Language-Independent Named Entity Analysis Using Parallel Projection and Rule-Based Disambiguation

James Mayfield and **Paul McNamee** and **Cash Costello**
Johns Hopkins University Applied Physics Laboratory
{james.mayfield, paul.mcnamee, cash.costello}@jhuapl.edu

Abstract

The 2017 shared task at the Balto-Slavic NLP workshop requires identifying coarse-grained named entities in seven languages, identifying each entity's base form, and clustering name mentions across the multilingual set of documents. The fact that no training data is provided to systems for building supervised classifiers further adds to the complexity. To complete the task we first use publicly available parallel texts to project named entity recognition capability from English to each evaluation language. We ignore entirely the subtask of identifying non-inflected forms of names. Finally, we create cross-document entity identifiers by clustering named mentions using a procedure-based approach.

1 Introduction

The LITESABER project at Johns Hopkins University Applied Physics Laboratory is investigating techniques to perform analysis of named entities in low-resource languages. The tasks we are investigating include: named entity detection and coarse type classification, commonly referred to as named entity recognition (NER); linking of named entities to online databases such as Wikipedia; and clustering of entities across documents. We have applied some of our techniques to the BSNLP 2017 Shared Task. Specifically, we submitted results in two of the three categories: Named Entity Mention Detection and Classification (or NER), which asks systems to locate mentions of named entities in text and identify their types; and Entity Matching (also known as *cross-lingual identification*, or *cross-document coreference resolution*) which asks systems to determine when two entity mentions, either in the same document or in different documents, refer to the same real-world entity. We did not participate in the Name Normalization task, which asks systems to convert each entity mention to its lemmatized form. This paper describes our approach and results.

2 Approach to NER

Our approach to developing named entity recognizers for Balto-Slavic languages takes the following steps:

- Obtain parallel texts for the target language and English.
- Apply an English-language named entity recognizer to the English side of the corpus.
- Project the resulting annotations from English over to the target language by aligning tagged English words to their target language equivalents.
- Train a target language tagger off of the inferred named entity labels.

These steps are described further in the following subsections.

2.1 Parallel Collections

Exploitation of a parallel collection is at the heart of our method. English is a well-studied, high-resource language for which annotated NER corpora are available, therefore we used parallel collections with English on one side and the target Balto-Slavic language on the other.

Our parallel bitext comes from the OPUS archive[1] maintained by Tiedemann (2012). Over one million parallel sentences were available for six of the seven languages; Ukrainian was our least resourced language. Principal sources included Europarl (Koehn, 2005) and Open Subtitles. We

[1] http://opus.lingfil.uu.se

randomly sampled 250,000 sentences for each language, and after filtering for various quality issues we arrived at the data described in Table 1.

Language	Training # words	Test # words
Croatian	632,915	43,593
Czech	1,028,778	45,659
Polish	843,632	45,362
Russian	560,296	44,801
Slovak	1,081,397	45,611
Slovenian	966,431	45,444
Ukrainian	601,539	43,556

Table 1: Parallel collection sizes, in words.

2.2 English NER

Our first step was to identify the named entities on the English side of the parallel collections. There are many well-developed approaches to NER in English.[2] We chose to use the Illinois Named Entity Tagger from the Cognitive Computation Group at UIUC (Ratinov and Roth, 2009), which at the time of its publication had the highest reported NER score on the 2003 CoNLL English shared task (Tjong Kim Sang and De Meulder, 2003). It is a perceptron-based tagger that can take into consideration non-local features and external data sources.

