State-of-the-Art Multilingual NER Using Loosely Aligned Text
Whitepaper
Stanford University, 2012

Robert Munro
Stanford University
Stanford, CA
rmunro@stanford.edu

Christopher Manning
Stanford University
Stanford, CA
manning@stanford.edu

Abstract

We present a new approach to named entity recognition that uses only loosely aligned text for input, but can consistently out-perform state-of-the-art supervised approaches by more than 10%, despite the absence of labeled training data. The system generates seed candidates through local edit distance deviation across language pairs, and then bootstraps to make broad predictions, optimizing joint contextual, word-shape, subword and alignment models. Results from English, Finnish, French, German, Lithuanian and Slovene show that the method is robust across a variety of Latin-script languages, making it especially promising for a large number of languages that do not have labeled training data for named entity recognition but do have collections of loosely aligned sentences.

1 Introduction

Named Entity Recognition (NER) is a cornerstone of many Information Extraction and Question-Answering tasks, but state-of-the-art NER systems have historically relied on substantial labeled training data and supervised machine learning (Mikheev et al., 1999; Cucerzan and Yarowsky, 1999; Collins, 2002; Carreras et al., 2003; Sang, 2002; Florian et al., 2003; Sang and De Meulder, 2003; Huang et al., 2004; Finkel et al., 2005; Lee et al., 2006; Steinberger and Pouliquen, 2007; Ratinov and Roth, 2009; Piskorski et al., 2009; Finkel and Manning, 2010; Faruqui and Padó, 2010; Liu et al., 2011; Ritter et al., 2011; Kim et al., 2012; Munro and Manning, 2012).

It is expensive and time-consuming to create labeled training data in any language. Supervised approaches are also prone to domain dependence, meaning that building a corpus of labeled NER data will not necessarily have broad coverage, even in the same language. Conversely, loosely aligned text is one of our greatest language resources: how many books and articles have been translated into another language? We present a system that achieves state-of-the-art performance when identifying named entities, using only loosely aligned text. It builds on the observation that named entities are the least likely words to change form across languages (Gallippi, 1996). The following example from the *Europarl* corpus shows the cross-linguistic similarity, with a Finnish sentence and its French translation:

> Herra **Schiedermeierilla** on puheenvuoro menettelyä koskevaa esitystä varten.
> M. **Schiedermeier** a la parole pour une motion de procédure.

Even without knowledge of Finnish or French, it should be clear to a person comparing the two sentences that *Schiedermeier(-illa)* is a name, and that *Herra* and *M.* are probably titles. For a person, this evidence will come from the similarity of the name across the two languages, the corresponding sentence positions, and from the structure of the word itself, especially capitalization.

These patterns can also be leveraged by a named entity recognition system. Our approach builds on this: we take the phrases with the least local edit distance between the two languages as high-probability named entity candidates. We then use the candidates to seed a model that optimizes the predictions from features representing the context, word-shape, subword and alignment. In the example above, the context features will include the preceding *Herra* and *M.*; the word-shapes will include the capitalization; the subword models will capture the same entity with different suffixes (*Schiedermeier* also occurs with *-ia*, *-ille* and *-n*

suffixes in Finnish), as well as the suffixes themselves; and the alignment features will capture the relative sentence position of the candidate entities.

By exploiting this simple observation, we are able to bootstrap systems using only loosely aligned text that are better than state-of-the-art supervised results for German NER on the *Europarl* corpus, and on par with supervised English systems. We find similar accuracy for Finnish, French, Lithuanian, and Slovene, showing that the approach is broadly applicable and not confined to particular languages. We can further improve the accuracy by combining the supervised and unsupervised approaches, producing state-of-the-art performance across the board.

Outline: Section 2 reports on related work. Section 3 introduces the data. Section 4 defines the method for generating candidate entity seeds, introducing a new, efficient edit distance metric. Section 5 defines the method for building models based on the seeds, defining the features. Section 6.1 gives the results for identifying named entity versus non-entity spans, comparing the results to supervised systems. Section 6.2 gives the results for identifying types of named entities (Person, Location and Organization) combining the unsupervised and supervised approaches. Section 7 analyzes the results and finds common patterns for successes and errors. Section 8 concludes.

