Low-resource named entity recognition via multi-source projection: Not quite there yet?

Jan Vium Enghoff **Søren Harrison** **Željko Agić**
Department of Computer Science
IT University of Copenhagen
Rued Langgaards Vej 7, 2300 Copenhagen S, Denmark
zeag@itu.dk

Abstract

Projecting linguistic annotations through word alignments is one of the most prevalent approaches to cross-lingual transfer learning. Conventional wisdom suggests that annotation projection "just works" regardless of the task at hand. We carefully consider multi-source projection for named entity recognition. Our experiment with 17 languages shows that to detect named entities in true low-resource languages, annotation projection may not be the right way to move forward. On a more positive note, we also uncover the conditions that do favor named entity projection from multiple sources. We argue these are infeasible under noisy low-resource constraints.

1 Motivation

Annotation projection plays a crucial role in cross-lingual NLP. For instance, the state of the art approaches to low-resource part-of-speech tagging (Das and Petrov, 2011; Täckström et al., 2013) and dependency parsing (Ma and Xia, 2014; Rasooli and Collins, 2015) all make use of parallel corpora under the source-target language dichotomy in some way or another. Beyond syntactic tasks, aligned corpora facilitate cross-lingual transfer through multilingual embeddings (Ruder et al., 2017) across diverse tasks.

What about named entity recognition (NER)? This sequence labeling task with ample source languages appears like an easy target for projection. However, as recently argued by Mayhew et al. (2017), the issue is more complex:

> "For NER, the received wisdom is that parallel projection methods work very well, although there is no consensus on the necessary size of the parallel corpus. Most approaches require millions of sentences, with a few exceptions which require thousands. Accordingly, the drawback to this approach is the difficulty of finding any parallel data, let alone millions of sentences. Religious texts (such as the Bible and the Koran) exist in a large number of languages, but the domain is too far removed from typical target domains (such as newswire) to be useful. As a simple example, the Bible contains almost no entities tagged as organization."

Our paper is a thorough empirical assessment of the quoted conjecture for named entity (NE) tagging in true low-resource languages. In specific, we ask the following questions:

– Are there conditions under which the projection of named entity labels from multiple sources yields feasible NE taggers?
– If yes, do these conditions scale down to real low-resource languages?

To answer these questions, we conduct an extensive study of annotation projection from multiple sources for low-resource NER. It includes 17 diverse languages with heterogeneous datasets, and 2 massive parallel corpora. In terms of cross-lingual breadth, ours is one of the largest NER experiments to date,[1] and the only one that focuses on standalone annotation projection. We uncover that the specific conditions that do make NER projection work are not trivially met at a feasibly large scale by true low-resource languages.

2 Multilingual projection

We project NE labels from multiple sources into multiple targets through sentence and word align-

[1] Cross-lingual NER is typically tested on 4-10 languages, predominantly the four CoNLL shared task languages (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003): Dutch, English, German, and Spanish. We discuss some recent notable exceptions as related work.

Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, pages 195–201
Brussels, Belgium, Nov 1, 2018.

