Multi-Task MERT

Simianer, Wäschle, Riezler

Multi-Task Minimum Error Rate Training for SMT

Patrick Simianer, Katharina Wäschle, Stefan Riezler

Department of Computational Linguistics University of Heidelberg, Germany

Multi-Task MERT

Simianer, Wäschle, Riezler

 Multi-task learning aims at learning several different tasks simultaneously,

Multi-Task MERT

- Multi-task learning aims at learning several different tasks simultaneously,
 - addressing commonalities through shared parameters

Multi-Task

- Multi-task learning aims at learning several different tasks simultaneously,
 - addressing commonalities through shared parameters
 - and modeling differences through task-specific parameters.

Multi-Task

- Multi-task learning aims at learning several different tasks simultaneously,
 - addressing commonalities through shared parameters
 - and modeling differences through task-specific parameters.
- Predestined application: Patent translation over classes of patents w.r.t. International Patent Classification (IPC)

Multi-Task

- Multi-task learning aims at learning several different tasks simultaneously,
 - addressing commonalities through shared parameters
 - and modeling differences through task-specific parameters.
- Predestined application: Patent translation over classes of patents w.r.t. International Patent Classification (IPC)
 - commonalities: highly specialized legal jargon not found in everyday language, rigid textual structure including highly formulaic language.

Multi-Task

- Multi-task learning aims at learning several different tasks simultaneously,
 - addressing commonalities through shared parameters
 - and modeling differences through task-specific parameters.
- Predestined application: Patent translation over classes of patents w.r.t. International Patent Classification (IPC)
 - commonalities: highly specialized legal jargon not found in everyday language, rigid textual structure including highly formulaic language.
 - differences: technological terminology specific to IPC class.

IPC Sections

Multi-Task MERT

- 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

Goal and Approach

Multi-Task MERT

Simianer, Wäschle, Riezler

Goal: 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.

Goal and Approach

Multi-Task MERT

Simianer, Wäschle, Riezler

Goal: 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.

Approach: Machine learning approach to trading off optimality of parameter vectors for each task-specific model and closeness of these model parameters to average parameter vector across models.

Multi-Task Minimum Error Rate Training

Multi-Task MERT

Simianer, Wäschle, Riezler

> Assume specific setting: Not enough data for training generative SMT pipeline on all tasks, however, enough data for tuning for each specific task.

Multi-Task Minimum Error Rate Training

Multi-Task MERT

- Assume specific setting: Not enough data for training generative SMT pipeline on all tasks, however, enough data for tuning for each specific task.
- In other words: How much gain is there in extending the standard tuning technique of minimum error rate training (MERT) to multi-task MERT for SMT.

Multi-Task Minimum Error Rate Training

Multi-Task MERT

- Assume specific setting: Not enough data for training generative SMT pipeline on all tasks, however, enough data for tuning for each specific task.
- In other words: How much gain is there in extending the standard tuning technique of minimum error rate training (MERT) to multi-task MERT for SMT.
- Also apply techniques for parameter averaging from distributed learning to a version of averaged MERT.

Multi-Task MERT

Simianer, Wäschle, Riezler

 MAREC: 19 million patent applications and granted patents, 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.

¹http://www.statmt.org/europarl/

Multi-Task MERT

- MAREC: 19 million patent applications and granted patents, 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.
- Extract bilingual abstract and claims sections from the EP and WO parts for German-to-English translation.

¹http://www.statmt.org/europarl/

Multi-Task MERT

- MAREC: 19 million patent applications and granted patents, 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.
- Extract bilingual abstract and claims sections from the EP and WO parts for German-to-English translation.
- Sentence splitting and tokenizing with Europarl tools¹.

¹http://www.statmt.org/europarl/

Multi-Task MERT

- MAREC: 19 million patent applications and granted patents, 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.
- Extract bilingual abstract and claims sections from the EP and WO parts for German-to-English translation.
- Sentence splitting and tokenizing with Europarl tools¹.
- Sentence alignment with Gargantua 1.0b².

¹http://www.statmt.org/europarl/

Distribution of IPC sections for de-en abstracts and claims

Multi-Task

Α	266,521	21.81%
В	384,517	31.47%
C	372,903	30.52%
D	50,579	4.14%
Ε	54,396	4.45%
F	149,370	12.22%
G	291,671	23.87%
Н	228,147	18.67%

Parallel data for de-en patent translation

Multi-Task

	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

Multi-task learning objective

Multi-Task MFRT

Simianer, Wäschle, Riezler

Objective: Minimize task-specific loss functions I_d under regularization of task-specific parameter vectors w_d towards an average parameter vector $w_{\rm avg}$.