2 Background and Related Work

Despite the (we believe) intuitive nature of our approach, there is no previous work reporting cross-linguistic named entity recognition that similarly exploits the similarity of entities across languages.

The exception is the work that we build on here (Munro and Manning, 2012; Munro, 2012b), where a Levenshtein-based edit distance was used to generate seeds candidates from 3,000 short messages in English and Haitian Krèyol, concentrating on the entity/non-entity division. Here, we have simplified the edit distance measure with a new $O(N)$ algorithm, applied the method across millions of examples in five languages, and investigated the more complex problem of predicting entity types. In addition to the gain in speed, the new approach gives a substantial F=0.20+ gain in accuracy when identifying entity types (Micro-F = 0.825, as compared to 0.619 in Munro and Manning (2012)).

Cross-linguistic syntax has boostrapped parsers into new languages (Cohen et al., 2011; McDonald et al., 2011; Zeman and Resnik, 2008), building on Yarowsky et al.'s induction of projections (Yarowsky et al., 2001; Hwa et al., 2005). Kim and Toutanova have recently shown how this can be extended to NER (Kim et al., 2012) by employing edit distance on Wikipedia data, a common source of data (Richman and Schone, 2008; Lin et al., 2011; Simon and Nemeskey, 2012).

Language	Sentences	Words
Finnish	1,924,943	79,856,773
French	2,007,724	102,788,003
German	1,920,210	92,494,398
Lithuanian	635,147	26,736,933
Slovene	623,491	27,547,502
Total	**7,111,515**	**329,423,609**

Table 1: Word and sentence counts for the five data sets. Each sentence is aligned with an English translation in the *Europarl* corpus. The word counts include both the English and non-English sentences. For each language, a held-out evaluation set of about 1,000 examples was manually annotated, creating 10,000 evaluation items.

Most unsupervised entity detection has been within one language (Pedersen et al., 2006; Nadeau et al., 2006) and have underperformed supervised systems.

Past cross-linguistic NER has drawn from cross-linguistic gazetteers (Sang, 2002; Sang and De Meulder, 2003). While the goal of these systems is also broad cross-linguistic coverage for NER, this is *not* the same use of the term 'cross-linguistic', as these systems assumed substantial labeled data in the target language.

Most earlier work used a tagger in one language in combination with machine translation-style alignments models. For example, Huang et al. translated rare named entities in a similar low-resource context (Huang et al., 2003a; Huang et al., 2003b; Huang et al., 2004), and Màrquez et al. also used unsupervised resources in cross-language NER, utilizing word-level mappings between Catalan and Spanish to improve NER in the low-resource Catalan (Màrquez et al., 2003).

Klementiev and Roth exploited a different property of named entities across languages: compared to other words, the use of named entities in the news changes over time in correlation to the reporting of certain people, organizations and loca-

Language	Example Sentence, Seed and Translation
Finnish	Herra **Schiedermeierilla** on puheenvuoro menettelyä koskevaa esitystä varten. Mr **Schiedermeier** has the floor on a point of order.
French	Ceci étant, nous aboutirons à un accord à **Copenhague**, j'en suis convaincu. Nonetheless, I believe that agreement will be reached in **Copenhagen**.
German	Ich stimme dem Vorschlag der **Kommission** ohne Einschränkung zu. I fully agree with the proposal of the **Commission**.
Lithuanian	Kitas klausimas - diskusija dėl **Madagaskaro**. The next item is the debate on **Madagascar**.
Slovene	Zahvala pa je namenjena tudi gospe **Zdravkovi** za njeno poročilo. Thanks are also due to Mrs **Zdravkova** for her report.

Table 2: Example seeds. The phrases in bold are the most similar across the sentence pairs. Relative to the average similarity between randomly selected pairs of phrases in the sentences, the phrases in bold have z-scores, from top to bottom, of 178.79, 128.37, 138.33, 99.42 and 100.72, meaning that they were strong matches in otherwise dissimilar sentences. The top 50,000 were selected from each language as seed *Entities*. The bottom 50,000 were chosen as seed *Non-Entities*.

tions (Klementiev and Roth, 2008).