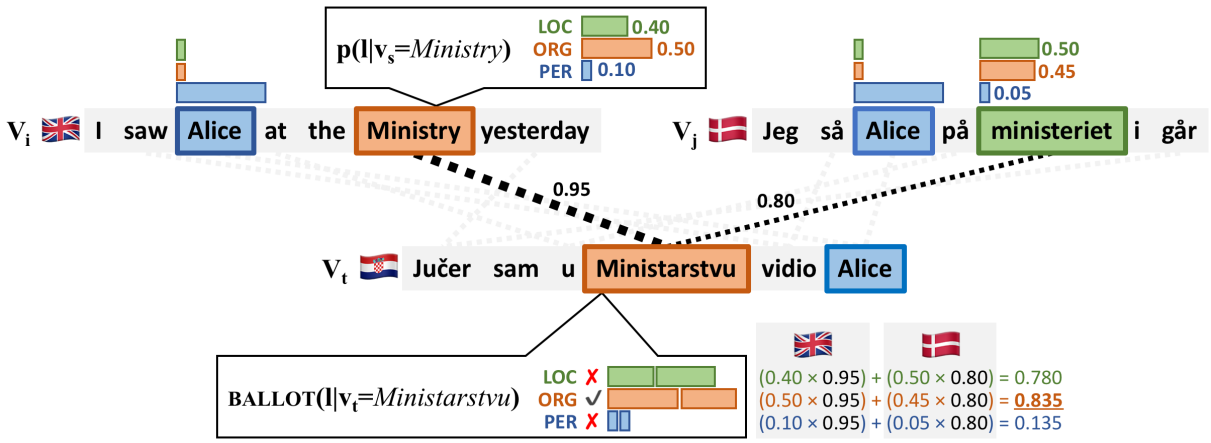


Figure 1: An illustration of named entity projection from two source sentences (Danish, English) to one target (Croatian). In this example, the voting of entity labels is weighted by tagger confidence and alignment probability. The outside label (O) is omitted for simplicity.

Algorithm 1: Multi-source label projection

Data: Multilingual sentence graph $G = (V_s \cup V_t, A)$; sequential labels L; source label distributions $p(l|v_s)$

Result: A labeling of target words $v_t \in V_t$

BALLOT $\leftarrow$ *empty voting table*

LABELING $\leftarrow$ *empty label-to-vertex mapping*

for $v_t \in V_t$ **do**

  for $l \in L$ **do**

    $\text{BALLOT}(l|v_t) \leftarrow \sum_{v_s \in V_s} p(l|v_s) \cdot a(v_s, v_t)$

  $\text{LABELING}(v_t) = \arg\max_l \text{BALLOT}(l|v_t)$

return BALLOT, LABELING

ments. Our projection requires source NE taggers and parallel corpora that are ideally large in both breadth (across many languages) and depth (number of parallel sentences). Evidently, we require that i) the source language texts in the corpus are tagged for named entities, and that ii) the parallel corpora are aligned. Both conditions are typically met under some noise: by applying source-language NE taggers, and unsupervised sentence and word aligners, respectively.

We view a parallel corpus as a large collection of multilingual sentences. A multilingual sentence is a graph $G = (V, A)$ comprising a target sentence t and n source sentences. The vertice sets $V = V_0 \cup \cdots \cup V_n$ represent words in sentences, where the words $v_t \in V_0$ belong to the target sentence $V_0 = V_t$, while all other words $v_s \in V_i$ belong to their respective source sentences $V_i, i \in \{1, ..., n\}$. The graph is bipartite between source vertices $V_s = V \setminus V_t$ and target vertices V_t, where the edges are word alignments with aligner confidences $a(v_s, v_t) \in (0, 1)$ as weights. Each source token v_s is associated with a label distribution $p(l|v_s)$ that comes from a respective source-language tagger and indicates its confidence over labels $l \in L$. Here, the labels L are NE tags, but elsewhere they could instantiate other sequence labeling such as POS or shallow parses.

Under these assumptions, we implement projection as weighted voting of source contributions to target words, such that for each target word v_t we collect votes into a ballot:

$$\text{BALLOT}(l|v_t) = \sum_{v_s \in V_s} p(l|v_s) \cdot a(v_s, v_t).$$

Here, each source token v_s gets to cast a vote for the future label of v_t. Each vote is weighted by its own tagger confidence and reliability of its alignment to target token v_t: $p(l|v_s) \cdot a(v_s, v_t)$. The individual votes are then summed and the tags for the target tokens are elected. We can train a NE tagger directly from BALLOT provided some normalization to $(0, 1)$, or we can decode a single majority tag for each target word:

$$\text{LABELING}(v_t) = \arg\max_l \text{BALLOT}(l|v_t).$$

The process is further detailed as Algorithm 1 and also depicted in Figure 1 for two source vertice sets V_i and V_j, and one target set V_t. This simple procedure was proven to be markedly robust and effective in massively multilingual transfer of POS taggers especially for truly low-resource languages by Agić et al. (2015; 2016).

