Topic Modeling-based Domain Adaptation for System Combination

Tsuyoshi Okita[1] *Antonio Toral*[1] *Josef van Genabith*[1]
(1) Dublin City University, Glasnevin, Dublin 9
tokita@computing.dcu.ie, atoral@computing.dcu.ie, josef@computing.dcu.ie

ABSTRACT
This paper gives the system description of the domain adaptation team of Dublin City University for our participation in the system combination task in the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT (ML4HMT-12). We used the results of unsupervised document classification as meta information to the system combination module. For the Spanish-English data, our strategy achieved 26.33 BLEU points, 0.33 BLEU points absolute improvement over the standard confusion-network-based system combination. This was the best score in terms of BLEU among six participants in ML4HMT-12.

KEYWORDS: Statistical Machine Translation, Topic Model, System Combination.

1 Introduction

This paper describes a new extension to our system combination module developed in Dublin City University (Du and Way, 2010a,b; Okita and van Genabith, 2012). We have added a domain adaptation technique (Foster and Kuhn, 2007; Koehn and Schroeder, 2007; Daumé III, 2007) to our system combination module and tested it in the system combination task at the ML4HMT-2012 workshop.

The study of translation outputs obtained by systems trained on out-of-domain training data has contributed to the advance of domain adaptation techniques for statistical machine translation (SMT) (Foster and Kuhn, 2007; Koehn and Schroeder, 2007; Daumé III, 2007; Pecina et al., 2012). The literature shows that the performance gain obtained by using in-domain data (compared to out-of-domain data) is, in most cases, rather significant. Although it is often the case in the SMT literature that genre classification is done in a supervised setting (Jiang et al., 2012), analogous to genre-specific dictionaries in rule-based machine translation (RBMT) systems, a cache-based approach (Tiedemann, 2010) further investigates this on a fine-grained level of context, which does not need the notion of genre. Therefore, one idea worth exploring is to employ unsupervised document classification to cluster the documents (Blei et al., 2003; Steyvers and Griffiths, 2007; Blei, 2011; Sontag and Roy, 2011; Murphy, 2012).

In the context of system combination, the effect of out-of-domain training data is slightly different. Unlike the training of SMT systems, system combination essentially handles only the translation outputs, which can be considered to be in-domain. However, if we consider a training procedure which takes two steps (Du and Way, 2010a; Okita and van Genabith, 2012), these two steps are possible candidates that have a connection with the out-of-domain data. This two step approach to system combination tunes parameters in the first step over the development set and subsequently produces a final translation combining fragments obtained by translating the test set with different MT systems using such parameters.

Apart from this line of motivation, a number of times we have encountered obstacles to deploy a system combination module whose origin is difficult to trace. Although the system combination strategy works effectively in most cases, with some particular datasets we have experienced difficulties trying to achieve better performance than the single best system. Such cases include the ZH–EN translation task (Ma et al., 2009) and the EN–FR direction in the system combination task at WMT09[1].

In order to investigate this issue, we need to hypothesise what the cause might be. The super confusion network approach of Du and Way (2010a) assumed that the cause was related to the alignment metric. The strategy was then to incorporate not only one alignment metric but multiple metrics. The current paper hypothesises that the genre of the test and tuning sets exhibit variance, hence out-of-domain effects, and that this causes some variance in the performance of each MT system. If this is indeed the case, as is our assumption, the two methods explored in this paper should be effective: to identify and remove the out-of-domain data from the tuning set and to train on in-domain partitioned data.

The remainder of this paper is organized as follows. Section 2 describes our algorithm. In Section 3, our experimental results are presented. We conclude in Section 4.

[1] http://www.statmt.org/wmt09

2 Our Algorithm

Our algorithm consists of the following two steps in Algorithm 1.

Algorithm 1 Our Algorithm

Step 1: Run the out-of-domain data cleaning.
Step 2: Run the in-domain data partitioning.

This algorithm applies unsupervised document classification on the source side. The classification results of the source side are naturally linked to the target side since any parallel corpus forms translation pairs. Obviously another possibility would be to apply the unsupervised document classification jointly both of the source and the target sides.

The details of these two steps are explained in the following subsections.