Piskorski et al.'s work on NER for inflectional languages (Piskorski et al., 2009) relied on similarities in edit distance between the *intra*-language variation of names that results from complex morphology, and could be used to improve the results for Lithuanian and Finnish.

3 Data and assumptions

We used 7,000,000 sentences from the proceedings of the European Parliament (*Europarl Corpus*) (Koehn, 2005), drawn from German, Finnish French, Lithuanian and Slovene. The sentences have a translation in English, making five large, parallel data sets (exact numbers in Table 1). For each of the five languages, approximately 1,000 phrases were labeled by bilingual speakers by *Idibon* annotators fluent in both English and the aligned language, resulting in 10,000 labeled evaluation items. We provide the annotations freely.

The five languages from Europarl were chosen for maximum diversity among Latin-script languages, with one each from the *Germanic, Romance, Slavic, Uralic*, and *Baltic* language families/sub-divisions. This was to ensure that the success was not a quirk of a particular language, and also to investigate language specific features. For example, Finnish and Lithuanian are more morphologically complex languages, meaning that subwords features should play a bigger role as they capture common stems and affixes.

All partial matches were treated as errors with the lone exception of closed-set titles like "Mr", "Dr", "Inc.", (for our system and those used for comparison). To an extent, it is a coding convention as to whether titles are included as part of the entity, and we assume that closed set words could be trivially be recaptured or dropped post-processing. For open-set titles or phrasal entities, like "Baroness Nicholson of Winterbourne", it was an error unless fully identified.

4 Identifying entity candidate seeds

The first step is to create a selection of candidate named entities: the "*seeds*". The seeds will become the data for building the models so it is important to have broad coverage while ensuring that there are as few false-positives as possible.

For each aligned sentence, we calculate the average edit distance between all possible pairs of phrases in the aligned sentences, using a new and efficient local edit distance deviation metric (see below) along with the standard deviation. We then extract the phrase-pair with the lowest edit distance as the most likely entity candidate in that sentence. The final score for the aligned phrase, the *DIST*, is its z-score: that is, the number of standard deviations above the average edit distance between candidate alignments. The phrases were an average of 1.3 words long. See Table 2 for examples of sentences and seeds.

4.1 A faster edit-distance calculation

It was necessary to create a new edit-distance measure for this task. The large number of potential alignments between two sentences meant thou-

sands of comparisons per aligned-sentence and over a corpus of seven million sentences, this meant tens of billions of comparisons.

Levenshtein edit-distance has $O(N^2)$ complexity (or $O(Nlog(N))$ with Vantage-Point Trees) and turned out to be too expensive for our large corpus. As Munro and Manning (2012) pointed out, Levenshtein, Jaro-Winkler, and other edit-distance measures performed equally on this task, with the real difference in accuracy coming from the z-scores rather than the raw edit distance.

We therefore created a new edit distance measure with $O(N)$ complexity. Given two strings, every character bigram is hashed in the first string, indexed to their position in the string. The number of matching bigrams in the second string is calculated from this hash. The sequences of matching bigrams are calculated in a further pass, allowing a good approximation for the length of each matching substring. The edit distance is then calculated as the fraction of the length of the substrings divided by the total possible length. This is calculated in both directions for the two strings, and averaged across the two. While it requires multiple passes across both strings (8 in our current implementation), the number of passes is fixed so the cost is $O(N)$.

