Cross-lingual Model Transfer Using Feature Representation Projection

Mikhail Kozhevnikov
MMCI, University of Saarland
Saarbrücken, Germany
mkozhevn@mmci.uni-saarland.de

Ivan Titov
ILLC, University of Amsterdam
Amsterdam, Netherlands
titov@uva.nl

Abstract

We propose a novel approach to cross-lingual model transfer based on *feature representation projection*. First, a compact feature representation relevant for the task in question is constructed for either language independently and then the mapping between the two representations is determined using parallel data. The target instance can then be mapped into the source-side feature representation using the derived mapping and handled directly by the source-side model. This approach displays competitive performance on model transfer for semantic role labeling when compared to direct model transfer and annotation projection and suggests interesting directions for further research.

1 Introduction

Cross-lingual model transfer approaches are concerned with creating statistical models for various tasks for languages poor in annotated resources, utilising resources or models available for these tasks in other languages. That includes approaches such as *direct model transfer* (Zeman and Resnik, 2008) and *annotation projection* (Yarowsky et al., 2001). Such methods have been successfully applied to a variety of tasks, including POS tagging (Xi and Hwa, 2005; Das and Petrov, 2011; Täckström et al., 2013), syntactic parsing (Ganchev et al., 2009; Smith and Eisner, 2009; Hwa et al., 2005; Durrett et al., 2012; Søgaard, 2011), semantic role labeling (Padó and Lapata, 2009; Annesi and Basili, 2010; Tonelli and Pianta, 2008; Kozhevnikov and Titov, 2013) and others.

Direct model transfer attempts to find a shared feature representation for samples from the two languages, usually generalizing and abstracting away from language-specific representations. Once this is achieved, instances from both languages can be mapped into this space and a model trained on the source-language data directly applied to the target language. If parallel data is available, it can be further used to enforce model agreement on this data to adjust for discrepancies between the two languages, for example by means of *projected transfer* (McDonald et al., 2011).

The shared feature representation depends on the task in question, but most often each aspect of the original feature representation is handled separately. Word types, for example, may be replaced by cross-lingual word clusters (Täckström et al., 2012) or cross-lingual distributed word representations (Klementiev et al., 2012). Part-of-speech tags, which are often language-specific, can be converted into universal part-of-speech tags (Petrov et al., 2012) and morpho-syntactic information can also be represented in a unified way (Zeman et al., 2012; McDonald et al., 2013; Tsarfaty, 2013). Unfortunately, the design of such representations and corresponding conversion procedures is by no means trivial.

Annotation projection, on the other hand, does not require any changes to the feature representation. Instead, it operates on translation pairs, usually on sentence level, applying the available source-side model to the source sentence and transferring the resulting annotations through the word alignment links to the target one. The quality of predictions on source sentences depends heavily on the quality of parallel data and the domain it belongs to (or, rather, the similarity between this domain and that of the corpus the source-language model was trained on). The transfer itself also introduces errors due to translation shifts (Cyrus, 2006) and word alignment errors, which may lead to inaccurate predictions. These issues are generally handled using heuristics (Padó and Lapata, 2006) and filtering, for example based on alignment coverage (van der Plas et al., 2011).

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 579–585,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

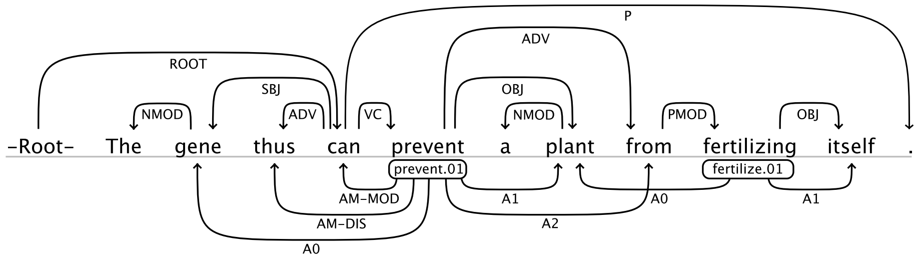

Figure 1: Dependency-based semantic role labeling example. The top arcs depict dependency relations, the bottom ones – semantic role structure. Rendered with `https://code.google.com/p/whatswrong/`.

