Treebank Translation for Cross-Lingual Parser Induction

Jörg Tiedemann
Dep. of Linguistics and Philology
Uppsala University
jorg.tiedemann@lingfil.uu.se

Željko Agić
Linguistics Department
University of Potsdam
zagic@uni-potsdam.de

Joakim Nivre
Dep. of Linguistics and Philology
Uppsala University
joakim.nivre@lingfil.uu.se

Abstract

Cross-lingual learning has become a popular approach to facilitate the development of resources and tools for low-density languages. Its underlying idea is to make use of existing tools and annotations in resource-rich languages to create similar tools and resources for resource-poor languages. Typically, this is achieved by either projecting annotations across parallel corpora, or by transferring models from one or more source languages to a target language. In this paper, we explore a third strategy by using machine translation to create synthetic training data from the original source-side annotations. Specifically, we apply this technique to dependency parsing, using a cross-lingually unified treebank for adequate evaluation. Our approach draws on annotation projection but avoids the use of noisy source-side annotation of an unrelated parallel corpus and instead relies on manual treebank annotation in combination with statistical machine translation, which makes it possible to train fully lexicalized parsers. We show that this approach significantly outperforms delexicalized transfer parsing.

1 Introduction

The lack of resources and tools is a serious problem for the majority of the world's languages (Bender, 2013). Many applications require robust tools and the development of language-specific resources is expensive and time consuming. Furthermore, many tasks such as data-driven syntactic parsing require strong supervision to achieve reasonable results for real-world applications, since the performance of fully unsupervised methods lags behind by a large margin in comparison with the state of the art. Cross-lingual learning has been proposed as one possible solution to quickly create initial tools for languages that lack the appropriate resources (Ganchev and Das, 2013). By and large, there are two main strategies that have been proposed in the literature: annotation projection and model transfer.

1.1 Previous Cross-Lingual Approaches

Annotation projection relies on the mapping of linguistic annotation across languages using parallel corpora and automatic alignment as basic resources (Yarowsky et al., 2001; Hwa et al., 2005; Täckström et al., 2013a). Tools that exist for the source language are used to annotate the source side of the corpus and projection heuristics are then applied to map the annotation through word alignment onto the corresponding target language text. Target language tools can then be trained on the projected annotation assuming that the mapping is sufficiently correct. 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, and expensive and time-consuming to produce.

Model transfer instead relies on universal features and model parameters that can be transferred from one language to another. Abstracting away from all language-specific parameters makes it possible to train, e.g., delexicalized parsers that ignore lexical information. This approach has been used with success for a variety of languages, drawing from a harmonized POS tagset (Petrov et al., 2012) that is used as the main source of information. One advantage compared to annotation projection is that no parallel data is required. In addition, training can be performed on gold standard annotation. However, model transfer assumes a common fea-

Proceedings of the Eighteenth Conference on Computational Language Learning, pages 130–140,
Baltimore, Maryland USA, June 26-27 2014.

ture representation across languages (McDonald et al., 2013), which can be a strong bottleneck. Several extensions have been proposed to make the approach more robust. First of all, multiple source languages can be involved to increase the statistical basis for learning (McDonald et al., 2011; Naseem et al., 2012), a strategy that can also be used in the case of annotation projection. Cross-lingual word clusters can be created to obtain additional universal features (Täckström et al., 2012). Techniques for target language adaptation can be used to improve model transfer with multiple sources (Täckström et al., 2013b).

1.2 The Translation Approach

In this paper, we propose a third strategy, based on automatically translating training data to a new language in order to create annotated resources directly from the original source. Recent advances in statistical machine translation (SMT) combined with the ever-growing availability of parallel corpora are now making this a realistic alternative. The relation to annotation projection is obvious as both involve parallel data with one side being annotated. However, the use of direct translation brings two important advantages. First of all, using SMT, we do not accumulate errors from two sources: the tool – e.g., tagger or parser – used to annotate the source language of a bilingual corpus and the noise coming from alignment and projection. Instead, we use the gold standard annotation of the source language which can safely be assumed to be of much higher quality than any automatic annotation obtained by using a tool trained on that data. Moreover, using SMT may help in bypassing domain shift problems, which are common when applying tools trained (and evaluated) on one resource to text from another domain. Secondly, we can assume that SMT will produce output that is much closer to the input than manual translations in parallel texts usually are. Even if this may seem like a short-coming in general, in the case of annotation projection it should rather be an advantage, because it makes it more straightforward and less error-prone to transfer annotation from source to target. Furthermore, the alignment between words and phrases is inherently provided as an output of all common SMT models. Hence, no additional procedures have to be performed on top of the translated corpus. Recent research (Zhao et al., 2009; Durrett et al., 2012) has attempted to address synthetic data creation for syntactic parsing via bilingual lexica. We seek to build on this work by utilizing more advanced translation techniques.