Let s be a sequence of bigram substrings that are also present in the aligned phrase S. For each bigram, the score for the bigram, b is the longest sequence of matching bigrams that occur in either direction:

$$SC(b) = \arg\max_{s} f(s), b \in S \quad (1)$$

The raw score per token (word), $SC(W)$, is the average score of its constituent bigrams:

$$SC(W) = \overline{SC(b)}, b \in W \quad (2)$$

The edit distance $DIST(W)$ is the deviation of $SC(W)$ from the average across all bigrams other than $SC(w)$, where $!b = b, \notin W$ (that is, where the z-score is calculated against all sequences *except* those in the candidate named entity):

$$DIST(W) = \frac{\sigma_{!b}}{SC(W) - \overline{!b}} \quad (3)$$

When the token with the greatest $DIST()$ is found, the method iterates out to calculate the adjacent tokens and add them to the candidate entity if they also satisfy the condition of being substantially divergent from the average for the remaining non-tokens.

We provide the source code along with the paper for the full algorithm and replicability.

Compared to Levenshtein and Jaro-Winkler, it gives higher scores to long sequences of transpositions, like "Lake Malawi" becoming "Malawi Lake", only penalizing the word-gap sequences. For this particular task, where Adjective-Noun word order changes across languages, this is probably a positive feature. Our metric also favors longer sequences, and therefore gives a smaller penalty to differences near the edge of the strings. Most of the differences between entities across languages occurred at word boundaries, as in all the examples in Table 2, so this was also a positive feature of our new edit distance measure.

5 Learning joint alignment, context and word-shape models

Taking the top and bottom 50,000 seeds from the previous step as entities and non-entities, a model is created to predict entities across the entire aligned corpus. We used the Stanford Maximum Entropy Classifier (Klein and Manning, 2003).

The model is built on features that include the words themselves, the subword models, the context (the preceding and following words), word-shape features (capitalization, punctuation, segmentation, and numerical patterns), and alignment (absolute and relative character offsets between the candidates in the messages and translation).

5.1 Features

The features are jointly learned over both languages:

Word features: simply captures the words of the named entities. A model with just words is used as the baseline for the feature analysis, as it is a reasonable approximation for the accuracy of the seed-generation step.

Subword features: were used to capture consistent patterns below the word-level. Sentences contained variable spellings due to prefixing, suffixing and compounds, so subword features capture the common properties. They also capture features like '-ia', which is submorphemic (in English) but a good predictor for locations. We selected the most predictive sequences through L_1 regression of all sequences of characters, up to four.

Word-shape features: are a form of subword modeling that is sensitive only to the differences between alphabetic, numeric, punctuation and delimiting characters, and upper/lower case distinctions (Collins, 2002). The template converts all upper-case letters to 'C', all lower-case letters to 'c', all punctuation to 'p', all spaces to 's', and all numbers to 'n'. A set of features was also created using a digest, where consecutive characters of the same form were ignored. The digest allows the system to generalize over similar words/phrases.

Contextual features: capture the preceding and following words.

Alignment features: model the relative locations of the candidate entities in the two sentences. Two types of alignment features were used: raw and relative. The raw alignment was the difference between the offsets. For example, if one entity begins at the 20th character and the candidate alignment begins at the 22nd character, than the difference is '2'. The relative is the percentage, to take into account that some sentences are inherently longer in some languages.

5.2 Restrictions on training

We limited the seeds to 100 training examples per token to ensure that the positive seeds were representative. This was because some very common tokens dominated the first 50,000 examples in each language when the corpus was ordered by *DIST*. We also made one concession to language-specific processes, by removing all tokens from the seed entities that occurred capitalized more often than not in the English data. This removed a few common non-entity cognates from this part of the data. We removed these words for the seed entities only: no language-specific processes were used on the seed non-entities or the evaluation data.

Edit distance was not included from the features: it would simply get all the weight as it was used to derive the seeds.

6 Results

We report results identifying entity spans (the entity/non-entity) division, and then the more complicated task of identifying entity types (Person, Location and Organization). In both cases we compare our results to supervised systems.

