Rediscovering Annotation Projection for Cross-Lingual Parser Induction

Jörg Tiedemann
Uppsala University
Department of Linguistics and Philology
firstname.lastname@lingfil.uu.se

Abstract

Previous research on annotation projection for parser induction across languages showed only limited success and often required substantial language-specific post-processing to fix inconsistencies and to lift the performance onto a useful level. Model transfer was introduced as another quite successful alternative and much research has been devoted to this paradigm recently. In this paper, we revisit annotation projection and show that the previously reported results are mainly spoiled by the flaws of evaluation with incompatible annotation schemes. Lexicalized parsers created on projected data are especially harmed by such discrepancies. However, recently developed cross-lingually harmonized annotation schemes remove this obstacle and restore the abilities of syntactic annotation projection. We demonstrate this by applying projection strategies to a number of European languages and a selection of human and machine-translated data. Our results outperform the simple direct transfer approach by a large margin and also pave the road to cross-lingual parsing without gold POS labels.

1 Introduction

Linguistic resources and tools exist only for a minority of the world's languages. However, many NLP applications require robust tools and the development of language-specific resources is expensive and time consuming. Many of the common tools are based on data-driven techniques and they often require strong supervision to achieve reasonable results for real world applications. Fully unsupervised techniques are not a good alternative yet for tasks like data-driven syntactic parsing and, therefore, cross-lingual learning has been proposed as a possible solution to quickly create initial tools for otherwise unsupported languages (Ganchev and Das, 2013).

In syntactic parsing, two main strategies have been explored in cross-lingual learning: annotation projection and model transfer. The first strategy relies on parallel corpora and automatic word alignment that make it possible to map linguistics annotation from a source language to a new target language (Yarowsky et al., 2001; Hwa et al., 2005; Täckström et al., 2013a). The basic idea is that existing tools and models are used to process the source side of a parallel corpus and that projection heuristics guided by alignment can be used to transfer the automatic annotation to the target language text. Using the projected annotation assuming that it is sufficiently correct, models can then be trained for the target language. However, directly projecting syntactic structure results in a rather poor performance when applied to resources that were developed separately for individual languages (Hwa et al., 2005). Extensive additional post-processing in form of transformation rules is required to achieve reasonable scores. Furthermore, incompatible tagsets make it impossible to directly transfer labeled annotation to a new language and previous literature on cross-lingual parsing via annotation projection is, therefore, bound to the evaluation of unlabeled attachment scores (UAS). Less frequent, but also possible, is the scenario where the source side of the corpus contains manual annotation (Agić et al., 2012). This addresses the problem created by projecting noisy annotations, but it presupposes parallel corpora with manual annotation, which are rarely available. Additionally, the problem of incompatible annotation still remains.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1854–1864, Dublin, Ireland, August 23-29 2014.

The second strategy, model transfer instead relies on universal features and the transfer of model parameters from one language to another. The main idea is to reduce the need of language-specific information, e.g. using delexicalized parsers that ignore lexical information. Drawing from a harmonized POS tagset (Petrov et al., 2012), transfer models have been used for a variety of languages. The advantage over annotation projection approaches is that no parallel data is required (at least in the basic settings) and that training can be performed on gold standard annotation. However, it requires a common feature representation across languages (McDonald et al., 2013), which can be a strong bottleneck. There are also several extensions to improve the performance of transfer models. One idea is to use multiple source languages to increase the statistical ground for the learning process (McDonald et al., 2011; Naseem et al., 2012), a strategy that can also be used in the case of annotation projection. Another idea is to enhance models by cross-lingual word clusters (Täckström et al., 2012) and to use target language adaptation techniques with prior knowledge of language properties and their relatedness when using multiple sources in training (Täckström et al., 2013b). Based on the success of these techniques, model transfer has dominated recent research on cross-lingual learning.

In this paper, we return to annotation projection as a powerful tool for porting syntactic parsers to new languages. Building on the availability of cross-lingually harmonized data sets, we show that projection performs well and outperforms direct transfer models by a large margin in contrast to previous findings on projection with incompatible treebanks. In the following, we first revisit the projection algorithms proposed earlier and discuss issues with transferring labels across languages. After that we report experimental results with various settings using human translations and machine-translated data. Finally, we also look at parsing results without gold standard POS labeling, which is ultimately required when porting parsers to new languages that lack appropriate resources.

2 Syntactic Annotation Projection

Hwa et al. (2005) propose a direct projection algorithm for syntactic dependency annotation. The algorithm defines several heuristics to map source side annotations to target languages using word alignments in a parallel corpus. The main difficulties with the projection arise with none-one-to-one links and unaligned tokens. Each of the following alignment types are adressed by the algorithm separately:

one-to-one: Copy relations $R(s_i, s_j)$ between source words s_i and s_j to relations $R(t_x, t_y)$ if s_i is aligned to t_x and s_j is aligned to t_y and nothing else.

unaligned source: Create an empty (dummy) word in the target language sentence that takes all relations (incoming and outgoing arcs) of the unaligned source language word.

one-to-many: Create an empty target word t_z that acts as the parent of all aligned target words $t_x, .., t_y$. Remove the alignments between s_i and $t_x, .., t_y$ and align s_i to the new empty word t_z instead.

many-to-one: Delete all alignments between $s_i, .., s_j$ and t_x except the link between the head of $s_i, .., s_j$ and t_x.

many-to-many: Perform the rule for one-to-many alignments first and then perform the rule for many-to-one alignments.

unaligned target: Remove all unaligned target words.

