Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure

Oscar Täckström*
SICS / Uppsala University
Kista / Uppsala, Sweden
oscar@sics.se

Ryan McDonald
Google
New York, NY
ryanmcd@google.com

Jakob Uszkoreit
Google
Mountain View, CA
uszkoreit@google.com

Abstract

It has been established that incorporating word cluster features derived from large unlabeled corpora can significantly improve prediction of linguistic structure. While previous work has focused primarily on English, we extend these results to other languages along two dimensions. First, we show that these results hold true for a number of languages across families. Second, and more interestingly, we provide an algorithm for inducing cross-lingual clusters and we show that features derived from these clusters significantly improve the accuracy of cross-lingual structure prediction. Specifically, we show that by augmenting direct-transfer systems with cross-lingual cluster features, the relative error of delexicalized dependency parsers, trained on English treebanks and transferred to foreign languages, can be reduced by up to 13%. When applying the same method to direct transfer of named-entity recognizers, we observe relative improvements of up to 26%.

1 Introduction

The ability to predict the linguistic structure of sentences or documents is central to the field of natural language processing (NLP). Structures such as named-entity tag sequences (Bikel et al., 1999) or sentiment relations (Pang and Lee, 2008) are inherently useful in data mining, information retrieval and other user-facing technologies. More fundamental structures such as part-of-speech tag sequences (Ratnaparkhi, 1996) or syntactic parse trees (Collins, 1997; Kübler et al., 2009), on the other hand, comprise the core linguistic analysis for many important downstream tasks such as machine translation (Chiang, 2005; Collins et al., 2005). Currently, supervised data-driven methods dominate the literature on linguistic structure prediction (Smith, 2011). Regrettably, the majority of studies on these methods have focused on evaluations specific to English, since it is the language with the most annotated resources. Notable exceptions include the CoNLL shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003; Buchholz and Marsi, 2006; Nivre et al., 2007) and subsequent studies on this data, as well as a number of focused studies on one or two specific languages, as discussed by Bender (2011).

While annotated resources for parsing and several other tasks are available in a number of languages, we cannot expect to have access to labeled resources for all tasks in all languages. This fact has given rise to a large body of research on unsupervised (Klein and Manning, 2004), semi-supervised (Koo et al., 2008) and transfer (Hwa et al., 2005) systems for prediction of linguistic structure. These methods all attempt to benefit from the plethora of unlabeled monolingual and/or cross-lingual data that has become available in the digital age. Unsupervised methods are appealing in that they are often inherently language independent. This is borne out by the many recent studies on unsupervised parsing that include evaluations covering a number of languages (Cohen and Smith, 2009; Gillenwater et al., 2010; Naseem et al., 2010; Spitkovsky et al., 2011). However, the performance for most languages is still well below that of supervised systems and recent work has established that the performance is also below simple methods of linguistic transfer (McDonald et al., 2011).

In this study we focus on semi-supervised and linguistic-transfer methods for multilingual structure prediction. In particular, we pursue two lines of research around the use of word cluster features in discriminative models for structure prediction:

*The majority of this work was performed while the author was an intern at Google, New York, NY.

1. *Monolingual word cluster features induced from large corpora of text for semi-supervised learning (SSL) of linguistic structure.* Previous studies on this approach have typically focused only on a small set of languages and tasks (Freitag, 2004; Miller et al., 2004; Koo et al., 2008; Turian et al., 2010; Faruqui and Padó, 2010; Haffari et al., 2011; Tratz and Hovy, 2011). Here we show that this method is robust across 13 languages for dependency parsing and 4 languages for named-entity recognition (NER). This is the first study with such a broad view on this subject, in terms of language diversity.

2. *Cross-lingual word cluster features for transferring linguistic structure from English to other languages.* We develop an algorithm that generates cross-lingual word clusters; that is clusters of words that are consistent across languages. This is achieved by means of a probabilistic model over large amounts of monolingual data in two languages, coupled with parallel data through which cross-lingual word-cluster constraints are enforced. We show that by augmenting the delexicalized direct transfer system of McDonald et al. (2011) with cross-lingual cluster features, we are able to reduce its error by up to 13% relative. Further, we show that by applying the same method to direct-transfer NER, we achieve a relative error reduction of 26%.

By incorporating cross-lingual cluster features in a linguistic transfer system, we are for the first time combining SSL and cross-lingual transfer.

2 Monolingual Word Cluster Features

Word cluster features have been shown to be useful in various tasks in natural language processing, including syntactic dependency parsing (Koo et al., 2008; Haffari et al., 2011; Tratz and Hovy, 2011), syntactic chunking (Turian et al., 2010), and NER (Freitag, 2004; Miller et al., 2004; Turian et al., 2010; Faruqui and Padó, 2010). Intuitively, the reason for the effectiveness of cluster features lie in their ability to aggregate local distributional information from large unlabeled corpora, which aid in conquering data sparsity in supervised training regimes as well as in mitigating cross-domain generalization issues.

In line with much previous work on word clusters for tasks such as dependency parsing and NER, for which local syntactic and semantic constraints are of importance, we induce word clusters by means of a probabilistic class-based language model (Brown et al., 1992; Clark, 2003). However, rather than the more commonly used model of Brown et al. (1992), we use the predictive class bigram model introduced by Uszkoreit and Brants (2008). The two models are very similar, but whereas the former takes class-to-class transitions into account, the latter directly models word-to-class transitions. By ignoring class-to-class transitions, an approximate maximum likelihood clustering can be found efficiently with the distributed exchange algorithm (Uszkoreit and Brants, 2008). This is a useful property, as we later develop an algorithm for inducing cross-lingual word clusters that calls this monolingual algorithm as a subroutine.