1.1 Motivation

The approach proposed here, which we will refer to as *feature representation projection* (FRP), constitutes an alternative to direct model transfer and annotation projection and can be seen as a compromise between the two.

It is similar to direct transfer in that we also use a shared feature representation. Instead of designing this representation manually, however, we create compact monolingual feature representations for source and target languages separately and automatically estimate the mapping between the two from parallel data. This allows us to make use of language-specific annotations and account for the interplay between different types of information. For example, a certain preposition attached to a token in the source language might map into a morphological tag in the target language, which would be hard to handle for traditional direct model transfer other than using some kind of refinement procedure involving parallel data. Note also that any such refinement procedure applicable to direct transfer would likely work for FRP as well.

Compared to annotation projection, our approach may be expected to be less sensitive to parallel data quality, since we do not have to commit to a particular prediction on a given instance from parallel data. We also believe that FRP may profit from using other sources of information about the correspondence between source and target feature representations, such as dictionary entries, and thus have an edge over annotation projection in those cases where the amount of parallel data available is limited.

2 Evaluation

We evaluate feature representation projection on the task of dependency-based semantic role labeling (SRL) (Hajič et al., 2009).

This task consists in identifying predicates and their arguments in sentences and assigning each argument a semantic role with respect to its predicate (see figure 1). Note that only a single word – the syntactic head of the argument phrase – is marked as an argument in this case, as opposed to constituent- or span-based SRL (Carreras and Màrquez, 2005). We focus on the assignment of semantic roles to identified arguments.

For the sake of simplicity we cast it as a multiclass classification problem, ignoring the interaction between different arguments in a predicate. It is well known that such interaction plays an important part in SRL (Punyakanok et al., 2008), but it is not well understood which kinds of interactions are preserved across languages and which are not. Also, should one like to apply constraints on the set of semantic roles in a given predicate, or, for example, use a reranker (Björkelund et al., 2009), this can be done using a factorized model obtained by cross-lingual transfer.

In our setting, each *instance* includes the word type and part-of-speech and morphological tags (if any) of argument token, its parent and corresponding predicate token, as well as their dependency relations to their respective parents. This representation is further denoted ω_0.

2.1 Approach

We consider a pair of languages (L^s, L^t) and assume that an annotated training set $D^s_T = \{(x^s, y^s)\}$ is available in the source language as well as a parallel corpus of instance pairs $D^{st} = \{(x^s, x^t)\}$ and a target dataset $D^t_E = \{x^t\}$ that needs to be labeled.

We design a pair of intermediate compact monolingual feature representations ω^s_1 and ω^t_1 and models M_s and M_t to map source and target samples x^s and x^t from their original representations, ω^s_0 and ω^t_0, to the new ones. We use the parallel instances in the new feature representation

$$\bar{D}^{st} = \left\{ \left(x_1^s, x_1^t \right) \right\} = \left\{ \left(M_s(x^s), M_t(x^t) \right) \right\}$$

to determine the mapping M_{ts} (usually, linear) between the two spaces:

$$M_{ts} = argmax_M \sum_{(x_1^s, x_1^t \in \bar{D}^{st})} \left\| x_1^s - M(x_1^t) \right\|_2$$

Then a classification model M_y is trained on the source training data

$$\bar{D}_T^s = \{(x_1^s, y^s)\} = \{(M_s(x^s), y^s)\}$$

and the labels are assigned to the target samples $x^t \in D_E^t$ using a composition of the models:

$$y^t = M_y(M_{ts}(M_t(x^t)))$$

2.2 Feature Representation

Our objective is to make the feature representation sufficiently compact that the mapping between source and target feature spaces could be reliably estimated from a limited amount of parallel data, while preserving, insofar as possible, the information relevant for classification.

Estimating the mapping directly from raw categorical features (ω_0) is both computationally expensive and likely inaccurate – using one-hot encoding the feature vectors in our experiments would have tens of thousands of components. There is a number of ways to make this representation more compact. To start with, we replace word types with corresponding neural language model representations estimated using the skip-gram model (Mikolov et al., 2013a). This corresponds to M_s and M_t above and reduces the dimension of the feature space, making direct estimation of the mapping practical. We will refer to this representation as ω_1.