In contrast to Hwa et al. (2005), we are also interested in labeled attachment and the projection of POS annotation. Therefore, we copy labels through the alignment using the heuristics listed above. Figure 1 illustrates some of the cases discussed. There are some important implications due to the treatment of complex alignment types. The direct projection algorithm frequently creates dummy nodes and relations that have no correspondence in the source language. Here, we need to make some decisions on how to project the annotation from source to target sentences.

First of all, we decided to name all additional tokens created by the algorithm with the same string *DUMMY*. An alternative would be to invent unique names for each newly created token within each sentence but this would blow up the vocabulary and would not add useful information to the data.

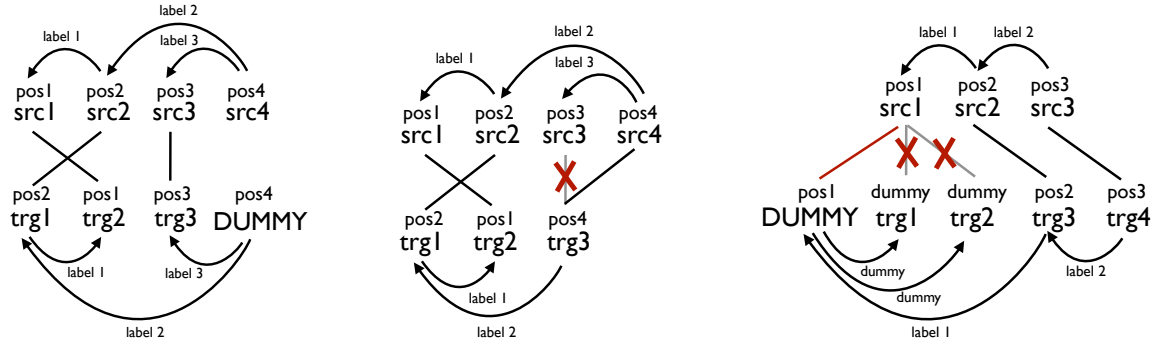

Figure 1: Annotation projection heuristics for special alignment types: Unaligned source words (left image), many-to-one alignments (center), one-to-many alignments (right image).

The second problem is related to the auxiliary relations that are created when treating one-to-many alignments. In these cases, multiple words are attached to newly created dummy nodes. However, no corresponding labels exist in the source language that would allow us to infer appropriate labels for these additional attachments. One possibility would be to use a specific label from the existing set of dependency relations, for example 'mwe'. However, one-to-many alignments do not always refer to proper multi-word expressions but often represent other grammatical or structural differences like the relation between the English preposition *'of'* which is linked together with the determiner *'the'* to the German determiner *'der'* in sentences like *'Resumption OF THE session'* translated to German *'Wiederaufnahme DER Sitzung'*. Therefore, we decided to label these additional dependency with a new unique label *dummy* instead of selecting an existing one.

Yet another problem arises with the projection of POS annotation. Similar to the labeling of dependency relations, we have to decide how to transfer POS tags to the target language in cases of one-to-many alignments. In our implementation, we transfer the source language label only to the newly created dummy node which dominates all target language words linked to the source language word in the projected dependency tree. The daughter nodes, however, obtain the label *dummy* even as their POS annotation. Alternatively, we may project the POS tag to all linked tokens according to the original alignment but our guiding principle is to resolve link ambiguity first using the heuristics in the direct projection algorithm and then to transfer annotation.

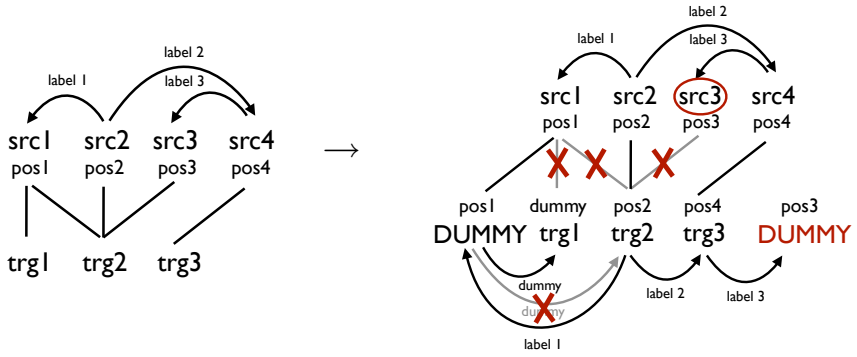

Figure 2: A complex example for annotation projection with many-to-many word alignments.

