Weakly Supervised Cross-Lingual Named Entity Recognition via Effective Annotation and Representation Projection

Jian Ni and **Georgiana Dinu** and **Radu Florian**
IBM T. J. Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY 10598, USA
{nij, gdinu, raduf}@us.ibm.com

Abstract

The state-of-the-art named entity recognition (NER) systems are supervised machine learning models that require large amounts of manually annotated data to achieve high accuracy. However, annotating NER data by human is expensive and time-consuming, and can be quite difficult for a new language. In this paper, we present two weakly supervised approaches for cross-lingual NER with no human annotation in a target language. The first approach is to create automatically labeled NER data for a target language via annotation projection on comparable corpora, where we develop a heuristic scheme that effectively selects good-quality projection-labeled data from noisy data. The second approach is to project distributed representations of words (word embeddings) from a target language to a source language, so that the source-language NER system can be applied to the target language without re-training. We also design two co-decoding schemes that effectively combine the outputs of the two projection-based approaches. We evaluate the performance of the proposed approaches on both in-house and open NER data for several target languages. The results show that the combined systems outperform three other weakly supervised approaches on the CoNLL data.

1 Introduction

Named entity recognition (NER) is a fundamental information extraction task that automatically detects named entities in text and classifies them into pre-defined entity types such as PERSON, ORGANIZATION, GPE (GeoPolitical Entities), EVENT, LOCATION, TIME, DATE, etc. NER provides essential inputs for many information extraction applications, including relation extraction, entity linking, question answering and text mining. Building fast and accurate NER systems is a crucial step towards enabling large-scale automated information extraction and knowledge discovery on the huge volumes of electronic documents existing today.

The state-of-the-art NER systems are supervised machine learning models (Nadeau and Sekine, 2007), including maximum entropy Markov models (MEMMs) (McCallum et al., 2000), conditional random fields (CRFs) (Lafferty et al., 2001) and neural networks (Collobert et al., 2011; Lample et al., 2016). To achieve high accuracy, a NER system needs to be trained with a large amount of manually annotated data, and is often supplied with language-specific resources (e.g., gazetteers, word clusters, etc.). Annotating NER data by human is rather expensive and time-consuming, and can be quite difficult for a new language. This creates a big challenge in building NER systems of multiple languages for supporting multilingual information extraction applications.

The difficulty of acquiring supervised annotation raises the following question: given a well-trained NER system in a source language (e.g., English), how can one go about extending it to a new language with decent performance and no human annotation in the target language? There are mainly two types of approaches for building weakly supervised cross-lingual NER systems.

The first type of approaches create weakly labeled NER training data in a target language. One way to create weakly labeled data is through annotation projection on aligned parallel corpora or translations between a source language and a target language, e.g., (Yarowsky et al., 2001; Zitouni and Florian, 2008; Ehrmann et al., 2011). Another way is to utilize the text and structure of

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1470–1480
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1135

Wikipedia to generate weakly labeled multilingual training annotations, e.g., (Richman and Schone, 2008; Nothman et al., 2013; Al-Rfou et al., 2015).

The second type of approaches are based on direct model transfer, e.g., (Täckström et al., 2012; Tsai et al., 2016). The basic idea is to train a single NER system in the source language with language independent features, so the system can be applied to other languages using those universal features.

In this paper, we make the following contributions to weakly supervised cross-lingual NER with no human annotation in the target languages. First, for the *annotation projection* approach, we develop a heuristic, language-independent data selection scheme that seeks to select good-quality projection-labeled NER data from comparable corpora. Experimental results show that the data selection scheme can significantly improve the accuracy of the target-language NER system when the alignment quality is low and the projection-labeled data are noisy.

Second, we propose a new approach for direct NER model transfer based on *representation projection*. It projects word representations in vector space (word embeddings) from a target language to a source language, to create a universal representation of the words in different languages. Under this approach, the NER system trained for the source language can be directly applied to the target language without the need for re-training.

Finally, we design two *co-decoding* schemes that combine the outputs (views) of the two projection-based systems to produce an output that is more accurate than the outputs of individual systems. We evaluate the performance of the proposed approaches on both in-house and open NER data sets for a number of target languages. The results show that the combined systems outperform the state-of-the-art cross-lingual NER approaches proposed in Täckström et al. (2012), Nothman et al. (2013) and Tsai et al. (2016) on the CoNLL NER test data (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003).

We organize the paper as follows. In Section 2 we introduce three NER models that are used in the paper. In Section 3 we present an annotation projection approach with effective data selection. In Section 4 we propose a representation projection approach for direct NER model transfer. In Section 5 we describe two co-decoding schemes that effectively combine the outputs of two projection-based approaches. In Section 6 we evaluate the performance of the proposed approaches. We describe related work in Section 7 and conclude the paper in Section 8.

2 NER Models

The NER task can be formulated as a sequence labeling problem: given a sequence of words $x_1, ..., x_n$, we want to infer the NER tag l_i for each word x_i, $1 \leq i \leq n$. In this section we introduce three NER models that are used in the paper.