2.1 Unsupervised Document Classification by Topic Model

We used Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Steyvers and Griffiths, 2007; Blei, 2011; Sontag and Roy, 2011; Murphy, 2012) to perform the (unsupervised) classification. LDA represents topics as multinomial distributions over the W unique word-types in the corpus and represents documents as a mixture of topics.

Let C be the number of unique labels in the corpus. Each label c is represented by a W-dimensional multinomial distribution ϕ_c over the vocabulary. For document d, we observe both the words in the document $w^{(d)}$ as well as the document labels $c^{(d)}$. Given the distribution over topics θ_d, the generation of words in the document is captured by the following generative model.

1. For each label $c \in \{1, \dots C\}$, sample a distribution over word-types $\phi_c \sim \textbf{Dirichlet}(\cdot|\beta)$
2. For each document $d \in \{1, \dots, D\}$
   (a) Sample a distribution over its observed labels $\theta_d \sim \textbf{Dirichlet}(\cdot|\alpha)$
   (b) For each word $i \in \{1, \dots, N_d^W\}$
      i. Sample a label $z_i^{(d)} \sim \textbf{Multinomial}(\theta_d)$
      ii. Sample a word $w_i^{(d)} \sim \textbf{Multinomial}(\phi_c)$ from the label $c = z_i^{(d)}$

The LDA model is represented as a graphical model in Figure 1. There are three levels in this figure: the corpus level, the document level and the within document level. The parameters α and β relate to the corpus level, the variables θ_d belong to the document level, and finally the variables z_{dn} and w_{dn} correspond to the word level, which are sampled once for each word in each document.

2.1.1 Out-of-domain Data Cleaning

Using topic modeling (or LDA) as described above, we propose to clean out-of-domain data from the tuning set as follows:

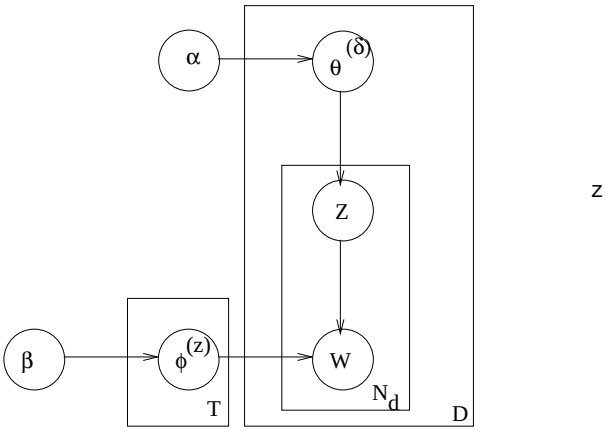

Figure 1: Figure shows the graphical model of LDA.

1. Fix the number of clusters C: choose a relatively big C.[2]

2. Do unsupervised document classification (or LDA) on the source side of the tuning and test sets.

3. Detect the classes that contain only data from the tuning set.

4. Discard the corresponding sentence pairs from the tuning set.

2.1.2 In-domain Data Partitioning

Using topic modeling (or LDA) as described above, we propose to perform in-domain data partitioning as follows:

1. Fix the number of clusters C, we explore values from small to big.[3]

2. Do unsupervised document classification (or LDA) on the source side of the tuning and test sets.

3. Separate each class of tuning and test sets (keep the original index and new index in the allocated separated dataset).

4. Run system combination on each class.

5. Reconstruct the system combined results of each class preserving the original index.

[2] C decides the size of clusters. In our case, 3,003 sentences will be clustered. If $C = 2$, the result cluster size will be 1,500 and we suggest this value of C is slightly too small. If $C = 3,000$, the result cluster size will be 1 and we suggest C is slightly too big. In this case, $C = 500 - 1,000$ would be the range considered and refereed as "relatively big".

[3] Currently, we do not have a definite recommendation on this. It needs to be studied more deeply.