Unsupervised			
	Prec.	**Recall**	**F-value**
Finnish	0.887	0.767	**0.822**
French	0.846	0.839	**0.842**
German	0.872	0.936	**0.903**
Lithuanian	0.810	0.763	0.797
Slovene	0.812	0.829	0.821
Supervised (from English)			
Finnish	0.947	0.724	0.821
French	0.937	0.760	0.838
German	0.949	0.830	0.885
Lithuanian	0.935	0.737	**0.824**
Slovene	0.876	0.789	**0.830**
Supervised (from German)			
HGC German	0.896	0.381	0.534
HGC German$^{\alpha}$	0.942	0.790	0.859
deWac German	0.889	0.392	0.544
deWav German$^{\alpha}$	0.938	0.801	0.864

Table 3: Results for identifying named entity spans (entity/non-entity division), showing that the unsupervised methods often performed with greater accuracy than the supervised systems.
$^{\alpha}$With a manual correction for "Kommission".

6.1 Identifying entity spans

Table 3 compares our unsupervised system to two top supervised NER systems on our evaluation data, showing competitive or better results for our approaches across the five languages.

The *Stanford Named Entity Recognizer* (Finkel et al., 2005) is used for English.

Faruqui and Padó's *German Named Entity Recognizer with Semantic Generalization* (Faruqui and Padó, 2010) is used for German, representing the state-of-the-art performance. They also base their system on the *Stanford Named Entity Recognizer*, extending it with a clustering scheme that allows out-of-domain unlabeled data to be clustered according to morphological similarity (Clark, 2003), which effectively extends the knowledge of unseen words by including their cluster membership. They extend a baseline system based on CoNLL data with two German corpora: the *Huge German Corpus* (HGC), consisting of approximately 175 million tokens from German newspapers, and the *deWac corpus* (deWac), taken from 1.7 billion tokens (Baroni et al., 2009). They evaluate on CoNLL, reporting the highest published results for German to-date, and also report on *Europarl* as an out-of-domain corpus.

	Combined (classified seeds)			Supervised (from English)		
	Prec.	**Recall**	**F-value**	**Prec.**	**Recall**	**F-value**
Finnish						
Person	0.866	0.839	**0.852**	0.957	0.395	0.559
Location	0.760	0.775	**0.767**	0.399	0.679	0.502
Organization	0.405	0.185	**0.254**	0.588	0.098	0.168
Micro-F			**0.712**			0.467
French						
Person	0.739	0.944	**0.829**	0	0	0
Location	0.881	0.963	0.920	0.956	0.896	**0.925**
Organization	0.804	0.854	**0.828**	0.963	0.481	0.642
Micro-F			**0.869**			0.667
German						
Person	0.824	0.923	**0.871**	0.960	0.615	0.750
Location	0.722	0.876	**0.792**	0.650	0.810	0.721
Organization	0.833	0.918	**0.874**	0.125	0.011	0.021
Micro-F			**0.857**			0.360
Lithuanian						
Person	0.727	0.873	**0.793**	0.923	0.655	0.766
Location	0.568	0.882	**0.691**	0.650	0.611	0.630
Organization	0.806	0.382	**0.518**	0.500	0.013	0.026
Micro-F			**0.626**			0.357
Slovene						
Person	0.765	0.788	0.776	0.879	0.773	**0.823**
Location	0.729	0.925	**0.815**	0.609	0.346	0.441
Organization	0.733	0.543	**0.624**	0.477	0.226	0.307
Micro-F			**0.740**			0.505
Overall (& English)						
Person	0.853	0.877	**0.864**	0.916	0.558	0.673
Location	0.806	0.898	**0.848**	0.685	0.721	0.683
Organization	0.796	0.776	**0.777**	0.513	0.321	0.360
Micro-F			**0.825**			0.554

Table 4: Results for the full Named Entity Recognition task, showing F=0.202 (French) to F=0.497 (German) improvement over a supervised system alone. The combined system uses the supervised learner to make the best possible prediction of the entity type for the English candidate entities seeds, and propagates the prediction via the candidate alignment. This produced a more accurate result than predicting directly from the supervised learners for all tasks except *Location* in French and *Person* in Slovene. The results for Overall are also the English results, as each other language was aligned with English and *DIST* produced no misalignments in the evaluation data. The overall accuracy of F=0.825 is competitive with any out-of-domain results for English in prior published research. For individual languages, the result is much higher, for example, F=0.857 for German is about F=0.20 higher than the previously reported results for out-of-domain *Europarl* of F=0.656 (Faruqui and Padó, 2010), and approximately F=0.25 higher than Pinnis' results for Lithuanian, which was reported as within F=0.01 of the Stanford NER system (Pinnis, 2012).