Multi-task learning objective

Multi-Task MERT

Simianer, Wäschle, Riezler

Objective: Minimize task-specific loss functions I_d under regularization of task-specific parameter vectors w_d towards an average parameter vector $w_{\rm avg}$.

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

Multi-task prediction

Multi-Task MERT

Simianer, Wäschle, Riezler

Prediction:

Task-specific weight vectors $w_d \in \{w_1, \dots, w_D\}$ that have been adjusted to trade off task-specificity (small λ) and commonality (large λ).

Multi-task prediction

Multi-Task MERT

Simianer, Wäschle, Riezler

Prediction:

Task-specific weight vectors $w_d \in \{w_1, \dots, w_D\}$ that have been adjusted to trade off task-specificity (small λ) and commonality (large λ).

or: Average weight vector w_{avg} as a global model.

Average MERT

Multi-Task MERT

Simianer Wäschle, Riezler

```
\begin{array}{l} \operatorname{AvgMERT}(w^{(0)},D,\{c_d\}_{d=1}^D) \colon \\ \text{for } d=1,\ldots,D \text{ parallel do} \\ \text{ for } t=1,\ldots,T \text{ do} \\ w_d^{(t)} = \operatorname{MERT}(w_d^{(t-1)},c_d(w_d)) \\ \text{ end for} \\ \text{end for} \\ \text{return } w_{\operatorname{avg}} = \frac{1}{D} \sum_{d=1}^D w_d^{(T)} \end{array}
```

 Apply ideas from distributed learning (Zinkevich et al. NIPS'10) by basing the distribution strategy on task-specific partitions of data.

Multi-Task MERT

Simianer, Wäschle, Riezler

regularization: Set $p{=}1$ in equation 1 to obtain an ℓ_1 regularizer.

Multi-Task MERT

Simianer, Wäschle, Riezler

regularization: Set p=1 in equation 1 to obtain an ℓ_1 regularizer.

clipping: Weight vector w_d is moved towards the average weight vector $w_{\rm avg}$ by adding or subtracting the penalty λ for each weight component $w_d[k]$, and clipped when it crosses the average.

Multi-Task MERT

Simianer, Wäschle, Riezler

regularization: Set p=1 in equation 1 to obtain an ℓ_1 regularizer.

clipping: 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.

code: Script wrapper around the MERT implementation of Bertoldi et al. 2009; licensed unter the LGPL; online at http://www.cl.uni-heidelberg.de/statnlpgroup/mmert/.

Multi-Task MERT

```
MMERT(w^{(0)}, D, \{c_d\}_{d=1}^D):
for t = 1, \ldots, T do
    w_{\text{avg}}^{(t)} = \frac{1}{D} \sum_{t=1}^{D} w_{t}^{(t-1)}
    for d = 1, \ldots, D parallel do
        w_d^{(t)} = MERT(w_d^{(t-1)}, c_d(w_d))
        for k = 1, \dots, K do
            if w[k]_{J}^{(t)} - w_{\text{avg}}^{(t)}[k] > 0 then
                w_d^{(t)}[k] = \max(w_{\text{avg}}^{(t)}[k], w_d^{(t)}[k] - \lambda)
            else if w_{J}^{(t)}[k] - w_{\text{avg}}^{(t)}[k] < 0 then
                w_{J}^{(t)}[k] = \min(w_{\text{avg}}^{(t)}[k], w_{J}^{(t)}[k] + \lambda)
            end if
        end for
    end for
end for
return w_1^{(T)}, ..., w_D^{(T)}, w_{\text{avg}}^{(T)}
```

Multi-Task MERT

Simianer, Wäschle, Riezler

Open-source Moses SMT system (Koehn et al. 2007);
 MERT implementation of Bertoldi et al. 2009.

Multi-Task MFRT

- Open-source Moses SMT system (Koehn et al. 2007);
 MERT implementation of Bertoldi et al. 2009.
- All systems use same phrase tables and language models, trained on 1M parallel data pooled from all IPC sections.

Multi-Task

- Open-source Moses SMT system (Koehn et al. 2007);
 MERT implementation of Bertoldi et al. 2009.
- All systems use same phrase tables and language models, trained on 1M parallel data pooled from all IPC sections.
- ind. systems are tuned on each IPC section separately.

