Summit XIV*, Nice, France.

Jia Xu, Yonggang Deng, Yuqing Gao, and Hermann Ney. 2007. Domain dependent statistical machine translation. In *Proceedings of the MT Summit*, pages 515–520, Copenhagen, Denmark, September.