2.1 CRFs and MEMMs

Conditional random fields (CRFs) are a class of discriminative probabilistic graphical models that provide powerful tools for labeling sequential data (Lafferty et al., 2001). CRFs learn a conditional probability model $p_\lambda(\mathbf{l}|\mathbf{x})$ from a set of labeled training data, where $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ is a random sequence of input words, $\mathbf{l} = (\mathbf{l}_1, ..., \mathbf{l}_n)$ is the sequence of label variables (NER tags) for $\mathbf{x}$, and $\mathbf{l}$ has certain Markov properties conditioned on $\mathbf{x}$. Specifically, a general-order CRF with order o assumes that label variable $\mathbf{l}_i$ is dependent on a fixed number o of previous label variables $\mathbf{l}_{i-1}, ..., \mathbf{l}_{i-o}$, with the following conditional distribution:

$$p_\lambda(\mathbf{l}|\mathbf{x}) = \frac{e^{\sum_{i=1}^{n}\sum_{k=1}^{K}\lambda_k f_k(\mathbf{l}_i, \mathbf{l}_{i-1}, ..., \mathbf{l}_{i-o}, \mathbf{x})}}{Z_\lambda(\mathbf{x})} \quad (1)$$

where f_k's are feature functions, λ_k's are weights of the feature functions (parameters to learn), and $Z_\lambda(\mathbf{x})$ is a normalization constant. When $o = 1$, we have a first-order CRF which is also known as a linear-chain CRF.

Given a set of labeled training data $\mathcal{D} = (\mathbf{x}^{(j)}, \mathbf{l}^{(j)})_{j=1,...,N}$, we seek to find an optimal set of parameters λ^* that maximize the conditional log-likelihood of the data:

$$\lambda^* = \arg\max_\lambda \sum_{j=1}^{N} \log p_\lambda(\mathbf{l}^{(j)}|\mathbf{x}^{(j)}) \quad (2)$$

Once we obtain λ^*, we can use the trained model $p_{\lambda^*}(\mathbf{l}|\mathbf{x})$ to decode the most likely label sequence $\mathbf{l}^*$ for any new input sequence of words $\mathbf{x}$ (via the Viterbi algorithm for example):

$$\mathbf{l}^* = \arg\max_{\mathbf{l}} p_{\lambda^*}(\mathbf{l}|\mathbf{x}) \quad (3)$$

A related conditional probability model, called *maximum entropy Markov model* (MEMM) (McCallum et al., 2000), assumes that $\mathbf{l}$ is a Markov

chain conditioned on $\mathbf{x}$:

$$p_\lambda(\mathbf{l}|\mathbf{x}) = \prod_{i=1}^{n} p_\lambda(\mathbf{l}_i|\mathbf{l}_{i-1},...,\mathbf{l}_{i-o},\mathbf{x}) = \prod_{i=1}^{n} \frac{e^{\sum_{k=1}^{K}\lambda_k f_k(\mathbf{l}_i,\mathbf{l}_{i-1},...,\mathbf{l}_{i-o},\mathbf{x})}}{Z_\lambda(\mathbf{l}_{i-1},...,\mathbf{l}_{i-o},\mathbf{x})} \quad (4)$$

The main difference between CRFs and MEMMs is that CRFs normalize the conditional distribution over the whole sequence as in (1), while MEMMs normalize the conditional distribution per token as in (4). As a result, CRFs can better handle the label bias problem (Lafferty et al., 2001). This benefit, however, comes at a price. The training time of order-o CRFs grows exponentially ($O(M^{o+1})$) with the number of output labels M, which is typically slow even for moderate-size training data if M is large. In contrast, the training time of order-o MEMMs is linear ($O(M)$) with respect to M independent of o, so it can handle larger training data with higher order of dependency. We have implemented both a linear-chain CRF model and a general-order MEMM model.

2.2 Neural Networks

With the increasing popularity of distributed (vector) representations of words, neural network models have recently been applied to tackle many NLP tasks including NER (Collobert et al., 2011; Lample et al., 2016).

We have implemented a feedforward neural network model which maximizes the log-likelihood of the training data similar to that of (Collobert et al., 2011). We adopt a locally normalized model (the conditional distribution is normalized per token as in MEMMs) and introduce context dependency by conditioning on the previously assigned tags. We use a target word and its surrounding context as features. We do not use other common features such as gazetteers or character-level representations as such features might not be readily available or might not transfer to other languages.

We have deployed two neural network architectures. The first one (called NN1) uses the word embedding of a word as the input. The second one (called NN2) adds a smoothing prototype layer that computes the cosine similarity between a word embedding and a fixed set of prototype vectors (learned during training) and returns a weighted average of these prototype vectors as the input. In our experiments we find that with the smoothing layer, NN2 tends to have a more balanced precision and recall than NN1. Both networks have one hidden layer, with sigmoid and softmax activation functions on the hidden and output layers respectively. The two neural network models are depicted in Figure 1.

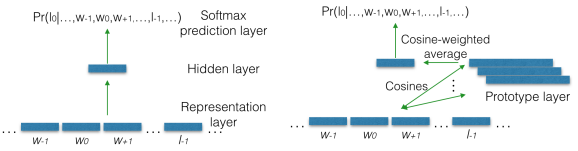

Figure 1: Architecture of the two neural network models: left-NN1, right-NN2.

3 Annotation Projection Approach

The existing annotation projection approaches require parallel corpora or translations between a source language and a target language with alignment information. In this paper, we develop a heuristic, language-independent data selection scheme that seeks to select good-quality projection-labeled data from noisy comparable corpora. We use English as the source language.

