Seeking to Reproduce "Easy Domain Adaptation"

Luís Gomes*, **Gertjan van Noord**†, **António Branco***, **Steven Neale***
*NLX – Natural Language and Speech Group
Faculdade de Ciências, Universidade de Lisboa, Portugal
{luis.gomes,steven.neale,antonio.branco}@di.fc.ul.pt

†University of Groningen, Netherlands
g.j.m.van.noord@rug.nl

Abstract

The *frustratingly easy domain adaptation* technique proposed by Daumé III (2007) is simple, easy to implement, and reported to be very successful in a range of NLP tasks (named entity recognition, part-of-speech tagging, and shallow parsing), giving us high hopes of successfully replicating and applying it to an English↔Portuguese hybrid machine translation system. While our hopes became 'frustration' in one translation direction – as the results obtained with the domain-adapted model do not improve upon the in-domain baseline model – our results are more encouraging in the opposite direction. This paper describes our replication of the technique and our application of it to machine translation, and offers a discussion on possible reasons for our mixed success in doing so.

1. Introduction

Frustratingly easy domain adaptation (EasyAdapt) is a technique put forward by Daumé III (2007) that allows learning algorithms that perform well across multiple domains to be easily developed. The technique is based on the principle of augmenting features from source language text in one domain – for which there might be more training data available – with features from text in a second, target domain in order that this domain is represented within a larger quantity of input data that can be fed to a learning algorithm. As well as purportedly being 'incredibly' easy to implement, the technique is shown to outperform existing results in a number of NLP tasks, including named entity recognition (NER), part-of-speech (POS) tagging, and shallow parsing.

Against this backdrop, we had high hopes of replicating the EasyAdapt technique and implementing it in the hybrid tree-to-tree machine translation (MT) system (Silva et al., 2015; Dušek et al., 2015) we have been developing as part of the QTLeap project[1]. While our system – based on the hybrid MT framework *TectoMT* (Zabokrtsky et al., 2008) – had been primarily trained on the much larger and broader-domained Europarl (EP) corpus (Silva et al., 2015; Dušek et al., 2015), we recently constructed a smaller, in-domain (IT) corpus of parallel questions and answers taken from real users' interactions with an information technology company's question answering (QA) system. Having initially obtained slightly improved results using this small in-domain corpus to train our MT system in the English↔Portuguese directions, we saw a potential benefit to replicating the EasyAdapt model and using it to produce larger, targeted training data encompassing both corpora.

In this paper, we report our results from replicating the EasyAdapt technique and applying it to the maximum entropy (MaxEnt) transfer models on which our hybrid tree-to-tree MT system is based, thus making use of an augmentation of features from both the larger, broader domain (EP) corpus and the smaller, in-domain (IT) corpus. Despite our initial results being slightly better when training our system on the IT corpus than with the much larger EP corpus, we were left frustrated after seeing encouraging results in only one translation direction when combining the two using the EasyAdapt model. As well as reporting our results using the EasyAdapt model, we also describe why – despite having been shown by Daumé III (2007) to be a successful technique for a range of NLP tasks – our attempts to replicate the model for our MT system did not lead to similarly improved results in both translation directions.

2. Hybrid Tree-to-Tree MT

The QTLeap project explores deep language engineering approaches to improving the quality of machine translation, and involves the creation of MT systems for seven languages paired with English: Basque, Bulgarian, Czech, Dutch, German, Portuguese, and Spanish. The systems are being used to translate the questions posed by users of an IT company's interactive QA system from their native language into English (the primary language of the QA system's database), using these machine-translations to retrieve the most similar questions and their respective answers from said database. The retrieved answers are then machine-translated back into the user's native language – if they have not already been translated before – to complete the cycle and provide users with an answer to their initial question.

For English↔Portuguese, our translation pipeline is based on the tectogrammatical hybrid-MT framework *TectoMT* (Zabokrtsky et al., 2008) and follows a classical analysis→transfer→synthesis structure with the transfer taking place at a deep syntactic (tectogrammatical) level of representation.