By Language		Added Feature				
	Words	**Context**	**Subwords**	**WShapes**	**Alignmt**	**All**
Finnish	0.564	9.75%	7.80%	8.51%	-0.35%	26.24%
French	0.866	-4.85%	-12.70%	-1.27%	-4.27%	0.35%
German	0.792	-5.18%	-2.40%	8.46%	-13.01%	8.08%
Lithuanian	0.388	5.67%	20.88%	43.30%	-0.52%	61.34%
Slovene	0.515	7.38%	4.85%	16.89%	-8.93%	43.69%
By Entity Type						
Person	0.698	17.77%	-13.04%	-3.44%	-3.30%	15.76%
Location	0.577	-11.44%	12.82%	33.28%	-16.12%	35.70%
Organization	0.518	-3.86%	11.97%	7.14%	0.97%	16.41%

Table 5: Feature analysis, showing the % improvement of adding features to a baseline model with only words. The results show that most features increased accuracy, but none consistently across all languages. This is an encouraging result from a research perspective: the methods applied here could be improved upon in a number of ways to increase overall accuracy.

The *Stanford Named Entity Recognizer* for English is probably not state-of-the-art, but is a high-performing publicly available supervised system. We cannot be certain which of all published supervised English systems would be more accurate than our unsupervised approach without implementing every single one, but this is our goal: we are presenting an approach that can be applied to languages where there is *not* a large amount of existing data, systems or research. The fact we are competitive with top English supervised systems is enough to demonstrate the utility for languages where no such resources exist.

For Faruqui and Padó's system, there was one token that was a frequent error: *Kommission*. Non-entity nouns are capitalized in German, so when short for *Europäische Kommission* ("European Commission") it was an entity, but the same word was also used to refer to "committees", in general. For English the lower case variant ("committee") trivially distinguishes the non-entity use. For German alone, contextual information is required to distinguish the use, sometimes from several sentences away. So as not to penalize the supervised system for a single token, we also report results for Faruqui and Padó with "*Kommission*" manually corrected (marked with α in Table 3, and used by default elsewhere.)

(Munro, 2012a)

6.2 Classifying types of entities

While the purely unsupervised system is competitive with the state-of-the-art supervised systems, it does not distinguish types of entities.

We combined the supervised and unsupervised approaches by applying the supervised English system to candidate seeds. We classify each candidate entity seed from the unsupervised output with its most probable label (Person, Location or Organization). The labels are then propagated across the candidate alignments, allowing us to build models that predict the entity types in Finnish, French, German, Lithuanian and Slovene.

Table 4 gives the results when predicting types of entities, broken down for each of the five languages, The results are even better than for the entity/non-entity division, showing between F=0.202 (French) and F=0.497 (German) improvement over a purely supervised system.

7 Analysis

Here, we analyze the patterns of errors and success. Table 5 gives the contribution of individual feature types to the model, broken down by language and entity type. The results for "Words" alone are indicative of the accuracy of the seed-generation step.

The second step added substantial accuracy from the seed stage, with up to $F = 0.228$ gain in accuracy, with the exception of the French data.

The alignment features did not produce any significant increases in accuracy and could be removed entirely. However, in unreported results we found that they did improve accuracy for the entity/non-entity distinction.