Suppose we have comparable[1] sentence pairs $(\mathbf{X},\mathbf{Y})$ between English and a target language, where $\mathbf{X}$ includes N English sentences $\mathbf{x}^{(1)},...,\mathbf{x}^{(N)}$, $\mathbf{Y}$ includes N target-language sentences $\mathbf{y}^{(1)},...,\mathbf{y}^{(N)}$, and $\mathbf{y}^{(j)}$ is aligned to $\mathbf{x}^{(j)}$ via an alignment model, $1 \le j \le N$. We use a sentence pair $(\mathbf{x},\mathbf{y})$ as an example to illustrate how the annotation projection procedure works, where $\mathbf{x} = (x_1, x_2, ..., x_s)$ is an English sentence, and $\mathbf{y} = (y_1, y_2, ..., y_t)$ is a target-language sentence that is aligned to $\mathbf{x}$.

Annotation Projection Procedure

1. Apply the English NER system on the English sentence $\mathbf{x}$ to generate the NER tags $\mathbf{l} = (l_1, l_2, ..., l_s)$ for $\mathbf{x}$.

2. Project the NER tags to the target-language sentence $\mathbf{y}$ using the alignment information. Specifically, if a sequence of English words $(x_i, ..., x_{i+p})$ is aligned to a sequence of target-language words $(y_j, ..., y_{j+q})$, and $(x_i, ..., x_{i+p})$ is recognized (by the English NER system) as an entity with NER tag l,

[1] Ideally, the sentences would be translations of each other, but we only require possibly parallel sentences.

then $(y_j, ..., y_{j+q})$ is labeled with l[2].
Let $\mathbf{l}' = (l'_1, l'_2, ..., l'_t)$ be the projected NER tags for the target-language sentence $\mathbf{y}$.

We can apply the annotation projection procedure on all the sentence pairs $(\mathbf{X}, \mathbf{Y})$, to generate projected NER tags $\mathbf{L}'$ for the target-language sentences $\mathbf{Y}$. ($\mathbf{Y}$, $\mathbf{L}'$) are automatically labeled NER data with no human annotation in the target language. One can use those projection-labeled data to train an NER system in the target language. The quality of such weakly labeled NER data, and consequently the accuracy of the target-language NER system, depend on both 1) the accuracy of the English NER system, and 2) the alignment accuracy of the sentence pairs.

Since we don't require actual translations, but only comparable data, the downside is that if some of the data are not actually parallel and if we use all for weakly supervised learning, the accuracy of the target-language NER system might be adversely affected. We are therefore motivated to design effective data selection schemes that can select good-quality projection-labeled data from noisy data, to improve the accuracy of the annotation projection approach for cross-lingual NER.

3.1 Data Selection Scheme

We first design a metric to measure the annotation quality of a projection-labeled sentence in the target language. We construct a frequency table T which includes all the entities in the projection-labeled target-language sentences. For each entity e, T also includes the projected NER tags for e and the relative frequency (empirical probability) $\hat{P}(l|e)$ that entity e is labeled with tag l. Table 1 shows a snapshot of the frequency table where the target language is Portuguese.

We use $\hat{P}(l|e)$ to measure the reliability of labeling entity e with tag l in the target language. The intuition is that if an entity e is labeled by a tag l with higher frequency than other tags in the projection-labeled data, it is more likely that the annotation is correct. For example, if the joint accuracy of the source NER system and alignment system is greater than 0.5, then the correct tag of a random entity will have a higher relative frequency than other tags in a large enough sample.

Based on the frequency scores, we calculate the quality score of a projection-labeled target-language sentence $\mathbf{y}$ by averaging the frequency scores of the projected entities in the sentence:

$$q(\mathbf{y}) = \frac{\Sigma_{e \in \mathbf{y}} \hat{P}(l'(e)|e)}{n(\mathbf{y})} \quad (5)$$

where $l'(e)$ is the projected NER tag for e, and $n(\mathbf{y})$ is the total number of entities in sentence $\mathbf{y}$.

We use $q(\mathbf{y})$ to measure the annotation quality of sentence $\mathbf{y}$, and $n(\mathbf{y})$ to measure the amount of annotation information contained in sentence $\mathbf{y}$. We design a *heuristic data selection scheme* which selects projection-labeled sentences in the target language that satisfy the following condition:

$$q(\mathbf{y}) \geq q;\ n(\mathbf{y}) \geq n \quad (6)$$

where q is a quality score threshold and n is an entity number threshold. We can tune the two parameters to make tradeoffs among the annotation quality of the selected sentences, the annotation information contained in the selected sentences, and the total number of sentence selected.

One way to select the threshold parameters q and n is via a development set - either a small set of human-annotated data or a sample of the projection-labeled data. We select the threshold parameters via *coordinate search* using the development set: we first fix $n = 3$ and search the best $\hat{q}$ in $[0, 0.9]$ with a step size of 0.1; we then fix $q = \hat{q}$ and select the best $\hat{n}$ in $[1, 5]$ with a step size of 1.

