# Nudging the Envelope of Direct Transfer Methods for Multilingual Named Entity Recognition

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

## Abstract

In this paper, we study direct transfer methods for multilingual named entity recognition. Specifically, we extend the method recently proposed by Täckström et al. (2012), which is based on cross-lingual word cluster features. First, we show that by using multiple source languages, combined with self-training for target language adaptation, we can achieve significant improvements compared to using only single source direct transfer. Second, we investigate how the direct transfer system fares against a supervised target language system and conclude that between 8,000 and 16,000 word tokens need to be annotated in each target language to match the best direct transfer system. Finally, we show that we can significantly improve target language performance, even after annotating up to 64,000 tokens in the target language, by simply concatenating source and target language annotations.

## 1 Introduction

Recognition of named entities in natural language text is an important subtask of information extraction and thus bears importance for modern text mining and information retrieval applications. The need to identify named entities such as persons, locations, organizations and places, arises both in applications where the entities are first class objects of interest, such as in *Wikification* of documents (Ratinov et al., 2011), and in applications where knowledge of named entities is helpful in boosting performance, e.g., machine translation (Babych and Hartley, 2003) and question answering (Leidner et al., 2003). The advent of massive machine readable factual databases, such as Freebase[1] and the proposed Wikidata[2], will likely push the need for automatic extraction tools further. While these databases store information about entity *types* and the relationships between those types, the named entity recognition (NER) task concerns finding occurrences of named entities *in context*. This view originated with the Message Understanding Conferences (MUC) (Grishman and Sundheim, 1996).

As with the majority of tasks in contemporary natural language processing, most approaches to NER have been based on supervised machine learning. However, although resources for a handful of languages have been created, through initiatives such as MUC, the Multilingual Entity Task (Merchant et al., 1996) and the CoNLL shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), coverage is still very limited in terms of both domains and languages. With fine-grained entity taxonomies such as that proposed by Sekine and Nobata (2004), who define over two hundred categories, we can expect an increase in the amount of annotated data required for acceptable performance, as well as an increased annotation cost for each entity occurrence. Although semi-supervised approaches have been shown to reduce the need for manual annotation (Freitag, 2004; Miller et al., 2004; Ando and Zhang, 2005; Suzuki and Isozaki, 2008; Lin and Wu, 2009; Turian et al., 2010; Dhillon et al., 2011; Täckström et al., 2012), these methods still require a substantial amount of manual annotation for each target language. Manually creating a sufficient amount of annotated resources for all entity types in all languages thus seems like an Herculean task.

In this study, we turn to direct transfer methods (McDonald et al., 2011; Täckström et al., 2012) as

[1] http://www.freebase.com

[2] http://meta.wikimedia.org/wiki/Wikidata

a way to combat the need for annotated resources in all languages. These methods allow one to train a system for a target language, using only annotations in some source language, as long as all source language features also have support in the target languages. Specifically, we extend the direct transfer method proposed by Täckström et al. (2012) in two ways. First, in §3, we use multiple source languages for training. We then propose a self-training algorithm, which allows for the inclusion of additional target language specific features, in §4. By combining these extensions, we achieve significant error reductions on all tested languages. Finally, in §5, we assess the viability of the different direct transfer systems compared to a supervised system trained on target language annotations, and conclude that direct transfer methods may be useful even in this scenario.

## 2 Direct Transfer for Cross-lingual NER

Rather than starting from scratch when creating systems that predict linguistic structure in one language, we should be able to take advantage of any corresponding annotations that are available in other languages. This idea is at the heart of both direct transfer methods (McDonald et al., 2011; Täckström et al., 2012) and of annotation projection methods (Yarowsky et al., 2001; Diab and Resnik, 2002; Hwa et al., 2005). While the aim of the latter is to transfer annotations across languages, direct transfer methods instead aim to transfer systems, trained on some source language, directly to other languages. In this paper, we focus on direct transfer methods, however, we briefly discuss the relationship between these approaches in §6.

Considering the substantial differences between languages at the grammatical and lexical level, the prospect of directly applying a system trained on one language to another language may seem bleak. However, McDonald et al. (2011) showed that a language independent dependency parser can indeed be created by training on a delexicalized treebank and by only incorporating features defined on universal part-of-speech tags (Das and Petrov, 2011).