To go further, one can, for example, apply dimensionality reduction techniques to obtain a more compact representation of ω_1 by eliminating redundancy or define auxiliary tasks and produce a vector representation useful for those tasks. In source language, one can even directly tune an intermediate representation for the target problem.

2.3 Baselines

As mentioned above we compare the performance of this approach to that of direct transfer and annotation projection. Both baselines are using the same set of features as the proposed model, as described earlier.

The shared feature representation for direct transfer is derived from ω_0 by replacing language-specific part-of-speech tags with universal ones (Petrov et al., 2012) and adding cross-lingual word clusters (Täckström et al., 2012) to word types. The word types themselves are left as they are in the source language and replaced with their gloss translations in the target one (Zeman and Resnik, 2008). In English-Czech and Czech-English we also use the dependency relation information, since the annotations are partly compatible.

The annotation projection baseline implementation is straightforward. The source-side instances from a parallel corpus are labeled using a classifier trained on source-language training data and transferred to the target side. The resulting annotations are then used to train a target-side classifier for evaluation. Note that predicate and argument identification in both languages is performed using monolingual classifiers and only aligned pairs are used in projection. A more common approach would be to project the whole structure from the source language, but in our case this may give unfair advantage to feature representation projection, which relies on target-side argument identification.

2.4 Tools

We use the same type of log-linear classifiers in the model itself and the two baselines to avoid any discrepancy due to learning procedure. These classifiers are implemented using Pylearn2 (Goodfellow et al., 2013), based on Theano (Bergstra et al., 2010). We also use this framework to estimate the linear mapping M_{ts} between source and target feature spaces in FRP.

The 250-dimensional word representations for ω_1 are obtained using Word2vec tool. Both monolingual data and that from the parallel corpus are included in the training. In Mikolov et al. (2013b) the authors consider embeddings of up to 800 dimensions, but we would not expect to benefit as much from larger vectors since we are using a much smaller corpus to train them. We did not tune the size of the word representation to our task, as this would not be appropriate in a cross-lingual transfer setup, but we observe that the classifier is relatively robust to their dimension when evaluated on source language – in our experiments the performance of the monolingual classifier does not improve significantly if the dimension is increased past 300 and decreases only by a small margin (less than one absolute point) if it is reduced to 100. It should be noted, however, that the dimension that is optimal in this sense is not necessarily the best choice for FRP, especially if the amount of available parallel data is limited.

2.5 Data

We use two language pairs for evaluation: English-Czech and English-French. In the first case, the data is converted from Prague Czech-English Dependency Treebank 2.0 (Hajič et al., 2012) using the script from Kozhevnikov and Titov (2013). In the second, we use CoNLL 2009 shared task (Hajič et al., 2009) corpus for English and the manually corrected dataset from van der Plas et al. (2011) for French. Since the size of the latter dataset is relatively small – one thousand sentences – we reserve the whole dataset for testing and only evaluate transfer from English to French, but not the other way around. Datasets for other languages are sufficiently large, so we take 30 thousand samples for testing and use the rest as training data. The validation set in each experiment is withheld from the corresponding training corpus and contains 10 thousand samples.

Parallel data for both language pairs is derived from Europarl (Koehn, 2005), which we preprocess using MATE-TOOLS (Björkelund et al., 2009; Bohnet, 2010).

3 Results

The classification error of FRP and the baselines given varying amount of parallel data is reported in figures 2, 3 and 4. The training set for each language is fixed. We denote the two baselines AP (annotation projection) and DT (direct transfer).

The number of parallel instances in these experiments is shown on a logarithmic scale, the values considered are 2, 5, 10, 20 and 50 thousand pairs.

Please note that we report only a single value for direct transfer, since this approach does not explicitly rely on parallel data. Although some of the features – namely, gloss translations and cross-lingual clusters – used in direct transfer are, in fact, derived from parallel data, we consider the effect of this on the performance of direct transfer to be indirect and outside the scope of this work.