Entity Name	NER Tag	Frequency
Estados Unidos	GPE	0.853
Estados Unidos	ORGANIZATION	0.143
Estados Unidos	PEOPLE	0.001
Estados Unidos	PRODUCT	0.001
Estados Unidos	TITLEWORK	0.001
Estados Unidos	EVENT	0.001

Table 1: A snapshot of the frequency table where the target language is Portuguese. *Estados Unidos* means *United States*. The correct NER tag for Estados Unidos is GPE which has the highest relative frequency in the weakly labeled data.

3.2 Accuracy Improvements

We evaluate the effectiveness of the data selection scheme via experiments on 4 target languages: Japanese, Korean, German and Portuguese. We use comparable corpora between English and each target language (ranging from 2M to 6M tokens) with alignment information. For each target language, we also have a set of manually annotated NER data (ranging from 30K to 45K tokens)

[2]If the IOB (Inside, Outside, Beginning) tagging format is used, then $(y_j, y_{j+1}, ..., y_{j+q})$ is labeled with (B-l, I-l,...,I-l).

Language	(q, n)	Training Size	F_1 Score
Japanese	(0, 0)	4.9M	41.2
	(0.7, 4)	1.3M	53.4
Korean	(0, 0)	4.5M	25.0
	(0.4, 2)	1.5M	38.7
German	(0, 0)	5.2M	67.2
	(0.4, 4)	2.6M	67.5
Portuguese	(0, 0)	2.1M	61.5
	(0.1, 4)	1.5M	62.7

Table 2: Performance comparison of weakly supervised NER systems trained without data selection ($(q, n) = (0, 0)$) and with data selection ($(\hat{q}, \hat{n})$ determined by coordinate search).

which are served as the test data for evaluating the target-language NER system.

The source (English) NER system is a linear-chain CRF model which achieves an accuracy of 88.9 F_1 score on an independent NER test set. The alignment systems between English and the target languages are maximum entropy models (Ittycheriah and Roukos, 2005), with an accuracy of 69.4/62.0/76.1/88.0 F_1 score on independent Japanese/Korean/German/Portuguese alignment test sets.

For each target language, we randomly select 5% of the projection-labeled data as the development set and the remaining 95% as the training set. We compare an NER system trained with all the projection-labeled training data with no data selection (i.e., $(q, n) = (0, 0)$) and an NER system trained with projection-labeled data selected by the data selection scheme where the development set is used to select the threshold parameters q and n via coordinate search. Both NER systems are 2nd-order MEMM models[3] which use the same template of features.

The results are shown in Table 2. For different target languages, we use the same source (English) NER system for annotation projection, so the differences in the accuracy improvements are mainly due to the alignment quality of the comparable corpora between English and different target languages. When the alignment quality is low (e.g., as for Japanese and Korean) and hence the projection-labeled NER data are quite noisy, the proposed data selection scheme is very effective in selecting good-quality projection-labeled data and the improvement is big: +12.2 F_1 score for Japanese and +13.7 F_1 score for Korean. Using a stratified shuffling test (Noreen, 1989), for a significance level of 0.05, data-selection is statistically significantly better than no-selection for Japanese, Korean and Portuguese.

[3]In our experiments, CRFs cannot handle training data with a few million words, since our NER system has over 50 entity types, and the training time of CRFs grows at least quadratically in the number of entity types.

4 Representation Projection Approach

In this paper, we propose a new approach for direct NER model transfer based on representation projection. Under this approach, we train a single English NER system that uses only word embeddings as input representations. We create mapping functions which can map words in any language into English and we simply use the English NER system to decode. In particular, by mapping all languages into English, we are using one universal NER system and we do not need to re-train the system when a new language is added.

4.1 Monolingual Word Embeddings

We first build vector representations of words (word embeddings) for a language using monolingual data. We use a variant of the Continuous Bag-of-Words (CBOW) word2vec model (Mikolov et al., 2013a), which concatenates the context words surrounding a target word instead of adding them (similarly to (Ling et al., 2015)). Additionally, we employ weights $w = \frac{1}{dist(x,x_c)}$ that decay with the distance of a context word x_c to a target word x. Tests on word similarity benchmarks show this variant leads to small improvements over the standard CBOW model.

We train 300-dimensional word embeddings for English. Following (Mikolov et al., 2013b), we use larger dimensional embeddings for the target languages, namely 800. We train word2vec for 1 epoch for English/Spanish and 5 epochs for the rest of the languages for which we have less data.

4.2 Cross-Lingual Representation Projection

We learn cross-lingual word embedding mappings, similarly to (Mikolov et al., 2013b). For a target language f, we first extract a small training dictionary from a phrase table that includes word-to-word alignments between English and the target language f. The dictionary contains English and target-language word pairs with weights: $(x_i, y_i, w_i)_{i=1,...,n}$, where x_i is an English word, y_i is a target-language word, and the weight $w_i = \hat{P}(x_i|y_i)$ is the relative frequency of x_i given y_i as extracted from the phrase table.