Recently, Täckström et al. (2012) developed an algorithm for inducing cross-lingual word clusters and proposed to use these clusters to enrich the feature space of direct transfer systems. The richer set of cross-lingual features was shown to substantially improve on direct transfer of both dependency parsing and NER from English to other languages.

Cross-lingual word clusters are clusterings of words in two (or more) languages, such that the clusters are adequate in each language and at the same time consistent across languages. For cross-lingual word clusters to be useful in direct transfer of linguistic structure, the clusters should capture cross-lingual properties on both the semantic and syntactic level. Täckström et al. (2012) showed that this is, at least to some degree, achievable by coupling monolingual class-based language models, via word alignments. The basic building block is the following simple monolingual class-based language model (Saul and Pereira, 1997; Uszkoreit and Brants, 2008):

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

where $L(\boldsymbol{w}; \mathcal{C})$ is the likelihood of a sequence of words, $\boldsymbol{w}$, and $\mathcal{C}$ is a (hard) clustering function, which maps words to cluster identities. These monolingual models are coupled through word alignments, which constrains the clusterings to be consistent across languages, and optimized by approximately maximizing the joint likelihood across languages. Just as monolingual word clusters are broadly applicable as features in monolingual models for linguistic structure prediction (Turian et al., 2010), the resulting cross-lingual word clusters can be used as features in various cross-lingual direct transfer models. We believe that the extensions that we propose are likely to be useful for other tasks as well, e.g., direct transfer dependency parsing, in this paper, we focus solely on discriminative direct transfer models for NER.

## 3 Multi-source Direct Transfer

Learning from multiple languages have been shown to be of benefit both in unsupervised learning of syntax and part-of-speech (Snyder et al., 2009; Berg-Kirkpatrick and Klein, 2010) and in transfer learning of dependency syntax (Cohen et al., 2011; McDonald et al., 2011). Here we perform a set of experiments where we investigate the potential of multi-source transfer for NER, in German (DE), English (EN), Spanish (ES) and Dutch (NL), using cross-lingual word clusters. For all experiments, we use the same

| Source | DE | ES | NL |
|---|---|---|---|
| EN | 39.7 | 62.0 | 63.7 |
| EN + DE | – | 61.8 | 65.5 |
| EN + ES | 39.3 | – | 65.6 |
| EN + NL | **41.0** | 62.5 | – |
| ALL | **41.0** | **63.6** | **66.4** |
| ↑ DEVELOPMENT SET ↓ TEST SET | | | |
| EN | 37.8 | 59.1 | 57.2 |
| EN + DE | – | 59.4 | 57.9 |
| EN + ES | 35.9 | – | 59.1 |
| EN + NL | **38.1** | 59.7 | – |
| ALL | 36.4 | **61.9** | **59.9** |

Table 1: Results of multi-source direct transfer, measured with $F_1$-score on the CoNLL 2002/2003 development and test sets. ALL: all languages except the target language are used as source languages.

256 cross-lingual word clusters and the same feature templates as Täckström et al. (2012), with the exception that the transition factors are not conditioned on the input.[3] The features used are similar to those used by Turian et al. (2010), but include cross-lingual rather than monolingual word clusters. We remove the capitalization features when transferring to German, but keep them in all other cases, even when German is included in the set of source languages. We use the training, development and test data sets provided by the CoNLL 2002/2003 shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). The multi-source training sets are created by concatenating each of the source languages' training sets. In order to have equivalent label sets across languages, we use the IO (inside/outside) encoding, rather than the BIO (begin/inside/outside) encoding, since the latter is available only for Spanish and Dutch. The models are trained using CRFSuite 0.12 (Okazaki, 2007), by running stochastic gradient descent for a maximum of 100 iterations.

Table 1 shows the result of using different source languages for different target languages. We see that multi-source transfer is somewhat helpful in general, but that the results are sensitive to the combination of source and target languages. On average, using all source languages only give a relative error reduction of about 3% on the test set. However, results for Spanish and Dutch are more promising, with relative reductions of 7% and 6%, respectively, when using all source languages. Using all available source languages gives the best results for both Spanish and Dutch, but slightly worse results for German. When transferring to Dutch, using more source languages consistently help, while Spanish and German are more sensitive to the choice of source languages. Based on the characteristics of these languages, this is not too surprising: while Dutch and German has the most similar vocabularies, Dutch uses similar capitalization rules to English and Spanish. Dutch should thus benefit from all the other languages, while Spanish may not bring much to the table for German and vice versa, given their lexical differences. Knowledge of such relationships between the languages, could potentially be used to give different weights to different source languages in the training objective, as was shown effective by Cohen et al. (2011) in the context of direct transfer of generative dependency parsing models. Although better results could be achieved by cherry-picking language combinations, since we do not have any general principled way of choosing/weighting source languages in discriminative models, we include all source languages with equal weight in all subsequent experiments where multiple source languages are used.

