Combining Domain and Topic Adaptation for SMT

Eva Hasler, Barry Haddow, Philipp Koehn

ILCC, School of Informatics
University of Edinburgh

October 25, 2014
Domain vs. Topic Adaptation

Cross-domain adaptation

- Small sample of parallel in-domain text is available
- Build translation models from different corpora
- Optimize mixture weights for texts from same domain
  [Foster and Kuhn, 2007, Sennrich, 2012]
  or learn corpus/instance weights
  [Matsoukas et al., 2009, Foster et al., 2010]
Domain vs. Topic Adaptation

**Cross-domain adaptation**

- Small sample of parallel in-domain text is available
- Build translation models from different corpora
- Optimize mixture weights for texts from same domain
  [Foster and Kuhn, 2007, Sennrich, 2012]
  or learn corpus/instance weights
  [Matsoukas et al., 2009, Foster et al., 2010]

**Dynamic domain adaptation**

- No domain information available ahead of time
- Adaptation based on current source text
  [Foster and Kuhn, 2007, Finch, 2008]
Domain vs. Topic Adaptation

*Topic* adaptation

- Learn **topical structure** of training data automatically
- Apply structural information to test data to infer topic mixture [Gong et al., 2010, Eidelman et al., 2012, Xiao et al., 2012, Hasler et al., 2014a]
- Few examples of non-dynamic adaptation [Su et al., 2012]
Domain vs. Topic Adaptation

**Topic** adaptation

- Learn **topical structure** of training data automatically
- Apply structural information to test data to infer topic mixture
  
  [Gong et al., 2010, Eidelman et al., 2012, Xiao et al., 2012, Hasler et al., 2014a]
- Few examples of non-dynamic adaptation [Su et al., 2012]

Advantages of dynamic topic adaptation

- No need for labelled domain boundaries
- No need for specific development set
Overview of the adaptation problem

Examples of wrong lexical choice

Input  le débit est en augmentation très rapide.
Reference these flows are increasing very rapidly.
MT output the speed is growing very rapidly.

le débit a augmenté.
the flows have increased.
the bitrate has increased.

Context

in the andes, this glacier is the source of drinking water for this city.
the flows have increased.
but when they go away, so does much of the drinking water.
Combining Domain and Topic Adaptation

Motivation

▶ Topic modelling useful for finding semantic structure in training data
▶ Domain labels of training documents/sentences available but not used

Questions

▶ Does it help to use both domain and topic information?
▶ Do they model different kinds of information, such as style vs. topic?
Combining Domain and Topic Adaptation

Approach: Building on previous work [Hasler et al., 2014b]

▶ Topic Adaptation with Distributional Profiles
▶ Extend with more features
▶ Adapt to each test document

Task: Prediction + Adaptation

▶ Old: Need to infer topic mixture of each test document
▶ New: Need to predict domain of test document
Phrase Pair Topic Model

How to learn semantic representations?

- Represent each phrase pair as distributional profile: pseudo document containing all context words
- Collect all source context words in local training contexts of a phrase pair
Phrase Pair Topic Model

How to learn semantic representations?

- Represent each phrase pair as *distributional profile*: pseudo document containing all context words
- Collect all *source context* words in local training contexts of a phrase pair
- Learn latent representation $\theta_p$ for each phrase pair
For each of $P$ phrase pairs $pp_i$ in the collection

1. Draw a topic distribution from an asymmetric Dirichlet prior,
   $\theta_p \sim \text{Dirichlet}(\alpha_0, \alpha \ldots \alpha)$.

2. For each position $c$ in the distributional profile of $pp_i$,
   draw a topic from that distribution,
   $z_{p,c} \sim \text{Multinomial}(\theta_p)$.