Suppose we have monolingual word embeddings for English and the target language f. Let $\mathbf{u_i} \in \mathcal{R}^{d_1}$ be the vector representation for English word x_i, $\mathbf{v_i} \in \mathcal{R}^{d_2}$ be the vector representation for target-language word y_i. We find a linear mapping $\mathbf{M}_{f \to e}$ by solving the following weighted least squares problem where the dictionary is used as the training data:

$$\mathbf{M}_{f \to e} = \arg\min_{\mathbf{M}} \sum_{i=1}^{n} w_i ||\mathbf{u_i} - \mathbf{M}\mathbf{v_i}||^2 \quad (7)$$

In (7) we generalize the formulation in (Mikolov et al., 2013b) by adding frequency weights to the word pairs, so that more frequent pairs are of higher importance. Using $\mathbf{M}_{f \to e}$, for any new word in f with vector representation $\mathbf{v}$, we can project it into the English vector space as the vector $\mathbf{M}_{f \to e}\mathbf{v}$.

The training dictionary plays a key role in finding an effective cross-lingual embedding mapping. To control the size of the dictionary, we only include word pairs with a minimum frequency threshold. We set the threshold to obtain approximately 5K to 6K unique word pairs for a target language, as our experiments show that larger-size dictionaries might harm the performance of representation projection for direct NER model transfer.

4.3 Direct NER Model Transfer

The source (English) NER system is a neural network model (with architecture NN1 or NN2) that uses only word embedding features (embeddings of a word and its surrounding context) in the English vector space. Model transfer is achieved simply by projecting the target language word embeddings into the English vector space and decoding these using the English NER system.

More specifically, given the word embeddings of a sequence of words in a target language f, $(\mathbf{v}_1, ..., \mathbf{v}_t)$, we project them into the English vector space by applying the linear mapping $\mathbf{M}_{f \to e}$: $(\mathbf{M}_{f \to e}\mathbf{v}_1, ..., \mathbf{M}_{f \to e}\mathbf{v}_t)$. The English NER system is then applied on the projected input to produce NER tags. Words not in the target-language vocabulary are projected into their English embeddings if they are found in the English vocabulary, or into an NER-trained UNK vector otherwise.

5 Co-Decoding

Given two weakly supervised NER systems which are trained with different data using different models (MEMM model for annotation projection and neural network model for representation projection), we would like to design a *co-decoding* scheme that can combine the outputs (views) of the two systems to produce an output that is more accurate than the outputs of individual systems.

Since both systems are statistical models and can produce confidence scores (probabilities), a natural co-decoding scheme is to compare the confidence scores of the NER tags generated by the two systems and select the tags with higher confidences scores. However, confidence scores of two weakly supervised systems may not be directly comparable, especially when comparing O tags with non-O tags (i.e., entity tags). We consider an *exclude-O confidence-based co-decoding scheme* which we find to be more effective empirically. It is similar to the pure confidence-based scheme, with the only difference that it always prefers a non-O tag of one system to an O tag of the other system, regardless of their confidence scores.

In our experiments we find that the annotation projection system tends to have a high precision and low recall, i.e., it detects fewer entities, but for the detected entities the accuracy is high. The representation projection system tends to have a more balanced precision and recall. Based on this observation, we develop the following *rank-based co-decoding scheme* that gives higher priority to the high-precision annotation projection system:

1. The combined output includes all the entities detected by the annotation projection system.

2. It then adds all the entities detected by the representation projection system that do not conflict[4] with entities detected by the annotation projection system (to improve recall).

Note that an entity X detected by the representation projection system does not conflict with the annotation projection system if the annotation projection system produces O tags for the entire span of X. For example, suppose the output tag sequence of annotation projection is (B-PER,O,O,O,O), of representation projection is (B-ORG,I-ORG,O,B-LOC,I-LOC), then the combined output under the rank-based scheme will be (B-PER,O,O,B-LOC,I-LOC).

[4]Two entities detected by two different systems conflict with each other if either 1) the two entities have different spans but overlap with each other; or 2) the two entities have the same span but with different NER tags.

Japanese	P	R	F_1
Annotation-Projection (AP)	69.9	43.2	53.4
Representation-Projection (NN1)	71.5	36.6	48.4
Representation-Projection (NN2)	59.9	42.4	49.7
Co-Decoding (Conf): AP+NN1	65.7	49.5	56.5
Co-Decoding (Rank): AP+NN1	68.3	51.6	**58.8**
Co-Decoding (Conf): AP+NN2	59.5	53.3	56.2
Co-Decoding (Rank): AP+NN2	61.6	54.5	57.8
Supervised (272K)	*84.5*	*80.9*	*82.7*
Korean	**P**	**R**	**F_1**
Annotation-Projection (AP)	69.5	26.8	38.7
Representation-Projection (NN1)	66.1	23.2	34.4
Representation-Projection (NN2)	68.5	43.4	53.1
Co-Decoding (Conf): AP+NN1	68.2	41.0	51.2
Co-Decoding (Rank): AP+NN1	71.3	42.8	53.5
Co-Decoding (Conf): AP+NN2	68.9	53.4	60.2
Co-Decoding (Rank): AP+NN2	70.0	53.3	**60.5**
Supervised (97K)	*88.2*	*74.0*	*80.4*
German	**P**	**R**	**F_1**
Annotation-Projection (AP)	76.5	60.5	67.5
Representation-Projection (NN1)	69.0	48.8	57.2
Representation-Projection (NN2)	63.7	66.1	64.9
Co-Decoding (Conf): AP+NN1	68.5	61.7	64.9
Co-Decoding (Rank): AP+NN1	72.7	65.0	68.6
Co-Decoding (Conf): AP+NN2	64.7	71.3	67.9
Co-Decoding (Rank): AP+NN2	67.1	72.6	**69.7**
Supervised (125K)	*77.8*	*68.1*	*72.6*
Portuguese	**P**	**R**	**F_1**
Annotation-Projection (AP)	84.0	50.1	62.7
Representation-Projection (NN1)	70.5	47.6	56.8
Representation-Projection (NN2)	66.0	63.4	64.7
Co-Decoding (Conf): AP+NN1	72.0	55.8	62.9
Co-Decoding (Rank): AP+NN1	77.5	59.7	67.4
Co-Decoding (Conf): AP+NN2	68.1	67.1	67.6
Co-Decoding (Rank): AP+NN2	70.9	68.3	**69.6**
Supervised (173K)	*79.8*	*71.9*	*75.6*