[3] This is due to limitations in the sequence labeling software used and gives slightly lower results, across the board, than those reported by Täckström et al. (2012).

| | DE | ES | NL | AVG |
|---|---|---|---|---|
| NATIVE CLUSTERS | 71.2 | 80.7 | 82.5 | *78.1* |
| X-LING CLUSTERS | 68.9 | 78.8 | 80.9 | *76.2* |
| NATIVE & X-LING CLUST. | 72.5 | 81.2 | 83.6 | *79.1* |
| ↑ DEVELOPMENT SET ↓ TEST SET | | | | |
| NATIVE CLUSTERS | 72.2 | 81.0 | 83.0 | *78.7* |
| X-LING CLUSTERS | 71.0 | 80.2 | 80.7 | *77.3* |
| NATIVE & X-LING CLUST. | 73.5 | 81.8 | 83.7 | *79.7* |

Table 2: The impact of different word clusters in the supervised monolingual setting. Results are measured with $F_1$-score on the CoNLL 2002/2003 development and test sets. NATIVE/X-LING CLUSTERS: The cross-lingual/monolingual clusters from Täckström et al. (2012).

## 4 Domain Adaptation via Self-Training

Thus far, we have not made use of any information specific to the target language, except when inducing the cross-lingual word clusters. However, as shown in Table 2, which lists the results of experiments on

**Algorithm 1** Self-Training for Domain Adaptation

$\mathcal{D}_s^l$: Labeled source domain data
$\mathcal{D}_t^l$: Labeled target domain data (possibly empty)
$\mathcal{D}_t^u$: Unlabeled target domain data
$\delta$: Dominance threshold
$T$: Number of iterations
**procedure** SELFTRAIN($\mathcal{D}_s^l, \mathcal{D}_t^l, \mathcal{D}_t^u, \delta, T$)
  $\theta^0 \leftarrow$ LEARN($\mathcal{D}_s^l \cup \mathcal{D}_t^l$) ▷ Train supervised model
  **for** $i \leftarrow 1$ **to** $T$ **do**
    $P^i \leftarrow$ PREDICT($\mathcal{D}_t^u, \theta^{i-1}$) ▷ Predict w/ curr. mod.
    $F^i \leftarrow$ FILTER($P^i, \delta$) ▷ Filter $p_{\theta^{i-1}}(\boldsymbol{y}^*|\boldsymbol{x}) \leq \delta$
    $S^i \leftarrow$ SAMPLE($F^i$) ▷ Pick $\sim p_{\theta^{i-1}}(\boldsymbol{y}|\boldsymbol{x})$. (†)
    $\theta^i \leftarrow$ LEARN($\mathcal{D}_s^l \cup \mathcal{D}_t^l \cup S^i$) ▷ Retrain
  **end for**
  **return** $\theta^T$ ▷ Return adapted model
**end procedure**

† If LEARN($\cdot$) supports instance weighting, we could weight each instance $(\boldsymbol{x}, \boldsymbol{y}^*) \in F^i$ by $p_{\theta^{i-1}}(\boldsymbol{y}^*|\boldsymbol{x})$ in the training objective, rather than performing sampling according to the same distribution.

supervised target language models trained with different cluster features,[4] these clusters are not optimally adapted to the target language, compared to the monolingual *native* clusters that are induced solely on the target language, without any cross-lingual constraints. This is to be expected, as the probabilistic model used to learn the cross-lingual clusters strikes a balance between two language specific models. On the other hand, this suggests an opportunity for adapting to target language specific features through self-training. In fact, since the direct transfer models are trained using cross-lingual features, the target language can be viewed as simply representing a different domain from the source language.