2.1. Analysis

The analysis of input sentences is performed in two stages: the first stage takes the raw string representation and produces a surface-level analytical tree (a-tree) representation; the second stage takes the a-tree and produces a deeper tectogrammatical tree representation (t-tree). Figure 1 shows

[1] http://qtleap.eu

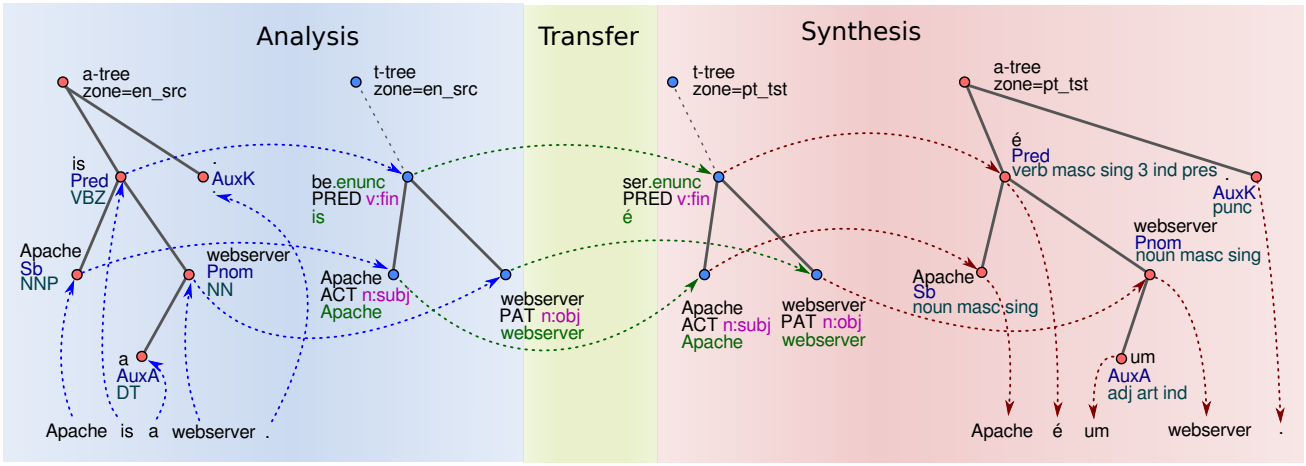

Figure 1: Trees at several stages of the analysis-transfer-synthesis of the pipeline.

the a-trees and t-trees of a short sentence as it is translated from English to Portuguese.

In the surface-level analytical tree representation of a sentence (a-tree), each word is represented by a node and the edges represent syntactic dependency relations. Nodes have several attributes, the most relevant being the lemma, part-of-speech tag, and morphosyntactic features (gender, number, tense, person, etc). By contrast, in the tectogrammatical tree representation (t-tree) only content words (nouns, pronouns, main verbs, adjectives and adverbs) are represented as nodes, but the t-tree may also contain nodes that have no corresponding word in the original sentence, such as nodes representing pro-dropped pronouns.

2.1.1. Surface-level analysis

For the initial surface-level analysis we use LX-Suite (Branco and Silva, 2006), a set of shallow processing tools for Portuguese that includes a sentence segmenter, a tokenizer, a PoS tagger, a morphological analyser and a dependency parser. All of the LX-Suite components have state-of-the-art performance. The trees produced by the dependecy parser are converted into the Universal Stanford Dependencies tagset (USD) proposed by de Marneffe et al. (2014). Additionally, the part-of-speech and morphological feature tags are converted into the Interset tagset (Zeman, 2008).

2.1.2. Deeper analysis

The second analysis stage transforms the surface-level a-trees into deeper t-trees. This transformation is purely rule-based; some rules are language-specific, others are language-independent. The two most important transformations are: (1) drop nodes corresponding to non-content words (articles, prepositions, auxiliary verbs, etc) and (2) add nodes for pro-dropped pronouns. Because all TectoMT-based translation pipelines adopt a universal representation for a-trees (USD and Interset), the language-independent rules are shareable across pipelines, reducing the amount of work needed to add support for new languages.

