Cross-lingual alignment transfer: a chicken-and-egg story?

Lauriane Aufrant, Guillaume Wisniewski, François Yvon

Cross-lingual alignment transfer: a chicken-and-egg story?

Lauriane Aufrant[1,2] and **Guillaume Wisniewski**[1] and **François Yvon**[1]
[1] LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91 403 Orsay, France
[2]DGA, 60 boulevard du Général Martial Valin, 75 509 Paris, France
{lauriane.aufrant,guillaume.wisniewski,francois.yvon}@limsi.fr

Abstract

In this paper, we challenge a basic assumption of many cross-lingual transfer techniques: the availability of word aligned parallel corpora, and consider ways to accommodate situations in which such resources do not exist. We show experimentally that, here again, weakly supervised cross-lingual learning techniques can prove useful, once adapted to transfer knowledge *across pairs of languages*.

1 Introduction

Supervised machine learning techniques lie at the core of many robust Natural Language Processing (NLP) systems and components. The dissemination of such methodologies is however hindered by the lack of appropriate supervision data, which are costly to produce and only available for a restricted number of genres, tasks, domains and languages. *Weakly supervised learning* techniques have emerged as an effective way to remedy, at least partially, to this unsatisfactory situation. Among them, *cross-lingual learning methods* enable to transfer useful supervision information from well-resourced to under-resourced languages, speeding up the development of NLP tools for new domains and tasks.

Many techniques for transferring knowledge across languages have been proposed in the literature (see § 2 for a brief overview). A widely-used methodology consists in generating automatic annotations for the resource-poor language by projecting linguistic information through word alignment links (see eg. (Yarowsky et al., 2001; Täckström et al., 2013) for PoS tagging, (Hwa et al., 2005; Lacroix et al., 2016a) for dependency parsing, (Ehrmann et al., 2011) for Named Entity Recognition, (Kozhevnikov and Titov, 2013) for Semantic Role Labeling, etc.). Implementing this methodology requires the existence of (a) parallel corpora aligned at the word level, and (b) annotation and/or tools on the resource-rich side. However, requirement (a) is somewhat paradoxical: reliable word alignments can only be computed for large-scale parallel corpora, a situation that is unlikely to happen for actual under-resourced languages.

In this study, we explore ways to overcome this paradox and consider techniques for *transferring alignment models or annotations across language pairs*, a task that has hardly been addressed in literature (see however (Wang et al., 2006; Levinboim and Chiang, 2015)). Based on a high-level typology of cross-lingual transfer methodologies (§ 2), our contribution is to formalize realistic scenarios (defined in § 3) as well as some basic methodologies for projecting knowledge about bilingual alignments cross-linguistically (§ 4). Experiments in § 5 show that, at least for some of these scenarios, simple-minded methods can be surprisingly effective and open a discussion on further prospects and perspectives for future work.

2 Techniques for cross-lingual transfer

In this section, we briefly review existing cross-lingual transfer techniques for various NLP applications, aiming at identifying techniques that could be adapted for transferring alignments. We will successively consider techniques that operate in the data space, then techniques that perform the transfer of

Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP, pages 35–44,
San Diego, California, June 17, 2016.

parameters. Note that for the sake of the presentation, the resource-rich is viewed as the source language, and the resource-poor is accordingly the target language.

2.1 Transfer in data space

This family of techniques seeks to automatically supply the annotations that are lacking on the target side, so that a model can be learned on these artificially generated data.