More formally, let $\mathcal{C} : \mathcal{V} \mapsto 1, \ldots, K$ be a (hard) clustering function that maps each word type from the vocabulary, $\mathcal{V}$, to one of K cluster identities. With the model of Uszkoreit and Brants (2008), the likelihood of a sequence of word tokens, $\boldsymbol{w} = \langle w_i \rangle_{i=1}^{m}$, with $w_i \in \mathcal{V} \cup \{S\}$, where S is a designated start-of-segment symbol, factors as

$$L(\boldsymbol{w}; \mathcal{C}) = \prod_{i=1}^{m} p(w_i | \mathcal{C}(w_i)) p(\mathcal{C}(w_i) | w_{i-1}) \,. \quad (1)$$

Compare this to the model of Brown et al. (1992):

$$L'(\boldsymbol{w}; \mathcal{C}) = \prod_{i=1}^{m} p(w_i | \mathcal{C}(w_i)) p(\mathcal{C}(w_i) | \mathcal{C}(w_{i-1})) \,.$$

While the use of class-to-class transitions can lead to more compact models, which is often useful for conquering data sparsity, when clustering large data sets we can get reliable statistics directly on the word-to-class transitions (Uszkoreit and Brants, 2008).

In addition to the clustering model that we make use of in this study, a number of additional word clustering and embedding variants have been proposed. For example, Turian et al. (2010) assessed the effectiveness of the word embedding techniques of Collobert and Weston (2008) and Mnih and Hinton (2007) along with the word clustering technique of Brown et al. (1992) for syntactic chunking and NER. Recently, Dhillon et al. (2011) proposed a word

Single words	$S_0c\{p\}$, $N_0c\{p\}$, $N_1c\{p\}$, $N_2c\{p\}$
Word pairs	$S_0c\{p\}N_0c\{p\}$, S_0pcN_0p, S_0pN_0pc, S_0wN_0c, S_0cN_0w, N_0cN_1c, N_1cN_2c
Word triples	$N_0cN_1cN_2c$, $S_0cN_0cN_1c$, $S_{0h}cS_0cN_0c$, $S_0cS_{0l}cN_0c$, $S_0cS_{0r}cN_0c$, $S_0cN_0cN_{0l}c$
Distance	S_0cd, N_0cd, S_0cN_0cd
Valency	S_0cv_l, S_0cv_r, $N_0cS_0v_l$
Unigrams	$S_{0h}c$, $S_{0l}c$, $S_{0r}c$, $N_{0l}c$
Third-order	$S_{0h2}c$, $S_{0l2}c$, $S_{0r2}c$, $N_{0l2}c$
Label set	$S_0cS_{0l}l$, $S_0cS_{0r}l$, $N_0cN_{0l}l$, $N_0cN_{0r}l$

Table 1: Additional cluster-based parser features. S_i and N_i: the i^{th} tokens in the stack and buffer. p: the part-of-speech tag, c: the cluster. v: the valence of the left (l) or right (r) set of children. l: the label of the token under consideration. d: distance between the words on the top of the stack and buffer. S_{ih}, S_{ir} and S_{il}: the head, right-most modifier and left-most modifier of the token at the top of the stack. $Gx\{y\}$ expands to Gxy and Gx.

embedding method based on canonical correlation analysis that provides state-of-the art results for word-based SSL for English NER. As an alternative to clustering words, Lin and Wu (2009) proposed a phrase clustering approach that obtained the state-of-the-art result for English NER.

3 Monolingual Cluster Experiments

Before moving on to the multilingual setting, we conduct a set of monolingual experiments where we evaluate the use of the monolingual word clusters just described as features for dependency parsing and NER. In the parsing experiments, we study the following thirteen languages:[1] Danish (DA), German (DE), Greek (EL), English (EN), Spanish (ES), French (FR), Italian (IT), Korean (KO), Dutch (NL), Portugese (PT), Russian (RU), Swedish (SV) and Chinese (ZH) – representing the Chinese, Germanic, Hellenic, Romance, Slavic, Altaic and Korean genera. In the NER experiments, we study three Germanic languages: German (DE), English (EN) and Dutch (NL); and one Romance language: Spanish (ES).

Details of the labeled and unlabeled data sets used are given in Appendix A. For all experiments we fixed the number of clusters to 256 as this performed well on held-out data. Furthermore, we only clustered the 1 million most frequent word types in each language for both efficiency and sparsity reasons. For languages in which our unlabeled data did not have at least 1 million types, we considered all types.

[1]The particular choice of languages was made purely based on data availability and institution licensing.