Finally, we also need to look at the interaction between the various projection heuristics. Figure 2 illustrates a complex case with many-to-many word alignments. Resolving the alignment ambiguity is not entirely straightforward. In our implementation, we start by looking at all one-to-many alignments and resolve them according to the definitions of the projection algorithm. In our example, this creates a *DUMMY* node that dominates target words *trg1* and *trg2* and links between *src1* and (*trg1*,*trg2*) are deleted. We label the new relations with *dummy*. The next step considers many-to-one alignments,

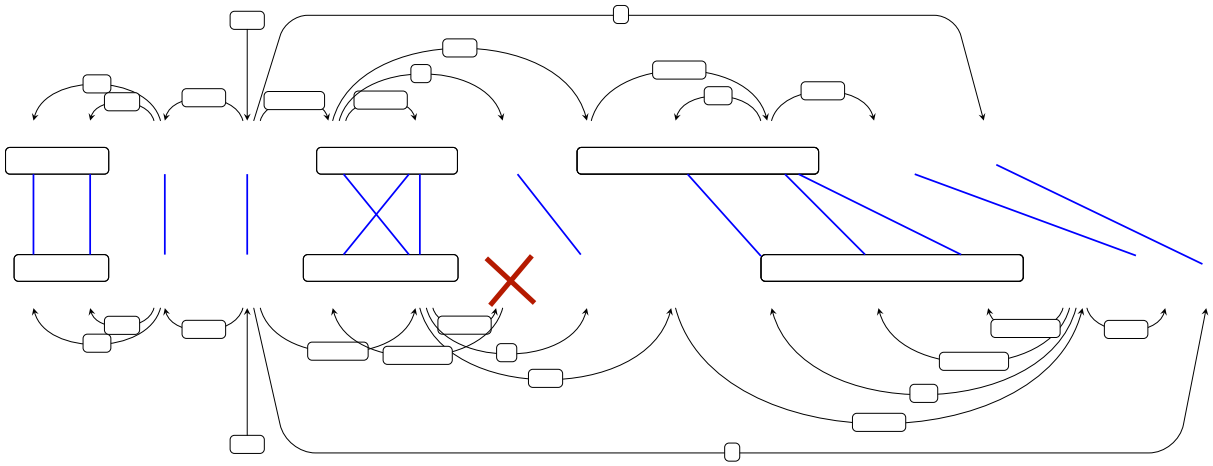

Figure 3: A complete projection example from a translated treebank including transitive relations over a *DUMMY* node that can safely be collapsed (which also removes the non-projectivity of the projected tree). The resulting relation between *quality* and *high-* will be labeled as *adpobj*. Note that projection errors appear due to the ambiguous alignments between *de qualité* and *high- quality*. Boxes indicate phrases that are translated as units by the SMT engine.

which, using the remaining links, is source words (*src2*,*src3*) aligned to *trg2*. According to the algorithm we delete the link between *src3* and *trg2* (because *src2* dominates *src3* in the source language tree) and proceed. This, however, creates an unaligned source language word (*src3*), which we treat in the next step. The unaligned token gives rise to the second *DUMMY* word, which is attached to *trg3* as the result of the alignment between *src4* and *trg3* and the relation between *src4* and *src3*. Finally, we can map all other relations according to the one-to-one alignment rule. This, however, creates a conflict with the already existing *dummy* relation between the first *DUMMY* word and *trg2*. Mapping according to the one-to-one rule turns the relation around and attaches the *DUMMY* word to *trg2* and labels the relation with *label 1*. Now, we could remove the second *DUMMY* node according to the rule about unaligned target language words. However, this rule should not apply to these special nodes as they may play a crucial role to keep elements connected in the final target language tree.

Another difficult case, which is not illustrated here, is when many-to-one alignments need to be resolved but the aligned source language words are siblings in the syntactic tree and no unique head can be identified. In our implementation, we randomly pick a node but more linguistically informed guesses would probably be better. Yet another difficult decision is the placement of the *DUMMY* nodes. We decided to put them next to the head node they attach to. Other heuristics are possible and all placements greatly influence the projectivity of the resulting tree structure. One final adjustment that we apply is the removal of unary productions over *DUMMY* nodes. We collaps all relations that run with single attachments via *DUMMY* nodes to reduce the number of these uninformative tokens. This may also have positive effects on projectivity as we can see in the example in Figure 3.