Table 3: In-house NER data: Precision, Recall and F_1 score on exact phrasal matches. The highest F_1 score among all the weakly supervised approaches is shown in bold. Same for Tables 4 and 5.

6 Experiments

In this section, we evaluate the performance of the proposed approaches for cross-lingual NER, including the 2 projection-based approaches and the 2 co-decoding schemes for combining them:
(1) The annotation projection (AP) approach with heuristic data selection;
(2) The representation projection approach (with two neural network architectures NN1 and NN2);
(3) The exclude-O confidence-based co-decoding scheme;
(4) The rank-based co-decoding scheme.

6.1 NER Data Sets

We have used various NER data sets for evaluation. The first group includes in-house human-annotated newswire NER data for four languages: Japanese, Korean, German and Portuguese, annotated with over 50 entity types. The main motivation of deploying such a fine-grained entity type set is to build cognitive question answering applications on top of the NER systems. The entity type set has been engineered to cover many of the frequent entity types that are targeted by naturally-phrased questions. The sizes of the test data sets are ranging from 30K to 45K tokens.

The second group includes open human-annotated newswire NER data for Spanish, Dutch and German from the CoNLL NER data sets (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). The CoNLL data have 4 entity types: PER (persons), ORG (organizations), LOC (locations) and MISC (miscellaneous entities). The sizes of the development/test data sets are ranging from 35K to 70K tokens. The development data are used for tuning the parameters of learning methods.

6.2 Evaluation for In-House NER Data

In Table 3, we show the results of different approaches for the in-house NER data. For annotation projection, the source (English) NER system is a linear-chain CRF model trained with 328K tokens of human-annotated English newswire data. The target-language NER systems are 2nd-order MEMM models trained with 1.3M, 1.5M, 2.6M and 1.5M tokens of projection-labeled data for Japanese, Korean, German and Portuguese, respectively. The projection-labeled data are selected using the heuristic data selection scheme (see Table 2). For representation projection, the source (English) NER systems are neural network models with architectures NN1 and NN2 (see Figure 1), both trained with 328K tokens of human-annotated English newswire data.

The results show that the annotation projection (AP) approach has a relatively high precision and low recall. For representation projection, neural network model NN2 (with a smoothing layer) is better than NN1, and NN2 tends to have a more balanced precision and recall. The rank-based co-decoding scheme is more effective for combining the two projection-based approaches. In particular, the rank-based scheme that combines AP and NN2 achieves the highest F_1 score among all the weakly supervised approaches for Korean, German and Portuguese (second highest F_1 score for Japanese), and it improves over the best of the two

Spanish	P	R	F_1
Annotation-Projection (AP)	65.5	59.1	62.1
Representation-Projection (NN1)	63.9	52.2	57.4
Representation-Projection (NN2)	55.3	51.8	53.5
Co-Decoding (Conf): AP+NN1	64.3	66.8	**65.5**
Co-Decoding (Rank): AP+NN1	63.7	65.3	64.5
Co-Decoding (Conf): AP+NN2	58.0	63.9	60.8
Co-Decoding (Rank): AP+NN2	60.8	64.5	62.6
Supervised (264K)	*81.3*	*79.8*	*80.6*
Dutch	**P**	**R**	**F_1**
Annotation-Projection (AP)	73.3	63.0	67.8
Representation-Projection (NN1)	82.6	47.4	60.3
Representation-Projection (NN2)	66.3	43.5	52.5
Co-Decoding (Conf): AP+NN1	72.3	66.5	**69.3**
Co-Decoding (Rank): AP+NN1	72.8	65.3	68.8
Co-Decoding (Conf): AP+NN2	65.3	64.7	65.0
Co-Decoding (Rank): AP+NN2	69.7	66.0	67.8
Supervised (199K)	*82.9*	*81.7*	*82.3*
German	**P**	**R**	**F_1**
Annotation-Projection (AP)	71.8	54.7	62.1
Representation-Projection (NN1)	79.4	41.4	54.4
Representation-Projection (NN2)	64.6	42.7	51.4
Co-Decoding (Conf): AP+NN1	70.1	59.5	64.4
Co-Decoding (Rank): AP+NN1	71.0	59.4	**64.7**
Co-Decoding (Conf): AP+NN2	64.2	59.9	62.0
Co-Decoding (Rank): AP+NN2	66.8	60.6	63.6
Supervised (206K)	*81.2*	*64.3*	*71.8*

Table 4: CoNLL NER development data.

projection-based systems by 2.2 to 7.4 F_1 score.