2.2. Transfer

The transfer step is statistical and is responsible for transforming a source-language t-tree into a target-language t-tree. It is assumed that the source and target t-trees are isomorphic, which is true most of the time, given that at the tectogrammatical representation level, most language-dependent features have been abstracted away. Thus, the transformation of a source-language t-tree into a target-language t-tree is done by statistically transferring node attributes (t-lemmas and formemes) and then reordering nodes as needed to meet the target-language word ordering rules. Some reorderings are encoded in formemes, as for example *adj:prenon* and *adj:postnom*, which represent prenominal and postnominal adjectives respectively.

2.2.1. Transfer Models

The transfer models are multi-label classifiers that predict an attribute (t-lemma or formeme) of each target-language t-node given as input a set of attributes of the corresponding source-language t-node and its immediate neighbours (parent, siblings and children). There are separate models for predicting t-lemmas and formemes, but other than the different output labels, the input feature sets are identical. Two kinds of statistical models are employed and interpolated: (1) a *static* model that predicts output t-lemmas (or formeme) based solely on the source-language t-lemma (or formeme), i.e. without taking into account any other contextual feature, and (2) a *MaxEnt* model[2] that takes all contextual features into account.

2.3. Synthesis

The synthesis step of the pipeline is rule-based and relies on two pre-existing tools for Portuguese synthesis: a verbal conjugator (Branco and Nunes, 2012) and a nominal inflector (Martins, 2006). Besides these synthesis tools, there are rules for adding auxiliary verbs, articles and prepositions as needed to transform the deep tectogrammatical representation into a surface-level tree representation, which is then converted into the final string representation by concatenating nodes (words) in depth-first left-to-right tree-traversal ordering (adding spaces as needed).

[2]there is one MaxEnt model for each distinct source-language t-lemma (or formeme) so, in fact, we have an ensemble of MaxEnt models

3. Frustratingly Easy Domain Adaptation

The 'frustratingly easy domain adaptation' (EasyAdapt) technique (Daumé III, 2007) is a simple feature augmentation technique that can be used in combination with many learning algorithms. The application of EasyAdapt for various NLP tasks, including Named Entity Recognition, Part-of-Speech Tagging, and Shallow Parsing was reported as successfull. Even if EasyAdapt is not directly applicable to the models typically used in Statistical Machine Translation, a similar approach has been shown to improve results for translation as well (Clark et al., 2012).

Although EasyAdapt has been developed in the context of domain adaptation, it is best described as a very simple, yet effective, multi-domain learning technique (Joshi et al., 2012). In EasyAdapt, each input feature is augmented with domain specific versions of it. If we have data from K domains, the augmented feature space will consist of $K + 1$ copies of the original feature space. Each training/testing instance is associated with a particular domain, and therefore two versions of each feature are present for a given instance: the original, general, version and the domain specific version.

The classifier may learn that a specific feature is always important, regardless of the domain (and thus it will rely more on the general version of the feature), or it may learn that a specific feature is relevant only for particular domain(s) and thus rely more on the relevant domain specific features. As a result, we obtain a single model which encodes both generic properties of the task as well as domain specific preferences.

We implemented EasyAdapt in our MaxEnt transfer models by adding, for each original feature f, a feature f_d if the training/testing instance is from domain d. In the experiments below, there are only two domains, the IT domain, which we regard as in-domain for the translation system, and the EP domain, which is out-of-domain for our translation system. [3].

4. Experiments

We performed a total of 24 experiments, evaluating four different models on three testsets (models and testsets outlined below) and in both translation directions – English→Portuguese and Portuguese→English.

4.1. Models

The smaller, in-domain parallel corpus we created for training the transfer models comprises 2000 questions and 2000 answers collected from real user interactions with the QA-based chat system used by an information technology company to provide technical assistance to its customers. The out-of-domain corpus is the English and Portuguese-aligned version of Europarl (Koehn, 2005).