3 Machine-Translated Treebanks

Another strategy for annotation projection is based on automatic translation. Machine translation models can be used to create synthetic parallel data for projecting annotations from one language to another (Tiedemann et al., 2014). Recent advances in machine translation (MT) are now making this a realistic alternative. The use of direct treebank translation instead of existing parallel corpora has several important advantages. First of all, we skip the use of an error-prone annotation step when producing the source language side of the training data. Starting with a noisy source language annotation, we accumulate two sources of errors in annotation projection. However, with direct translation we can start with the gold standard annotation provided in the original treebank. Furthermore, we avoid problems of domain shifts which is typically the case when applying a parser trained on one domain to texts (a parallel corpus in

DELEXICALIZED	DE	EN	ES	FR	SV
DE	62.71	43.20	46.09	46.09	50.64
EN	46.62	77.66	55.65	56.46	**57.68**
ES	44.03	46.73	68.21	**57.91**	53.82
FR	43.91	46.75	**59.65**	67.51	52.01
SV	**50.69**	**49.13**	53.62	51.97	70.22

MCDONALD ET AL. (2013)	DE	EN	ES	FR	SV
DE	64.84	47.09	48.14	49.59	53.57
EN	48.11	78.54	56.86	58.20	**57.04**
ES	45.52	47.87	70.29	**63.65**	53.09
FR	45.96	47.41	**62.56**	73.37	52.25
SV	**52.19**	**49.71**	54.72	54.96	70.90

Table 1: Baselines – labeled attachment score (LAS) for delexicalized transfer parsing; results of McDonald et al. (2013) included for reference.

our case) coming from another domain. Finally, we can also assume that machine translation produces output which is closer to the original text than most human translations will be in any parallel corpus. Even if this may sound as a disadvantage, for projection this is preferred. Being close to the original source makes it easier to map annotation from one language to another as we expect a lower degree of grammatical and structural divergences that originate in the linguistic freedom human translators can apply. Furthermore, common statistical MT models inherently provide alignments between words and phrases, which removes the requirement to apply yet another error-prone alignment step on the parallel data. In the experiments below we, therefore, explore the translation strategy as yet another way of applying annotation projection.

4 Experiments

In the following, we show our experimental results using annotation projection in several cross-lingual scenarios. However, we start by presenting a delexicalized baseline, which is, to our knowledge, the only previous model that has been presented for labeled dependency parsing across languages using the recently created Universal Treebank. We will use this baseline as reference point even though our projection models are not directly comparable with delexicalized direct transfer models. Note that all results below are computed on the held-out test data sections of the Universal Treebank if not stated otherwise.

4.1 Delexicalized Baselines

McDonald et al. (2013) present the Universal Treebank that comes with a harmonized syntactic annotation scheme across six languages. This data set enables cross-lingual learning of labeled dependency parsing models. McDonald et al. (2013) propose delexicalized models as a simple baseline for model transfer and present encouraging labeled attachment scores (LAS) especially for closely related languages. As a reference, we have created similar baseline models using the same data set but a slightly different setup, which is compatible with the experiments we present later. Table 1 summarizes the scores in terms of LAS for all language pairs in the data set.[1] In our setup, we apply MaltParser (Nivre et al., 2006) and optimize feature models and learning parameters using MaltOptimizer (Ballesteros and Nivre, 2012). For all cross-lingual experiments (columns represent target languages we test on), we always use the same feature model and parameters as we have found for the source language treebank. Contrasting our models with the scores from McDonald et al. (2013), we can see that they are comparable with some differences that are due to the tools and learning parameters they apply which are along the lines of Zhang and Nivre (2011).

4.2 Annotation Projection with Human Translations

Our first batch of projection experiments considers parallel data taken from the well-known Europarl corpus, which is frequently used in research on statistical machine translation (SMT). It contains large quantities of translated proceedings from the European Parliament for all but one language (namely

[1]Note that we include punctuation in our evaluation. Ignoring punctuation leads to slightly higher scores but we do not report those numbers here.

UAS on CoNLL data

	DE	EN	ES	SV
DE	–	41.60	47.89	**58.80**
EN	49.67	–	51.44	58.66
ES	46.14	37.78	–	52.53
SV	**57.99**	**51.57**	**57.25**	–

UAS on Universal Treebank data

	DE	EN	ES	SV
DE	–	56.21	65.18	70.27
EN	63.17	–	68.02	70.40
ES	61.98	56.16	–	**71.06**
SV	**64.78**	**58.93**	**69.15**	–

Table 2: Unlabeled attachment scores for projected treebank models; comparing CoNLL data to Universal Treebank data for evaluation.

Korean) that are included in the Universal Treebank v1. The entire corpus (version 7) contains over two million sentences in each language and we use increasing amounts of the corpus to investigate the impact on cross-lingual parser induction. The corpus comes with automatic sentence alignments and is quite clean with respect to translation quality and sentence alignment accuracy. It is, therefore, well suited for our initial experiments with annotation projection even though the domain does not necessarily match the one included in the treebank test sets.

Another important prerequisite for annotation projection is word alignment. Following the typical setup, we rely on automatic word alignment produced by models developed for statistical machine translation. Similar to Hwa et al. (2005), we apply GIZA++ (Och and Ney, 2003) to align the corpus for all language pairs in all translation directions using IBM model 4 Viterbi alignments. In contrast to Hwa et al. (2005), we then use symmetrization heuristics to combine forward and backward alignments, which is common practice in the SMT community. In particular, we apply the popular grow-diag-final-and heuristics as implemented in the Moses toolbox (Koehn et al., 2007).

