PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 96 OCTOBER 2011 99–108

Multi-Task Minimum Error Rate Training for SMT

Patrick Simianer, Katharina Wäschle, Stefan Riezler

Department of Computational Linguistics, University of Heidelberg, Germany

Abstract

We present experiments on multi-task learning for discriminative training in statistical machine translation (SMT), extending standard minimum-error-rate training (MERT) by techniques that take advantage of the similarity of related tasks. We apply our techniques to German-to-English translation of patents from 8 tasks according to the International Patent Classification (IPC) system. Our experiments show statistically significant gains over task-specific training by techniques that model commonalities through shared parameters. However, more fine-grained combinations of shared parameters with task-specific ones could not be brought to bear on models with a small number of dense features. The software used in the experiments is released as open-source tool.

1. Introduction

Multi-task learning aims at learning several different tasks simultaneously, addressing commonalities through shared parameters and modeling differences through task-specific parameters. This learning framework is advantageous if the tasks are not completely independent of each other, which would advocate to train a separate model for each task. Instead, they should be related and share some commonalities, yet be different enough to counter a simple pooling of training data.

A predestined application for multi-task learning in the area of statistical machine translation (SMT) is patent translation over several different classes of patents according to the International Patent Classification (IPC)[1]. Table 1 shows the eight top level

[1] http://www.wipo.int/classifications/ipc/en/

Corresponding author: riezler@cl.uni-heidelberg.de
Cite as: Patrick Simianer, Katharina Wäschle, Stefan Riezler. Multi-Task Minimum Error Rate Training for SMT. The Prague Bulletin of Mathematical Linguistics No. 96, 2011, pp. 99–108.
doi: 10.2478/v10108-011-0015-0.

sections of the IPC categorization. They aim at a distinction of main technological fields, each of which is characterized by its own technological terminology.

A Human Necessities
B Performing Operations; Transporting
C Chemistry; Metallurgy
D Textiles; Paper
E Fixed Constructions
F Mechanical Engineering; Lighting; Heating; Weapons; Blasting
G Physics
H Electricity

Table 1. Patent sections according to the IPC classification.

On the other hand, patents exhibit strong commonalities across IPC sections in sharing a highly specialized vocabulary, consisting of a legal jargon not found in everyday language, and a rigid textual structure including highly formulaic language. The goal of multi-task learning for SMT is thus to learn a translation system that performs well across several different patent sections, thus benefits from shared information, and yet is able to address the specifics of each patent section.

The machine learning community has developed several different formalizations of the central idea of trading off optimality of parameter vectors for each task-specific model and closeness of these model parameters to the average parameter vector across models. For example, Evgeniou and Pontil (2004) develop this idea in the framework of support vector machines (SVM) as finding a tradeoff between each task-specific SVM having a large margin and having each SVM close to the average SVM. They formalize this tradeoff via regularization of the task-specific parameter vectors and of the distance to the average parameter vector. The starting point of all this and related algorithms is a linear classifier (or a non-linear kernelized variant) with a fixed feature vector (or kernel) whose associated parameters are adjusted in multi-task learning.

(Multi-)domain adaptation[2] for SMT has so far been seen as a challenge of out-of-vocabulary (OOV) terms. Adaptation techniques thus have focused on gathering OOV information from various sources in order to feed the standard generative SMT pipeline of translation and language model with it. A recent approach is Daumé and Jagarlamudi (2011) who mine translations for OOV terms from comparable corpora.

Patent translation exhibits an even more severe OOV problem because of very specialized terminology in different IPC patent sections. Multi-task learning or domain

[2]We consider domain adaptation as a special case of multi-task learning for two tasks, and multi-domain adaptation as equivalent to multi-task learning. Other definitions are possible (see, e.g., Dredze et al. (2010)).

adaptation efforts in patent translation have so far been restricted to experimental combinations of translation and language models from different sets of IPC sections (Utiyama and Isahara, 2007; Tinsley et al., 2010; Ceauşu et al., 2011).