2.2 System Combination

The first step of system combination is to select a backbone by MBR decoding. Let E be the target language, F be the source language, and $M(\cdot)$ be an MT system which maps some sequence in the source language F into some sequence in the target language E. Let $\mathscr{E}$ be the translation outputs of all the participating MT systems. Given a loss function $L(E, E')$ between an automatic translation E' and the reference E, a set of translation outputs $\mathscr{E}$, and an underlying probability model $P(E|F)$, a MBR decoder is defined as in (1) (Kumar and Byrne, 2002):

$$\hat{E} \quad = \quad \arg\min_{E' \in \mathscr{E}} R(E') = \arg\min_{E' \in \mathscr{E}} \sum_{E' \in \mathscr{E}} L(E, E') P(E|F) \tag{1}$$

where $R(E')$ denotes the Bayes risk of candidate translation E' under the loss function L. We use BLEU (Papineni et al., 2002) as this loss function L. According to this selected backbone, other translation outputs are aligned to form a confusion network.

The second step is by the (monotonic) consensus decoding for the given confusion network. There are two cases when this consensus decoding is executed: one is with references (tuning phase) and one is without references (test phase). Let $E_{j,n}$ be the nth best confusion network hypothesis and F_j be the jth source sentence. The hypothesis confidence (Rosti et al., 2007) is given as follows:

$$\log p(E_{j,n}/F_j) \quad = \quad \sum_{i=1}^{N_j - 1} \log(\sum_{l=1}^{N_S} \lambda_l p(w|l,i)) + \nu L(E_{j,n}) + \mu N_{nulls}(E_{j,n}) + \xi N_{words}(E_{j,n}) \tag{2}$$

where ν is the language model weight, $L(E_{j,n})$ is the LM log-probability and $N_{words}(E_{j,n})$ is the number of words in the hypothesis $E_{j,n}$. In the tuning phase, the parameters in Equation (2) are tuned. Then, using these tuned parameters, the test phase will be carried out. In this respect, the partitioning of in-domain data is very important. If we partition the in-domain data, the partitioned data will be guaranteed to be in-domain data (if we partition the data in general, the partitioned data will not be guaranteed to be in-domain tuning data).

3 Experimental Results

ML4HMT-2012 provides four translation outputs (*s1* to *s4*) which are MT output by two RBMT systems, APERTIUM and LUCY, PB-SMT (MOSES) and HPB-SMT (MOSES), respectively. The tuning data consists of 20,000 sentence pairs, while the test data consists of 3,003 sentence pairs.

class 1	20000					3003				
class 2	10213	9787				1821	1182			
class 3	6752	6428	6820			838	962	1203		
class 4	4461	4766	5954	4819		785	432	776	1010	
class 5	3846	3669	3665	3978	4842	542	343	1311	404	403

Table 1: Unsupervised document classification by a fixed number of clusters. Each column shows the number of items classified in each class.

Our experimental setting is as follows. We use our system combination module (Du and Way, 2010a,b; Okita and van Genabith, 2012), which has its own language modeling tool, MERT process, and MBR decoding. We use the BLEU metric as loss function in MBR decoding. We

	NIST	BLEU	METEOR	WER	PER
cleaned	7.4945	0.2500	0.5499287	56.6991	42.3032
wo cleaning	7.6846	0.2600	0.5643944	56.2368	41.5399

Table 2: The results of out-of-domain data cleaning compared with without cleaning.

use TERp[4] as alignment metrics in monolingual word alignment.[5] We use MALLET[6] for topic modeling. Although topic modeling is often used to obtain unsupervised clustering, our interest is focused on unsupervised classification of documents.

Given a specified number of classes C, we run MALLET to train the model on the tuning set. In this process, we obtained the label distribution for each document. Then, we infer the class using the trained model which yields the label distribution for each document. Results are shown in Table 1.

	NIST	BLEU	METEOR	WER	PER
s1	6.7456	0.2016	0.5712806	67.2881	54.7614
s2	7.3982	0.2388	0.6195136	63.9684	51.6444
s3	9.4167	0.3400	0.6650655	49.9341	37.4271
s4	9.1167	0.3273	0.6744035	52.0578	38.9179
topic modeling (devset)					
2 class	9.3504	0.3292	0.6529581	50.2061	36.8001
3 class	9.3045	0.3268	0.6522747	50.7730	37.4164
4 class	9.3084	0.3267	0.6513981	50.7391	37.3968
5 class	9.3950	0.3302	0.6531211	50.1131	36.7148
system combination					
syscom	9.2912	0.3268	0.6531500	50.7681	37.2779

Table 3: Table shows the performance of translation outputs s1 to s4 and results of system combination on development set.