Word & bias	$w_{-1,0,1}$, $w_{-1:0}$, $w_{0:1}$, $w_{-1:1}$, b
Pre-/suffix	$w^{:1,:2,:3,:4,:5}_{-1,0,1}$, $w^{-5:,-4:,-3:,-2:,-1:}_{-1,0,1}$
Orthography	$\mathrm{Hyp}_{-1,0,1}$, $\mathrm{Cap}_{-1,0,1}$, $\mathrm{Cap}_{-1:0}$, $\mathrm{Cap}_{0:1}$, $\mathrm{Cap}_{-1:1}$
PoS	$p_{-1,0,1}$, $p_{-1:0}$, $p_{0:1}$, $p_{-1:1}$, $p_{-2:1}$, $p_{-1:2}$
Cluster	$c_{-1,0,1}$, $c_{-1:0}$, $c_{0:1}$, $c_{-1:1}$, $c_{-2:1}$, $c_{-1:2}$
Transition	$\rightarrow/p_{-1,0,1}$, $\rightarrow/c_{-1,0,1}$, $\rightarrow/\mathrm{Cap}_{-1,0,1}$, $\rightarrow$/b

Table 2: NER features. Hyp: Word contains hyphen. Cap: First letter is capitalized. $\rightarrow/f$ - Transition from previous to current label conjoined with feature f. $w^{:j}$: j-character prefix of w. $w^{-j:}$: j-character suffix of w. f_i: Feature f at relative position i. $f_{i,j}$: Union of features at positions i and j. $f_{i:j}$: Conjoined feature sequence between relative positions i and j (inclusive). b: Bias.

3.1 Cluster Augmented Feature Models

All of the parsing experiments reported in this study are based on the transition-based dependency parsing paradigm (Nivre, 2008). For all languages and settings, we use an arc-eager decoding strategy, with a beam of eight hypotheses, and perform ten epochs of the averaged structured perceptron algorithm (Zhang and Clark, 2008). We extend the state-of-the-art feature model recently introduced by Zhang and Nivre (2011) by adding an additional word cluster based feature template for each word based template. Additionally, we add templates where one or more part-of-speech feature is replaced with the corresponding cluster feature. The resulting set of additional feature templates are shown in Table 1. The expanded feature model includes all of the feature templates defined by Zhang and Nivre (2011), which we also use as the baseline model, whereas Table 1 only shows our new templates due to space limitations.

For all NER experiments, we use a sequential first-order conditional random field (CRF) with a unit variance Normal prior, trained with L-BFGS until ϵ-convergence ($\epsilon = 0.0001$, typically obtained after less than 400 iterations). The feature model used for the NER tagger is shown in Table 2. These are similar to the features used by Turian et al. (2010), with the main difference that we do not use any long range features and that we add templates that conjoin adjacent clusters and adjacent tags as well as templates that conjoin label transitions with tags, clusters and capitalization features.

	DA	DE	EL	EN	ES	FR	IT	KO	NL	PT	RU	SV	ZH	AVG
NO CLUSTERS	84.3	88.9	76.1	90.3	82.8	85.7	81.4	82.0	77.2	86.9	83.5	84.7	74.9	*83.0*
CLUSTERS	85.8	89.5	77.3	90.7	83.6	85.7	82.2	83.6	77.8	87.6	86.0	86.5	75.5	*84.0*

Table 3: Supervised parsing results measured with labeled attachment score (LAS) on the test set. All results are statistically significant at $p < 0.05$, except FR and NL.

	DE	EN	ES	NL	AVG
NO CLUSTERS	65.4	89.2	75.0	75.7	*76.3*
CLUSTERS	74.8	91.8	81.1	84.2	*83.0*
↑ DEVELOPMENT SET ↓ TEST SET					
NO CLUSTERS	69.1	83.5	78.9	79.6	*77.8*
CLUSTERS	74.4	87.8	82.0	85.7	*82.5*

Table 4: Supervised NER results measured with F_1-score on the CoNLL 2002/2003 development and test sets.

3.2 Results

The results of the parsing experiments, measured with labeled accuracy score (LAS) on all sentence lengths, excluding punctuation, are shown in Table 3. The baselines are all comparable to the state-of-the-art. On average, the addition of word cluster features yields a 6% relative reduction in error and upwards of 15% (for RU). All languages improve except FR, which sees neither an increase nor a decrease in LAS. We observe an average absolute increase in LAS of approximately 1%, which is inline with previous observations (Koo et al., 2008). It is perhaps not surprising that RU sees a large gain as it is a highly inflected language, making observations of lexical features far more sparse. Some languages, e.g., FR, NL, and ZH see much smaller gains. One likely culprit is a divergence between the tokenization schemes used in the treebank and in our unlabeled data, which for Indo-European languages is closely related to the Penn Treebank tokenization. For example, the NL treebank contains many multi-word tokens that are typically broken apart by our automatic tokenizer.

The NER results, in terms of F_1 measure, are listed in Table 4. Introducing word cluster features for NER reduces relative errors on the test set by 21% (39% on the development set) on average. Broken down per language, reductions on the test set vary from substantial for NL (30%) and EN (26%), down to more modest for DE (17%) and ES (12%). The addition of cluster features most markedly improve recognition of the PER category, with an average error reduction on the test set of 44%, while the reductions for ORG (19%), LOC (17%) and MISC (10%) are more modest, but still significant. Although our results are below the best reported results for EN and DE (Lin and Wu, 2009; Faruqui and Padó, 2010), the relative improvements of adding word clusters are inline with previous results on NER for EN (Miller et al., 2004; Turian et al., 2010), who report error reductions of approximately 25% from adding word cluster features. Slightly higher reductions where achieved for DE by Faruqui and Padó (2010), who report a reduction of 22%. Note that we did not tune hyper-parameters of the supervised learning methods and of the clustering method, such as the number of clusters (Turian et al., 2010; Faruqui and Padó, 2010), and that we did not apply any heuristic for data cleaning such as that used by Turian et al. (2010).