Self-training has previously been shown to be a simple and effective way to perform domain adaptation for syntactic parsers and other tasks (McClosky et al., 2006; Chen et al., 2011). The idea of self-training for domain adaptation is to first train a supervised predictor on labeled instances from a source domain. This predictor is then used to label instances from some unlabeled target domain. Those instances for which the predictor is confident are added to the source training set, and the process is repeated until some stopping criterion is met. Recently, Daumé et al. (2010) and Chen et al. (2011) proposed more complex domain adaptation techniques, based on co-training. In this work, however, we stick with the simple single-view self-training approach just outlined. In the self-training for domain adaptation method, described by Chen et al. (2011), the top-$k$ instances for which the predictor is most confident are added to the training set in each iteration. We instead propose to weight the target instances selected for self-training in each iteration proportional to the confidence of the classifier trained in the previous iteration.

In short, let $\boldsymbol{x} \in \mathcal{D}_t^u$ be an unlabeled target language input sequence (in our case a sentence) and $\boldsymbol{y}^* \in \mathcal{Y}_t(\boldsymbol{x})$ its top-ranked label sequence (in our case an IO sequence). In the first iteration, a predictor is trained on the labeled source language data, $\mathcal{D}_s^l$. In each subsequent iteration the sequences are scored according to the probabilities assigned by the predictor trained in the previous iteration, $p_{\theta^{i-1}}(\boldsymbol{y}^*|\boldsymbol{x})$. When constructing the training set for the next iteration, we first filter out all instances for which the top-ranked label sequence is not $\delta$-dominating. That is, we filter out all instances $\boldsymbol{x} \in \mathcal{D}_u^t$ such that $p_{\theta^{i-1}}(\boldsymbol{y}^*|\boldsymbol{x}) < \delta$, for some user-specified $\delta$. In this work, we set $\delta = 0.5$, since this guarantees that the output associated with each instance that is kept is assigned the majority of the probability mass. This is important, as we only consider the most likely output $\boldsymbol{y}^*$ for each input $\boldsymbol{x}$, so that sampling low-confidence instances will result in a highly biased sample. After filtering, we sample from the remaining instances, i.e. from the set of instances $\boldsymbol{x} \in \mathcal{D}_u^t$ such that $p_{\theta^{i-1}}(\boldsymbol{y}^*|\boldsymbol{x}) \geq \delta$, adding each instance $(\boldsymbol{x}, \boldsymbol{y}^*)$ to the training set with probability $p_{\theta^{i-1}}(\boldsymbol{y}^*|\boldsymbol{x})$. This procedure is repeated for $T$ iterations as outlined in Algorithm 1. By using instance weighting rather than a top-$k$ list, we remove the need to heuristically set the number of instances to be selected for self-training in each iteration. Further, although we have not verified this empirically, we hypothesize that using instance weighting is more robust than picking only the most confident instances, as it maintains diversity in the training set in the face of uncertainty. Note also that when we have access to target language test data during training, we can perform transductive learning by including the test set in the pool of unlabeled data. This gives the model the opportunity to adapt to the characteristics of the test domain.

Our use of self-training for exploiting features na-

[4] For these experiments, the same settings were used as in the multi-source transfer experiments in §3, with the difference that only target language training data was used.

| | DE | ES | NL | AVG |
|---|---|---|---|---|
| SINGLE | 39.7 | 62.0 | 63.7 | *55.2* |
| MULTI | 41.0 | 63.6 | 66.4 | *57.0* |
| SINGLE + SELF | 42.6 | 65.7 | 64.0 | *57.4* |
| SINGLE + SELF/NATIVE | 44.5 | **66.5** | 65.9 | *59.0* |
| MULTI + SELF | 48.4 | 64.7 | 68.1 | *60.4* |
| MULTI + SELF/NATIVE | **49.5** | **66.5** | **69.7** | ***61.9*** |
| ↑ DEVELOPMENT SET ↓ TEST SET | | | | |
| SINGLE | 37.8 | 59.1 | 57.2 | *51.4* |
| MULTI | 36.4 | 61.9 | 59.9 | *52.8* |
| SINGLE + SELF | 41.3 | 61.0 | 57.8 | *53.3* |
| SINGLE + SELF/NATIVE | 43.0 | 62.5 | 58.9 | *54.8* |
| MULTI + SELF | 45.3 | 62.3 | 61.9 | *56.5* |
| MULTI + SELF/NATIVE | **47.2** | **64.8** | **63.1** | ***58.4*** |

Table 3: Results of different extensions to direct transfer as measured with $F_1$-score on the CoNLL 2002/2003 development and test sets. SINGLE: single-source transfer, MULTI: multi-source transfer, SELF: self-training with only cross-lingual word clusters, SELF/NATIVE: self-training with cross-lingual and native word clusters.