In this paper, we consider the specific setting in which the generative SMT pipeline is not adaptable. Such situations arise if there are not enough parallel data to train generative models on the new tasks. However, we assume that there are enough parallel data available to perform discriminative training (Och, 2003) for each specific task. Our goal is to investigate how state-of-the-art multi-task learning techniques for linear classifiers can be applied to standard discriminative training for SMT. In other words, we would like to know how much gain there is in extending the standard tuning technique of minimum error rate training (MERT) to multi-task MERT for SMT. To this aim, we present a generic new algorithm to model commonalities by regularized parameter averaging, building upon Evgeniou and Pontil (2004), and apply it to multi-task MERT for SMT . Furthermore, we present a distributed implementation of MERT for multiple tasks that allows us to apply techniques for parameter averaging from distributed learning (Zinkevich et al., 2010) to a version of averaged MERT. Our experimental results show that averaged and multi-task MERT achieve statistically significant gains over training separate task-specific models. However, multi-task MERT's fine-grained combination of shared parameters with task-specific ones did not improve upon parameter averaging in our experiments on models with a small number of dense features.

2. Related Work

A central idea to learn common behaviors across related task is to learn task-specific models and to minimize their deviation from an average model. Starting from a separate SVM for each task, Evgeniou and Pontil (2004) present a regularization method that trades off optimization of the task-specific parameter vectors and the distance of each SVM to the average SVM. Equivalent formalizations replace parameter regularization by Bayesian prior distributions on the parameters (Finkel and Manning, 2009) or by augmentation of the feature space with domain independent features (Daumé, 2007). Besides SVMs, several learning algorithms have been extended to the multi-task scenario in a parameter regularization setting, e.g., perceptron-type algorithms (Dredze et al., 2010) or boosting (Chapelle et al., 2011). Further variants include different formalizations of norms for parameter regularization, e.g., $\ell_{1,2}$ regularization (Obozinski et al., 2010) or $\ell_{1,\infty}$ regularization (Quattoni et al., 2009), where only the features that are most important across all tasks are kept in the model.

While the standard machine learning approaches to multi-task learning are based on linear classifiers (or non-linear kernelized versions), SMT approaches to multi-task learning have concentrated on adapting unsupervised generative modules such as translation models or language models to new tasks. For example, transductive approaches have used automatic translations of monolingual corpora for self-training

modules of the generative SMT pipeline (Ueffing et al., 2007; Schwenk, 2008; Bertoldi and Federico, 2009). Other approaches have extracted parallel data from similar or comparable corpora (Zhao et al., 2004; Snover et al., 2008). Several approaches have been presented to train separate translation and language models on task-specific subsets of the data and combine them in different mixture models (Foster and Kuhn, 2007; Koehn and Schroeder, 2007).

Multi-task learning efforts in patent translation have so far been restricted to experimental combinations of translation and language models from different sets of IPC sections. For example, Utiyama and Isahara (2007) and Tinsley et al. (2010) investigate translation and language models trained on different sets of patent sections, with larger pools of parallel data improving results. Ceauşu et al. (2011) find that language models always and translation model mostly benefit from larger pools of data from different sections.[3]

3. Parallel Data from Patent Classes for Patent Translation

Our work on patent translation is based on the MAREC[4] patent data corpus. MAREC contains over 19 million patent applications and granted patents in a standardized format from four patent organizations (European Patent Office (EP), World Intellectual Property Organisation (WO), United States Patent and Trademark Office (US), Japan Patent Office (JP)), from 1976 to 2008.

Patent text is organized in 4 document sections, the patent title, abstract, description and claims. The patent title is usually a short noun phrase. The abstract contains a short summary of the invention. The description is a detailed explanation of the patent. The claims are a list of sentences that define the scope of protection granted by the patent with a standardized sentence structure. MAREC contains comparable text sections, mainly in English, French, and German. Patent titles are automatically parallel, since they only consist of one sentence and there is one title per document. Text in abstracts and claims must be split into sentences and aligned. There are no parallel descriptions.

For our experiments, we extracted bilingual abstract and claims sections from the EP and WO parts for German-to-English translation. The distribution over the sections mirrors the overall distribution of IPC sections in the corpus (see Table 2). For sentence splitting and tokenizing we used the Europarl tools[5]. Sentence alignment was done with Gargantua 1.0b[6]. The training data for the de-en language pair contains 1,000,000 sentences extracted from all top-level IPC sections from abstracts (5%)

[3] Ceauşu et al. (2011) report that including data from IPC section C (chemistry) in pooled training data is detrimental for translation models.