Direct Transfer Two main lines of reasoning have been considered: the first assumes that the source and target languages are sufficiently similar, to the point that source annotations can be readily used to train a model in the target language (Hana et al., 2004; Zeman and Resnik, 2008). When such assumption does not hold, a necessary preliminary step will be to map the source and target data in a shared representation space: *delexicalization*, i.e. the replacement of words with (universal) PoS (McDonald et al., 2013) readily yields such mappings (Wisniewski et al., 2014), but it is also conceivable to consider automatically inferred multilingual representations (Jagarlamudi et al., 2011; Kočiský et al., 2014; Gouws et al., 2015). This simple approach has one downside: learning can only use features based on this inter-lingual representation – in particular this makes it impossible to include powerful lexical features. Delexicalized training thus needs to be complemented by a relexicalization phase, where more informative features can then come into play (McDonald et al., 2011). In fact, in this situation, transfer is nothing but a specific case of domain adaptation (Blitzer, 2008), and can be handled with the same tools (semi-supervised learning, instance re-weighting, etc).

Annotation Projection The alternative is to preserve the lexical representations in each language, which are matched through word alignment in a parallel corpus. As mentioned above, this approach has been successful for many tasks. A very strong assumption is that words (or structures) that are mutual translations will carry identical annotations: this only holds for a restricted number of annotations and languages. For instance, it is commonly assumed that coarse-grained morphosyntactic tags can reasonably be projected between most West-European languages; such projections are less appropriate for fine-grained morphological information such as case or gender (as those distinctions greatly vary across languages), and would be even less so for pairs of languages having antagonist definitions of a word. Furthermore, its success will depend on the density and quality of the alignments (Lacroix et al., 2016b), meaning that it might be more suited to situations in which large bitexts are available. A possible workaround to the noisiness issue is to interpret transferred annotations as soft, rather than hard constraints: see e.g. (Ganchev et al., 2009; Das and Petrov, 2011; Li et al., 2014; Wang and Manning, 2014) for various implementations of this idea; or to combine it with another source of information (Täckström et al., 2013). Alignment projection is not only noisy: it also yields *incomplete annotations*, requiring methods that learn from partially annotated corpora (Wisniewski et al., 2014). A last strategy worth mentioning here for generating artificial annotations is to use Machine Translation (Tiedemann, 2014).

2.2 Transfer in parameter space

The second main family of techniques *use the same model for the source and target languages*: learned parameters in the former can then readily be used for the latter.

A first instance of model transfer has already been mentioned: indeed, taking source annotations to supervise the training in target can also be viewed as a (trivial) form of direct model transfer. This approach has been extended in many ways. Cohen et al. (2011) use several source languages and train one delexicalized model in each; the optimal convex combination of these models is used to process the target language. A variant of this strategy is to view the source parameter values as priors for the target model, an idea that has been used repeatedly in the context of domain adaptation. It has notably been used for transferring parsers (Cohen and Smith, 2009; Burkett et al., 2010; Berg-Kirkpatrick and Klein, 2010) and, more recently, to also transfer alignment models (Levinboim and Chiang, 2015).

This brief retrospective has demonstrated the variety of cross-lingual transfer techniques, many of which are borrowed from the domain adaptation literature. The applicability and success of these

methods depend on the task and of the available resources. We now explore ways to apply them for the word alignment task.

3 Word alignments: cross-lingual scenarios

After a quick review of standard algorithms for word alignment, we present situations in which they can be improved by cross-lingual knowledge.

3.1 Aligning words

The most popular models for statistical word alignment are the IBM models 1 to 6 (Brown et al., 1993; Och and Ney, 2003) and the HMM model of Vogel et al. (1996). These probabilistic generative models decompose the probability of an aligned sentence pair as the conjunction of a word translation model, a distortion model (models 2 and up) and a fertility model (models 3 and up). Distortion is absolute for models 2-3 and relative for models 4-6; in the HMM model, it is captured by Markovian dependencies between consecutive alignments links. Among these parameters, the translation model is lexicalized in both languages, fertility is lexicalized in the source side and distortion is unlexicalized but rely on word clusters for models 4-6. In all cases, parameters are learned in an unsupervised way using the EM algorithm.

Many refinements to these algorithms have been proposed, often to improve computational performance (Dyer et al., 2013). Another line of work tries to improve IBM and HMM models' low generalization power by using feature-based models (Moore, 2005; Berg-Kirkpatrick et al., 2010). However, the IBM models remain today the most widely used approach both because of their efficiency and because they do not require any annotated data. They will thus serve as our main baseline.