Further in the paper, we first describe the tools and resources used in our experiments (§2). We elaborate on our approach to translating treebanks (§3) and projecting syntactic annotations (§4) for a new language. Finally, we provide empirical evaluation of the suggested approach (§5) and observe a substantial increase in parsing accuracy over the delexicalized parsing baselines.

2 Resources and Tools

In our experiments, we rely on standard resources and tools for both dependency parsing and machine translation without any special enhancements. Since we are primarily trying to provide a proof of concept for the use of SMT-derived synthetic training data in dependency parsing, we believe it is more important to facilitate reproducibility than to tweak system components to obtain maximum accuracy.

We use the Universal Dependency Treebank v1 (McDonald et al., 2013) for annotation projection, parser training and evaluation. It is a collection of data sets with consistent syntactic annotation for six languages: English, French, German, Korean, Spanish, and Swedish.[1] The annotation is based on Stanford Typed Dependencies for English (De Marneffe et al., 2006) but has been adapted and harmonized to allow adequate annotation of typologically different languages. This is the first collection of data sets that allows reliable evaluation of labeled dependency parsing accuracy across multiple languages (McDonald et al., 2013). We use the dedicated training and test sets from the treebank distribution in all our experiments. As argued in (McDonald et al., 2013), most cross-lingual dependency parsing experiments up to theirs relied on heterogeneous treebanks such as the CoNLL datasets for syntactic dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007a), making it difficult to address challenges like consistent cross-lingual analysis for downstream applications and reliable cross-lingual evaluation of syntactic parsers. More specifically, none of the previous research could report full labeled parsing accuracies, but rather just unlabeled structural accuracies across different attachment schemes. Following the line of McDonald et al. (2013) regarding the

[1] https://code.google.com/p/uni-dep-tb/

emphasized importance of homogenous data and the assignment of labels, we only report labeled attachment scores (LAS) in all our experiments. As it is likely the first reliable cross-lingual parsing evaluation, we also choose their results as the baseline reference point for comparison with our experiments.

For dependency parsing, we use MaltParser (Nivre et al., 2006a)[2] due to its efficiency in both training and parsing, and we facilitate MaltOptimizer (Ballesteros and Nivre, 2012)[3] to bypass the tedious task of manual feature selection. MaltParser is a transition-based dependency parser that has been evaluated on a number of different languages with competitive results (Nivre et al., 2006b; Nivre et al., 2007b; Hall et al., 2007) and it is widely used for benchmarking and application development. Although more accurate dependency parsers exist for the task of monolingual supervised parsing, it is not clear that these differences carry over to the cross-lingual scenario, where baselines are lower and more complex models are more likely to overfit. The use of a transition-based parser also facilitates comparison with delexicalized transfer parsing, where transition-based parsers are dominant so far (McDonald et al., 2011; McDonald et al., 2013). We leave the exploration of additional parsing approaches for future research.

For machine translation, we select the popular Moses toolbox (Koehn et al., 2007) and the phrase-based translation paradigm as our basic framework. Phrase-based SMT has the advantage of being straightforward and efficient in training and decoding, while maintaining robustness and reliability for many language pairs. More details about the setup and the translation procedures are given in Section 3 below. The most essential ingredient for translation performance is the parallel corpus used for training the translation models. For our experiments we use the freely available and widely used Europarl corpus v7 (Koehn, 2005).[4] It is commonly used for training SMT models and includes parallel data for all languages represented in the Universal Treebank except Korean, which we will, therefore, leave out in our experiments. For tuning we apply the newstest 2012 data provided by the annual workshop on statistical machine translation.[5] For language modeling, we use a combination of Europarl and News data provided from the same source. The statistics of the corpora are given in Table 1.

	DE	EN	ES	FR	SV
DE		94 M	94 M	96 M	81 M
EN	2.0 M		103 M	105 M	89 M
ES	1.9 M	2.0 M		104 M	89 M
FR	1.9 M	2.0 M	2.0 M		91 M
SV	1.8 M	1.9 M	1.8 M	1.9 M	
mono	22.9 M	17.1 M	6.3 M	6.3 M	2.3 M

Table 1: Parallel data and monolingual data used for training the SMT models. Lower-left triangle = number of sentence pairs; upper-right triangle = number of tokens (source and target language together); bottom row = number of sentences in monolingual corpora.