From these two corpora, we created four models: **IT** (trained with what we consider to be our in-domain data only, the 4000 sentences from the user interactions with the QA system), **EP** (trained with what we consider to be our out-of-domain data only, the Europarl corpus), **IT+EP** (a trivially domain-adapted model obtained by concatenating both the IT and the EP corpora), and finally the **EasyAdapt** model (a domain-adapted model created by using the EasyAdapt technique to combine features from both corpora).

4.2. Testsets

We have used three testsets for evaluation: **IT** (an in-domain testset composed of 1000 questions and 1000 answers collected from the same real user interactions with the QA-based chat system as the previously described model, but with no overlap with the corpus used for training), **News** (an out-of-domain testset with 604 sentences from the news domain, created by manually translating a subset of the testset used in the WMT12 tasks[4] into Portuguese in the context of the QTLeap project), and **EP** (the first 1000 parallel sentences in English and Portuguese from the Europarl corpus).

Note that the **News** testset is from a different domain than either of the other two corpora (IT and EP) used for training – we wanted to experiment with this additional testset to see whether or not the EasyAdapt model is more general than the model obtained by simply concatenating both corpora (IT+EP).

4.3. Results

Tables 1 and 2 show the BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) scores obtained with the four models on the three testsets for the English→Portuguese and Portuguese→English translation directions respectively. Frustratingly, the EasyAdapt model did not outperform the baseline in-domain model (IT) on the in-domain (IT) testset for English→Portuguese. However, the EasyAdapt model was the best performing model in the Portuguese→English direction. In this context, our initial goal of improving in-domain translation by learning from a larger out-of-domain corpus augmented with features from a smaller, targeted in-domain corpus using the EasyAdapt model has been met with mixed success.

For the better performing direction of Portuguese→English, the scores obtained using the EasyAdapt model outperform other models on all but the EP testset, for which they are only slightly behind. This suggests that in the scenario that we need to use a single model to translate two domains instead of a separate model for each domain – to ease memory concerns, perhaps – the EasyAdapt model would likely be a much better option than simply concatenating the corpora from both domains. Furthermore, the EasyAdapt model is the best performing model when translating texts from a third (News) domain.

5. Discussion

Although the EasyAdapt model was effective in the Portuguese→English direction, we were disappointed that the same good results were not obtained for in-domain translation in the English→Portuguese direction. Considering possible reasons for this, we note that the development

[3] Below, we also apply our models to a third domain, News, but since we do not train on that domain, there is no point in having News-specific features

[4] `http://www.statmt.org/wmt12/test.tgz`

of our hybrid MT system has so far been more heavily concentrated on the English→Portuguese direction given that – as described in section 2. – this is the direction whose translation will be presented to end users. As a result, the system was weaker in the Portuguese→English direction to begin with.

With this in mind, we expect that there is more room for the EasyAdapt model to impact on results in a positive manner in the Portuguese→English than in the English→Portuguese translation direction, and that this is why we see improvements in translation quality when using the model in this direction. This is also likely when we consider that the kinds of tasks for which Daumé III (2007) reported improved performance – pos tagging, shallow parsing etc. – are surface-level in nature, as are both the shallow-processing tasks used in the analysis phase and the rule-based components used in the synthesis phase of the hybrid MT system.

Considering the synthesis steps in particular, the fact that the Portuguese components used in the English→Portuguese direction are more matured and have received more attention than their Portuguese→English counterparts may simply mean that these components already perform well enough that no real improvements can be seen from using the EasyAdapt model, in contrast to the equivalent components in the opposite direction which are less mature and therefore improved by adopting the EasyAdapt model.

Model	Testset					
	IT		News		EP	
	BLEU	NIST	BLEU	NIST	BLEU	NIST
IT	**22.81**	**6.47**	4.10	3.21	4.25	2.72
EP	18.73	5.60	8.03	**4.46**	**8.00**	4.39
IT+EP	21.25	6.09	7.84	4.43	7.89	4.36
EasyAdapt	22.63	6.44	**8.13**	4.40	7.82	**4.43**

Table 1: BLEU and NIST scores obtained with four transfer models (rows) in three different domain testsets (columns) for the English→Portuguese direction.