3.2 Real-world situations for alignment transfer

Scenarios for improving word alignment with cross-lingual transfer fall into two categories, depending on whether the source and target languages play a symmetric role.

We first consider the standard symmetric BRIDGE scenario: it involves two languages S and T, for which large bitexts with a 'bridge' language B exist, readily yielding reasonably-good alignment models for S-B and B-T. We are however specifically interested in the S-T pair, for which we only possess a small parallel corpus. This can happen either in the context of a bilingual task like translation or because word alignments are needed for cross-lingual transfer of a monolingual model. In this case, the purpose is to annotate the S-T data thanks to information contained in the high quality S-B and B-T models. Two variants provide interesting refinement opportunities: MULTIPARALLEL, in which part or all the S-T parallel data is also aligned with sentences in B, and RELATED, when all three languages belong to the same linguistic family. Taking advantage of such similarity however requires more expressive models than the IBM series, which only operate at the level of word forms and are therefore agnostic to lexical similarities.

We can illustrate the RELATED scenario on the example of morphosyntactic model transfer from Italian (it) to Romanian (ro), using annotation projection. For lack of a large it:ro corpus to compute robust word alignments, an option is to use French as a bridge, collect large bitexts for Italian-French and for French-Romanian, and improve the quality of the Italian-Romanian word alignment model thanks to the it:fr and fr:ro models. Even though the resulting it:ro bitext is noisier and smaller than the fr:ro one, which could also be used for transferring PoS labels, transfer to Romanian may still be more accurate when using Italian as an additional source, or even as a better source than French.

In the second type of scenarios, source and target languages play asymmetric roles. Bitexts for S-T (eg. English and Ukrainian) come with small parallel data, but there exists a language $\tilde{T}$ related to T (and unrelated to S) for which large parallel data with S are available (eg. Russian). We consequently consider transfer of S-$\tilde{T}$ word alignments to the S-T pair. This can be interpreted as standard cross-lingual transfer from $\tilde{T}$ to T, but with the difference that the transferred knowledge is not monolingual but bilingual because of the interactions with S. With large data in both S-$\tilde{T}$ and $\tilde{T}$-T, we call this scenario DIRECTED BRIDGE. This is for instance the context of Wang et al. (2006)'s works on English-Japanese, using Chinese as a bridge lan-

guage, but their cross-language word similarity does not exploit Chinese-Japanese linguistic similarity.

Finally, in the DIALECT scenario, T is a dialect of $\tilde{T}$, and even though parallel $\tilde{T}$-T data is not necessarily available, the transfer process can rely on the large number of common word forms. This would, for instance, be the case with the alignment of English with MS Arabic and dialects. Thanks to the large linguistic overlap, and contrarily to the previous scenarios, here again methods from the domain adaptation literature (Hua et al., 2005) may also successfully apply.

Before closing this section, we would finally like to stress the fact that the motivations for transferring alignments can be many: one might want to get alignments for a small parallel bitext, to then transfer other annotations, or one might want to bootstrap an alignment model with transferred parameters, or even to train a small SMT, etc. Each such motivation may call for different strategies.

4 Methods for transferring alignment

In this section, we exemplify with simple systems how general transfer methods can be instantiated for alignment transfer.

From now on, we focus on a DIRECTED BRIDGE scenario, further assuming that the task is to annotate a very small parallel corpus. We will simulate this situation, in Section 5, for the English-Swedish pair and choose Danish as a bridge language, as it is closely related to Swedish. Both English-Danish and Danish-Swedish pairs will be considered well-resourced.

We start with a straightforward baseline (CAT-DA), consisting in concatenating the English-Danish and the English-Swedish data before training. The underlying assumption is that both Danish and Swedish are subsumed by a Scandinavian meta-language. Despite their similarity, these languages only have few common word forms (Zeman and Resnik, 2008) and their vocabularies overlap mostly on function words, numbers and proper nouns. Consequently, the resulting translation model will in fact consist in two quite separate sub-models that hardly interact. A unique model is however trained for the unlexicalized parameters like distortion.