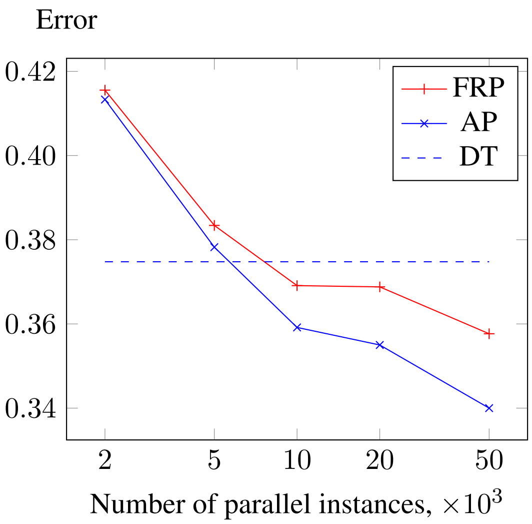

Figure 2: English-Czech transfer results

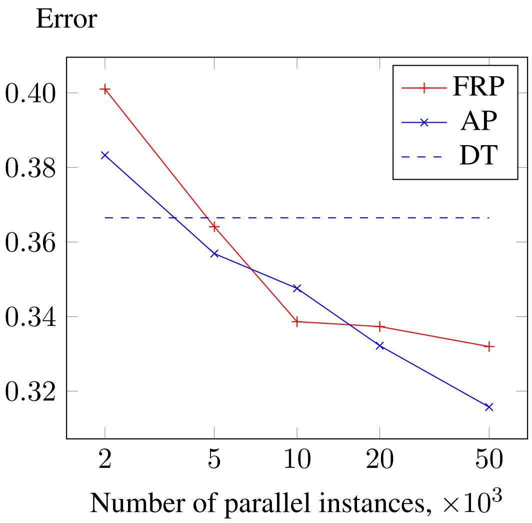

Figure 3: Czech-English transfer results

The rather inferior performance of direct transfer baseline on English-French may be partially attributed to the fact that it cannot rely on dependency relation features, as the corpora we consider make use of different dependency relation inventories. Replacing language-specific dependency annotations with the universal ones (McDonald et al., 2013) may help somewhat, but we would still expect the methods directly relying on parallel data to achieve better results given a sufficiently large parallel corpus.

Overall, we observe that the proposed method with ω_1 representation demonstrates performance competitive to direct transfer and annotation projection baselines.

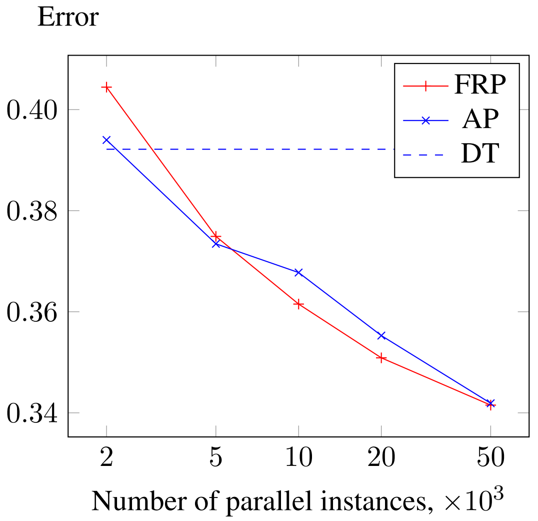

Figure 4: English-French transfer results

4 Additional Related Work

Apart from the work on direct/projected transfer and annotation projection mentioned above, the proposed method can be seen as a more explicit kind of domain adaptation, similar to Titov (2011) or Blitzer et al. (2006).

It is also somewhat similar in spirit to Mikolov et al. (2013b), where a small number of word translation pairs are used to estimate a mapping between distributed representations of words in two different languages and build a word translation model.

5 Conclusions

In this paper we propose a new method of cross-lingual model transfer, report initial evaluation results and highlight directions for its further development.

We observe that the performance of this method is competitive with that of established cross-lingual transfer approaches and its application requires very little manual adjustment – no heuristics or filtering and no explicit shared feature representation design. It also retains compatibility with any refinement procedures similar to projected transfer (McDonald et al., 2011) that may have been designed to work in conjunction with direct model transfer.

6 Future Work

This paper reports work in progress and there is a number of directions we would like to pursue further.