3 Translating Treebanks

The main contribution of this paper is the empirical study of automatic treebank translation for parser transfer. We compare three different translation approaches in order to investigate the influence of several parameters. All of them are based on automatic word alignment and subsequent extraction of translation equivalents as common in phrase-based SMT. In particular, word alignment is performed using GIZA++ (Och and Ney, 2003) and IBM model 4 as the final model for creating the Viterbi word alignments for all parallel corpora used in our experiments. For the extraction of translation tables, we use the Moses toolkit with its standard settings to extract phrase tables with a maximum of seven tokens per phrase from a symmetrized word alignment. Symmetrization is done using the grow-diag-final-and heuristics (Koehn et al., 2003). We tune phrase-based SMT models using minimum error rate training (Och, 2003) and the development data for each language pair. 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)).

Our first translation approach is based on a very simple word-by-word translation model. For this, we select the most reliable translations of single words from the phrase translation tables extracted from the parallel corpora as described above. We restrict the model to tokens with alphabetic characters only using pre-defined Unicode character

[2] http://www.maltparser.org/
[3] http://nil.fdi.ucm.es/maltoptimizer/
[4] http://www.statmt.org/europarl/
[5] http://www.statmt.org/wmt14

sets. The selection of translation alternatives is based on the Dice coefficient, which combines the two essential conditional translation probabilities given in the phrase table. The Dice coefficient is in fact the harmonic mean of these two probabilities and has successfully been used for the extraction of translation equivalents before (Smadja et al., 1996):

$$Dice(s,t) = \frac{2\,p(s,t)}{p(s)+p(t)} = 2\left(\frac{1}{p(s|t)} + \frac{1}{p(t|s)}\right)^{-1}$$

Other association measures would be possible as well but Smadja et al. (1996) argue that the Dice coefficient is more robust with respect to low frequency events than other common metrics such as pointwise mutual information, which can be a serious issue with the unsmoothed probability estimations in standard phrase tables. Our first translation model then applies the final one-to-one correspondences to monotonically translate treebanks word by word. We refer to it as the LOOKUP approach. Note that any bilingual dictionary could have been used to perform the same procedure.

The second translation approach (WORD-BASED MT) is slightly more elaborate but still restricts the translation model to one-to-one word mappings. For this, we extract all single word translation pairs from the phrase tables and apply the standard beam-search decoder implemented in Moses to translate the original treebanks to all target languages. The motivation for this model is to investigate the impact of reordering and language models while still keeping the projection of annotated data as simple as possible. Note that the language model may influence not only the word order but also the lexical choice as we now allow multiple translation options in our phrase table.

The final model implements translation based on the entire phrase table using the standard approach to PHRASE-BASED SMT. We basically run the Moses decoder with default settings and the parameters and models trained on our parallel corpora. Note that it is important for the annotation transfer to keep track of the alignment between phrases and words of the input and output sentences. The Moses decoder provides both, phrase segmentation and word alignment (if the latter is coded into the phrase tables). This will be important as we will see in the annotation projection discussed below.

ORIGINAL					
	DE	EN	ES	FR	SV
	14.0	0.00	7.90	13.3	4.20
WORD-BASED MT					
	DE	EN	ES	FR	SV
DE	–	49.1	62.6	52.8	60.4
EN	43.3	–	27.6	34.8	0.00
ES	54.9	25.1	–	12.3	18.3
FR	68.2	39.6	32.8	–	57.8
SV	34.1	5.20	21.6	33.7	–
PHRASE-BASED MT					
	DE	EN	ES	FR	SV
DE	–	51.5	57.3	58.8	46.8
EN	49.3	–	50.3	61.7	14.6
ES	65.9	66.7	–	62.8	49.0
FR	58.0	53.7	44.7	–	38.2
SV	43.9	43.6	49.6	57.1	–

Table 2: Non-projectivity in synthetic treebanks.

4 Transferring Annotation

The next step in preparing synthetic training data is to project the annotation from the original treebank to the target language. Given the properties of a dependency tree, where every word has exactly one syntactic head and dependency label, the annotation transfer is trivial for the two initial translation models. All annotation can simply be copied using the dictionary LOOKUP in which we enforce a monotonic one-to-one word mapping between source and target language.