4 Cross-lingual Word Cluster Features

All results of the previous section rely on the availability of large quantities of language specific annotations for each task. Cross-lingual transfer methods and unsupervised methods have recently been shown to hold promise as a way to at least partially sidestep the demand for labeled data. Unsupervised methods attempt to infer linguistic structure without using any annotated data (Klein and Manning, 2004) or possibly by using a set of linguistically motivated rules (Naseem et al., 2010) or a linguistically informed model structure (Berg-Kirkpatrick and Klein, 2010). The aim of transfer methods is instead to use knowledge induced from labeled resources in one or more *source* languages to construct systems for *target* languages in which no or few such resources are available (Hwa et al., 2005). Currently, the performance of even the most simple direct transfer systems far exceeds that of unsupervised systems (Cohen et al., 2011; McDonald et al., 2011; Søgaard, 2011).

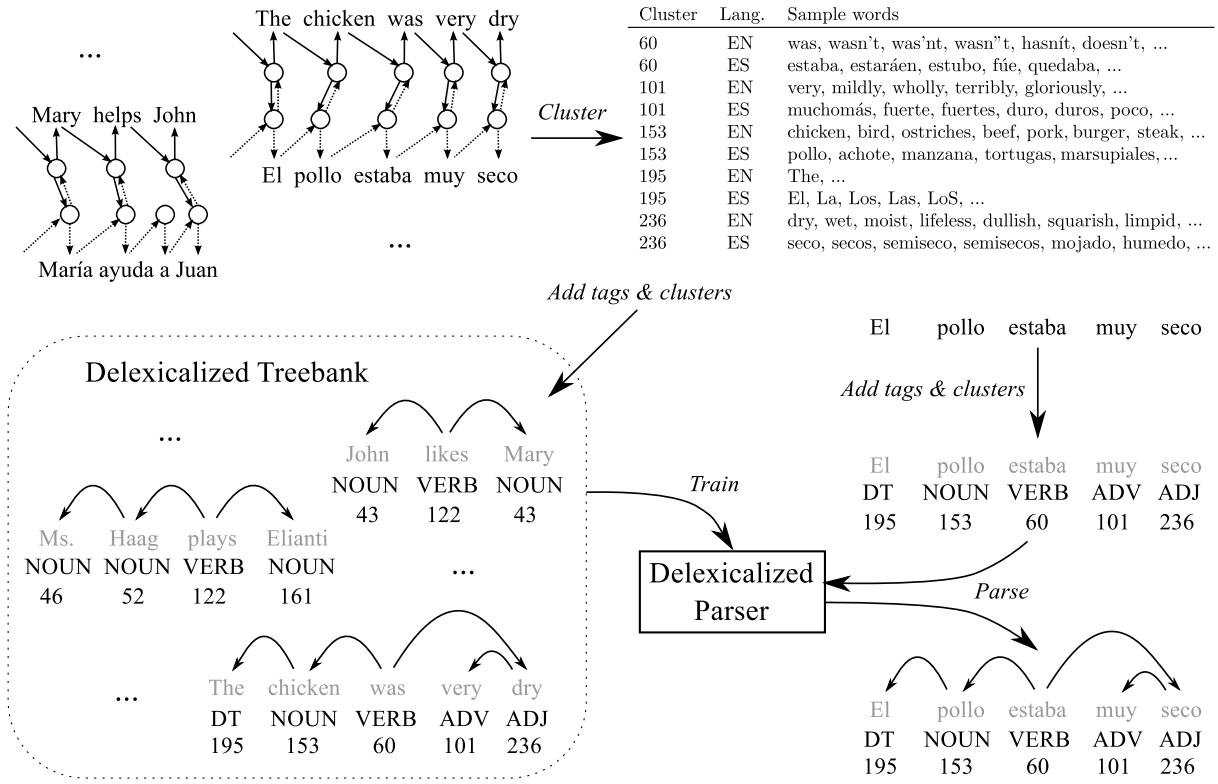

Figure 1: Cross-lingual word cluster features for parsing. Top-left: Cross-lingual (EN-ES) word clustering model. Top-right: Samples of some of the induced cross-lingual word clusters. Bottom-left: Delexicalized cluster-augmented source (EN) treebank for training transfer parser. Bottom-right: Parsing of target (ES) sentence using the transfer parser.

4.1 Direct Transfer of Discriminative Models

Our starting point is the delexicalized direct transfer method proposed by McDonald et al. (2011) based on work by Zeman and Resnik (2008). This method was shown to outperform a number of state-of-the-art unsupervised and transfer-based baselines. The method is simple; for a given training set, the learner ignores all lexical identities and only observes features over other characteristics, e.g., part-of-speech tags, orthographic features, direction of syntactic attachment, etc. The learner builds a model from an annotated source language data set, after which the model is used to directly make target language predictions.

There are three basic assumptions that drive this approach. First, that high-level tasks, such as syntactic parsing, can be learned reliably using coarse-grained statistics, such as part-of-speech tags, in place of fine-grained statistics such as lexical word identities. Second, that the parameters of features over coarse-grained statistics are in some sense language independent, e.g., a feature that indicates that adjectives modify their closest noun is useful in all languages. Third, that these coarse-grained statistics are robustly available across languages. The approach proposed by McDonald et al. (2011) relies on these three assumptions. Specifically, by replacing fine-grained language specific part-of-speech tags with universal part-of-speech tags, generated with the method described by Das and Petrov (2011), a universal parser is achieved that can be applied to any language for which universal part-of-speech tags are available.