CoNLL 2002 (Tjong Kim Sang, 2002)	es nl	news
CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003)	en de	news
OntoNotes 5.0 (Weischedel et al., 2011)	en ar	news
NER FIRE 2013 (Rao and Devi, 2013)	ta hi	wiki
ANERCorp (Benajiba et al., 2007)	ar	news
BSNLP 2017 (Piskorski et al., 2017)	cs hu pl sk sl	news
Estonian NER (Tkachenko et al., 2013)	et	news
Europeana NER (Neudecker, 2016)	fr	news
I-CAB (Magnini et al., 2006)	it	news
HAREM (Santos et al., 2006)	pt	–
Stockholm Internet Corpus (Östling, 2013)	sv	blogs

Table 1: The NER datasets in our exepriment. We indicate the languages[2] and domains they cover.

We take into account a set of additional design choices in multi-source NER projection beyond what the algorithm itself encodes.

Sentence selection. We compare two ways to sample the target sentences for training: at random vs. through word-alignment coverage ranking. A target word covered if it has an incoming alignment edge from at least one source word. We mark the target sentences by percentage of covered words from each source, and rank them by mean coverage across sources. We then select the top k ranked sentences to train a tagger. We optimize this parameter for maximum NER scores on development data.

Language similarity. Some source languages arguably help some targets more than others. We model this relation through language similarity between source and target WALS feature vectors (Dryer and Haspelmath, 2013): $\mathbf{v_s}$ and $\mathbf{v_t}$. We implement language similarity as inverse normalized Hamming distance between the two vectors: $1 - d_h(\mathbf{v_s}, \mathbf{v_t})$. Only the non-null fields are taken into account. Similarity is contrasted to random selection in our experiment.

Tagger performance. Some source NE taggers perform better than the others monolingually. We thus consider the option to weigh the source contributions not just by language similarity but also through their monolingual NER accuracy, so that the contributions by more accurate source taggers are selected more often.

3 Experiment setup

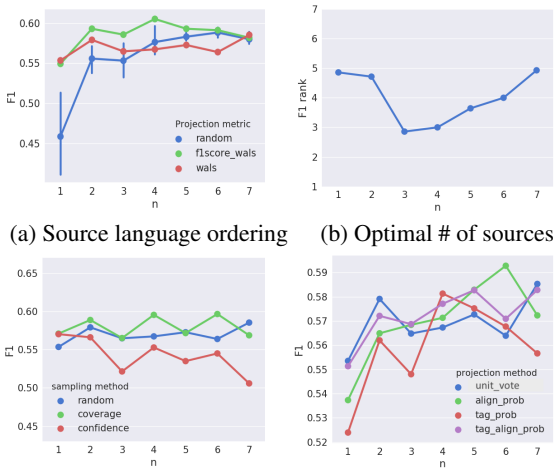
Sources and targets. Table 1 shows the NER-annotated datasets we used. These datasets adhere to various differing standards of NE encoding. In a non-trivial effort, we semi-automatically normalize the data into 3-class CoNLL IO encoding (Tjong Kim Sang and De Meulder, 2003), as the common denominator for the widely heterogeneous datasets. We thus detect names of locations (LOC), organizations (ORG), and persons (PER). Languages with more than 5k monolingual training sentences serve as sources and development languages for parameter tuning, while the remainder pose as low-resource targets; see Table 2. For languages that have multiple datasets, we concatenate the data. We end up with typologically diverse sets of sources and targets. We use the predefined train-dev-test splits if available; if not, we split the data at 70-10-20%.

Parallel text. We contrast two sources of parallel data: Europarl (Koehn, 2005) and Watchtower (Agić et al., 2016). The former covers only 21 resource-rich languages but with 400k-2M parallel sentences for each language pair, while the latter currently spans over 300 languages, but with only 10-100k sentences per pair. Europarl comes with near-perfect sentence alignment and tokenization, and we align its words using IBM2 (Dyer et al., 2013). For Watchtower we inherit the original noisy preprocessing: simple whitespace tokenization, automatic sentence alignment, and IBM1 word alignments by Agić et al. (2016) as they show that IBM1 in particular helps debias for low-resource languages.