Multi-Task MFRT

- Open-source Moses SMT system (Koehn et al. 2007);
 MERT implementation of Bertoldi et al. 2009.
- All systems use same phrase tables and language models, trained on 1M parallel data pooled from all IPC sections.
- ind. systems are tuned on each IPC section separately.
- pooled system is tuned on 2K sentences pooled from 250 sentences from each IPC section.

Multi-Task MERT

- Open-source Moses SMT system (Koehn et al. 2007);
 MERT implementation of Bertoldi et al. 2009.
- All systems use same phrase tables and language models, trained on 1M parallel data pooled from all IPC sections.
- ind. systems are tuned on each IPC section separately.
- pooled system is tuned on 2K sentences pooled from 250 sentences from each IPC section.
- AvgMERT and MMERT are algorithms described above.

Multi-Task MERT

- Open-source Moses SMT system (Koehn et al. 2007);
 MERT implementation of Bertoldi et al. 2009.
- All systems use same phrase tables and language models, trained on 1M parallel data pooled from all IPC sections.
- ind. systems are tuned on each IPC section separately.
- pooled system is tuned on 2K sentences pooled from 250 sentences from each IPC section.
- AvgMERT and MMERT are algorithms described above.
- ullet $w_{
 m avg}$ is global model produced as by-product in multi-task learning.

Multi-Task MERT

Simianer, Wäschle, Riezler

 All systems evaluated on 8 test sets, each consisting of 2K sentences from a separate IPC domain.

Multi-Task MERT

- All systems evaluated on 8 test sets, each consisting of 2K sentences from a separate IPC domain.
- Statistical significance of pairwise result differences assessed by p-values smaller than 0.05 using Approximate Randomization test (Riezler & Maxwell2005).

Multi-Task MERT

- All systems evaluated on 8 test sets, each consisting of 2K sentences from a separate IPC domain.
- Statistical significance of pairwise result differences assessed by p-values smaller than 0.05 using Approximate Randomization test (Riezler & Maxwell2005).
- statistically significant improvement over *ind*. indicated by

Multi-Task MERT

- All systems evaluated on 8 test sets, each consisting of 2K sentences from a separate IPC domain.
- Statistical significance of pairwise result differences assessed by p-values smaller than 0.05 using Approximate Randomization test (Riezler & Maxwell2005).
- statistically significant improvement over *ind*. indicated by
- \bullet statistically significant improvement over pooled indicated by +

Multi-Task MERT

- All systems evaluated on 8 test sets, each consisting of 2K sentences from a separate IPC domain.
- Statistical significance of pairwise result differences assessed by p-values smaller than 0.05 using Approximate Randomization test (Riezler & Maxwell2005).
- statistically significant improvement over *ind*. indicated by
- \bullet statistically significant improvement over pooled indicated by +
- statistically significant improvement over AvgMERT indicated by #

Experimental Results

Multi-Task MERT

section	ind.	pooled	AvgMERT	MMERT	$w_{ m avg}$
Α	0.5187	0.5199	0.5213*	$0.5195^{\#}$	$0.5196^{\#}$
В	0.4877	0.4885	0.4908*+	0.4911*	0.4921*#
С	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*
Н	0.4573	0.4560	0.4578	0.4581	0.4581

Multi-Task MERT

Simianer, Wäschle, Riezler

• pooled shows no s.s. improvement over ind.

Multi-Task MERT

- pooled shows no s.s. improvement over ind.
- Best results (**bold face**) achieved by AvgMERT, MMERT, or w_{avg} .

Multi-Task MERT

- pooled shows no s.s. improvement over ind.
- Best results (**bold face**) achieved by AvgMERT, MMERT, or w_{avg} .
- Best results are small, but statistically significant improvements over *ind*. and *pooled*.

Multi-Task MERT

- pooled shows no s.s. improvement over ind.
- Best results (**bold face**) achieved by AvgMERT, MMERT, or w_{avg} .
- Best results are small, but statistically significant improvements over ind. and pooled.
- Significant degradation on section C ("chemistry") by averaging techniques due to expeptional character of chemical formulae and compound names.

Multi-Task MERT

- pooled shows no s.s. improvement over ind.
- Best results (**bold face**) achieved by AvgMERT, MMERT, or w_{avg} .
- Best results are small, but statistically significant improvements over ind. and pooled.
- Significant degradation on section C ("chemistry") by averaging techniques due to expeptional character of chemical formulae and compound names.
- Interpretation of small improvements with a grain of salt, however, hope for larger improvements with larger feature sets.