2.3 Parallel Projection

Once we have tagged an English document we need to map those tags onto words in the corresponding target language document. Yarowsky *et al.* pioneered this style of parallel projection (2001), using it to induce part of speech taggers and noun phrase bracketers in addition to named entity recognizers. We use the Giza++ tool (Och and Ney, 2003) to align words in our parallel corpora. In most cases, a single English word will align with a single target language word. In these cases, the tag assigned to the English word is also assigned to the aligned target language word. In some cases, the alignment will be one-to-many, many-to-one, or many-to-many. For one-to-many alignments, the tag of the English word is applied to all of the aligned target language words. For many-to-one and many-to-many alignments, if any English word is tagged with an entity tag, then all aligned target language words are tagged with the first such tag. Because Balto-Slavic languages are more heavily inflected than English, most alignments from English are one-to-one or many-to-one. In Czech, for example, our parallel collection produced 71M one-to-one and many-to-one alignments, but only 13M one-to-many alignments. We believe this favors the above heuristics for the BSNLP 2017 task, because one-to-many alignments are likely to be due to inflections in the Balto-Slavic language that encode English function words.

2.4 Supervised Tagging and Classification

Projection of named entity tags onto the Balto-Slavic side of the parallel collection gives us a training collection for a supervised NER system. Because we are training many recognizers, we prefer to rely on language-independent techniques. Features that work well for one language (*e.g.,* capitalization) will not necessarily work well for another. Thus, we prefer an NER system that can consider many different features, selecting those that work well for a particular language without overtraining. To this end, we use the *SVMLattice* named entity recognizer (Mayfield et al., 2003). *SVMLattice* uses support vector machines (SVMs) at its core. Like other discriminatively trained systems, support vector machines can handle large numbers of features without overtraining. *SVMLattice* trains a separate SVM for each possible transition from label to label. It then uses Viterbi decoding to identify the best path through the lattice of transitions for a given input sentence.

We did not include gazetteers as features, though their use has been shown to be beneficial in statistically trained NER systems. But we intend to investigate their use in future research.

3 Cross-Document Entity Coreference Resolution

We used the *Kripke* system (Mayfield et al., 2014) to identify co-referential mentions of the same named entity across the multilingual document collection. *Kripke* is an unsupervised agglomerative clusterer that produces equivalence sets of entities using a combination of procedural rules. We used the *uroman* transliterator[3] to convert Cyrillic names to the Roman alphabet to support cross-script clustering.

[2] See (Nadeau and Sekine, 2007) for a survey of approaches.

[3] http://www.isi.edu/projects/nlg/software_1

To avoid the customary quadratic-time complexity required for brute-force pairwise comparisons, *Kripke* maintains an inverted index of names used for each entity. Only entities matching by full name, or some shared words or character n-grams are considered as potentially coreferential. Related indexing techniques are variously known as blocking (Whang et al., 2009) or canopies (McCallum et al., 2000).

Approximate name matching is accomplished using techniques such as: Dice scores of padded character tri-grams, recursive longest common subsequence, and expanding abbreviations. Christen (2006) gives a nice survey of related methods.

Contextual matching is accomplished by comparing named entities that co-occur in the same document. Between candidate clusters, the intersection of names occurring in the clusters is computed. Names are weighted by normalized Inverse Document Frequency, so that rarer (*i.e.,* discriminating) names have greater weights. The top-k (*i.e.,* k=10) highest weighted names in common are examined, and if the sum of their weights exceeds a cutoff, then the contextual similarity is deemed adequate.

A series of five clustering passes was performed. In early iterations matching criteria are strict, and merges have both good name string and context matching. This builds high-precision clusters in the beginning, using relaxed conditions in successive rounds to elevate entity recall.

For the BSNLP shared task the documents in the evaluation corpora are based on a focal entity. As a result the same name string found in different documents almost surely refers to the same entity. *Kripke* was designed for more diverse corpora, where this is less often the case.

4 NER Experiments

We had no collections with ground truth for six of the seven BSNLP languages. To gauge performance, we divided the induced label collection (*i.e.,* the Balto-Slavic side of the parallel collection) into training and test sets (Table 1). We then built an *SVMLattice* tagger using the training set, and applied it to the test set, assuming that the projected tags were entirely accurate. The results are shown in Table 2.

Digging slightly deeper into these results (Table 3), we see that in general, performance is highest on locations, and lowest for the miscellaneous class. The organization class is inconsistent, being high in some languages and low in others.