This approach corresponds to joint learning with parameter sharing. It can however be interpreted from another point of view: as the purpose of transfer at the data level is to produce noisy artificial training data, here we produce approximate English-Swedish sentence pairs, with the Danish set considered as a proxy to Swedish.

We continue along these lines and propose a second method to produce fake Swedish data: first train an IBM 1 model on the Danish-Swedish corpus to extract a translation model, then replace every Danish token of the English-Danish data with its most likely translation (leaving OOV tokens unchanged). Then again the resulting bitext is concatenated with the English-Swedish corpus and standard training ensues. We notice that a symmetrical procedure can be straightforwardly implemented, by using a Swedish-Danish IBM 1 model to translate the test set in Danish: alignment is performed in the Danish domain, but since there is an exact Swedish-Danish token correspondence, the resulting word alignment can be directly used for the English-Swedish part. This is an example where English-Swedish IBM models will not be delivered, and can hardly be re-estimated from so few sentence pairs. We denote these methods TR-DA and DA-TR respectively.

The same intuitions can finally be applied for transferring in the parameter space: direct transfer is achieved by training an English-Danish model, then using it directly to annotate the English-Swedish pairs (E step of the EM algorithm). This approach (denoted DA) is rather naive and is expected to perform poorly, but it can be improved in a similar manner to the 'glosses' method of (Zeman and Resnik, 2008), by translating the Swedish tokens into Danish with a Swedish-Danish IBM 1 model (a method denoted GLOSSES-DA). Using the converse translation model to translate the English-Danish lexicalized model parameters and thus produce a full English-Swedish model yields slightly different results (PARAM-DA). The six methods are summarized in Table 1.

Deriving other methods The purpose of those strategies is mostly to set baselines and qualitative analyses, and more complex alignment transfer methods can be designed, following the typology of § 2. Two restrictions apply however.

First, annotation projection can not be entertained

DATA SPACE	
CAT-L	concatenate en:L and test data; train
TR-L	word-for-word translate en:L data; concatenate with test data; train
L-TR	word-for-word translate test data in L; concatenate with en:L data; train
PARAMETER SPACE	
L	train an en:L model; apply on test data
GLOSSES-L	train an en:L model; apply on test data word-for-word translated in L
PARAM-L	train an en:L model; translate the parameters; apply on test data

Table 1: Summary of proposed methods, for bridge language L.

in any scenario: while annotation projection for a monolingual task needs parallel data, for a bilingual model it would require multiparallel data. The converse is not true however, and the MULTIPARALLEL scenario can be successfully exploited without annotation projection (Kumar et al., 2007).

Second, the delexicalized approach causes a chicken-and-egg situation in real-life scenarios. Indeed, when the target language is under-resourced, one cannot assume the availability of a PoS tagger that is needed to compute delexicalized representations. Conversely, methods like (Wisniewski et al., 2014)'s cross-lingual PoS tagger projection and (Täckström et al., 2012)'s clusters are not applicable without a word aligned corpus. Finding common, even coarse-grained, representations then becomes a huge obstacle in many scenarios where alignment transfer is needed, which makes this approach less relevant.

As a final note, we point out that the RELATED scenario can be simulated with two symmetric instances of DIRECTED BRIDGE interpolated in the data or parameter space. The straightforward strategies described here can therefore be extended to other scenarios.

5 Experiments

In this section, we experiment with the methods introduced above and compare them with standard unsupervised models of varying sizes.

Experimental setup We evaluate the proposed alignment transfer methods on the English-Swedish test set provided by Holmqvist and Ahrenberg (2011), which consists of 192 word aligned sentence pairs extracted from the English-Swedish part of Europarl (Koehn,). We score the methods according to the intrinsic Alignment Error Rate (AER) metric proposed by Och and Ney (2000).