Below, we extend this approach to universal parsing by adding cross-lingual word cluster features. A cross-lingual word clustering is a clustering of words in two languages, in which the clusters correspond to some meaningful cross-lingual property. For example, prepositions from both languages should be in the same cluster, proper names from both languages in another cluster and so on. By adding features defined over these clusters, we can, to some degree,

re-lexicalize the delexicalized models, while maintaining the "universality" of the features. This approach is outlined in Figure 1. Assuming that we have an algorithm for generating cross-lingual word clusters (see Section 4.2), we can augment the delexicalized parsing algorithm to use these word cluster features at training and testing time.

In order to further motivate the proposed approach, consider the accuracy of the supervised English parser. A parser with lexical, part-of-speech and cluster features achieves 90.7% LAS (see Table 3). If we remove all lexical and cluster features, the same parser achieves 83.1%. However, if we add back just the cluster features, the accuracy jumps back up to 89.5%, which is only 1.2% below the full system. Thus, if we can accurately learn cross-lingual clusters, there is hope of regaining some of the accuracy lost due to the delexicalization process.

4.2 Inducing Cross-lingual Word Clusters

Our first method for inducing cross-lingual clusters has two stages. First, it clusters a source language (S) as in the monolingual case, and then projects these clusters to a target language (T), using word alignments. Given two aligned word sequences $\boldsymbol{w}^S = \langle w_i^S \rangle_{i=1}^{m_S}$ and $\boldsymbol{w}^T = \langle w_i^T \rangle_{j=1}^{m_T}$, let $\mathcal{A}^{T|S}$ be a set of scored alignments from the source language to the target language, where $(w_j^T, w_{a_j}^S, s_{j,a_j}) \in \mathcal{A}^{T|S}$ is an alignment from the a_jth source word to the jth target word, with score $s_{j,a_j} \geq \delta$.[2] We use the shorthand $j \in \mathcal{A}^{T|S}$ to denote those target words w_j^T that are aligned to some source word $w_{a_j}^S$. Provided a clustering $\mathcal{C}^S$, we assign the target word $t \in \mathcal{V}^T$ to the cluster with which it is most often aligned:

$$\mathcal{C}^T(t) = \operatorname*{argmax}_k \sum_{\substack{j \in \mathcal{A}^{T|S} \\ \text{s.t. } w_j^T = t}} s_{j,a_j} \Big[\mathcal{C}^S(w_{a_j}^S) = k \Big] , \quad (2)$$

where $[\cdot]$ is the indicator function. We refer to the cross-lingual clusters induced in this way as PROJECTED CLUSTERS.

This simple projection approach has two potential drawbacks. First, it only provides a clustering of those target language words that occur in the word aligned data, which is typically smaller than our monolingual data sets. Second, the mapped clustering may not necessarily correspond to an acceptable target language clustering in terms of monolingual likelihood. In order to tackle these issues, we propose the following more complex model. First, to find clusterings that are good according to both the source and target language, and to make use of more unlabeled data, we model word sequences in each language by the monolingual language model with likelihood function defined by equation (1). Denote these likelihood functions respectively by $L^S(\boldsymbol{w}^S; \mathcal{C}^S)$ and $L^T(\boldsymbol{w}^T; \mathcal{C}^T)$, where we have overloaded notation so that the word sequences denoted by $\boldsymbol{w}^S$ and $\boldsymbol{w}^T$ include much more plentiful non-aligned data when taken as an argument of the monolingual likelihood functions. Second, we couple the clusterings defined by these individual models, by introducing additional factors based on word alignments, as proposed by Och (1999):

$$L^{T|S}(\boldsymbol{w}^T; \mathcal{A}^{T|S}, \mathcal{C}^T, \mathcal{C}^S) =$$

$$\prod_{j \in \mathcal{A}^{T|S}} p(w_j^T | \mathcal{C}^T(w_j^T)) p(\mathcal{C}^T(w_j^T) | \mathcal{C}^S(w_{a_j}^S)) \, .$$

and the symmetric $L^{S|T}(\boldsymbol{w}^S; \mathcal{A}^{S|T}, \mathcal{C}^S, C^T)$. Note that the simple projection defined by equation (2) correspond to a hard assignment variant of this probabilistic formulation when the source clustering is fixed. Combining all four factors results in the joint monolingual and cross-lingual objective function

$$L^{S,T}(\boldsymbol{w}^S, \boldsymbol{w}^T; \mathcal{A}^{T|S}, \mathcal{A}^{S|T}, \mathcal{C}^S, \mathcal{C}^T) =$$

$$L^S(\ldots) \cdot L^T(\ldots) \cdot L^{T|S}(\ldots) \cdot L^{S|T}(\ldots) \, . \quad (3)$$

The intuition of this approach is that the clusterings $\mathcal{C}^S$ and $\mathcal{C}^T$ are forced to jointly explain the source and target data, treating the word alignments as a form of soft constraints. We approximately optimize (3) with the alternating procedure in Algorithm 1, in which we iteratively maximize L^S and L^T, keeping the other factors fixed. In this way we can generate cross-lingual clusterings using all the monolingual data while forcing the clusterings to obey the word alignment constraints. We refer to the clusters induced with this method as X-LINGUAL CLUSTERS.