In the second approach, we only have to keep track of reordering, which is reported by the decoder when translating with our model. Note that the mapping is strictly one-to-one (bijective) as phrase-based SMT does not allow deletions or insertions at any point. This also ensures that we will always maintain a tree structure even though reordering may have a strong impact on projectivity (see Table 2). An illustration of this type of annotation transfer is shown in the left image of Figure 1.

The third model, full PHRASE-BASED SMT, requires the most attention when transferring annotation across languages. Here we have to rely on the alignment information and projection heuristics similar to the ones presented in related work (Hwa et al., 2005). In their work, Hwa et al. (2005) define a direct projection algorithm that transfers automatic annotation to a target language via word alignment. The algorithm defines a number of

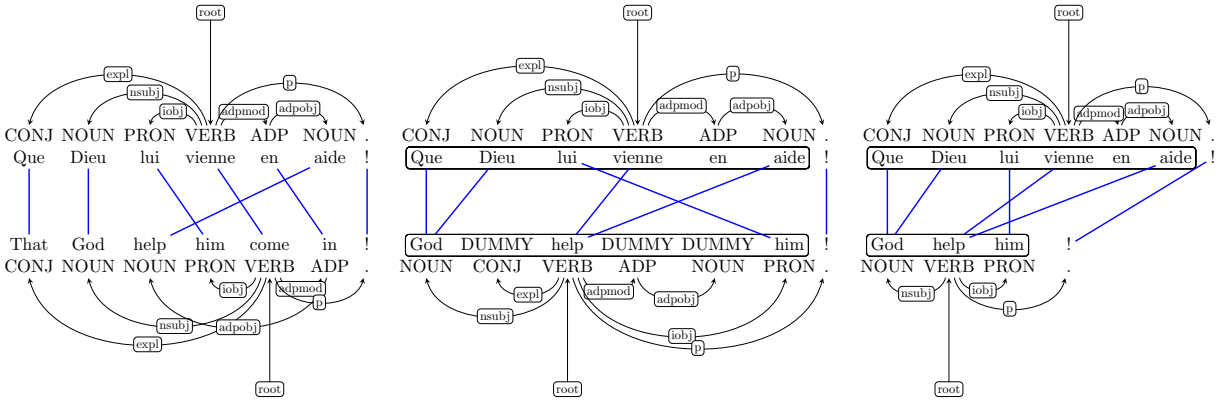

Figure 1: Transferring annotation from French to an English translation with a WORD-BASED translation model (left) and with a PHRASE-BASED translation model (middle and right). Annotation projection using the Direct Projection Algorithm by Hwa et al. (2005) (middle) and our approach (right).

heuristics to handle unaligned, one-to-many, many-to-one and many-to-many alignments. As a side effect, this approach produces several dummy-nodes in the target language to ensure a complete projection of the source language tree (see Hwa et al. (2005) for more details).

In our approach, we try to make use of the additional information provided by the SMT decoder to avoid dummy-nodes and relations that may negatively influence the induced target language parser. Compared to the annotation projection approach of Hwa et al. (2005), the situation in our PHRASE-BASED SMT setting is slightly different. Here, we have two types of alignments that can be considered when relating source and target language items: (i) the alignment between phrases (pairs of consecutive n-grams) and (ii) the phrase-internal word alignment on which phrase extraction is based. The primary information used for annotation transfer is still the latter which has the same properties as described by Hwa et al. (2005) (except that we have truly many-to-many alignments in our data which were not available in their experiments).

Note that words may be unaligned in phrase-based SMT as the phrase extraction algorithm used in Moses includes unaligned adjacent tokens. However, for these unaligned words, we know to which phrase they belong and can also identify the corresponding phrase in the other language using phrase alignment information. This makes it possible to avoid the creation of dummy-nodes altogether and instead to link unaligned words to existing nodes based on the given phrase segmentation.

Similarly, we define heuristics for handling one-to-many, many-to-one and many-to-many alignments that avoid the creation of dummy-nodes. The main procedure is illustrated in Figure 2.

The key feature of this projection algorithm is that ambiguous alignments are handled by attaching words to the nodes that are highest up in the dependency tree (the procedure find_highest() returns the node with minimum distance to the root of the tree). This ensures that we avoid cycles and isolated cliques in the graph. Furthermore, unaligned words are attached to the head of the target phrase they belong to, which seems to be the most appropriate place without further knowledge. The procedures in_trg_phrase() and in_src_phrase() make use of the phrase segmentation used in the translation process.