Better Monolingual Representations The representation we used in the initial evaluation does not discriminate between aspects that are relevant for the assignment of semantic roles and those that are not. Since we are using a relatively small set of features to start with, this does not present much of a problem. In general, however, retaining only relevant aspects of intermediate monolingual representations would simplify the estimation of mapping between them and make FRP more robust.

For source language, this is relatively straightforward, as the intermediate representation can be directly tuned for the problem in question using labeled training data. For target language, however, we assume that no labeled data is available and auxiliary tasks have to be used to achieve this.

Alternative Sources of Information The amount of parallel data available for many language pairs is growing steadily. However, cross-lingual transfer methods are often applied in cases where parallel resources are scarce or of poor quality and must be used with care. In such situations an ability to use alternative sources of information may be crucial. Potential sources of such information include dictionary entries or information about the mapping between certain elements of syntactic structure, for example a known part-of-speech tag mapping.

The available parallel data itself may also be used more comprehensively – aligned arguments of aligned predicates, for example, constitute only a small part of it, while the mapping of vector representations of individual tokens is likely to be the same for all aligned words.

Multi-source Transfer One of the strong points of direct model transfer is that it naturally fits the multi-source transfer setting. There are several possible ways of adapting FRP to such a setting. It remains to be seen which one would produce the best results and how multi-source feature representation projection would compare to, for example, multi-source projected transfer (McDonald et al., 2011).

Acknowledgements

The authors would like to acknowledge the support of MMCI Cluster of Excellence and Saarbrücken Graduate School of Computer Science and thank the anonymous reviewers for their suggestions.

References

Paolo Annesi and Roberto Basili. 2010. Cross-lingual alignment of FrameNet annotations through hidden Markov models. In *Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing*, CICLing'10, pages 12–25. Springer-Verlag.

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, Austin, TX.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, Boulder, Colorado, June. Association for Computational Linguistics.

John Blitzer, Ryan McDonal, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proc. Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of CoNLL-2005*, Ann Arbor, MI USA.

Lea Cyrus. 2006. Building a resource for studying translation shifts. *CoRR*, abs/cs/0606096.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609, Portland, Oregon, USA, June. Association for Computational Linguistics.

Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1–11, Jeju Island, Korea, July. Association for Computational Linguistics.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 369–377, Suntec, Singapore, August. Association for Computational Linguistics.

Ian J. Goodfellow, David Warde-Farley, Pascal Lamblin, Vincent Dumoulin, Mehdi Mirza, Razvan Pascanu, James Bergstra, Frédéric Bastien, and Yoshua Bengio. 2013. Pylearn2: a machine learning research library. *CoRR*, abs/1308.4214.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English dependency treebank 2.0. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel text. *Natural Language Engineering*, 11(3):311–325.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Bombay, India.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

Mikhail Kozhevnikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1200, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 62–72, Edinburgh, United Kingdom. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Sebastian Padó and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proc. 44th Annual Meeting of Association for Computational Linguistics and 21st International Conf. on Computational Linguistics*, ACL-COLING 2006, pages 1161–1168, Sydney, Australia.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*, May.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287.

David A Smith and Jason Eisner. 2009. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 822–831. Association for Computational Linguistics.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 2 of *HLT '11*, pages 682–686, Portland, Oregon. Association for Computational Linguistics.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proc. of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*, pages 477–487, Montréal, Canada.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Ivan Titov. 2011. Domain adaptation by constraining inter-domain variability of latent feature representation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 62–71, Portland, Oregon, USA, June. Association for Computational Linguistics.

Sara Tonelli and Emanuele Pianta. 2008. Frame information transfer from English to Italian. In *Proceedings of LREC 2008*.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of stanford dependencies. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 578–584, Sofia, Bulgaria, August. Association for Computational Linguistics.

Lonneke van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 299–304, Portland, Oregon, USA. Association for Computational Linguistics.

Chenhai Xi and Rebecca Hwa. 2005. A backoff model for bootstrapping resources for non-english languages. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 851–858, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42, Hyderabad, India, January. Asian Federation of Natural Language Processing.

Daniel Zeman, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2012. Hamledt: To parse or not to parse? In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).