We also provide the performance of *supervised learning* where the NER system is trained with human-annotated data in the target language (with size shown in the bracket). While the performance of the weakly supervised systems is not as good as supervised learning, it is important to build weakly supervised systems with decent performance when supervised annotation is unavailable. Even if supervised annotation is feasible, the weakly supervised systems can be used to pre-annotate the data, and we observed that pre-annotation can improve the annotation speed by 40%-60%, which greatly reduces the annotation cost.

6.3 Evaluation for CoNLL NER Data

For the CoNLL data, the source (English) NER system for annotation projection is a linear-chain CRF model trained with the CoNLL English training data (203K tokens), and the target-language NER systems are 2nd-order MEMM models trained with 1.3M, 7.0M and 1.2M tokens of projection-labeled data for Spanish, Dutch and German, respectively. The projection-labeled data are selected using the heuristic data selection scheme, where the threshold parameters q and n are determined via coordinate search based on the CoNLL development sets. Compared with no data selection, the data selection scheme improves the annotation projection approach by 2.7/2.0/2.7 F_1 score on the Spanish/Dutch/German development data. In addition to standard NER features such as n-gram word features, word type features, prefix and suffix features, the target-language NER systems also use the multilingual Wikipedia entity type mappings developed in (Ni and Florian, 2016) to generate dictionary features and as decoding constraints, which improve the annotation projection approach by 3.0/5.4/7.9 F_1 score on the Spanish/Dutch/German development data.

For representation projection, the source (English) NER systems are neural network models (NN1 and NN2) trained with the CoNLL English training data. Compared with the standard CBOW word2vec model, the concatenated variant improves the representation projection approach (NN1) by 8.9/11.4/6.8 F_1 score on the Spanish/Dutch/German development data, as well as by 2.0 F_1 score on English. In addition, the frequency-weighted cross-lingual word embedding projection (7) improves the representation projection approach (NN1) by 2.2/6.3/3.7 F_1 score on the Spanish/Dutch/German development data, compared with using uniform weights on the same data. We do observe, however, that using uniform weights when keeping only the most frequent translation of a word instead of all word pairs above a threshold in the training dictionary, leads to performance similar to that of the frequency-weighted projection.

In Table 4 we show the results for the CoNLL development data. For representation projection, NN1 is better than NN2. Both the annotation projection approach and NN1 tend to have a high precision. In this case, the exclude-O confidence-based co-decoding scheme that combines AP and NN1 achieves the highest F_1 score for Spanish and Dutch (second highest F_1 score for German), and improves over the best of the two projection-based systems by 1.5 to 3.4 F_1 score.

In Table 5 we compare our top systems (confidence or rank-based co-decoding of AP and NN1, determined by the development data) with the best results of the cross-lingual NER approaches proposed in Täckström et al. (2012), Nothman et al. (2013) and Tsai et al. (2016) on the CoNLL test data. Our systems outperform the previous state-of-the-art approaches, closing more of the gap to

Spanish	P	R	F_1
Täckström et al. (2012)	x	x	59.3
Nothman et al. (2013)	x	x	61.0
Tsai et al. (2016)	x	x	60.6
Co-Decoding (Conf): AP+NN1	64.9	65.2	**65.1**
Co-Decoding (Rank): AP+NN1	64.6	63.9	64.3
Supervised (264K)	*82.5*	*82.3*	*82.4*
Dutch	**P**	**R**	**F_1**
Täckström et al. (2012)	x	x	58.4
Nothman et al. (2013)	x	x	64.0
Tsai et al. (2016)	x	x	61.6
Co-Decoding (Conf): AP+NN1	69.1	62.0	**65.4**
Co-Decoding (Rank): AP+NN1	69.3	61.0	64.8
Supervised (199K)	*85.1*	*83.9*	*84.5*
German	**P**	**R**	**F_1**
Täckström et al. (2012)	x	x	40.4
Nothman et al. (2013)	x	x	55.8
Tsai et al. (2016)	x	x	48.1
Co-Decoding (Conf): AP+NN1	68.5	51.0	**58.5**
Co-Decoding (Rank): AP+NN1	68.3	50.4	58.0
Supervised (206K)	*79.6*	*65.3*	*71.8*

Table 5: CoNLL NER test data.

supervised learning.

7 Related Work

The traditional annotation projection approaches (Yarowsky et al., 2001; Zitouni and Florian, 2008; Ehrmann et al., 2011) project NER tags across language pairs using parallel corpora or translations. Wang and Manning (2014) proposed a variant of annotation projection which projects expectations of tags and uses them as constraints to train a model based on generalized expectation criteria. Annotation projection has also been applied to several other cross-lingual NLP tasks, including word sense disambiguation (Diab and Resnik, 2002), part-of-speech (POS) tagging (Yarowsky et al., 2001) and dependency parsing (Rasooli and Collins, 2015).