One complication is the search for the corresponding target head word in cases where the source language head is not aligned or aligned to multiple target language words. Figure 3 shows the head alignment procedure that we define in our projection algorithm. Procedure find_aligned() returns the rightmost word of all words aligned to the given source language word s. Other heuristics or linguistically motivated rules based on POS tags and general language properties would be possible here as well. If s is not aligned, we move up in the dependency tree until we hit ROOT or find an aligned word. If we are at the root position we return ROOT as this does not require further mappings. The effect of this algorithm is illustrated by the right-hand side image in Figure 1.

5 Parsing Across Languages

In this section, we present the results of two experimental batches. First, we establish the base-

```
Input: source tree S, target sentence T,
word alignment A, phrase segmentation P
Output: syntactic heads head[],
word attributes attr[]
treeSize = max_distance_to_root(S) ;
attr = [] ;
head = [] ;
for t ∈ T do
    if is_unaligned_trg(t,A) then
        for t' ∈ in_trg_phrase(t,P) do
            [s_x,..,s_y] = aligned_to(t') ;
            ŝ = find_highest([s_x,..,s_y],S) ;
            t̂ = find_aligned(ŝ,S,T,A) ;
            attr[t] = DUMMY ;
            head[t] = t̂ ;
        end
    else
        [s_x,..,s_y] = aligned_to(t) ;
        s = find_highest([s_x,..,s_y],S) ;
        attr[t] = attr(s) ;
        ŝ = head_of(s,S) ;
        t̂ = find_aligned(ŝ,S,T,A) ;
        if t̂ == t then
            [s_x,..,s_y] = in_src_phrase(s,P) ;
            s* = find_highest([s_x,..,s_y],S) ;
            ŝ = head_of(s*,S) ;
            t̂ = find_aligned(ŝ,S,T,A) ;
            head[t] = t̂ ;
        end
    end
end
```

Figure 2: Annotation projection algorithm.

lines by comparing monolingual supervised parsing to delexicalized transfer parsing following the approach of McDonald et al. (2013). Second, we present the results obtained with parsers trained on target language treebanks produced using machine translation and annotation projection. Here, we also look at delexicalized models trained on translated treebanks to show the effect of machine translation without additional lexical features.

5.1 Baseline Results

First we present the baseline parsing scores. The baselines we explore are: (i) the monolingual baseline, i.e., training and testing using the same language data from the Universal Dependency Treebank and (ii) the delexicalized baseline, i.e., applying delexicalized parsers across languages.

For the monolingual baseline, MaltParser models are trained on the original treebanks with universal POS labels and lexical features but leaving out other language-specific features if they exist in the original treebanks. The delexicalized parsers are trained on universal POS labels only for each language and are then applied to all other languages

```
Input: node s, source tree S with root ROOT,
target sentence T, word alignment A
Output: node t*
if s == ROOT then
    return ROOT ;
end
while is_unaligned_src(s,A) do
    s = head_of(s,S) ;
    if s == ROOT then
        return ROOT ;
    end
end
p = 0 ;
t* = undef ;
for t' ∈ aligned(s,A) do
    if position(t',T) > p then
        t* = t' ;
        p = position(t',T) ;
    end
end
return t* ;
```

Figure 3: Procedure find_aligned().

without modification. For all models, features and options are optimized using MaltOptimizer. The accuracy is given in Table 3 as a set of labeled attachment scores (LAS). We include punctuation in our evaluation. Ignoring punctuation generally leads to slightly higher scores as we have noted in our experiments but we do not report those numbers here. Note also that the columns represent the target languages (used for testing), while the rows denote the source languages (used in training), as in McDonald et al. (2013).

From the table, we can see that the baseline scores are compatible with the ones in the original experiments presented by (McDonald et al., 2013), included in Table 3 for reference. The differences are due to parser selection, as they use a transition-based parser with beam search and perceptron learning along the lines of Zhang and Nivre (2011) whereas we rely on greedy transition-based parsing with linear support vector machines. In the following, we will compare results to our baseline as we have a comparable setup in those experiments. However, most improvements shown below also apply in comparison with (McDonald et al., 2013).

5.2 Translated Treebanks

Now we turn to the experiments on translated treebanks. We consider two setups. First, we look at the effect of translation when training delexicalized parsers. In this way, we can perform a direct comparison to the baseline performance presented

MONOLINGUAL					
	DE	EN	ES	FR	SV
	72.13	87.50	78.54	77.51	81.28
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 3: Baselines – labeled attachment score (LAS) for monolingual and delexicalized transfer parsing. Delexicalized transfer parsing results of McDonald et al. (2013) included for reference.