Tagger. We implement a bi-LSTM NE tagger inspired by Lample et al. (2016) and Plank et al. (2016). We tune it on English development data at two bi-LSTM layers ($d = 300$), a final dense layer ($d = 4$), 10 training epochs with SGD, and regular and recurrent dropout at $p = 0.5$. We use pretrained fastText embeddings (Bojanowski et al., 2017). Currently fastText supports 294 languages and is superior to random initialization in our tagger. Other than through fastText, we don't make explicit use of sub-word embeddings. Our monolingual F_1 score on English is 86.35 under the more standard IOB2 encoding. We do not aim to produce a state-of-the-art model, but to contrast the scores for various annotation projection parameters. We use our tagger both to annotate the source sides of parallel corpora, and to train projected target language NER models. All reported NE tagging results are means over 4 runs.

4 Results

Europarl sweet spots. With Europarl we show that the combination of monolingual F_1 source



(a) Source language ordering (b) Optimal # of sources

(c) Parallel sentence sampling (d) Weights in label voting

Figure 2: Projection tuning on Europarl: a) Ordering the sources by their monolingual F_1 scores $\times$ WALS similarity works best; b) At $n = 3$ sources the average rank of F_1 scores across development languages is lowest, which indicates that $n = 3$ is the optimal number of sources in Europarl projection; c) Parallel sentences are best selected by mean word alignment coverage, in contrast to tagger confidence or random sampling; d) Weighted voting for LABELING performs best when weights are word alignment weights $\times$ tagger confidences. Results under (b), (c), and (d) all use the best source ordering approach from (a). For random sampling under (a), the sources were randomly selected 5 times for each n.

scores and WALS similarities is the optimal source language ordering. The respective optimal number of sources is $n = 3$ for Europarl. We show that the best way to sample parallel sentences is through mean word alignment coverage, where we find $k = 70000$ to roughly be the optimal number of target sentences. Of the different weighting schemes in voting, we select the product of word alignment probability and NER tagger confidence as best. We visualize these experiments in Figure 2. Table 2 shows stable performance on Europarl across the languages, with mean F_1 at 60.7 for $n = 3$ and only +1.53 higher for $n_{\max}$ which is in fact lower than 3.

Moving to Watchtower. Table 2 shows that the performance plunges across languages when Watchtower religious text replaces Europarl, with a mean F_1 of 16.3. There, the gap between $n = 3$ and mean $n_{\max} = 4.82$ is much larger: Watchtower needs more sources, and even then the benefits are low, as the +4.82 increase gets us to an infeasible mean F_1 of 21.12. In target sentence selection we find $k = 20000$ to be roughly optimal for Watchtower, but we also observe very little change in F_1 when moving to its full size of around 120 thousand target sentences.

		Europarl			Watchtower		
Sources	sup.	$F_1^{n=3}$	$F_1^{n_{\max}}$	$n_{\max}$	$F_1^{n=3}$	$F_1^{n_{\max}}$	$n_{\max}$
Arabic	78.21	–	–	–	05.50	09.84	5
Dutch	82.26	63.37	63.79	3	12.80	22.02	6
English	91.03	59.96	60.13	2	18.23	21.83	6
Estonian	85.77	63.20	63.82	3	13.14	21.63	7
French	67.98	50.10	50.10	4	10.24	14.12	2
German	80.82	61.44	62.81	2	06.26	09.62	6
Hindi	67.15	–	–	–	00.00	00.00	1
Hungarian	94.13	58.84	61.11	5	39.85	39.85	4
Italian	80.63	64.71	65.20	3	18.30	25.94	6
Spanish	82.91	63.26	65.67	3	21.02	31.36	7
Targets							
Czech	–	63.38	69.90	1	20.52	21.98	7
Polish	–	71.00	71.86	3	32.42	32.42	4
Portuguese	–	59.38	59.38	4	20.99	29.59	7
Slovak	–	64.98	64.98	4	–	–	
Slovene	–	66.63	67.86	1	30.14	35.11	6
Swedish	–	39.48	44.54	1	02.38	13.02	5
Tamil	–	–	–	–	09.04	09.64	3
Means	81.09	60.70	62.23	2.29	16.30	21.12	4.82

Table 2: F_1 scores for NER tagging in the experiment languages, shown separately for Europarl and Watchtower, also for fixed number of source languages $n = 3$ and optimal $n_{\max}$. Full supervision scores are reported for the source languages. All scores are given for 3-class IO encoding.