As documented in a large body of literature (Lopez and Resnik, 2006; Fraser and Marcu, 2007; Lambert et al., 2009; Lambert et al., 2010), AER poorly correlates with translation quality of the systems trained on the evaluated alignments, especially for large corpora, and extrinsic metrics like the BLEU score should be preferred, were MT training the final goal of alignment.

The concatenation methods proposed here are intended for very small data, with large unbalance between the target and the bridge sets, a data size for which the SMT application is not relevant. Consequently, we use the PoS accuracy of a cross-lingual tagger (Wisniewski et al., 2014) weakly supervised by the word alignments as the extrinsic evaluation metric of our methods. In such a system, each extra sentence pair brings valuable knowledge, while incorrect links strongly noise the system, making the accuracy a direct indicator of alignment quality. Besides, this step completes a realistic low-resource scenario where word alignments are needed for an intended cross-lingual use.

We use the English-Swedish bitext both as a test set for intrinsic evaluation and as projection data to train cross-lingual taggers. PoS accuracies are computed on the coarse PoS tags of the Swedish test treebank of the Universal Dependencies 1.2 (McDonald et al., 2013) and the source English tagger is trained on the training portion of the same corpus. In every method where additional parallel data is required, we use Europarl without the Q4-2000 section, which is reserved for tests.

We evaluate AER and PoS accuracy for the

three concatenation methods presented in § 4, with transfer through Danish: CAT-DA, TR-DA, DA-TR, and the three parameter transfer methods: DA, GLOSSES-DA, PARAM-DA. To evaluate the benefits of using a related bridge language, we also run the concatenation experiments with transfer through Greek, that is only distantly related to Swedish and also uses a different alphabet (methods CAT-EL, TR-EL, EL-TR). Finally, the alignment performance is confronted with that of concatenation with English-Swedish data of various sizes, from no added pair (BASELINE) to full concatenation of the 1.8M sentence pairs in Europarl (CAT-SV).

Results Table 2 reports the AERs of the various methods when using IBM 1, HMM or IBM 4 models. The methods involving test set translation (DA-TR and GLOSSES-DA) consistently yield the best cross-lingual accuracies for each model, with a relative error reduction of 45%, 52% and 59% respectively for DA-TR and scores comparable to the full English-Swedish ones. Figure 1 reports those AERs along the learning curves of English-Swedish models for increasing data sizes. It shows that the DA-TR method yields alignment quality comparable to unsupervised learning on 0.1M to 0.5M sentence pairs.

The PoS accuracy measures are reported in Table 3. They show a clear correlation of the most effective (in AER) transfer methods with high PoS accuracies. However, this measure does not allow to clearly rank the top few models (CAT-SV, TR-DA, DA-TR and GLOSSES-DA). Notably, here CAT-DA and DA-TR respectively outperform BASELINE and CAT-SV. It may be that word alignments obtained by transfer are less accurate but focus more on general cross-lingual structures which, in turn, enables a better annotation projection, or that these score differences are simply not significant. Coming up with a reliable interpretation of this issue however will require further experiments that are beyond the scope of this paper.

Discussion Unsurprisingly, direct data transfer methods CAT-DA and CAT-EL perform at best in par with the baseline, and often worse. Indeed, because of mostly disjoint vocabularies, the translation model is globally not improved by the external knowledge, while training the Swedish parameters on few pairs compared to the whole data produces weak parameters that are easily subject to any noise coming from the few common word forms (see IBM 1 results). This approach could however perform better in a DIALECT scenario, assuming that the $\tilde{T}$ and T languages have large enough common vocabularies to structure the alignment sets, and that shared word forms are not accidental but actually correspond to common words that can be reliably exploited to transfer knowledge. The absolute AER gap reduces when adding shared distortion and fertility models, but this does not allow us to conclude on a positive or negative effect of unlexicalized parameter sharing. This suggests however the need for experimenting with more selective parameter sharing.