3. Conditioned on topic $z_{p,c}$,
   choose a context word
   $w_{p,c} \sim \text{Multinomial}(\psi_{z_{p,c}})$. 
Learned topic representations

- Some ambiguity remains: both *kernel* and *core* occur in *IT* contexts as translations of *noyau*
Phrase Pair Topic Model with additional features

**Conditional translation probability**

\[ p(t|s, \text{context}) = \sum_k p(t, k|s, \text{context}) \]

\[ p(t, k|s, \text{context}) \propto p(t, s, k|\text{context}) \]

\[ = p(t|s, k) \cdot p(s|k) \cdot p(k|\text{context}) \]

**Joint-conditional probability**

\[ p(t, \text{context}|s) = p(\text{context}|t, s) \cdot p(t|s) \]

\[ \approx p(\theta_{\text{context}}|\theta_{pp}) \cdot p(t|s) \]

\[ \approx \cos(\theta_{\text{context}}|\theta_{pp}) \cdot p(t|s) \]

\( k \): topic

\( \theta \): topic vector
Phrase Pair Topic Model with additional features

**Target-unigrams**

\[
\text{trgUnigrams}_t = \prod_{i=1}^{|t|} f\left( \frac{P_{doc}(w_i)}{P_{baseline}(w_i)} \right) \cdot f\left( \frac{P_{doc}(w_i)}{P_{topic0}(w_i)} \right)
\]

**Sim-phrasePair**

similarity = \( \cos(\theta_{pp}, \theta_{context}) \)

**Sim-targetPhrase**

similarity = \( \cos(\theta_{tp}, \theta_{context}) \)

**Sim-targetWord**

similarity = \( \cos(\theta_{tw}, \theta_{context}) \)
Dealing with multiple output domains

Multi-domain adaptation

▶ Adapt model to each of several (known) target domains

Domain classification for multi-domain adaptation

▶ Use perplexity of in-domain LMs [Xu et al., 2007]
▶ Use stemmed word bigrams + SVM [Banerjee et al., 2010]
▶ Use phrase pair provenance + perceptron [Wang et al., 2012]
Our approach to document classification

- Build domain classifiers using topic representations
Our approach to document classification

- For each test document:
Our approach to document classification

▶ Infer topic mixture → adapt features to topical context

Train document 3
Le noyau atomique désigne la région située au centre d'un atome constituée de protons et de neutrons (les nucléons). La taille du noyau (10^-15 mètre) est environ 100 000 fois plus petite que celle de l'atome et concentre quasiment toute sa masse. Les forces nucléaires qui s'exercent entre les nucléons sont à peu près un million de fois plus grandes.

Test document
En effet, l'écriture en espace noyau suppose l'absence de mécanismes tels que la protection de la mémoire. Il est donc plus complexe d'écrire un logiciel fonctionnant dans l'espace noyau que dans l'espace utilisateur, les bugs et failles de sécurité sont bien plus dangereux.

Train document 5
Comme chaque individu accepte un échange uniquement s'il préfère le nouveau stock à l'ancien, la solution choisie sur la courbe de contrat sera délimitée par les courbes d'indifférence qui passent par le stock. Selon la terminologie de la théorie des jeux coopératifs, les points entre ces deux limites constituent le noyau ou le cœur de l'économie ...

Train document 6
Le noyau d'un système d'exploitation est lui-même un logiciel, mais ne peut cependant utiliser tous les mécanismes d'abstraction qu'il fournit aux autres logiciels. Son rôle central impose par ailleurs des performances élevées. Cela fait du noyau la partie la plus critique d'un système d'exploitation et rend sa conception et sa ...

Train document 8
Comme chaque individu accepte un échange uniquement s'il préfère le nouveau stock à l'ancien, la solution choisie sur la courbe de contrat sera délimitée par les courbes d'indifférence qui passent par le stock. Selon la terminologie de la théorie des jeux coopératifs, les points entre ces deux limites constituent le noyau ou le cœur de l'économie ...

Train document 9
Comme chaque individu accepte un échange uniquement s'il préfère le nouveau stock à l'ancien, la solution choisie sur la courbe de contrat sera délimitée par les courbes d'indifférence qui passent par le stock. Selon la terminologie de la théorie des jeux coopératifs, les points entre ces deux limites constituent le noyau ou le cœur de l'économie ...
Our approach to document classification

- Predict domain → load domain-adapted translation features
Our approach to document classification

- Apply trained Phrase Pair Topic model to all training documents $\rightarrow$ one topic vector per document

**Single-prototype**
Average document vectors of same training domain ($\rightarrow$ domain vectors), max cosine similarity of test doc with domain vectors.