[4] http://www.ir-facility.org/prototypes/marec

[5] http://www.statmt.org/europarl/

[6] http://sourceforge.net/projects/gargantua/

A	266,521	21.81%
B	384,517	31.47%
C	372,903	30.52%
D	50,579	4.14%

E	54,396	4.45%
F	149,370	12.22%
G	291,671	23.87%
H	228,147	18.67%

Table 2. Distribution of IPC sections for comparable de-en abstracts and claims.

and claims (95%) from 1993 to 1995. Furthermore, we extracted for each top-level IPC section 2,000 randomly sampled sentences from abstracts (5%) and claims (95%) from 2007 (development) and 2008 (development-testing and final-testing). Table 3 gives an overview over the processed data.

	train	dev	devtest	test
# parallel sents	1M	2K	2K	2K
avg. # tokens de	32,329,745	59,376	60,061	59,930
avg. # tokens en	36,005,763	69,584	70,700	70,331
year	1993-1995	2007	2008	2008

Table 3. Statistics on parallel de-en data extracted from MAREC patent corpus.

4. Distributed Multi-Task Parameter Regularization

Multi-task learning assumes learning tasks or domains $d = 1, \ldots, D$, each coming with a separate sample of $n(d)$ training points from the same space. Evgeniou and Pontil (2004)'s idea of trading off optimal parameter weights for each task-specific model and closeness to an average parameter vector can be stated in a more general form as follows. We aim at minimization of task-specific loss functions l_d under a regularization of task-specific parameter vectors w_d towards an average parameter vector w_{avg}.

$$\min_{w_1,\ldots,w_D} \sum_{d=1}^{D} l_d(w_d) + \lambda \sum_{d=1}^{D} |w_d - w_{avg}|_p^p \quad (1)$$

For prediction, one can use task-specific weight vectors $w_d \in \{w_1, \ldots, w_D\}$ that have been adjusted to trade off task-specificity (small λ) and commonality (large λ), or the average weight vector w_{avg} as a global model.

An average or global model can be estimated directly by applying ideas from distributed learning (Zinkevich et al., 2010). The idea is to base the distribution strategy on task-specific partitions of data. An algorithm for distributed average learning will

take a loss function $c_d(w_d)$ for data and weights specific to task d, parameter initializations $w^{(0)}$, and return an averaged weight vector w_{avg}, for D tasks. An instantiation of such an algorithm to our problem, called `AvgMERT`, calls one iteration of a MERT implementation, denoted by `MERT`, that continues from parameter vector $w_d^{(t-1)}$ and optimizes translation loss $c_d(w_d)$ on the data from task d.

```
AvgMERT(w^(0), D, {c_d}_{d=1}^D):
for d = 1,...,D parallel do
  for t = 1,...,T do
    w_d^(t) = MERT(w_d^(t-1), c_d(w_d))
  end for
end for
return w_avg = (1/D) Σ_{d=1}^D w_d^(T)
```

For multi-task learning, we set p=1 to obtain an ℓ_1 regularizer, and apply the penalty term λ to the parameter weights the extent that they do not cross the average weights. That is, the weight vector w_d is moved towards the average weight vector w_{avg} by adding or subtracting the penalty λ for each weight component $w_d[k]$, and clipped when it crosses the average. This strategy can be motivated in a stochastic gradient descent framework (Tsuruoka et al., 2009), however, we apply it to regularized loss minimization in general, and to regularized MERT in specific. As stopping criterion we used a threshold on the maximal change in the average parameter vector.

```
MMERT(w^(0), D, {c_d}_{d=1}^D):
for t = 1,...,T do
  w_avg^(t) = (1/D) Σ_{d=1}^D w_d^(t-1)
  for d = 1,...,D parallel do
    w_d^(t) = MERT(w_d^(t-1), c_d(w_d))
    for k = 1,...,K do
      if w[k]_d^(t) − w_avg^(t)[k] > 0 then
        w_d^(t)[k] = max(w_avg^(t)[k], w_d^(t)[k] − λ)
      else if w_d^(t)[k] − w_avg^(t)[k] < 0 then
        w_d^(t)[k] = min(w_avg^(t)[k], w_d^(t)[k] + λ)
      end if
    end for
  end for
end for
return w_1^(T),...,w_D^(T), w_avg^(T)
```

The code described in this section is written as script wrapper around the MERT implementation of Bertoldi et al. (2009). The code is licensed unter the LGPL and can be found online[7].