Wikipedia has been exploited to generate weakly labeled multilingual NER training data. The basic idea is to first categorize Wikipedia pages into entity types, either based on manually constructed rules that utilize the category information of Wikipedia (Richman and Schone, 2008) or Freebase attributes (Al-Rfou et al., 2015), or via a classifier trained with manually labeled Wikipedia pages (Nothman et al., 2013). Heuristic rules are then developed in these works to automatically label the Wikipedia text with NER tags. Ni and Florian (2016) built high-accuracy, high-coverage multilingual Wikipedia entity type mappings using weakly labeled data and applied those mappings as decoding constrains or dictionary features to improve multilingual NER systems.

For direct NER model transfer, Täckström et al. (2012) built cross-lingual word clusters using monolingual data in source/target languages and aligned parallel data between source and target languages. The cross-lingual word clusters were then used to generate universal features. Tsai et al. (2016) applied the cross-lingual wikifier developed in (Tsai and Roth, 2016) and multilingual Wikipedia dump to generate language-independent labels (FreeBase types and Wikipedia categories) for n-grams in a document, and those labels were used as universal features.

Different ways of obtaining cross-lingual embeddings have been proposed in the literature. One approach builds monolingual representations separately and then brings them to the same space typically using a seed dictionary (Mikolov et al., 2013b; Faruqui and Dyer, 2014). Another line of work builds inter-lingual representations simultaneously, often by generating mixed language corpora using the supervision at hand (aligned sentences, documents, etc.) (Vulić and Moens, 2015; Gouws et al., 2015). We opt for the first solution in this paper because of its flexibility: we can map all languages to English rather than requiring separate embeddings for each language pair. Additionally we are able to easily add a new language without any constraints on the type of data needed. Note that although we do not specifically create inter-lingual representations, by training mappings to the common language, English, we are able to map words in different languages to a common space. Similar approaches for cross-lingual model transfer have been applied to other NLP tasks such as document classification (Klementiev et al., 2012), dependency parsing (Guo et al., 2015) and POS tagging (Gouws and Søgaard, 2015).

8 Conclusion

In this paper, we developed two weakly supervised approaches for cross-lingual NER based on effective annotation and representation projection. We also designed two co-decoding schemes that combine the two projection-based systems in an intelligent way. Experimental results show that the combined systems outperform three state-of-the-art cross-lingual NER approaches, providing a strong baseline for building cross-lingual NER systems with no human annotation in target languages.

References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, Vancouver, British Columbia, Canada. https://doi.org/10.1137/1.9781611974010.66.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12:2493–2537. http://dl.acm.org/citation.cfm?id=1953048.2078186.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, ACL'02, pages 255–262. https://doi.org/10.3115/1073083.1073126.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of Recent Advances in Natural Language Processing*. Association for Computational Linguistics, pages 118–124. http://aclweb.org/anthology/R11-1017.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, pages 462–471. http://www.aclweb.org/anthology/E14-1049.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*. JMLR Workshop and Conference Proceedings, pages 748–756. http://jmlr.org/proceedings/papers/v37/gouws15.pdf.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1386–1390. http://www.aclweb.org/anthology/N15-1157.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Beijing, China, pages 1234–1244. http://www.aclweb.org/anthology/P15-1119.

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. pages 89–96. http://aclweb.org/anthology/H05-1012.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, pages 1459–1474. http://www.aclweb.org/anthology/C12-1089.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML'01, pages 282–289. http://dl.acm.org/citation.cfm?id=645530.655813.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 260–270. https://doi.org/10.18653/v1/N16-1030.

Wang Ling, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1299–1304. http://www.aclweb.org/anthology/N15-1142.

Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML'00, pages 591–598. http://dl.acm.org/citation.cfm?id=645529.658277.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR* abs/1309.4168. http://arxiv.org/abs/1309.4168.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1):3–26. Publisher: John Benjamins Publishing Company. https://doi.org/10.1075/li.30.1.03nad.

Jian Ni and Radu Florian. 2016. Improving multilingual named entity recognition with wikipedia entity type mapping. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1275–1284. https://doi.org/10.18653/v1/D16-1135.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. John Wiley & Sons, Inc., New York, NY, USA.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Journal of Artificial Intelligence* 194:151–175. https://doi.org/10.1016/j.artint.2012.03.006.

Sadegh Mohammad Rasooli and Michael Collins. 2015. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 328–338. https://doi.org/10.18653/v1/D15-1039.

E. Alexander Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, pages 1–9. http://aclweb.org/anthology/P08-1001.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 477–487. http://aclweb.org/anthology/N12-1052.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the Sixth Conference on Natural Language Learning - Volume 20*. Association for Computational Linguistics, Stroudsburg, PA, USA, CONLL'02, pages 1–4. https://doi.org/10.3115/1118853.1118877.

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 the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*. Association for Computational Linguistics, Stroudsburg, PA, USA, CONLL'03, pages 142–147. https://doi.org/10.3115/1119176.1119195.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pages 219–228. https://doi.org/10.18653/v1/K16-1022.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 589–598. https://doi.org/10.18653/v1/N16-1072.

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Beijing, China, pages 719–725. http://www.aclweb.org/anthology/P15-2118.

Mengqiu Wang and D. Christopher Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association of Computational Linguistics* 2:55–66. http://aclweb.org/anthology/Q14-1005.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT'01, pages 1–8. https://doi.org/10.3115/1072133.1072187.

Imed Zitouni and Radu Florian. 2008. Mention detection crossing the language barrier. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 600–609. http://aclweb.org/anthology/D08-1063.