In practice we found that each unconstrained monolingual run of the exchange algorithm (lines

[2] In our case, the alignment score corresponds to the conditional alignment probability $p(w_j^T | w_{a_j}^S)$. All ϵ-alignments are ignored and we use $\delta = 0.95$ throughout.

Algorithm 1 Cross-lingual clustering.

Randomly initialize source/target clusterings $\mathcal{C}^S$ and $\mathcal{C}^T$.
for $i = 1 \ldots N$ **do**
1. Find $\tilde{\mathcal{C}}^S \approx \text{argmax}_{\mathcal{C}^S} L^S(\boldsymbol{w}^S; \mathcal{C}^S)$. (†)
2. Project $\tilde{\mathcal{C}}^S$ to $\mathcal{C}^T$ using equation (2).
   - keep cluster of non-projected words in $\mathcal{C}^T$ fixed.
3. Find $\tilde{\mathcal{C}}^T \approx \text{argmax}_{\mathcal{C}^T} L^T(\boldsymbol{w}^T; \mathcal{C}^T)$. (†)
4. Project $\tilde{\mathcal{C}}^T$ to $\mathcal{C}^S$ using equation (2).
   - keep cluster of non-projected words in $\mathcal{C}^S$ fixed.

end for

† Optimized via the exchange algorithm keeping the cluster of projected words fixed and only clustering additional words not in the projection.

1 and 3) moves the clustering too far from those that obey the word alignment constraints, which causes the procedure to fail to converge. However, we found that fixing the clustering of the words that are assigned clusters in the projection stages (lines 2 and 4) and only clustering the remaining words works well in practice. Furthermore, we found that iterating the procedure has little effect on performance and set $N = 1$ for all subsequent experiments.

5 Cross-lingual Experiments

In our first set of experiments on using cross-lingual cluster features, we evaluate direct transfer of our EN parser, trained on Stanford style dependencies (De Marneffe et al., 2006), to the the ten non-EN Indo-European languages listed in Section 3. We exclude KO and ZH as initial experiments proved direct transfer a poor technique when transferring parsers between such diverse languages. We study the impact of using cross-lingual cluster features by comparing the strong delexicalized baseline model of McDonald et al. (2011), which only has features derived from universal part-of-speech tags, projected from English with the method of Das and Petrov (2011), to the same model when adding features derived from cross-lingual clusters. In both cases the feature models are the same as those used in Section 3.1, except that they are delexicalized by removing all lexical word-identity features. We evaluate both the PROJECTED CLUSTERS and the X-LINGUAL CLUSTERS.

For these experiments we train the perceptron for only five epochs in order to prevent over-fitting, which is an acute problem due to the divergence between the training and testing data sets in this setting. Furthermore, in accordance to standard practices we only evaluate unlabeled attachment score (UAS) due to the fact that each treebank uses a different – possibly non-overlapping – label set.

In our second set of experiments, we evaluate direct transfer of a NER system trained on EN to DE, ES and NL. We use the same feature models as in the monolingual case, with the exception that we use universal part-of-speech tags for all languages and we remove the capitalization feature when transferring from EN to DE. Capitalization is both a prevalent and highly predictive feature of named-entities in EN, while in DE, capitalization is even more prevalent, but has very low predictive power. Interestingly, while delexicalization has shown to be important for direct transfer of dependency-parsers (McDonald et al., 2011), we noticed in preliminary experiments that it substantially degrades performance for NER. We hypothesize that this is because word features are predictive of common proper names and that these are often translated directly across languages, at least in the case of newswire text. As for the transfer parser, when training the source NER model, we regularize the model more heavily by setting $\sigma = 0.1$.

Appendix A contains the details of the training, testing, unlabeled and parallel/aligned data sets.

5.1 Results

Table 5 lists the results of the transfer experiments for dependency parsing. The baseline results are comparable to those in McDonald et al. (2011) and thus also significantly outperform the results of recent unsupervised approaches (Berg-Kirkpatrick and Klein, 2010; Naseem et al., 2010). Importantly, cross-lingual cluster features are helpful across the board and give a relative error reduction ranging from 3% for DA to 13% for PT, with an average reduction of 6%, in terms of unlabeled attachment score (UAS). This shows the utility of cross-lingual cluster features for syntactic transfer. However, X-LINGUAL CLUSTERS provides roughly the same performance as PROJECTED CLUSTERS suggesting that even simple methods of cross-lingual clustering are sufficient for direct transfer dependency parsing.

We would like to stress that these results are likely to be under-estimating the parsers' actual ability to predict Stanford-style dependencies in the target languages. This is because the target language annotations that we use for evaluation differ from the

	DA	DE	EL	ES	FR	IT	NL	PT	RU	SV	AVG
NO CLUSTERS	36.7	48.9	59.5	60.2	70.0	64.6	52.8	66.8	29.7	55.4	*54.5*
PROJECTED CLUSTERS	38.9	50.3	61.1	62.6	71.6	68.6	54.5	70.7	32.9	57.0	*56.8*
X-LINGUAL CLUSTERS	38.7	50.7	63.0	62.9	72.1	68.8	54.3	71.0	34.4	56.9	*57.3*
↑ ALL DEPENDENCY RELATIONS ↓ ONLY SUBJECT/OBJECT RELATIONS											
NO CLUSTERS	44.6	56.7	67.2	60.7	77.4	64.6	59.5	53.3	29.3	57.3	*57.1*
PROJECTED CLUSTERS	49.8	57.1	72.2	65.9	80.4	70.5	67.0	62.6	34.6	65.0	*62.5*
X-LINGUAL CLUSTERS	49.2	59.0	72.5	65.9	80.9	72.7	65.7	62.5	37.2	64.4	*63.0*

Table 5: Direct transfer dependency parsing from English. Results measured by unlabeled attachment score (UAS). ONLY SUBJECT/OBJECT RELATIONS – UAS measured only over words marked as subject/object in the evaluation data.