Let us first look at unlabeled attachment scores to compare results that can be achieved with harmonized annotation in contrast to the ones that we can see on the cross-lingually incompatible data from the CoNLL shared task (Buchholz and Marsi, 2006). Table 2 lists the scores that we obtain when applying our implementation of the direct projection algorithm.[2] As expected, the performance on the CoNLL data is rather poor, which confirms the findings of Hwa et al. (2005) even though our scores are significantly above their results without post-correction. The scores on the Universal Treebank data, however, are up to about 20 UAS points higher than the corresponding results on CoNLL data but without any of the extensive post-processing transformations proposed by Hwa et al. (2005).

LAS on Universal Treebank data

	DE	EN	ES	FR	SV
DE	–	49.44	56.58	58.75	61.04
EN	**56.59**	–	60.07	62.78	**62.15**
ES	54.04	47.90	–	**65.50**	61.45
FR	53.93	**51.23**	**65.03**	–	58.71
SV	56.13	49.18	60.82	62.00	–

data set: 40,000 sentences

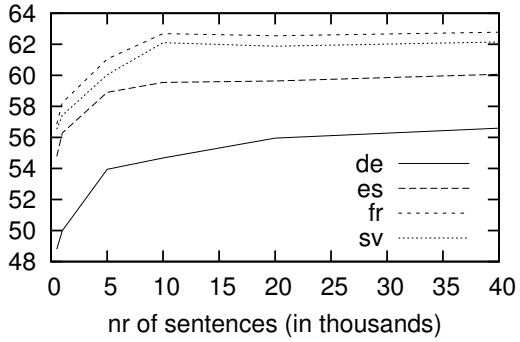

Figure 4: Annotation projection on Europarl data: LAS for induced parser models. The Figure to the right plots the learning curves for increasing training data for projections from English to the other languages.

Moreover, the real power of the harmonized annotation in the Universal Treebank comes from the possibility to obtain attachment labels. The table in Figure 4 shows the labeled attachment scores obtained for training on 40,000 sentences[3] of each language pair. Next to the table in Figure 4 we also show the

[2]We leave out French in this comparison as there is no French treebank in the CoNLL data.

[3]Note that there may be repeated sentences in the data.

with original source side annotation					
	DE	EN	ES	FR	SV
DE	–	53.02	54.96	58.20	59.65
EN	52.93	–	61.25	64.58	**63.82**
ES	50.88	50.28	–	**66.17**	60.48
FR	50.46	**53.95**	**65.46**	–	59.05
SV	**53.69**	51.51	60.58	60.19	–

jackknifing for source side annotation					
	DE	EN	ES	FR	SV
DE	–	50.27	54.91	56.00	57.91
EN	52.65	–	61.28	63.86	**63.72**
ES	49.19	50.04	–	**64.43**	59.65
FR	49.37	**53.25**	**64.41**	–	57.78
SV	**54.83**	50.25	60.27	60.04	–

Table 3: Cross-lingual parsing results (LAS) using translated treebanks (phrase-based model) and DCA-based annotation projection. The table to the left contrasts the result with two-sample jackknifing experiments where the source side dependencies are created by automatically parsing each half of the treebank using a model trained on the other half of the training data.

learning curves for increasing amounts of training data using the example of data projected from English to other languages. The figure illustrates that the LAS levels out at around 10,000 - 20,000 sentences and this trend is essentially the same for all other languages as well.

4.3 Annotation Projection with Synthetic Machine-Translated Data

The next possibility we would like to explore is the use of synthetic parallel data. Annotating parallel data with a statistical parser may lead to quite a lot of noise especially when the domain does not match the original training data. Starting with noisy source language annotations, the projection algorithm may transfer errors to the target language that can cause problems for the target language parsing model induced from that data. Using machine translation and the original source language treebanks, we avoid this kind of error propagation. Furthermore, we suspect that human translations are more difficult to align on the word level then machine translated data which are inherently based on word alignments and, therefore, tend to be more literal and consistent (Carpuat and Simard, 2012). Using statistical MT as our translation model, we can also obtain such alignment as a given output from the decoding process, which makes it unnecessary to run yet another error-prone process such as automatic word alignment. Furthermore, the treebank data is too small to be used alone with generative statistical alignment models. Concatenating the data with larger parallel data would help but domain mismatches may, again, negatively influence the alignment performance.

In the following, we show the cross-lingual scores obtained by translating all treebanks in the Universal Treebank to all other languages. We leave out Korean here again, because no SMT training data is included in Europarl for that language. The translation models are trained on the entire Europarl corpus using a standard setup for phrase-based SMT and the Moses toolbox for training, tuning and decoding (Koehn et al., 2007). For tuning we use MERT (Och, 2003) and the newstest 2012 data provided by the annual workshop on statistical machine translation.[4] and for language modeling, we use a combination of Europarl and News data provided from the same source. The language model is a standard 5-gram model estimated from the monolingual data using modified Kneser-Ney smoothing without pruning (applying KenLM tools (Heafield et al., 2013)).