Compared to the Danish ones, we notice that the Greek systems yield smaller but still significant improvements over the baseline, even if (a) Greek and Swedish are only distantly related and have therefore less common structures (b) the impact of untranslated OOV words is higher in Greek because of the change of alphabet: indeed, it is quite unlikely to find identical word forms in Swedish and Greek. The Greek systems have also been trained on slightly smaller bitexts (1.2M pairs, compared to 1.9M for English-Danish), but this ratio corresponds to a loss of at most 1 AER point in standard learning, which remains negligible in comparison to the cost of choosing Greek over Danish. All in all, using Greek as a bridge language is comparable to training with 10,000 English-Swedish pairs, instead of 400,000 when using Danish. This suggests that our methods can be useful even for unrelated languages in the BRIDGE scenario, even though the return ratio is much lesser than when languages are related: for Greek, 1.2M parallel sentences used cross-linguistically yield the same accuracy as 10,000 parallel sentences of the targeted pair (a ratio of approximately 1%); for Danish the ratio is closer to 20% (1.9M sentences providing the same information as 400,000 en:sv sentence pairs).

Finally, the comparison of the TR-DA and DA-TR columns shows that English-Swedish alignment biased by the English-Danish alignment is more accurate when performed in the Danish domain than in the Swedish one. Intuitively, the noisy application of a valid model performs better than a valid application of a noisy model. Columns PARAM-DA and

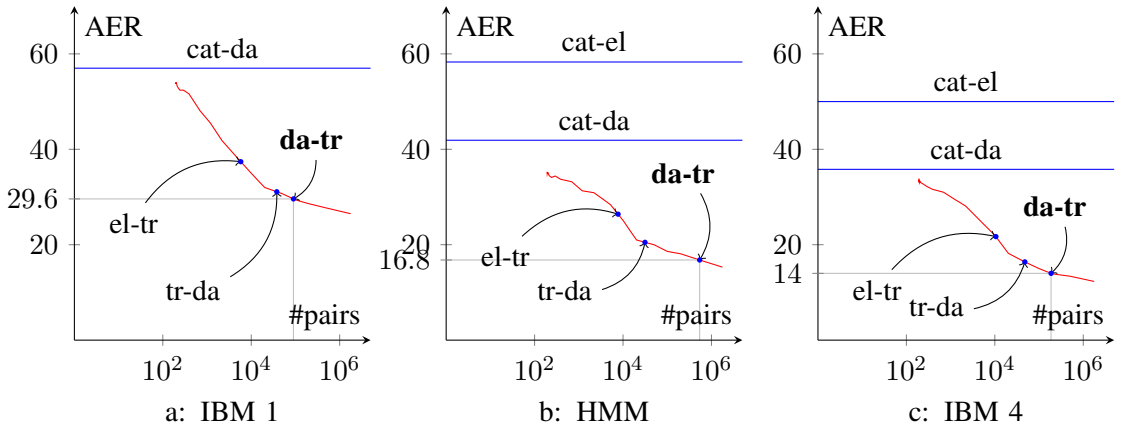

Figure 1: AER on the English-Swedish test set for increasing data sizes (red curve) and some cross-lingual methods (reported in blue along the curve). The number of English-Swedish sentence pairs includes the 192 test pairs.

	Swedish only		Danish data			Greek data			Danish parameters		
	baseline	cat-sv	cat-da	tr-da	da-tr	cat-el	tr-el	el-tr	da	glosses-da	param-da
IBM 1	53.9	**26.5**	57.0	31.1	**29.6**	74.3	**35.9**	37.4	65.94	**28.3**	33.29
HMM	35.3	**15.3**	41.9	20.5	**16.8**	58.3	26.9	**26.4**	46.74	**16.4**	25.79
IBM 4	33.9	**12.3**	35.8	16.4	**14.0**	50.0	**20.6**	21.7	49.08	**14.8**	24.34

Table 2: AER achieved by the proposed cross-lingual methods with Danish and Greek as bridge languages, compared to the baseline (unsupervised alignment on test data only) and the cat-sv higher bound (addition of large English-Swedish data).