Table 2 shows the performance on standard system combination, with and without data cleaning. In this out-of-domain data cleaning, we removed 2,207 sentences (11.0%) from the tuning data. The remaining 17,793 sentences are considered to be in-domain data from the point of view of the test set. However, this out-of-domain data cleaning did not quite work as expected.

Table 3 shows the performance on the development set. The performance of s1 and s2 is radically lower than that of s3 and s4 across all evaluation metrics considered. Although it may be that the performance of s1 and s2 is always inferior to that of the other systems, it may also be that s1 and s2 do not work well for some particular genre (the results shown in Table 4 seem to corroborate this hypothesis, particularly for s2).

We also performed the in-domain partitioning with the out-of-domain tuning set and without using the out-of-domain tuning set. Table 4 shows our results when we partitioned into 2, 3, 4, and 5 clusters.

The results show that 4 class classification achieved the best result, namely 26.33 BLEU points.

[4] `http://www.cs.umd.edu/~snover/terp`

[5] For example, Du and Way (2010a) explains various monolingual alignment methods such as TER alignment, HMM alignment and IHMM alignment.

[6] `http://mallet.cs.umass.edu/`

This is an improvement of 0.33 BLEU points absolute over system combination without topic modeling. Note that the baseline achieved 26.00 BLEU points, the best single system in terms of BLEU was s4 which achieved 25.31 BLEU points, and the best single system in terms of METEOR was s2 which achieved 0.5853.

	NIST	BLEU	METEOR	WER	PER
s1	6.4996	0.2248	0.5458641	64.2452	49.9806
s2	6.9281	0.2500	0.5853446	62.9194	48.0065
s3	7.4022	0.2446	0.5544660	58.0752	44.0221
s4	7.2100	0.2531	0.5596933	59.3930	44.5230
topic modeling (testset)					
2 class	7.7036	0.2620	0.5626187	55.8092	41.7783
3 class	7.7134	0.2628	0.5645200	55.8865	41.7171
4 class	7.7146	0.2633	0.5647685	55.8612	41.7264
5 class	7.6245	0.2592	0.5620755	56.9575	42.6229
system combination without topic modeling					
syscom	7.6846	0.2600	0.5643944	56.2368	41.5399

Table 4: Table includes our results on testset (the row 4 to 7).

Conclusion and Perspectives

This paper deployed domain adaptation via unsupervised document clustering through topic modeling and applied it to system combination. On the one hand, the out-of-domain data cleaning lost 1 BLEU point compared to the results of standard system combination. On the other hand, the in-domain data partitioning improved 1.02 BLEU points absolute compared to the single best MT system, and improved 0.33 BLEU points absolute compared to the results of the standard system combination approach.

Further studies will be carried out to explore this topic. First, this paper only handled the partition size of at most 5. We would like to apply our method to a larger dataset. It is also interesting to seek a method to find the optimal number of clusters automatically by hierarchical clustering methods with non-parametric Baysian methods (Okita and Way, 2010, 2011a,b). Alternatively, we have an interest on the reason why the out-of-domain data cleaning did not work in connection with noise if there is a link (Okita, 2009; Okita et al., 2010a,b; Okita, 2012).

Second, although we described only the method that uses domain adaptation, we explored also the correction of the output based on corresponding tokens and PoS tags from the source and target sides (e.g. if a token in the source side is a singular noun and the corresponding target token is a plural noun, overwrite that token by its singular form). This is related to techniques we have explored for diagnostic evaluation using checkpoints (Naskar et al., 2011; Toral et al., 2012) and a more detailed study is necessary to apply them in system combination.

Acknowledgments

This research is partly supported by the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the T4ME (Grant agreement No. 249119) and the PANACEA (contract no. 248064) projects as well as by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University.

References

Blei, D., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blei, D. M. (2011). Introduction to probabilistic topic models. *Communications of the ACM*.

Daumé III, H. (2007). Frustratingly easy domain adaptation. In *Conference of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic.

Du, J. and Way, A. (2010a). An incremental three-pass system combination framework by combining multiple hypothesis alignment methods. *International Journal of Asian Language Processing*, 20(1):1–15.