Table 3 summarizes the labeled attachment scores obtained with our projection approach on synthetic machine-translated data. The main observation we can make here is that this approach is very robust with respect to the noise introduced by the translation engine. Automatic translation is a difficult task on its own but we still achieve results that are similar to the ones from the projection approach on human translated data. Note that our training data is now much smaller[5] compared to the data sizes used in Section 4.2 and, still, we outperform those models in several cases. This seems to prove that it can be a clear advantage to start with gold annotations in the source language and to have a close alignment between source and target language. An indication for this effect is illustrated by the contrastive jackknifing experiments shown in Table 3. The scores are generally lower with two minor exceptions. Note

[4] http://www.statmt.org/wmt14

[5] Most treebanks includes 2,000-5,000 sentences, except English with about 40,000 sentences.

that this experiment does not cover domain shift problems. Another trend that can be seen in our results is that some languages such as German are more difficult to translate to (which can be confirmed by the SMT literature) leading to lower cross-lingual parsing performance.

4.4 The Impact of Word Alignment

Crucial for the success of annotation projection is the quality of the word alignment used to map information from the source to the target language. Not only alignment errors cause problems but also ambiguous alignments can lead to projection difficulties as we have discussed before. In the previous sections, we relied on symmetrized word alignments that are common in the SMT community, which are based on Viterbi alignments created by the final IBM model 4 in the typical training pipeline. Even though this is a reasonable setup for training phrase-based SMT models (as presented in the previous section), the chosen symmetrization heuristics (grow-diag-final-and) may not be well suited for accurate annotation projection. In particular, it is known that these heuristics focus on recall and tend to add many additional links that may not be useful for our projection task and even lead to some confusion as depicted in the example in Figure 3.

In order to investigate the impact of word alignment, we, therefore, decided to look at other sym-