Turning to entity types, Table 5 also shows that different features played important roles for different entities: context was most important for Person

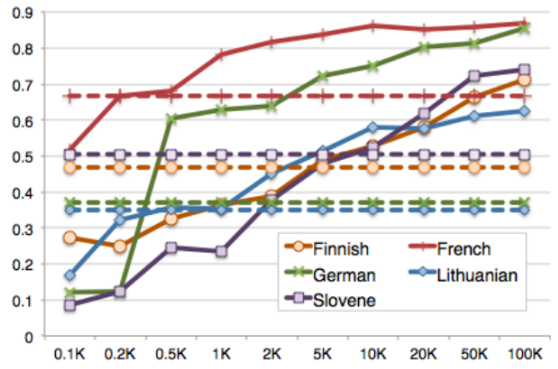

Figure 1: Learning rates, showing Micro-f accuracy with different volumes of candidate seeds, up to 100,000 items. Solid lines are the combined systems in this paper. Dashed lines are the supervised results for that language. Most languages pass the supervised results with 1K seed items and all pass the supervised results with 10K seed items, indicating that this is an accurate process with relatively small sets of aligned data.

because of titles ("Mrs", "Mr") while word shapes were the most effective for Locations because of features like numbers in addresses.

There were some items that the supervised system correctly predicted that the unsupervised approaches did not. The unsupervised systems were less accurate at identifying entity boundaries, and had more false positives. For example, "Internet", "September", and "Titanic" were false positives, as they are not Locations, Organizations or Persons. As a closed set group, dates could be filtered, but the other types could remain a problem without introducing labeled data or external resources.

There were many examples of entities that the unsupervised methods identified correctly but the supervised methods missed. Examples include "Kosovo", "Seychellen", and "Mugabe". These were all the result of domain-dependence, present in *Europarl* but probably not in the corpora that the supervised systems were trained on.

The learning curve is given in Figure 1. Among the languages, the French data was the most accurate, but analysis shows this to be the least interesting: most of the accuracy is from capitalization alone, which is a simple concept that is easy to learn at low volumes of data.

For German, the capitalization was not so important (using capitalization alone only returns about $F = 0.5$ for this data set). However, the word-shape features do play an important role for this set, but for the aligned English, not the German itself. This was why we out-performed Faruqui and Padó: their system was inherently better than the English system, but NER is much easier in English and with aligned data this proved to be the deciding factor.

Turning around the focus, the method introduced here can also be thought of as a form of domain adaptation: we can use translations of the out-of-domain evaluation data in German (or Finnish, French, Lithuanian & Slovene) to bootstrap a more accurate English named entity tagger.

The Slovene errors contained the largest number of entities that were not of the correct type, like the disease *H1N1* and the currency *Euro*. This seems to be a chance occurrence of the random selection of evaluation items and not property of the language. A larger evaluation corpus would be needed to confirm this, but the results were otherwise consistent with the output that we would expect from a supervised system.

For Lithuanian and Finnish, the most common error resulted from a misclassification of "Parliament", which occurs in the data with more than 40 different suffixes or compounds, including: *a, ą, -ai, -ais, -ams, -arai, -aras, -are, -arė, -arei, -arės, -aro, -arų, -arui, -arus, -as, -e, -i, -in, -inę, -inei, -inėje, -ins, -inėse, -iniai, -iniame, -inio, -inis, -iniu, -inių, -o, -os, -s, -u, -ų -ui, -uose*. Finnish had more than *400* different suffixes or compounds for "Parliament". Even with the subword models, this was too ambiguous for the models, accounting for 20% of the errors for Lithuanian and 10% for Finnish, most often misclassifying the string as a Location due to an adjacent variation on "Europe". Lithuanian and Finnish showed substantial improvement from the subword models, with $F = 0.081$ and 0.044 improvement over words alone in Table 5, but more gains are no doubt possible.

8 Concluding remarks

The system reported here out-performs state-of-the-art supervised approaches by up to $F = 0.20$, despite the absence of labeled training data, or by $F = 0.40$ when projecting labeled training from a high resource language.

By leveraging loosely aligned sentences, we are able to take advantage of one of the most abundant language resources that exists: translations. While not explicitly labeling the entities, the translations

encode the translators' cross-linguistic knowledge of names, keeping them more constant than their surrounding words. From this, we are able to bootstrap a method that could be applied to any Latin-script language with considerable accuracy.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2003. A simple named entity extractor using adaboost. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 152–155. Association for Computational Linguistics.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 59–66. Association for Computational Linguistics.

Shay B. Cohen, Dipanjan Das, and Noah A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 50–61, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Michael Collins. 2002. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 489–496. Association for Computational Linguistics.