**Single-prototype-threshold**
Like single-prototype but with prediction threshold of 0.35. For similarities below threshold, predict *unknown* and fall back to baseline model.
Experimental setup (French-English)

| Data       | Mixed    | CC     | NC     | TED    | Europarl |
|------------|----------|--------|--------|--------|----------|
| Train (condition 1) | 354K (6450) | 110K   | 103K   | 140K   | -        |
| Train (condition 2) | 2.3M     | 110K   | 103K   | 140K   | 1.9M     |
| Dev        | 2453 (39) | 818    | 817    | 818    | -        |
| Test       | 5664 (112)| 1892   | 1878   | 1894   | -        |

Baseline systems

- Unadapted system
- DA-TM: linear PT interpolation [Sennrich, 2012]
- DA-LM: linear LM interpolation
- DA-TM+LM: both TM and LM adaptation

Automatic domain prediction

- Applied whenever we combine domain + topic adaptation
Training condition 2

- 2.3M training sentences
  → many more training contexts per phrase pair
- Sample up to 50 contexts per phrase pair
- Exclude singletons and frequent phrase pairs
  (> 20K occurrences)
Results: Single-prototype-threshold classifier

| Model            | CC    | NC  | TED |
|------------------|-------|-----|-----|
| # dev+test docs  | 88    | 39  | 24  |
| k=10             | 0.68  | 0.30| 0.02|
| k=20             | 0.76  | 0.15| 0.09|
| k=50             | 0.60  | 0.19| 0.21|
| k=100            | 0.55  | 0.12| 0.33|

Accuracy of domain prediction

▶
## Results: training condition 1 (three domains)

| Model       | Mixed  | CC     | NC     | TED    |
|-------------|--------|--------|--------|--------|
| Baseline    | 26.86  | 19.61  | 29.42  | 31.88  |
| DA-TM       | 27.24  | 19.61  | 29.87  | 32.73  |
| DA-LM       | 27.16  | 19.71  | 29.77  | 32.46  |
| DA-TM+LM    | 27.34  | 19.59  | 29.92  | 33.02  |

Total gain over baseline:
- 0.87
- 0.72
- 0.46
- 1.67

▶ Best system: DA-TM + topics (+ domain prediction)
Results: training condition 1 (three domains)

| Model          | Mixed | CC    | NC    | TED    |
|---------------|-------|-------|-------|--------|
| Baseline      | 26.86 | 19.61 | 29.42 | 31.88  |
| + topics      | **27.57** | 20.35 | 29.68 | 33.22  |
| DA-TM         | 27.24 | 19.61 | 29.87 | 32.73  |
| + topics      | **27.73** | 20.33 | 29.88 | 33.55  |
| DA-LM         | 27.16 | 19.71 | 29.77 | 32.46  |
| + topics      | **27.60** | 20.37 | 29.80 | 33.20  |
| DA-TM+LM      | 27.34 | 19.59 | 29.92 | 33.02  |
| + topics      | **27.63** | 20.22 | 29.90 | 33.33  |

▶ Best system: DA-TM + topics (+ domain prediction)
### Results: training condition 1 (three domains)

| Model          | Mixed  | CC     | NC     | TED    |
|----------------|--------|--------|--------|--------|
| Baseline       | 26.86  | 19.61  | 29.42  | 31.88  |
| + topics       | **27.57** | 20.35  | 29.68  | 33.22  |
|                | +0.74  | +0.26  | +1.34  |        |
| DA-TM          | 27.24  | 19.61  | 29.87  | 32.73  |
| + topics       | **27.73** | 20.33  | 29.88  | 33.55  |
|                | +0.69  | +0.01  | +0.82  |        |
| DA-LM          | 27.16  | 19.71  | 29.77  | 32.46  |
| + topics       | **27.60** | 20.37  | 29.80  | 33.20  |
|                | +0.63  | +0.03  | +0.74  |        |
| DA-TM+LM       | 27.34  | 19.59  | 29.92  | 33.02  |
| + topics       | **27.63** | 20.22  | 29.90  | 33.33  |
|                | +0.60  | -0.02  | +0.31  |        |