tive to the target language resembles the way McDonald et al. (2011) re-lexicalize a delexicalized direct transfer parser. Both methods allow the model to move weights from shared parameters to more predictive target language specific parameters. However, rather than using the direct transfer parser's own predictions through self-training, these authors project head-modifier relations to the target language through loss-augmented learning (Hall et al., 2011). The bootstrapping methods for language independent NER of Cucerzan and Yarowsky (1999) have a similar effect. Our self-training approach is largely orthogonal to these approaches. We therefore believe that combining these methods could be fruitful.

### 4.1 Experiments

In these experiments we combine direct transfer with self-training using unlabeled target data. This is the transductive setting, as we include the test data (with labels removed, of course) in the unlabeled target data. We investigate the effect of adding self-training (SELF) to the single-source and multi-source transfer settings of §3, where only cross-lingual features are used (SINGLE and MULTI, respectively). We further study the effect of including native monolingual word cluster features in addition to the cross-lingual features (SELF/NATVE). The experimental settings and datasets used are the same as those described in §3. We performed self-training for $T = 5$ iterations for all languages, as preliminary experiments indicated that the procedure converges to a stable solution after this number of iterations. CRFSuite was used to compute all the required probabilities for the filtering and sampling steps.

The results of these experiments are shown in Table 3. By itself, self-training without target specific features result in an average relative error reduction of less than 4%, compared to the baseline direct transfer system. This is only slightly better than the improvement achieved with multi-source transfer. However, when adding target specific features, self-training works better, with a 7% reduction. Combining multi-source transfer with self-training, without target specific features, performs even better with a 10% reduction. Finally, combining multi-source transfer and self-training with target specific features, gives the best result across all three languages, with an average relative error reduction of more than 14%.

The results for German are particularly interesting, in that they highlight a rather surprising general trend. The relative improvement achieved by combining multi-source transfer and self training with native clusters is almost twice as large as that achieved when using only self-training with native clusters, despite the fact that multi-source transfer is not very effective on its own – in the case of German, multi-source transfer actually hurts results when used in isolation. One explanation for this behavior could be that the regularization imposed by the use of multiple source languages is beneficial to self-training, in that it generates better confidence estimates. Another, perhaps more speculative, explanation could be that each source language shares different characteristics with the target language. Even though the predictions on the target language are not much better on average in this case, as long as a large enough subset of the confident predictions are better than with single-source transfer, these predictions can be exploited during self-training.

In addition to using self-training with native word cluster features, we also experimented with creating target language specific versions of the cross-lingual features by means of the feature duplication trick (Daumé, 2007). However, preliminary experiments suggested that this is not an effective strategy in the

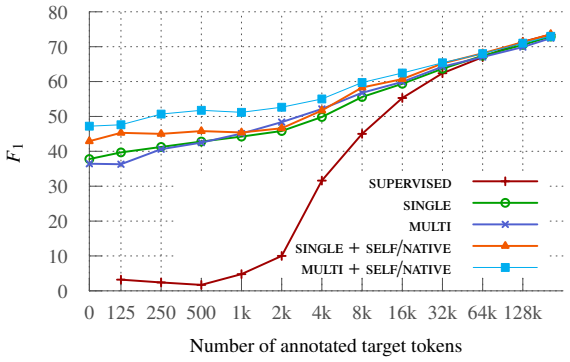


Figure 1: Learning curves for German.

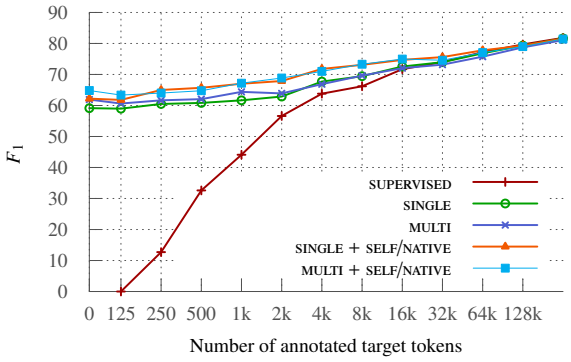


Figure 2: Learning curves for Spanish.

cross-lingual direct transfer scenario. It thus seems likely that the significant improvements that we observe are at least in part explained by the fact that the native features are distinct from the cross-lingual features and not mere duplicates.