To put the Watchtower results into perspective, we implement another simple baseline. Namely, we train a new monolingual English NER system, but instead of using monolingual fastText embeddings, we create simple cross-lingual embeddings following Søgaard et al. (2015) over Europarl for Dutch, German, and Spanish. In effect, the change to cross-lingual embeddings yields a multilingual tagger for these four languages. The respective F_1 scores of this tagger are low (27-28%), but they still surpass Watchtower projection.

5 Discussion

We further depict the breakdown of Watchtower projection in two figures. Figure 3 shows precision, recall, and F_1 learning curves for the best projection setup on both parallel corpora. For Europarl, adding more sources always increases recall at the cost of precision: new weaker

[2]ISO 639-1 language codes were used: https://www.iso.org/iso-639-language-codes.html.

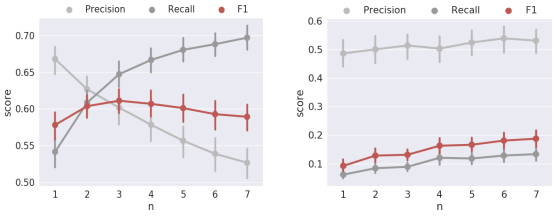
sources increase the noise, but also improve coverage. For Watchtower, precision slightly increases with more sources, but the recall stays very low throughout, at around 5-12%. The distribution of labels in the source sides of the two parallel corpora (see Figure 4) clarifies the learning curves issue of Watchtower. Namely, for both corpora the optimal word alignment coverage cutoff for selecting target sentences is around 80% covered words (best $k = 70000$ for Europarl, while $k = 20000$ for Watchtower). However, these cutoffs result in Europarl projections with nearly two orders of magnitude more named entities than in Watchtower (LOC: 65 times more, ORG: 60, PER: 15), and with different distributions.

To summarize, our results show that there exists a setup in which standalone annotation projection from multiple sources does work for cross-lingual NER. Europarl is an instance of such setup, with its large data volume per language, high-quality preprocessing, and domain rich in named entities. Arguably, there are no parallel corpora of such volume and quality that cover a multitude of true low-resource languages, and we have to do with more limited resources such as Watchtower. In turn, our experiment shows that in such setup standalone projection yields infeasible NE taggers, while it still may yield workable POS taggers or dependency parsers (cf. Agić et al. 2016).

Alternatives. In search for feasible alternatives, we conducted a proof-of-concept replication of the work by Mayhew et al. (2017), who rely on "cheap translation" of training data from multiple sources using bilingual lexicons. The replication involved only one language, Dutch, and we limited the time investment in the effort. We used three translation sources: German, English, and Spanish. Together with instance selection through alignment coverage, we reach a top F_1 score of 69.35 (with 3-class IO encoding), which surpasses even our best Europarl projection for Dutch by 4.56 points.

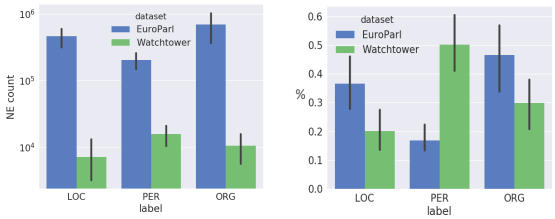
6 Related work

There is ample work in cross-lingual NER that exploits cross-lingual representations, comparable or parallel corpora together with entity dictionaries, translation, and the like (Täckström et al., 2012; Kim et al., 2012; Wang et al., 2013; Nothman et al., 2013; Tsai et al., 2016; Ni and Florian, 2016; Ni et al., 2017). We highlight a set of contributions that boast a larger cross-linguistic breadth.