5. Experiments

For training a German-to-English baseline model on the 1 million parallel patent data described in Section 3, we used the open-source Moses[8] SMT system. Parallel sentences were filtered to sentences of at most 80 tokens. For development, development-testing and final-testing data we additionally ensured that the random sample contained no duplicates.

BLEU scores on test set are shown on Table 4. Columns show an evaluation on test sets consisting of 2,000 parallel sentences from each of IPC sections A-H. All systems use the same phrase-table and language model trained on 1 million parallel data from all IPC sections. Different rows show results for systems that differ only in the approaches to discriminative optimization of the BLEU metric (Papineni et al., 2001). All models use the MERT implementation of (Bertoldi et al., 2009) for the 14 standard features of the Moses system. Best results are indicated by **bold face** type.

The baseline systems perform individual tuning for each IPC section, and tuning on a development set pooled from all sections. All MERT runs start from default hand-tuned weight vectors for each model. The first column (*ind.*) shows results for a system that is tuned on each individual IPC section separately, i.e., each system is tuned on a development set of 2,000 sentences from section X and evaluated on a test set of 2,000 sentences from the same section X. BLEU scores for this baseline system are already quite high, due to the repetitive nature of patents where many long and specific sub-sentential expressions are reused. The second column(*pooled*) shows a system that is tuned on a development set consisting of 2,000 sentences pooled from 250 sentences from each patent section. Result differences to *ind.* are not statistically significant.[9]

The distributed average learner AvgMERT produces some small, but statistically significant improvements over *ind.* (indicated by $*$) and *pooled* (indicated by $+$). The multi-task learner MMERT and the global model w_{avg} produced as by-product in multi-task learning show some improvements over *ind.* and AvgMERT (indicated by #). Meta-parameters for multi-task learning were set to a regularization parameter of λ=0.0001 and a convergence threshold of 0.001, resulting in convergence after 13 MERT iterations. The average weight vector w_{avg} was initialized to the zero vector.

[7] http://www.cl.uni-heidelberg.de/statnlpgroup/mmert/

[8] http://www.statmt.org/moses/

[9] Statistical significance of pairwise result differences is assessed by p-values smaller than 0.05 using an Approximate Randomization test (Riezler and Maxwell, 2005).

section	*ind.*	*pooled*	AvgMERT	MMERT	w_{avg}
A	0.5187	0.5199	**0.5213***	0.5195$^{\#}$	0.5196$^{\#}$
B	0.4877	0.4885	0.4908^{*+}	0.4911*	**0.4921**$^{*\#}$
C	0.5214	0.5175	0.5199^{*+}	**0.5218**$^{\#}$	0.5162$^{*\#}$
D	0.4724	0.4730	0.4733	**0.4736**	0.4734
E	0.4666	0.4661	0.4679^{*+}	0.4669	**0.4685***
F	0.4794	0.4801	0.4811*	0.4821*	**0.4830**$^{*\#}$
G	0.4596	0.4576	0.4607^{+}	0.4606	**0.4610***
H	0.4573	0.4560	0.4578	**0.4581**	**0.4581**

Table 4. BLEU scores on 2K parallel sentences for each of 8 patent sections.

6. Discussion

An interpretation of the results presented in Section 5 can be given as follows. The distributed average learner AvgMERT shows small, but statistically significant improvements over individual tuning for most IPC sections. This is consistent with theoretical and empirical results on distributed weight averaging for linear models (see, e.g., Zinkevich et al. (2010)). The evaluation on section C ("chemistry") shows a significant degradation. This confirms the intuition that averaging parameter weights over sections with commonalities is helpful, but not so for exceptional domains containing complex chemical formulae and compound names. Furthermore, this result is consistent with Ceauşu et al. (2011) who find that section C is best omitted when extracting a phrase table pooled across sections.

Similar results are found for the global model w_{avg} produced as by-product of multi-task learning. The multi-task learner MMERT is the only system that is able to improve results for section C over the results of individual tuning.