## 5 Direct Transfer vs. Supervised Learning

Finally, we look at the relative performance of the different direct transfer methods and a target language specific supervised system trained with native and cross-lingual word cluster features. For these experiments we use the same settings as for the experiments in §3 and §4.1.

Figures 1–3 show the learning curves for the supervised system, as more and more target language annotations, selected by picking sentences at random from the full training set, are added to the training set, compared to the same system when combined with different direct transfer methods. From these curves, we can see that the purely supervised model

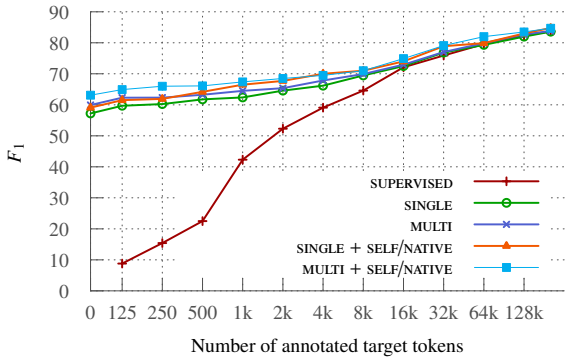


Figure 3: Learning curves for Dutch.

requires between 8,000 and 16,000 annotated word tokens (roughly corresponding to between 430 and 860 sentences) in each target language to match the best direct transfer system. The learning curves also show that adding source language data improves performance with as many as 64,000 annotated target language tokens.

Although we believe that the results on combining source and target data are interesting, in practice the marginal cost of annotation is typically quite low compared to the initial cost. Therefore, the cost of going from 125 to 64,000 annotated tokens is likely not too high, so that the benefit of cross-lingual transfer is small on the margin in this scenario. However, we believe that direct transfer methods can reduce the initial cost as well, especially when a larger label set is used, since a larger label set implies a larger cognitive load throughout annotation, but especially in the initial phase of the annotation.

Another aspect, which we were unable to investigate is the relative performance of these methods on domains other than news text. It is well known that the performance of supervised NER systems drop significantly when applied to data outside of the training domain (Nothman et al., 2008). Although the direct transfer systems in these experiments are also trained on news data, we suspect that the advantage of these methods will be more pronounced when applied to other domains, since the supervised target system runs a higher risk of overfitting to the characteristics of the target language training domain compared to the direct transfer system, which has already to some degree overfitted to the source language.

## 6 Discussion

We have focused on direct transfer methods that exploit cross-lingual word clusters, which are induced with the help of word alignments. A more common use of word alignments for cross-lingual linguistic structure prediction is for projecting annotations across languages (Yarowsky et al., 2001; Diab and Resnik, 2002; Hwa et al., 2005).

Apart from the algorithmic differences between these approaches, there are more fundamental differences in terms of the assumptions they make. Annotation projection relies on the construction of a mapping from structures in the source language to structures in the target language, $\mathcal{Y}_s \mapsto \mathcal{Y}'_t$. Based on the *direct correspondence assumption* (Diab and Resnik, 2002; Hwa et al., 2005), word alignments are assumed to be a good basis for this mapping. When projecting annotations, no consideration is taken to the source language input space, $\mathcal{X}_s$, nor to the target language input space, $\mathcal{X}_t$, except implicitly in the construction of the word alignments. The learning algorithm is thus free to use any parameters when training on instances from $\mathcal{X}_t \times \mathcal{Y}'_t$, but can at the same time not exploit any additional information that may be present in $\mathcal{X}_s \times \mathcal{Y}_s$ about $\mathcal{X}_t \times \mathcal{Y}_t$. Furthermore, word alignments are noisy and often only provide partial information about the target side annotations.

Direct transfer, on the other hand, makes a stronger assumption, as it relies on a mapping from the joint space of source inputs and output structures to the target language, $\mathcal{X}_s \times \mathcal{Y}_s \mapsto \mathcal{X}'_t \times \mathcal{Y}'_t$. Actually, the assumption is even stronger, since in order to achieve low error on the target language with a discriminative model, we must further assume that the conditional distribution $P(\mathcal{Y}'_t|\mathcal{X}'_t)$ does not diverge too much from $P(\mathcal{Y}_t|\mathcal{X}_t)$ in regions where $P(\mathcal{X}_t)$ is large. This suggests that direct transfer might be preferable when source and target languages are sufficiently similar so that a good mapping can be found.