Silviu Cucerzan and David Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence. In *Joint SIGDAT conference on EMNLP and VLC*.

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

Jenny Rose Finkel and Christopher D. Manning. 2010. Hierarchical joint learning: Improving joint parsing and named entity recognition with non-jointly labeled data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ACL 2010*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 168–171. Association for Computational Linguistics.

A.F. Gallippi. 1996. Learning to recognize names across languages. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 424–429. Association for Computational Linguistics.

Fei Huang, Stephan Vogel, and Alex Waibel. 2003a. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition-Volume 15*, pages 9–16. Association for Computational Linguistics.

Fei Huang, Stephan Vogel, and Alex Waibel. 2003b. Extracting named entity translingual equivalence with limited resources. *ACM Transactions on Asian Language Information Processing*, 2(2):124–129.

Fei Huang, Stephan Vogel, and Alex Waibel. 2004. Improving named entity translation combining phonetic and semantic similarities. In *Proceedings of HLT-NAACL*, pages 281–288.

R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 11(3):311–325.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. 2012. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*.

Dan. Klein and Christopher C. Manning. 2003. Maxent model, conditional estimation, and optimization. *HLT-NAACL 2003 Tutorial*.

A. Klementiev and D. Roth. 2008. Named entity transliteration and discovery in multilingual corpora. In Cyril Goutte, Nicola Cancedda, Marc Dymetman, and George Foster, editors, *Learning Machine Translation*. MIT Press.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5.

Chun-Jen Lee, Jason S. Chang, and Jyh-Shing R. Jang. 2006. Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5(2):121–145.

W.P. Lin, M. Snover, and H. Ji. 2011. Unsupervised Language-Independent Name Translation Mining from Wikipedia Infoboxes. *Proceedings of the EMNLP Workshop on Unsupervised Learning in NLP*, page 43.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pages 359–367.

Lluís Màrquez, Adrià de Gispert, Xavier Carreras, and Lluís Padró. 2003. Low-cost named entity classification for catalan: exploiting multilingual resources and unlabeled data. In *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition-Volume 15*, pages 25–32. Association for Computational Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics.

Robert Munro and Christopher D. Manning. 2012. Accurate unsupervised joint named-entity extraction from unaligned parallel text. In *Proceedings of the Named Entities Workshop (NEWS 2012)*, Korea.

Robert Munro. 2012a. Crowdsourcing and the crisis-affected population. *Information retrieval*, 16(2):210–266.

Robert Munro. 2012b. *Short Message Communications In Low-Resource Languages.* Ph.D. thesis, Stanford University, Stanford, CA.

D. Nadeau, P. Turney, and S. Matwin. 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. *Advances in Artificial Intelligence*, pages 266–277.

T. Pedersen, A. Kulkarni, R. Angheluta, Z. Kozareva, and T. Solorio. 2006. An unsupervised language independent method of name discrimination using second order co-occurrence features. *Computational Linguistics and Intelligent Text Processing*, pages 208–222.

Mārcis Pinnis. 2012. Latvian and lithuanian named entity recognition with tildener. *Proceedings of LREC-2012*, 40:37.

J. Piskorski, K. Wieloch, and M. Sydow. 2009. On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages. *Information retrieval*, 12(3):275–299.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Alexander E Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08)*.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

E.F Tjong Kim Sang and F. 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*, pages 142–147. Association for Computational Linguistics.

E.F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: language-independent named entity recognition, proceedings of the 6th conference on natural language learning. *August*, 31:1–4.

Eszter Simon and Dávid Márk Nemeskey. 2012. Automatically generated ne tagged corpora for english and hungarian. In *Proceedings of the Named Entities Workshop (NEWS 2012)*, Korea.

Ralf Steinberger and Bruno Pouliquen. 2007. Cross-lingual named entity recognition. *Lingvisticæ Investigationes*, 30(1):135–162.

D. Yarowsky, G. Ngai, and R. 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*, pages 1–8. Association for Computational Linguistics.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, page 35.