	Swedish only		Danish data			Greek data			Danish parameters		
	baseline	cat-sv	cat-da	tr-da	da-tr	cat-el	tr-el	el-tr	da	glosses-da	param-da
IBM 1	68.7	**73.3**	58.7	73.8	**74.0**	47.4	**71.9**	71.5	66.97	**72.20**	71.07
HMM	69.9	**73.8**	71.9	73.5	**73.6**	66.6	**73.4**	71.9	69.54	**73.42**	72.43
IBM 4	73.0	**74.7**	74.0	73.9	**74.9**	72.0	73.4	**73.5**	66.67	**73.56**	71.96

Table 3: Extrinsic cross-lingual PoS accuracies achieved by the proposed cross-lingual methods with Danish and Greek as bridge languages, compared to the baseline (unsupervised alignment on test data only) and the cat-sv higher bound (addition of large English-Swedish data).

GLOSSES-DA also support that interpretation, and the fact that among the evaluated strategies, the most refined data transfer models outperform the parameter ones shows that even from a very small piece of target data, it is still possible to extract valuable knowledge to guide model adaptation to a new language.

6 Conclusion

In this work, we have presented a brief typology of general cross-lingual transfer methods and have shown how it can apply on a poorly addressed task, the transfer of bilingual knowledge. We present a few realistic scenarios where transfer of word alignment is needed and focus on one of them in the frame of unsupervised word alignment, to propose six cross-lingual methods that are easy to set up.

Experiments on an English-Swedish test set re-

veal that even straightforward methods can extract valuable information for weak supervision: from five sentence pairs in the bridge language, they are able to extract knowledge equivalent to one English-Swedish pair. Altogether we achieve up to 59% relative error reduction. Further analyses also provide precious hints for accurate designs of alignment transfer methods.

In future work, we intend to further explore the benefits of language similarity in the RELATED, DIRECTED BRIDGE and DIALECT scenarios, along two tracks: weighting based on linguistic similarity during the EM training and selective transfer at the sub-model level.

Acknowledgments

This work has been partly funded by the French *Direction générale de l'armement*. We thank the anonymous reviewers for their detailed comments on the paper.

References

Taylor Berg-Kirkpatrick and Dan Klein. 2010. Phylogenetic Grammar Induction. In *The 48th Annual Meeting of the Association for Computational Linguistics*, pages 1288–1297, Uppsala, Sweden.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590.

John Blitzer. 2008. *Domain Adaptation of Natural Language Processing Systems*. Ph.D. thesis, University of Pennsylvania.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, pages 263–311.

David Burkett, Slav Petrov, John Blitzer, and Dan Klein. 2010. Learning Better Monolingual Models with Unannotated Bilingual Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 46–54.

Shay Cohen and Noah A. Smith. 2009. Shared Logistic Normal Distributions for Soft Parameter Tying in Unsupervised Grammar Induction. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 74–82, Boulder, Colorado.

Shay B. Cohen, Dipanjan Das, and Noah A. Smith. 2011. Unsupervised Structure Prediction with Non-Parallel Multilingual Guidance. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 50–61, Edinburgh, Scotland, UK., July.

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 - Volume 1*, HLT '11, pages 600–609.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of NAACL 2013, the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a Multilingual Named Entity-Annotated Corpus Using Annotation Projection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 118–124, Hissar, Bulgaria.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, pages 293–303.

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: Volume 1 - Volume 1*, ACL '09, pages 369–377.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In David Blei and Francis Bach, editors, *Proceedings of the 32nd International Conference on Machine Learning*, pages 748–756. JMLR Workshop and Conference Proceedings.

Jiri Hana, Anna Feldman, and Chris Brew. 2004. A Resource-light Approach to Russian Morphology: Tagging Russian using Czech resources. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 222–229, Barcelona, Spain, July.