Stanford dependency annotation. Some of these differences are warranted in that certain target language phenomena are better captured by the native annotation. However, differences such as choice of lexical versus functional head are more arbitrary.

To highlight this point we run two additional experiments. First, we had linguists, who were also fluent speakers of German, re-annotate the DE test set so that unlabeled arcs are consistent with Stanford-style dependencies. Using this data, NO CLUSTERS obtains 60.0% UAS, PROJECTED CLUSTERS 63.6% and X-LINGUAL CLUSTERS 64.4%. When compared to the scores on the original data set (48.9%, 50.3% and 50.7%, respectively) we can see that not only is the baseline system doing much better, but that the improvements from cross-lingual clustering are much more pronounced. Next, we investigated the accuracy of subject and object dependencies, as these are often annotated in similar ways across treebanks, typically modifying the main verb of the sentence. The bottom half of Table 5 gives the scores when restricted to such dependencies in the gold data. We measure the percentage of modifiers in subject and object dependencies that modify the correct word. Indeed, here we see the difference in performance become clearer, with the cross-lingual cluster model reducing errors by 14% relative to the non-cross-lingual model and upwards of 22% relative for IT.

We now turn to the results of the transfer experiments for NER, listed in Table 6. While the performance of the transfer systems is very poor when no word clusters are used, adding cross-lingual word clusters give substantial improvements across all languages. The simple PROJECTED CLUSTERS work well, but the X-LINGUAL CLUSTERS provide even larger improvements. On average the latter reduce errors on the test set by 22% in terms of F_1 and up to 26% for ES. We also measure how well the direct transfer NER systems are able to detect entity boundaries (ignoring the entity categories). Here, on average, the best clusters provide a 24% relative error reduction on the test set (75.8 vs. 68.1 F_1).

	DE	ES	NL	AVG
NO CLUSTERS	25.4	49.5	49.9	*41.6*
PROJECTED CLUSTERS	39.1	62.1	61.8	*54.4*
X-LINGUAL CLUSTERS	43.1	62.8	64.7	*56.9*
↑ DEVELOPMENT SET ↓ TEST SET				
NO CLUSTERS	23.5	45.6	48.4	*39.1*
PROJECTED CLUSTERS	35.2	59.1	56.4	*50.2*
X-LINGUAL CLUSTERS	40.4	59.3	58.4	*52.7*

Table 6: Direct transfer NER results (from English) measured with average F_1-score on the CoNLL 2002/2003 development and test sets.

To our knowledge there are no comparable results on transfer learning of NER systems. Based on the results of this first attempt at this scenario, we believe that transfer learning by multilingual word clusters could be developed into a practical way to construct NER systems for resource poor languages.

6 Conclusion

In the first part of this study, we showed that word clusters induced from a simple class-based language model can be used to significantly improve on state-of-the-art supervised dependency parsing and NER for a wide range of languages and even across language families. Although the improvements vary between languages, the addition of word cluster features never has a negative impact on performance.

This result has important practical consequences as it allows practitioners to simply plug in word cluster features into their current feature models. Given previous work on word clusters for various linguistic structure prediction tasks, these results are not too surprising. However, to our knowledge this is the first study to apply the same type of word cluster features across languages and tasks.

In the second part, we provided two simple methods for inducing cross-lingual word clusters. The first method works by projecting word clusters, induced from monolingual data, from a source language to a target language directly via word alignments. The second method, on the other hand, makes use of monolingual data in both the source and the target language, together with word alignments that act as constraints on the joint clustering. We then showed that by using these cross-lingual word clusters, we can significantly improve on direct transfer of discriminative models for both parsing and NER. As in the monolingual case, both types of cross-lingual word cluster features yield improvements across the board, with the more complex method providing a significantly larger improvement for NER. Although the performance of transfer systems is still substantially below that of supervised systems, this research provides one step towards bridging this gap. Further, we believe that it opens up an avenue for future work on multilingual clustering methods, cross-lingual feature projection and domain adaptation for direct transfer of linguistic structure.

Acknowledgments

We thank John DeNero for help with creating the word alignments; Reut Tsarfaty and Joakim Nivre for rewarding discussions on evaluation; Slav Petrov and Kuzman Ganchev for discussions on cross-lingual clustering; and the anonymous reviewers, along with Joakim Nivre, for valuable comments that helped improve the paper. The first author is grateful for the financial support of the Swedish National Graduate School of Language Technology (GSLT).