Clearly, all presented results have to be interpreted with a grain of salt because of the small, even if statistically significant, result differences. We conjecture that this is due to the small number of features deployed in MERT training, and can be overcome by moving to discriminative training with millions of sparse, lexicalized features. We believe that especially the fine-tuning between task-specific and average feature weights addressed by multi-task learning can be brought to bear on large-scale lexicalized models. This is due to future work.

Bibliography

Bertoldi, Nicola and Marcello Federico. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, Athens, Greece, 2009.

Bertoldi, Nicola, Barry Haddow, and Jean-Baptiste Fouet. Improved minimum error rate training in moses. *The Prague Bulletin of Mathematical Linguistics*, (91):7–16, 2009.

Ceauşu, Alexandru, John Tinsley, Jian Zhang, and Andy Way. Experiments on domain adaptation for patent machine translation in the PLuTO project. In *Proceedings of the 15th Conference of the European Assocation for Machine Translation (EAMT 2011)*, Leuven, Belgium, 2011.

Chapelle, Olivier, Pannagadatta Shivaswamy, Srinivas Vadrevu, Kilian Weinberger, Ya Zhang, and Belle Tseng. Boosted multi-task learning. *Machine Learning*, 2011.

Daumé, Hal. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic, 2007.

Daumé, Hal and Jagadeesh Jagarlamudi. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, Portland, OR, 2011.

Dredze, Mark, Alex Kulesza, and Koby Crammer. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79:123–149, 2010.

Evgeniou, Theodoros and Massimiliano Pontil. Regularized multi-task learning. In *Proceedings of the 10th ACM SIGKDD conference on knowledge discovery and data mining (KDD'04)*, Seattle, WA, 2004.

Finkel, Jenny Rose and Christopher D. Manning. Hierarchical bayesian domain adaptation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT'09)*, Boulder, CO, 2009.

Foster, George and Roland Kuhn. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, 2007.

Koehn, Philipp and Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, 2007.

Obozinski, Guillaume, Ben Taskar, and Michael I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20:231–252, 2010.

Och, Franz Josef. Minimum error rate training in statistical machine translation. In *Proceedings of the Human Language Technology Conference and the 3rd Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'03)*, Edmonton, Cananda, 2003.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. Technical Report IBM Research Division Technical Report, RC22176 (W0190-022), Yorktown Heights, N.Y., 2001.

Quattoni, Ariadna, Xavier Carreras, Michael Collins, and Trevor Darrell. An efficient projection for $l_{1,\infty}$ regularization. In *Proceedings of the 26th International Conference on Machine Learning (ICML'09)*, Montreal, Canada, 2009.

Riezler, Stefan and John Maxwell. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL-05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, MI, 2005.

Schwenk, Holger. Investigations on large-scale lightly-supervised training for statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT'08)*, Hawaii, 2008.

Snover, Matthew, Bonnie Dorr, and Richard Schwartz. Language and translation model adaptation using comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, Honolulu, Hawaii, 2008.

Tinsley, John, Andy Way, and Paraic Sheridan. PLuTO: MT for online patent translation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, Denver, CO, 2010.

Tsuruoka, Yoshimasa, Jun'ichi Tsujii, and Sophia Ananiadou. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP'09)*, Singapore, 2009.

Ueffing, Nicola, Gholamreza Haffari, and Anoop Sarkar. Transductive learning for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, Prague, Czech Republic, 2007.

Utiyama, Masao and Hitoshi Isahara. A japanese-english patent parallel corpus. In *Proceedings of MT Summit XI*, Copenhagen, Denmark, 2007.

Zhao, Bing, Matthias Eck, and Stephan Vogel. Language model adaptation for statistical machine translation with structured query models. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, Geneva, Switzerland, 2004.

Zinkevich, Martin A., Markus Weimer, Alex Smola, and Lihong Li. Parallelized stochastic gradient descent. In *Proceedings of the 24th Annual Conference on Neural Information Processing Sytems (NIPS'10)*, Vancouver, Canada, 2010.

Address for correspondence:
Stefan Riezler
riezler@cl.uni-heidelberg.de
Department of Computational Linguistics, University of Heidelberg,
Im Neuenheimer Feld 325, 69120 Heidelberg, Germany