- **Total gain over baseline**: +0.87 +0.72 +0.46 +1.67
- **Best system**: DA-TM + topics (+ domain prediction)
### Results: training condition 1 (three domains)

| Model               | Mixed   | CC      | NC      | TED     |
|---------------------|---------|---------|---------|---------|
| Baseline + topics   | **27.57 | 20.35   | 29.68   | 33.22   |
| **Total gain over baseline** |        |         |         |         |
| DA-TM + topics      | 27.73   | 20.33   | 29.88   | 33.55   |
| DA-LM + topics      | 27.60   | 20.37   | 29.80   | 33.20   |
| DA-TM+LM + topics   | 27.63   | 20.22   | 29.90   | 33.33   |

- Best system: DA-TM + topics (+ domain prediction)
Results: training condition 1 (three domains)

What do we gain from domain adaptation?

| Model                        | Mixed | CC  | NC  | TED |
|------------------------------|-------|-----|-----|-----|
| DA-TM                        | 27.24 | 19.61 | 29.87 | 32.73 |
| Baseline+Sim-combine         | 27.29 | 20.10 | 29.49 | 32.60 |

Topic similarity features + domain-adapted features yield similar performance to using all features.
### What do we gain from domain adaptation?

| Model                        | Mixed  | CC    | NC    | TED   |
|------------------------------|--------|-------|-------|-------|
| DA-TM                        | 27.24  | 19.61 | 29.87 | 32.73 |
| Baseline + Sim-combine       | 27.29  | 20.10 | 29.49 | 32.60 |
| + DA-TM                      | **27.69** | **20.13** | **29.90** | **33.37** |

- Topic similarity features + domain-adapted features yield similar performance to using all features.
**Results: training condition 1 (three domains)**

What do we gain from domain adaptation?

| Model                        | Mixed | CC  | NC  | TED  |
|------------------------------|-------|-----|-----|------|
| DA-TM                        | 27.24 | 19.61 | 29.87 | 32.73 |
| Baseline + Sim-combine       | 27.29 | 20.10 | 29.49 | 32.60 |
| + DA-TM                      | **27.69** | **20.13** | **29.90** | **33.37** |
|                              | +0.40 | +0.03 | +0.41 | +0.77 |

- **Topic similarity** features + domain-adapted features yield similar performance to using all features.
## Results: training condition 2 (three domains + Europarl)

| Model               | Mixed | CC    | NC    | TED   |
|---------------------|-------|-------|-------|-------|
| Baseline            | 25.74 | 20.01 | 29.01 | 27.82 |
| DA-TM               | 26.74 | 20.13 | 29.53 | 30.86 |
| DA-LM               | 27.01 | 20.26 | 30.48 | 30.43 |
| DA-TM+LM            | 27.70 | 20.10 | 30.68 | 32.70 |

Total gain over baseline: +2.17, +0.37, +1.79, +5.16

▶ Best model: DA-TM + DA-LM + topics (+ domain prediction)
## Results: training condition 2 (three domains + Europarl)

| Model               | Mixed | CC    | NC    | TED    |
|---------------------|-------|-------|-------|--------|
| Baseline            | 25.74 | 20.01 | 29.01 | 27.82  |
| + topics            | **26.54** | 20.30 | 29.55 | 29.97  |
| DA-TM               | 26.74 | 20.13 | 29.53 | 30.86  |
| + topics            | **27.21** | 20.35 | 29.74 | 31.96  |
| DA-LM               | 27.01 | 20.26 | 30.48 | 30.43  |
| + topics            | *27.36* | 20.34 | 30.62 | 31.34  |
| DA-TM+LM            | 27.70 | 20.10 | 30.68 | 32.70  |
| + topics            | *27.91* | 20.38 | 30.80 | 32.98  |

Total gain over baseline:

- **2.17**
- **0.37**
- **1.79**
- **5.16**

▶ Best model: DA-TM + DA-LM + topics (+ domain prediction)
### Results: training condition 2 (three domains + Europarl)

| Model           | Mixed  | CC     | NC     | TED     |
|-----------------|--------|--------|--------|---------|
| Baseline        | 25.74  | 20.01  | 29.01  | 27.82   |
| + topics        | **26.54** | 20.30  | 29.55  | 29.97   |
|                 |        | +0.29  | +0.54  | +2.15   |
| DA-TM           | 26.74  | 20.13  | 29.53  | 30.86   |
| + topics        | **27.21** | 20.35  | 29.74  | 31.96   |
|                 |        | +0.22  | +0.21  | +1.10   |
| DA-LM           | 27.01  | 20.26  | 30.48  | 30.43   |
| + topics        | *27.36 | 20.34  | 30.62  | 31.34   |
|                 |        | +0.08  | +0.14  | +0.91   |
| DA-TM+LM        | 27.70  | 20.10  | 30.68  | 32.70   |
| + topics        | *27.91 | 20.38  | 30.80  | 32.99   |
|                 |        | +0.28  | +0.12  | +0.28   |

▶ Best model: DA-TM + DA-LM + topics (+ domain prediction)
### Results: training condition 2 (three domains + Europarl)

| Model          | Mixed  | CC     | NC     | TED    |
|----------------|--------|--------|--------|--------|
| Baseline       | 25.74  | 20.01  | 29.01  | 27.82  |
| + topics       | **26.54** | 20.30  | 29.55  | 29.97  |
|                | +0.29  | +0.54  | +2.15  |        |
| DA-TM          | 26.74  | 20.13  | 29.53  | 30.86  |
| + topics       | **27.21** | 20.35  | 29.74  | 31.96  |
|                | +0.22  | +0.21  | +1.10  |        |
| DA-LM          | 27.01  | 20.26  | 30.48  | 30.43  |
| + topics       | *27.36 | 20.34  | 30.62  | 31.34  |
|                | +0.08  | +0.14  | +0.91  |        |
| DA-TM+LM       | 27.70  | 20.10  | 30.68  | 32.70  |
| + topics       | *27.91 | 20.38  | 30.80  | 32.98  |
|                | +0.28  | +0.12  | +0.28  |        |
| Total gain over baseline | +2.17  | +0.37  | +1.79  | +5.16  |

- **Best model:**
  - DA-TM + DA-LM + topics (+ domain prediction)
Comparison of training conditions

| Best Model | Mixed | CC   | NC   | TED  |
|------------|-------|------|------|------|
| Train condition 1 | 27.73 | 20.33 | 29.88 | 33.55 |
| Train condition 2  | 27.91 | 20.38 | **30.80** | 32.98 |

- Both domain and topic adaptation could be improved to deal better with unbalanced data
| Input | le débit est en augmentation très rapide. | le débit a augmenté. |
|------|------------------------------------------|---------------------|
| Reference | these flows are increasing very rapidly. | the flows have increased. |
| Baseline | the speed is growing very rapidly. | the bitrate has increased. |

|  | speed | bitrate | throughput | flow |
|---|------|---------|------------|------|
| Baseline | 0.830 | 0.770 | 0.700 | 0.700 |
|  | 0.652 | 0.606 | 0.892 | 0.803 |
|  | 0.960 | 0.918 | 0.919 | 0.979 |
|  | 1.031 | 1.000 | 1.026 | 1.058 |

22/24
| Input       | le débit est en augmentation très rapide. | le débit a augmenté. |
|-------------|------------------------------------------|----------------------|
| Reference   | these **flows** are increasing very rapidly. | the **flows** have increased. |
| Baseline    | the **speed** is growing very rapidly.    | the **bitrate** has increased. |
| +DA-TM      | the **throughput** is rising very fast.    | the **throughput** has increased. |

|          | Speed | Time 1 | Time 2 | Time 3 | Time 4 |
|----------|-------|--------|--------|--------|--------|
| Baseline | 0.830 | 0.652  | 0.960  | 1.031  |        |
| +DA-TM   | 0.770 | 0.606  | 0.918  | 1      |        |
|          | 0.700 | 0.892  | 0.919  | 1.026  |        |
|          | 0.700 | 0.803  | 0.979  | 1.058  |        |
le débit est en augmentation très rapide.