WORD-BASED MT					
	DE	EN	ES	FR	SV
DE	–	48.12 (4.92)	50.84 (4.75)	52.92 (6.83)	55.52 (4.88)
EN	49.53 (2.91)	–	57.41 (1.76)	**58.53** (2.07)	**57.82** (0.14)
ES	45.48 (1.45)	48.46 (1.73)	–	58.29 (0.38)	55.25 (1.43)
FR	46.59 (2.68)	47.88 (1.13)	**59.72** (0.07)	–	52.31 (0.30)
SV	**52.16** (1.47)	**49.14** (0.01)	56.50 (2.88)	56.71 (4.74)	–
PHRASE-BASED MT					
	DE	EN	ES	FR	SV
DE	–	45.43 (2.23)	47.26 (1.17)	49.14 (3.05)	53.37 (2.73)
EN	49.16 (2.54)	–	57.12 (1.47)	**58.23** (1.77)	**58.23** (0.55)
ES	46.75 (2.72)	46.82 (0.09)	–	58.22 (0.31)	54.14 (0.32)
FR	48.02 (4.11)	**49.06** (2.31)	**60.23** (0.58)	–	55.24 (3.23)
SV	**50.96** (0.27)	46.12 $^{-3.01}$	55.95 (2.33)	54.71 (2.74)	–

Table 4: Translated treebanks: labeled attachment score (LAS) for *delexicalized* parsers trained on synthetic data created by translation. Numbers in superscript show the absolute improvement over our delexicalized baselines.

above. The second setup then considers fully lexicalized models trained on translated treebanks. The main advantage of the translation approach is the availability of lexical information and this final setup represents the real power of this approach. In it, we compare lexicalized parsers trained on translated treebanks with their delexicalized counterparts and avoid a direct comparison with the delexicalized baselines as they involve different types of features.

5.3 Delexicalized Parsers

Table 4 presents the scores obtained by training delexicalized parsing models on synthetic data created by our translation approaches presented earlier. Feature models and training options are the same as for the delexicalized source language models when training and testing on the target language data. Note that we exclude the simple dictionary LOOKUP approach here, because this approach leads to identical models as the basic delexicalized models. This is because words are translated one-to-one without any reordering which leads to exactly the same annotation sequences as the source language treebank after projecting POS labels and dependency relations.

From the table, we can see that all but one model improve the scores obtained by delexicalized baseline models. The improvements are quite substantial up to +6.38 LAS. The boost in performance is especially striking for the simpleWORD-BASED translation model considering that the only difference to the baseline model is word order. The impact of the more complex PHRASE-BASED translation model is, however, difficult to judge. In 14 out of 20 models it actually leads to a drop in LAS when applying phrase-based translation instead of single-word translation. This is somewhat surprising but is probably related to the additional ambiguity in annotation projection introduced by many-to-many alignments. The largest drop can be seen for Swedish translated to English, which even falls behind the baseline performance when using the PHRASE-BASED translation model.

5.4 Lexicalized Parsers

The final experiment is concerned with lexical parsers trained on translated treebanks. The main objective here is to test the robustness of fully lexicalized models trained on noisy synthetic data created by simple automatic translation engines. Table 5 lists the scores obtained by our models when trained on treebanks translated with our three approaches (dictionary LOOKUP, WORD-BASED MT and full PHRASE-BASED translation). Again, we use the same feature model and training options as for the source language model when training models for the target languages. This time, of course, this refers to the features used by the lexicalized baseline models.

The capacity of the parsing models increases due to the lexical information which is now included. In order to see the effect of lexicalization, we com-