These differences suggest that it may be fruitful to combine direct transfer with annotation projection. For example, direct transfer could be used to first map $\mathcal{X}_s \times \mathcal{Y}_s \mapsto \mathcal{X}'_t \times \mathcal{Y}'_t$, while annotation projection could be used to derive constraints on the target output space by means of a mapping $\mathcal{Y}_s \mapsto \mathcal{Y}''_t$. These constraints could perhaps be exploited in self-training, e.g., through posterior regularization (Ganchev et al., 2010), or be used for co-training (Blum and Mitchell, 1998).

## 7 Conclusions

We investigated several open questions regarding the use of cross-lingual word clusters for direct transfer named entity recognition. First, we looked at the scenario where no annotated resources are available in the target language. We showed that multi-source direct transfer and self-training with additional features, exclusive to the target language, both bring benefits in this setting, but that combining these methods provide an even larger advantage. We then examined the rate with which a supervised system, trained with cross-lingual and native word cluster features, approaches the performance of the direct transfer system. We found that on average between 8,000 and 16,000 word tokens need to be annotated in each target language to match our best direct transfer system. We also found that combining native and cross-lingual word clusters leads to improved results across the board. Finally, we showed that direct transfer methods can aid even in the supervised target language scenario. By simply mixing annotated source language data with target language data, we can significantly reduce the annotation burden required to reach a given level of performance in the target language, even with up to 64,000 tokens annotated in the target language. We hypothesize that more elaborate domain adaptation techniques, such as that proposed by Chen et al. (2011), can lead to further improvements in these scenarios.

Our use of cross-lingual word clusters is orthogonal to several other approaches discussed in this paper. We therefore suggest that such clusters could be of general use in multilingual learning of linguistic structure, in the same way that monolingual word clusters have been shown to be a robust way to bring improvements in many monolingual applications (Turian et al., 2010; Täckström et al., 2012).

## Acknowledgments

This work benefited from discussions with Ryan McDonald and from comments by Joakim Nivre and three anonymous reviewers. The author is grateful for the financial support of the Swedish National Graduate School of Language Technology (GSLT).

## References


Rie Kubota Ando and Tong Zhang. 2005. A high-performance semi-supervised learning method for text chunking. In *Proceedings of ACL*.

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the EAMT workshop on Improving MT through other Language Technology Tools: Resources and Tools for Building MT*.

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

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of COLT*, COLT' 98, New York, NY, USA. ACM.

Minmin Chen, John Blitzer, and Kilian Q. Weinberger. 2011. Co-training for domain adaptation. In *Proceedings of NIPS*.

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

Silviu Cucerzan and David Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence. In *Proceedings of EMNLP-Very Large Corpora*.

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

Hal Daumé, III, Abhishek Kumar, and Avishek Saha. 2010. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*.

Hal Daumé, III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*.

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

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of ACL*.

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

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*.

Keith Hall, Ryan McDonald, Jason Katz-Brown, and Michael Ringgaard. 2011. Training dependency parsers by jointly optimizing multiple objectives. In *Proceedings of EMNLP*.

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.

Jochen L. Leidner, Gail Sinclair, and Bonnie Webber. 2003. Grounding spatial named entities for information extraction and question answering. In *Proceedings of HLT-NAACL-GEOREF*.

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

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of NAACL-HLT*.

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

Roberta Merchant, Mary Ellen Okurowski, and Nancy Chinchor. 1996. The multilingual entity task (met) overview. In *Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996*.

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

Joel Nothman, James R Curran, and Tara Murphy. 2008. Transforming wikipedia into named entity training data. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, pages 124–132, Hobart, Australia, December.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of ACL-HLT*.

Lawrence Saul and Fernando Pereira. 1997. Aggregate and mixed-order markov models for statistical language processing. In *Proceedings of EMNLP*, pages 81–89.

Satoshi Sekine and Chikashi Nobata. 2004. Definition, dictionaries and tagger for extended named entity hierarchy. In *Proceedings of LREC*.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2009. Adding more languages improves unsupervised multilingual part-of-speech tagging: A bayesian non-parametric approach. In *Proceedings of NAACL*.

Jun Suzuki and Hideki Isozaki. 2008. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *Proceedings of ACL-HLT*.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of NAACL-HLT*.

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*.

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*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT*.