Du, J. and Way, A. (2010b). Using TERp to augment the system combination for SMT. *In Proceedings of the Ninth Conference of the Association for Machine Translation (AMTA2010)*.

Foster, G. and Kuhn, R. (2007). Mixture-model adaptation for SMT. *In Proceedings of the Second ACL Workshop on Statistical Machine Translation*, page 128–135.

Jiang, J., Way, A., Ng, N., Haque, R., Dillinger, M., and Lu, J. (2012). Monolingual data optimisation for bootstrapping SMT engines. *In the Proceedings of the MONOMT-2012 Workshop at AMTA*.

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for Statistical Machine Translation. *In Proceedings of the ACL Workshop on Statistical Machine Translation*.

Kumar, S. and Byrne, W. (2002). Minimum Bayes-Risk word alignment of bilingual texts. *In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 140–147.

Ma, Y., Okita, T., Cetinoglu, O., Du, J., and Way, A. (2009). Low-resource Machine Translation using MaTrEx: the DCU Machine Translation system for IWSLT 2009. *In Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2009)*, pages 29–36.

Murphy, K. P. (2012). Machine learning: A probabilistic perspective. *The MIT Press*.

Naskar, S. K., Toral, A., Gaspari, F., and Way, A. (2011). A framework for diagnostic evaluation of MT based on linguistic checkpoints. *In the Proceedings of the Machine Translation Summit XIII*, pages 529–536.

Okita, T. (2009). Data cleaning for word alignment. *In Proceedings of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009) Student Research Workshop*, pages 72–80.

Okita, T. (2012). Annotated corpora for word alignment between Japanese and English and its evaluation w ith MAP-based word aligner. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evalu ation (LREC-2012)*, pages 3241–3248, Istanbul, Turkey. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1655.

Okita, T., Graham, Y., and Way, A. (2010a). Gap between theory and practice: Noise sensitive word alignment in Machine Translation. *In Proceedings of the Workshop on Applications of Pattern Analysis (WAPA2010). Cumberland Lodge, England.*

Okita, T., Guerra, A. M., Graham, Y., and Way, A. (2010b). Multi-Word Expression sensitive word alignment. *In Proceedings of the Fourth International Workshop On Cross Lingual Information Access (CLIA2010, collocated with COLING2010), Beijing, China.*, pages 1–8.

Okita, T. and van Genabith, J. (2012). Minimum Bayes risk decoding with enlarged hypothesis space in system combination. *In Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012). LNCS 7182 Part II. A. Gelbukh (Ed.)*, pages 40–51.

Okita, T. and Way, A. (2010). Hierarchical Pitman-Yor Language Model in Machine Translation. *In Proceedings of the International Conference on Asian Language Processing (IALP 2010)*.

Okita, T. and Way, A. (2011a). Given bilingual terminology in Statistical Machine Translation: MWE-sensitve word alignment and hierarchical Pitman-Yor process-based translation model smoothing. *In Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-24)*, pages 269–274.

Okita, T. and Way, A. (2011b). Pitman-Yor process-based language model for Machine Translation. *International Journal on Asian Language Processing*, 21(2):57–70.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Pecina, P., Toral, A., Papavassiliou, V., Prokopidis, P., and van Genabith, J. (2012). Domain adaptation of Statistical Machine Translation using web-crawled resources: a case study. In *EAMT 2012: Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 145–152, Trento, Italy.

Rosti, A.-V. I., Matsoukas, S., and Schwartz, R. (2007). Improved word-level system combination for Machine Translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 312–319.

Sontag, D. and Roy, D. M. (2011). The complexity of inference in Latent Dirichlet Allocation. *In Advances in Neural Information Processing Systems 24 (NIPS)*.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis. Psychology Press*.

Tiedemann, J. (2010). Context adaptation in Statistical Machine Translation using models with exponentially decaying cache. *In Proceedings of the ACL Workshop on Domain Adaptation for Natural Language Processing*.

Toral, A., Naskar, S. K., Gaspari, F., and Groves, D. (2012). DELiC4MT: A Tool for Diagnostic MT Evaluation over User-defined Linguistic Phenomena. *The Prague Bulletin of Mathematical Linguistics*, pages 121–132.