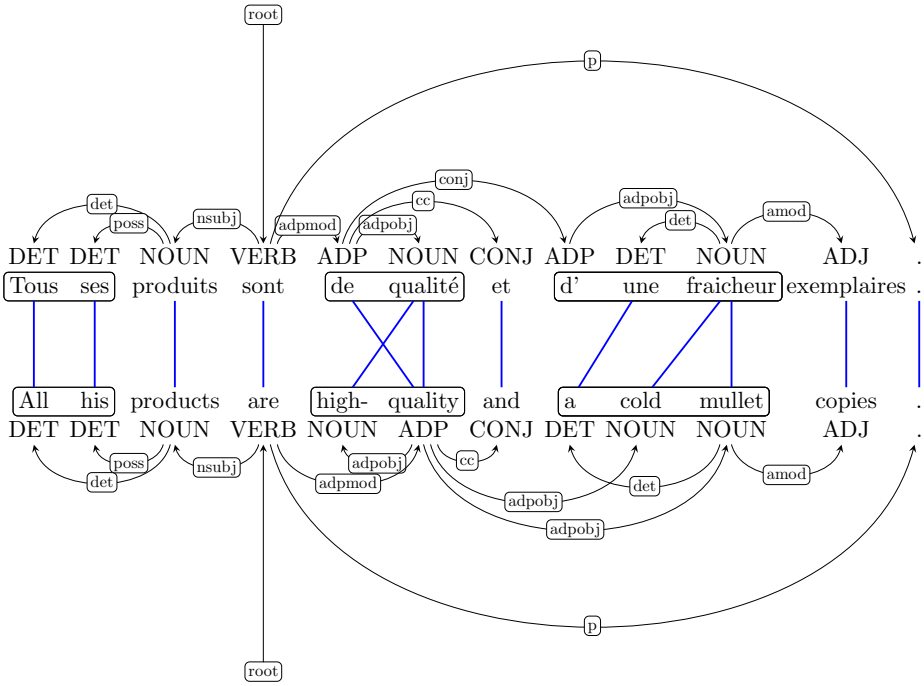

Figure 4: Problematic annotation projection with ambiguous word alignment.

pare the performance now with the corresponding delexicalized models. Note that the LOOKUP approach relates to the delexicalized baseline models without any translation.

As we can see, all models outperform their corresponding delexicalized version (with one exception), which demonstrates the ability of the training procedure to pick up valuable lexical information from the noisy translations. Again, we can see substantial absolute improvements of up to +7.31 LAS showing the effectiveness of the translation approach. Note that this also means that we outperform the delexicalized baselines in all cases by a large margin, even if we should not directly compare these models as they draw on different feature sets. Once again, we can also see that the very simple methods are quite successful. Even the very basic LOOKUP approach leads to significant improvements with one minor exception. Surprisingly, no gain can be seen with the PHRASE-BASED translation approach. The translation quality is certainly better when manually inspecting the data. However, the increased complexity of annotation projection seems to pull down the parsers induced on that kind of data. A question for future work is whether the performance of those models can be improved by better projection algorithms and heuristics that lead to cleaner annotations of otherwise better translations of the original treebanks.

One possible reason for this disappointing result could be the unreliable mapping of POS labels across many-to-many alignments. Figure 4 illustrates a typical case of link ambiguity that leads to erroneous projections. For example, the mapping of the label ADP onto the English word *quality* is due to the left-to-right procedure applied in our projection algorithm and the mapping of the NOUN label to the English adjective *cold* is due to the link to *fraicheur*. How much these errors effect our parsing models trained on the projected treebanks is difficult to estimate and further investigations are required to pinpoint these issues and to find ways of addressing problems that may occur in various contexts.

Nevertheless, the overall results are very positive. The experiments clearly show the potentials of the translation approach. Note that this paper presents the first attempt to study the effect of translation on cross-lingual parser induction. Further optimization of the translation process and the connected annotation projection procedures should lead to further improvements over our basic models.

6 Conclusions and Future Work

In this paper, we have addressed the problem of cross-lingual parser induction by using statistical machine translation to create synthetic training data. Our SMT approach avoids the noisy source-side

LOOKUP	DE	EN	ES	FR	SV
DE	–	48.63 (5.43)	52.66 (6.57)	52.06 (5.97)	58.78 (8.14)
EN	48.59 (1.97)	–	57.79 (2.14)	57.80 (1.34)	**62.21** (4.53)
ES	47.36 (3.33)	49.13 (2.40)	–	**62.24** (4.33)	57.50 (3.68)
FR	47.57 (3.66)	**54.06** (7.31)	**66.31** (6.66)	–	57.73 (5.72)
SV	**51.88** (1.19)	48.84 (0.29)	54.74 (1.12)	52.95 (0.98)	–
WORD-BASED MT					
	DE	EN	ES	FR	SV
DE	–	51.86 (3.74)	55.90 (5.06)	57.77 (4.85)	61.65 (6.13)
EN	**53.80** (4.27)	–	60.76 (3.35)	63.32 (4.79)	**62.93** (5.11)
ES	49.94 (4.46)	49.93 (1.47)	–	**65.60** (7.31)	59.22 (3.97)
FR	52.07 (5.48)	**54.44** (6.56)	**65.63** (5.91)	–	57.67 (5.36)
SV	53.18 (1.02)	50.91 (1.77)	60.82 (4.32)	59.14 (2.43)	–
PHRASE-BASED MT					
	DE	EN	ES	FR	SV
DE	–	50.89 (5.46)	52.54 (5.28)	54.99 (5.85)	59.46 (6.09)
EN	**53.71** (4.55)	–	60.70 (3.58)	62.89 (4.66)	**64.01** (5.78)
ES	49.59 (2.84)	48.35 (1.53)	–	**64.88** (6.66)	58.99 (4.85)
FR	51.83 (3.81)	**53.81** (4.75)	**65.55** (5.32)	–	59.01 (3.77)
SV	53.22 (2.26)	49.06 (2.94)	58.41 (2.46)	58.04 (3.33)	–