Maria Holmqvist and Lars Ahrenberg. 2011. A Gold Standard for English–Swedish Word Alignment. In *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA*, pages 106–113.

Wu Hua, Wang Haifeng, and Liu Zhanyi. 2005. Alignment model adaptation for domain-specific word alignment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 467–474.

Rebecca Hwa, Philip Resnik, A.Weinberg, C. Cabezas, and O. Kolak. 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural language engineering*, 11:311–325.

Jagadeesh Jagarlamudi, Raghavendra Udupa, Hal Daume III, and Abhijit Bhole. 2011. Improving Bilingual Projections via Sparse Covariance Matrices. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 930–940, Edinburgh, Scotland, UK.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation, year = 2005. In *2nd Workshop on EBMT of MT-Summit X*, pages 79–86, Phuket, Thailand.

Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning Bilingual Word Representations by Marginalizing Alignments. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–229, Baltimore, Maryland, June.

Mikhail Kozhevnikov and Ivan Titov. 2013. Bootstrapping Semantic Role Labelers from Parallel Data. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 317–327, Atlanta, Georgia, USA.

Shankar Kumar, Franz Josef Och, and Wolfgang Macherey. 2007. Improving Word Alignment with Bridge Languages. In *EMNLP-CoNLL*, pages 42–50.

Ophélie Lacroix, Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2016a. Frustratingly Easy Cross-Lingual Transfer for Transition-Based Dependency Parsing. In *The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL 2016, San Diego, California, USA.

Ophélie Lacroix, Guillaume Wisnewski, and François Yvon. 2016b. Cross-lingual Dependency Transfer: What Matters? Assessing the Impact of Pre- and Post-processing. In *Proceedings of the NAACL-16 Workshop on Multilingual and Crosslingual Methods in NLP*, MLCL 2016, San Diego, CA, USA. Association for Computational Linguistics.

Patrik Lambert, Yanjun Ma, Sylwia Ozdowska, and Andy Way. 2009. Tracking relevant alignment characteristics for machine translation. In *Proceedings of Machine Translation Summit XII*, pages 268–275.

Patrik Lambert, Simon Petitrenaud, Yanjun Ma, and Andy Way. 2010. Statistical analysis of alignment characteristics for phrase-based machine translation. In *Proceedings of the 14th European Association for Machine Translation*.

Tomer Levinboim and David Chiang. 2015. Multi-Task Word Alignment Triangulation for Low-Resource Languages. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1221–1226.

Zhenghua Li, Min Zhang, and Wenliang Chen. 2014. Soft Cross-lingual Syntax Projection for Dependency Parsing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 783–793, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Adam Lopez and Philip Resnik. 2006. Word-Based Alignment, Phrase-Based Translation: What's the Link? In *Proceedings of AMTA*, pages 90–99.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source Transfer of Delexicalized Dependency Parsers. In *Proceedings of EMNLP 2011, the Conference on Empirical Methods in Natural Language Processing*, pages 62–72.

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.

Robert C. Moore. 2005. A discriminative framework for bilingual word alignment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 81–88.

Franz Josef Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 1086–1090.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, pages 19–51.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 477–487, Stroudsburg, PA, USA.

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.

Jörg Tiedemann. 2014. Rediscovering Annotation Projection for Cross-Lingual Parser Induction. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING): Technical Papers*, pages 1854–1864, Dublin, Ireland.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-Based Word Alignment in Statistical Translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 836–841.

Mengqiu Wang and Christopher D. Manning. 2014. Cross-lingual Projected Expectation Regularization for Weakly Supervised Learning. *Transactions of the ACL*, 2(1):55–66.

Haifeng Wang, Hua Wu, and Zhanyi Liu. 2006. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 874–881.

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. 2014. Cross-Lingual Part-of-Speech Tagging through Ambiguous Learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1779–1785.

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*, HLT '01, pages 1–8.

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.