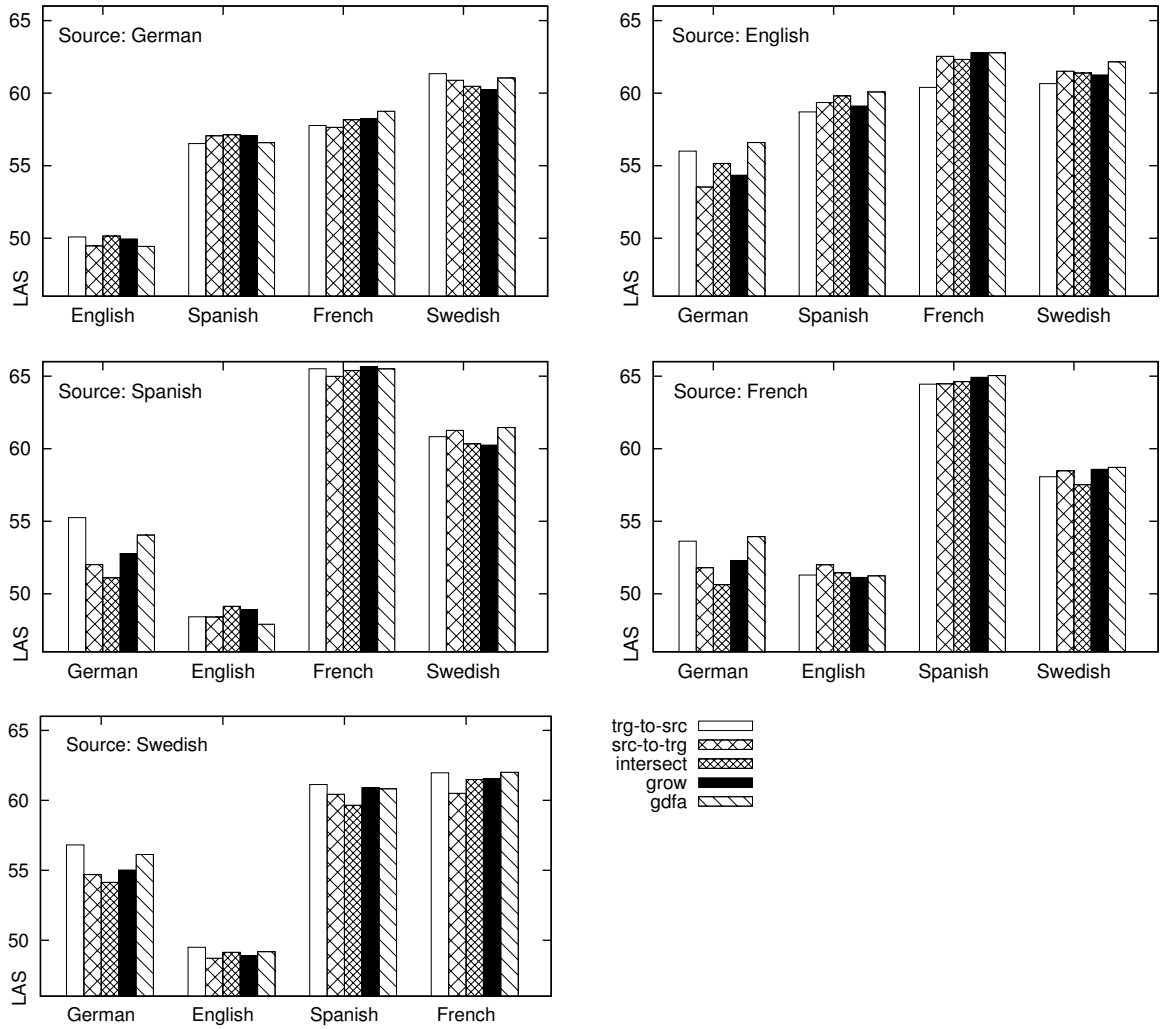

Figure 5: The impact of word alignment symmetrization on projection and parsing accuracy. *src-to-trg* and *trg-to-src* refer to the original directional Viterbi word alignments created by IBM model 4 in both directions; *intersect* refers to the intersection of both IBM 4 alignments; *grow* and *gdfa* (grow-diag-final-and) refer to popular symmetrization heuristics used in the SMT community.

metrization heuristics and their effect on projection and the quality of the parser model trained on the projected data. For this, we return to the setup of projecting annotations on human translations using the Europarl corpus with the same settings as described in Section 4.2 (using 40,000 sentences for the projection). We now compare five different word alignments based on IBM model 4 trained on the entire corpus for each language pair. First of all, we look at the original directional word alignment from source to target language and vice versa. We then include the intersection of these two directional link sets to represent a symmetrization heuristics that produces very sparse but high precision word alignments. Finally, we also consider the *grow* heuristics that adds adjacent alignment points coming from the union of directional alignment links to the sparse intersection of the same. In this way, the resulting word alignment covers most words while keeping precision at a rather high level. All of these alignment types are then contrasted with the *grow-diag-final-and* heuristics that we use in our default setup.

Figure 5 plots the parsing performance across languages based on the projection with the various alignment techniques listed above. A general observation is that the differences are rather small in most cases. Projecting annotation using the direct correspondence assumption seems to be quite robust with respect to alignment noise. In our experiments, no specific tendencies can be identified that would allow to draw immediate conclusions and to give clear recommendations for our task. Somewhat surprisingly we can see that the recall-oriented alignment heuristics (*grow-diag-final-and*) actually perform quite well in many cases, leading either to the best performing model or to one that is very close to the best result. However, in some cases, these models fall behind the ones based on alignment intersections (for instance Spanish-English) or directional word alignments (for example for Spanish-German, French-English, Swedish-German). A striking difference can be seen in the annotations projected to German. There, the target-to-source alignment performs pretty well and outperforms in two cases all other alignment types in the down-stream task. Furthermore, the intersection falls far behind in three of these cases, which indicates that both alignment directions are probably very different from each other leading to a very sparse word alignment when intersecting them. One possible reason for the success of the directional alignment might be that it favors the mapping to a compounding language such as German that frequently requires many-to-one links. However, the same effect cannot be seen for the other compounding language in our test set, Swedish.

4.5 Parsing Without Golden POS Labels

For a truly unsupported language, it does not make sense to assume a high quality POS tagger. Nevertheless, most cross-lingual experiments test their performance on data with human annotated golden POS labels. This is similar to the tradition of monolingual parsing where test accuracy is measured with perfect tokenization and completely correct POS annotation. In practice, this would not be realistic where new data needs to be parsed without proper tagging and unambiguous tokenization.

Direct transfer models are even more dependent on POS labels as those are the only source of information they can work with when making attachment decisions. Annotation projection approaches, on the other hand, are able to transfer POS information as well, which allows to train tagger models on projected data. In this section, we would like to test the feasibility of such an idea to see if we can truly port a parser to a new language without additional assumptions.

The first step is to train tagger models on our projected data sets. For this, we use the translated treebanks and a simple word-by-word translation approach in which we translate single-word-phrases only in our standard SMT model. The word-by-word translation model assures that we do not contaminate the data with DUMMY nodes and labels even though the translation quality lags behind the more powerful phrase-based models with larger translation options. We train standard Markov taggers with suffix backoff using HunPos (Halácsy et al., 2007) on each of the projected training data sets from the Universal Treebank. Table 4 summarizes the performance of all tagger models tested on the test sets in the treebank. The tagger all use the same universal POS tagset with its 12 labels as used in the Universal Treebank (Petrov et al., 2012). As we can see, the performance of those taggers is not great but still rather informative with overall accuracy values around 80%. The drop from source data to projected data is about 10-15 absolute points, which is, however, quite dramatic. Assuming that this is the best we can

POS	DE	EN	ES	FR	SV
DE	**95.24**	73.15	69.31	72.41	79.01
EN	82.04	**97.56**	79.91	81.23	84.44
ES	77.27	77.43	**95.37**	83.97	78.26
FR	80.99	78.74	88.47	**95.08**	79.62
SV	78.40	71.45	70.11	66.77	**95.86**