these flows are increasing very rapidly.

the speed is growing very rapidly.

the throughput is rising very fast.

the flow is growing very rapidly.

le débit a augmenté.

the flows have increased.

the bitrate has increased.

the throughput has increased.

the flow has increased.
| Input | le débit est en augmentation très rapide. | le débit a augmenté. |
|-------|--------------------------------------|---------------------|
| Reference | these flows are increasing very rapidly. | the flows have increased. |
| Baseline | the speed is growing very rapidly. | the bitrate has increased. |
| +DA-TM | the throughput is rising very fast. | the throughput has increased. |
| +topics | the flow is growing very rapidly. | the flow has increased. |

| débit | Baseline $P(t|s)$ | DA-TM $P(t|s)$ | Topic-adapted Sim-trgWord | TrgUnigrams |
|-------|-------------------|----------------|---------------------------|-------------|
| speed   | 0.830             | 0.652           | 0.960                     | 1.031       |
| bitrate | 0.770             | 0.606           | 0.918                     | 1           |
| throughput | 0.700           | 0.892           | 0.919                     | 1.026       |
| flow    | 0.700             | 0.803           | 0.979                     | 1.058       |
Conclusions

- Measured relative benefit of domain adaptation and topic adaptation
- Methods are complementary, depending on text type/domain
- Provide adaptation at different levels of granularity
- Domains can be accurately predicted with domain vectors

Future work

- Direct integration of domain information into topic modelling
Thank you!
Banerjee, P., Du, J., Li, B., Naskar, S. K., Way, A., and Genabith, J. V. (2010). Combining Multi-Domain Statistical Machine Translation Models using Automatic Classifiers. In *Proceedings of AMTA*.

Eidelman, V., Boyd-Graber, J., and Resnik, P. (2012). Topic Models for Dynamic Translation Model Adaptation. In *Proceedings of ACL*.

Finch, A. (2008). Dynamic Model Interpolation for Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 208–215.

Foster, G., Goutte, C., and Kuhn, R. (2010). Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In *Proceedings of EMNLP*. 
Foster, G. and Kuhn, R. (2007). Mixture-Model Adaptation for SMT. In Proceedings of the Second Workshop on Statistical Machine Translation.

Gong, Z., Zhang, Y., and Zhou, G. (2010). Statistical Machine Translation based on LDA. In 4th International Universal Communication Symposium (IUCS).

Hasler, E., Blunsom, P., Koehn, P., and Haddow, B. (2014a). Dynamic Topic Adaptation for Phrase-based MT. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics.

Hasler, E., Haddow, B., and Koehn, P. (2014b). Dynamic Topic Adaptation for SMT using Distributional Profiles. In Proceedings of the 9th Workshop on Statistical Machine Translation.
Matsoukas, S., Rosti, A.-V. I., and Zhang, B. (2009). Discriminative corpus weight estimation for machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Sennrich, R. (2012). Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. In *Proceedings of EACL*.

Su, J., Wu, H., Wang, H., Chen, Y., Shi, X., Dong, H., and Liu, Q. (2012). Translation Model Adaptation for Statistical Machine Translation with Monolingual Topic Information. In *Proceedings of ACL*.

Wang, W., Macherey, K., Macherey, W., Och, F., and Xu, P. (2012).
Improved Domain Adaptation for Statistical Machine Translation.
In *Proceedings of AMTA*.

Xiao, X., Xiong, D., Zhang, M., Liu, Q., and Lin, S. (2012). A Topic Similarity Model for Hierarchical Phrase-based Translation.
In *Proceedings of ACL*, pages 750–758.

Xu, J., Deng, Y., Gao, Y., and Ney, H. (2007). Domain Dependent Statistical Machine Translation.
In *Proceedings of MT Summit XI*, pages 2–7.