Model	Testset					
	IT		News		EP	
	BLEU	NIST	BLEU	NIST	BLEU	NIST
IT	13.78	4.97	2.77	2.90	2.41	2.50
EP	12.24	4.43	6.57	4.13	**7.25**	4.24
IT+EP	13.30	4.78	6.46	4.11	7.09	4.18
EasyAdapt	**14.13**	**5.13**	**6.81**	**4.18**	7.13	**4.25**

Table 2: BLEU and NIST scores obtained with four transfer models (rows) in three different domain testsets (columns) for the Portuguese→English direction.

6. Conclusions

We have presented the results of our replication of the EasyAdapt (*frustatingly easy domain adaptaion*) technique and our integration of it into an English↔Portuguese hybrid machine translation system. We had high hopes that by replicating the technique, we would be able to combine features from the large out-of-domain (EP) corpus we had previously used to train our system with features from a small in-domain (IT) corpus constructed within the scope of the QTLeap project and see improved results by feeding this combination of features to our maxent-based transfer models during the training of the system.

Our efforts to reproduce the improvements of the EasyAdapt technique reported by Daumé III (2007) have been of mixed success in the context of machine translation. While we were able to improve the Portuguese→English translation of in-domain texts using the EasyAdapt technique compared to the in-domain trained baseline, the EasyAdapt model did not outperform the in-domain trained baseline in the English→Portuguese direction, which is currently our best performing of the two directions. Among other possible reasons for this, it may simply be the case that as Portuguese→English is our weaker translation direction, the EasyAdapt model has more room to make an impact on translations and less so in the more matured and refined pipeline for the English→Portuguese direction.

The EasyAdapt technique was reported to lead to better results in a number of NLP tasks by preparing domain-adapted training data (Daumé III, 2007) but we have found it difficult to fully reproduce that success across the board in the machine translation context. These results highlight the importance of replicating techniques in different contexts to truly assess their suitability to and reproducibility of results across different scenarios.

7. Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7) under grant agreement nº 610516 (project QTLeap), and from the Portuguese Foundation for Science and Technology (FCT) under grant PTDC/EEI-SII/1940/2012 (project DP4LT).

8. References

Branco, A. and Nunes, F. (2012). Verb analysis in a highly inflective language with an mff algorithm. In *Computational Processing of the Portuguese Language*, pages 1–11. Springer.

Branco, A. and Silva, J. (2006). A suite of shallow processing tools for Portuguese: LX-Suite. In *Proceedings of the 11th European Chapter of the Association for Computational Linguistics*, pages 179–182.

Clark, J., Lavie, A., and Dyer, C. (2012). One system, many domains: Open-domain statistical machine translation via feature augmentation. In *AMTA 2012, Conference of the Association for Machine Translation in the Americas*, San Diego, October 28 – November 1.

Daumé III, H. (2007). Frustratingly easy domain adaptation. *ACL 2007*, page 256.

de Marneffe, M.-C., Silveira, N., Dozat, T., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the 9th Language Resources and Evaluation Conference*.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Dušek, O., Gomes, L., Novák, M., Popel, M., and Rosa, R. (2015). New language pairs in tectomt. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 98–104.

Joshi, M., Cohen, W. W., Dredze, M., and Rosé, C. P. (2012). Multi-domain learning: When do domains matter? In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1302–1312, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5.

Martins, P. (2006). LX-Inflector: Implementation Report and User Manual, University of Lisbon. Technical report, TagShare Project.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Silva, J., Rodrigues, J., Gomes, L., and Branco, A. (2015). Bootstrapping a hybrid deep MT system. In *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 1–5, Beijing, China, July. Association for Computational Linguistics.

Zabokrtsky, Z., Ptacek, J., and Pajas, P. (2008). TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, June. Association for Computational Linguistics.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *LREC*.