Table 5: Translated treebanks: labeled attachment score (LAS) for *lexicalized* parsers trained on synthetic data. Numbers in superscript show the absolute improvements over the delexicalized models based on the same translation strategy.

annotations of traditional annotation projection and makes it possible to train fully lexicalized target language models that significantly outperform delexicalized transfer parsers. We have also demonstrated that translation leads to better delexicalized models that can directly be compared with each other as they are based on the same feature space.

We have compared three SMT methods for synthesizing training data: LOOKUP-based translation, WORD-BASED translation and full PHRASE-BASED translation. Our experiments show that even noisy data sets and simple translation strategies can be used to achieve positive results. For all three approaches, we have recorded substantial improvements over the state of the art in labeled cross-lingual parsing (McDonald et al., 2013). According to our results, simple word-by-word translations are often sufficient to create reasonable translations to train lexicalized parsers on. More elaborated phrase-based models together with advanced annotation projection strategies do not necessarily lead to any improvements.

As future work, we want to improve our model by (i) studying the impact of other SMT properties and improve the quality of treebank translation, (ii) implementing more sophisticated methods for annotation projection and (iii) using n-best lists provided by SMT models to introduce additional synthetic data using a single resource. We also aim at (iv) applying our approach to transfer parsing for closely related languages (see Agić et al. (2012) and Zeman and Resnik (2008) for related work), (v) testing it in a multi-source transfer scenario (McDonald et al., 2011) and, finally, (vi) comparing different dependency parsing paradigms within our experimental framework.

Multi-source approaches are especially appealing using the translation approach. However, initial experiments (which we omit in this presentation) revealed that simple concatenation is not sufficient to obtain results that improve upon the single-best translated treebanks. A careful selection of appropriate training examples and their weights given to the training procedure seems to be essential to benefit from different sources.

7 Acknowledgements

This work was supported by the Swedish Research Council (Vetenskapsrådet) through the project on Discourse-Oriented Machine Translation (2012-916).

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.

Emily M. Bender. 2013. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Morgan & Claypool Publishers.

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

Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC 2006*, pages 449–454.

Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic Transfer Using a Bilingual Lexicon. In *Proceedings of EMNLP-CoNLL 2012*, pages 1–11.

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

Johan Hall, Jens Nilsson, Joakim Nivre, Gülsen Eryiğit, Beáta Megyesi, Mattias Nilsson, and Markus Saers. 2007. Single Malt or Blended? A Study in Multilingual Parser Optimization. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 933–939.

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, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase Based Translation. In *Proceedings of NAACL-HLT 2003*, pages 48–54.

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.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit 2005*, pages 79–86.

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. 2006a. MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of LREC 2006*, pages 2216–2219.

Joakim Nivre, Johan Hall, Jens Nilsson, Gülsen Eryiğit, and Svetoslav Marinov. 2006b. Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines. In *Proceedings of CoNLL 2006*, pages 221–225.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007a. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007b. MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering*, 13(2):95–135.

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.

Frank Smadja, Vasileios Hatzivassiloglou, and Kathleen R. McKeown. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22(1):1–38.

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.

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.

Daniel Zeman and Philip Resnik. 2008. Cross-Language Parser Adaptation between Related Languages. In *Proceedings of IJCNLP 2008*, pages 35–42.

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

Hai Zhao, Yan Song, Chunyu Kit, and Guodong Zhou. 2009. Cross Language Dependency Parsing Using a Bilingual Lexicon. In *Proceedings of ACL-IJCNLP 2009*, pages 55–63.