DELEXICALIZED MODELS

LAS	DE	EN	ES	FR	SV
DE	–	33.38	34.37	36.59	39.15
EN	36.55	–	45.53	47.71	**48.92**
ES	35.07	39.87	–	**51.40**	42.95
FR	35.89	**40.40**	**51.55**	–	40.30
SV	**37.87**	39.80	43.62	41.61	–

TRANSLATED TREEBANK MODELS

LAS	DE	EN	ES	FR	SV
DE	–	41.29	42.16	46.26	46.79
EN	42.24	–	50.54	53.63	**53.78**
ES	38.61	43.70	–	**57.58**	47.01
FR	**42.65**	**48.37**	**57.78**	–	45.55
SV	41.37	42.34	49.38	46.00	–

Table 4: Top (POS): Accuracy of POS tagging models trained on translated treebanks (word-by-word model). Bottom (LAS): Cross-lingual parser models tested on automatically POS tagged test sets. The delexicalized baseline (left) and the translated treebank model using word-by-word translation (right).

achieve for the target language, we now have to look at the parsing performance when relying on such noisy annotation.

Firstly, we look at the delexicalized baselines. The bottom-left part of Table 4 lists the labeled attachment scores when gold POS labels are replaced with automatic tags created by the corresponding projection tagger. The drop is huge and the original scores that were well above 50-70% go down to not more than 30-40% LAS. Clearly, this was to be expected as proper POS labeling is crucial for these models. Let us now look at the annotation projection approach using a translated treebank as our parallel data set. Table 4 on the bottom-right lists the corresponding labeled attachment scores with automatic POS tags. As expected, the performance is considerably lower than with golden POS labels, which are still the most informative features in those models. However, the performance remains in a range of above 40-50% LAS. Clearly, the lexical features help to keep the performance up at a higher level than the delexicalized baselines. We believe, that this difference can be crucial when porting language tools to new languages and that the models can be further optimized to rely less on golden POS tags.

5 Conclusions

In this paper we revisit annotation projection for cross-lingual parser induction. We show that annotation can successfully be transfered to target languages if the annotation is harmonized across languages. Despite previous negative results on diverse treebanks we demonstrate that direct projection works very well for a number of languages and outperforms direct delexicalized transfer models by a large margin. The approach is also quite robust with respect to word alignment. Furthermore, we show that machine translation can be a useful alternative for this strategy and that projected data can also be used to induce basic information such as POS labels in combination with syntactic parser models.

Acknowledgements

This work was supported by the Swedish Research Council (Vetenskapsrådet), project 2012-916. I would also like to thank Joakim Nivre, Željko Agić and the anonymous reviewers for helpful comments and suggestions.

References

Željko Agić, Danijela Merkler, and Daša Berović. 2012. Slovene-Croatian Treebank Transfer Using Bilingual Lexicon Improves Croatian Dependency Parsing. In *Proceedings of IS-LTC 2012*, pages 5–9.

Miguel Ballesteros and Joakim Nivre. 2012. MaltOptimizer: An Optimization Tool for MaltParser. In *Proceedings of EACL 2012*, pages 58–62.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of CoNLL 2006*, pages 149–164.

Marine Carpuat and Michel Simard. 2012. The trouble with smt consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449, Montréal, Canada.

Kuzman Ganchev and Dipanjan Das. 2013. Cross-Lingual Discriminative Learning of Sequence Models with Posterior Regularization. In *Proceedings of EMNLP 2013*, pages 1996–2006.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. Poster paper: Hunpos – an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 209–212, Prague, Czech Republic.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of ACL 2013*, pages 690–696.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural Language Engineering*, 11(3):311–325.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL 2007*, pages 177–180.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-Source Transfer of Delexicalized Dependency Parsers. In *Proceedings of EMNLP 2011*, 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 ACL 2013*, pages 92–97.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective Sharing for Multilingual Dependency Parsing. In *Proceedings of ACL 2012*, pages 629–637.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of LREC 2006*, pages 2216–2219.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–52.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL 2003*, pages 160–167.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of LREC 2012*, pages 2089–2096.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure. In *Proceedings of NAACL 2012*, pages 477–487.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013a. Token and Type Constraints for Cross-lingual Part-of-speech Tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013b. Target Language Adaptation of Discriminative Transfer Parsers. In *Proceedings of NAACL 2013*, pages 1061–1071.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Proceedings of the 18th Conference Natural Language Processing and Computational Natural Language Learning (CoNLL)*, Baltimore, Maryland, USA.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing Multilingual Text Analysis Tools via Robust Projection Across Aligned Corpora. In *Proceedings of HLT 2011*, pages 1–8.

Yue Zhang and Joakim Nivre. 2011. Transition-based Dependency Parsing with Rich Non-local Features. In *Proceedings of ACL 2011*, pages 188–193.