A Data Sets

In the parsing experiments, we use the following data sets. For DA, DE, EL, ES, IT, NL, PT and SV, we use the predefined training and evaluation data sets from the CoNLL 2006/2007 data sets (Buchholz and Marsi, 2006; Nivre et al., 2007). For EN we use sections 02-21, 22, and 23 of the Penn WSJ Treebank (Marcus et al., 1993) for training, development and evaluation. For FR we used the French Treebank (Abeillé and Barrier, 2004) with splits defined in Candito et al. (2010). For KO we use the Sejong Korean Treebank (Han et al., 2002) randomly splitting the data into 80% training, 10% development and 10% evaluation. For RU we use the SynTagRus Treebank (Boguslavsky et al., 2000; Apresjan et al., 2006) randomly splitting the data into 80% training, 10% development and 10% evaluation. For ZH we use the Penn Chinese Treebank v6 (Xue et al., 2005) using the proposed data splits from the documentation. Both EN and ZH were converted to dependencies using v1.6.8 of the Stanford Converter (De Marneffe et al., 2006). FR was converted using the procedure defined in Candito et al. (2010). RU and KO are native dependency treebanks. For the CoNLL data sets we use the part-of-speech tags provided with the data. For all other data sets, we train a part-of-speech tagger on the training data in order to tag the development and evaluation data.

For the NER experiments we use the training, development and evaluation data sets from the CoNLL 2002/2003 shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) for all four languages (DE, EN, ES and NL). The data set for each language consists of newswire text annotated with four entity categories: Location (LOC), Miscellaneous (MISC), Organization (ORG) and Person (PER). We use the part-of-speech tags supplied with the data, except for ES where we instead use universal part-of-speech tags (Petrov et al., 2011).

Unlabeled data for training the monolingual cluster models was extracted from one year of newswire articles from multiple sources from a news aggregation website. This consists of 0.8 billion (DA) to 121.6 billion (EN) tokens per language. All word alignments for the cross-lingual clusterings were produced with the dual decomposition aligner described by DeNero and Macherey (2011) using 10.5 million (DA) to 12.1 million (FR) sentences of aligned web data.

References

Anne Abeillé and Nicolas Barrier. 2004. Enriching a french treebank. In *Proceedings of LREC*.

Juri Apresjan, Igor Boguslavsky, Boris Iomdin, Leonid Iomdin, Andrei Sannikov, and Victor Sizov. 2006. A syntactically and semantically tagged corpus of russian: State of the art and prospects. In *Proceedings of LREC*.

Emily M. Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26.

Taylor Berg-Kirkpatrick and Dan Klein. 2010. Phylogenetic grammar induction. In *Proceedings of ACL*.

Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34(1):211–231.

Igor Boguslavsky, Svetlana Grigorieva, Nikolai Grigoriev, Leonid Kreidlin, and Nadezhda Frid. 2000. Dependency treebank for Russian: Concept, tools, types of information. In *Proceedings of COLING*.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL*.

Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical french dependency parsing: treebank conversion and first results. In *Proceedings of LREC*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL*.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of EACL*.

Shay B. Cohen and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *Proceedings of NAACL*.

Shay B. Cohen, Dipanjan Das, and Noah A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of EMNLP*.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL*.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of ACL*.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of ICML*.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL-HLT*.

Marie-Catherine De Marneffe, Bill MacCartney, and Chris D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*.

John DeNero and Klaus Macherey. 2011. Model-based aligner combination using dual decomposition. In *Proceedings of ACL-HLT*.

Paramveer Dhillon, Dean Foster, and Lyle Dean. 2011. Multi-view learning of word embeddings via cca. In *Proceedings of NIPS*.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS*.

Dayne Freitag. 2004. Trained named entity recognition using distributional clusters. In *Proceedings of EMNLP*.

Jennifer Gillenwater, Kuzman Ganchev, João Graça, Fernando Pereira, and Ben Taskar. 2010. Sparsity in dependency grammar induction. In *Proceedings of ACL*.

Gholamreza Haffari, Marzieh Razavi, and Anoop Sarkar. 2011. An ensemble model that combines syntactic and semantic clustering for discriminative dependency parsing. In *Proceedings of ACL*.

Chung-hye Han, Na-Rare Han, Eon-Suk Ko, and Martha Palmer. 2002. Development and evaluation of a korean treebank and its application to nlp. In *Proceedings of LREC*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(03):311–325.

Dan Klein and Chris D. Manning. 2004. Corpus-based induction of syntactic structure: models of dependency and constituency. In *Proceedings of ACL*.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-HLT*.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency parsing*. Morgan & Claypool Publishers.

Dekang Lin and Xiaoyun Wu. 2009. Phrase clustering for discriminative learning. In *Proceedings of ACL-IJCNLP*, pages 1030–1038.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of EMNLP*.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proceedings of HLT-NAACL*.

Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of ICML*.

Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of EMNLP*.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of EMNLP-CoNLL*.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *Proceedings of EACL*.

Bo Pang and Lillian Lee. 2008. *Opinion mining and sentiment analysis*. Now Publishers Inc.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. In *ArXiv:1104.2086*.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of EMNLP*.

Noah A. Smith. 2011. *Linguistic Structure Prediction*. Morgan & Claypool Publishers.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of ACL*.

Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2011. Lateen EM: Unsupervised training with multiple objectives, applied to dependency grammar induction. In *Proceedings of EMNLP*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL*.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL*.

Stephen Tratz and Eduard Hovy. 2011. A fast, effective, non-projective, semantically-enriched parser. In *Proceedings of EMNLP*.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*.

Jakob Uszkoreit and Thorsten Brants. 2008. Distributed word clustering for large scale class-based language modeling in machine translation. In *Proceedings of ACL-HLT*.

Naiwen Xue, Fei Xia, Fu-dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(02):207–238.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP Workshop: NLP for Less Privileged Languages*.

Yue Zhang and Stephen Clark. 2008. A tale of two parsers: Investigating and combining graph-based and transition-based dependency parsing. In *Proceedings of EMNLP*.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of ACL-HLT*.