(a) Europarl (b) Watchtower

Figure 3: Cross-lingual NER learning curves for precision, recall, and F_1 in relation to the number n of source languages in projection. Means for all experiment languages.



(a) Absolute (b) Relative

Figure 4: Absolute and relative counts for NE labels in Europarl and Watchtower for overlapping source languages.

Al-Rfou et al. (2015) work with 40 languages where NE annotations are derived from Wikipedia and Freebase, while they use a mix of human-annotated and machine-translated data for evaluation. Similarly, Pan et al. (2017) build and evaluate Wikipedia-based models for 282 languages; out of those, 20 are evaluated for NE linking and 9 for NER on human annotations that are not from Wikipedia. Cotterell and Duh (2017) jointly predict NE for high- and low-resource languages with a character-level neural CRF model. Their evaluation involves 15 diverse languages across 5 language families. The DARPA LORELEI program (Christianson et al., 2018) features challenges in low-resource NER development for "surprise" languages under time constraints.

7 Conclusions

Our work addresses an important gap in cross-lingual NER research. In an experiment with 17 languages, we show that while standalone multi-source annotation projection for NER can work when resources are rich in both quality and quantity, it is infeasible at a larger scale due to parallel corpora constraints. For NER in true low-resource languages, our results suggest it is better to choose an alternative approach.

References

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the bible: Learning pos taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272, Beijing, China. Association for Computational Linguistics.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Alonso Martínez, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

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*, pages 586–594. SIAM.

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007. Anersys: An Arabic named entity recognition system based on maximum entropy. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 143–153. Springer.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Caitlin Christianson, Jason Duncan, and Boyan Onyshkevych. 2018. Overview of the darpa lorelei program. *Machine Translation*, 32(1):3–9.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96. Asian Federation of Natural Language Processing.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609, Portland, Oregon, USA. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 694–702, Jeju Island, Korea. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, pages 79–86.

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*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Xuezhe Ma and Fei Xia. 2014. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1337–1348, Baltimore, Maryland. Association for Computational Linguistics.

Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. I-CAB: The italian content annotation bank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 963–968.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark. Association for Computational Linguistics.

Clemens Neudecker. 2016. An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada. Association for Computational Linguistics.

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, pages 1275–1284, Austin, Texas. Association for Computational Linguistics.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Robert Östling. 2013. Stagger: An open-source part of speech tagger for swedish. *Northern European Journal of Language Technology*, 3:1–18.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958. Association for Computational Linguistics.

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 2017. The first cross-lingual challenge on recognition, normalization, and matching of named entities in slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 76–85, Valencia, Spain. Association for Computational Linguistics.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Pattabhi RK Rao and Sobha Lalitha Devi. 2013. NER-IL: Named entity recognition for Indian languages, 1st edition @ FIRE 2013 - an overview. In *Forum for Information Retrieval and Evaluation*.

Mohammad Sadegh 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*, pages 328–338, Lisbon, Portugal. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Sögaard. 2017. A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902*.

Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. 2006. Harem: An advanced NER evaluation contest for Portuguese. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual nlp. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1713–1722, Beijing, China. Association for Computational Linguistics.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Oscar Täckström, Ryan McDonald, and 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*, pages 477–487, Montréal, Canada. Association for Computational Linguistics.

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*.

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*, pages 142–147.

Alexander Tkachenko, Timo Petmanson, and Sven Laur. 2013. Named entity recognition in estonian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 78–83, Sofia, Bulgaria. Association for Computational Linguistics.

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*, pages 219–228.

Mengqiu Wang, Wanxiang Che, and Christopher D. Manning. 2013. Joint word alignment and bilingual named entity recognition using dual decomposition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1082, Sofia, Bulgaria. Association for Computational Linguistics.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. Ontonotes: A large training corpus for enhanced processing. *Handbook of Natural Language Processing and Machine Translation*.