	Precision	Recall	F_1
Croatian	70.75	53.44	60.89
Czech	74.89	61.43	67.49
Polish	75.68	60.07	66.98
Russian	68.19	36.94	47.92
Slovak	76.97	63.30	69.47
Slovenian	78.44	61.03	68.65
Ukrainian	73.98	40.80	52.59

Table 2: NER results using projected labels.

	PER	ORG	LOC	MISC
Croatian	65.82	39.10	63.45	53.87
Czech	51.11	70.26	71.57	56.74
Polish	48.30	72.28	71.57	48.48
Russian	50.39	35.99	54.93	35.38
Slovak	61.19	70.53	75.27	58.96
Slovenian	57.50	73.00	71.75	54.26
Ukrainian	63.94	17.63	50.74	32.53

Table 3: F_1 Scores for the Four Entity Categories.

The one language for which we have some curated ground truth is Russian. The LDC collection LDC2016E95 (LORELEI Russian Representative Language Pack) contains, among other things, named entity annotations for 239 Russian documents.[4] We built a named entity recognizer for Russian using the methodology described above, and applied it to 10% of these LDC data. We used the CoNLL evaluation script to score the run. The results are shown in Table 4. Note that the label set for the LDC data is slightly different than the BSNLP label set; in particular, there is no MISC category (although the overall scores count all MISC labels as incorrect).

	Precision	Recall	F_1
Overall	52.13	22.69	31.61
PER	40.43	33.33	36.54
ORG	16.00	3.45	5.67
LOC	77.02	26.11	38.99

Table 4: Results on annotated Russian text.

We note from these results that the tagger is doing much more poorly on ORGs than is suggested by the experiments on projected labels. Thus, we

[4]We did not include the 765 annotated Tweets in our tests.

must view the results on ORGs for the other languages with a degree of skepticism. Possible reasons include wider variation in organization names than the other categories, the use of acronyms and abbreviations, or greater difficulty in aligning organization names.

5 Phase I Shared Task Results

Table 5 reports NER precision, recall, and F_1 scores for the seven languages.[5] Examining gross trends in the data, wee see that higher scores are obtained on the trump corpus. Performance is relatively consistent across language. However, recall is lower-than average in Polish and Russian, and dramatically lower for Ukrainian, particularly on the ec test set.

	trump			ec		
	P	R	F_1	P	R	F_1
ces	51.6	41.7	46.1	48.8	45.7	47.2
hrv	52.0	49.0	50.4	48.1	44.4	46.2
pol	66.8	29.7	41.1	58.1	36.6	44.9
rus	56.2	33.3	41.8	51.3	42.7	46.6
slk	56.6	40.2	47.0	47.9	44.6	46.2
slv	54.1	40.4	46.3	49.3	46.5	47.8
ukr	47.7	25.5	33.3	27.4	6.80	10.9
all	55.0	37.4	44.5	47.7	32.2	38.4

Table 5: NER results for the strict matching condition, by language.

Looking at performance by entity type (Table 6), we see best results for the PER and LOC classes, similar to our findings in Table 3 above. The ORG and MISC classes are substantially worse; scores for MISC are approximately zero.

	PER	ORG	LOC	MISC
ces	53.30	21.77	68.12	0.00
hrv	60.10	29.36	63.19	3.39
pol	35.29	13.19	68.73	0.00
rus	41.77	14.55	65.03	0.00
slk	57.52	18.67	63.20	2.94
slv	55.92	18.18	65.63	0.00
ukr	29.56	6.45	56.83	0.00
all	49.26	18.16	64.80	1.08

Table 6: F_1 scores by type and language for the trump test set with strict matching.

[5]Note, the task only permits reporting unique mentions in a document, unlike the CoNLL evaluations were every mention must be identified.

We have not had sufficient time to perform an in-depth analysis of the data. One reason for low performance on ORG and MISC classes may be that these entity mentions contain more words on average than PER and LOC entities, and our projected alignments may be less reliable for longer spanning entities. Additionally, our trained English model is based on the CoNLL dataset, and those tagging guidelines may be inconsistent with the BSNLP 2017 shared task guidelines. For example, demonyms and nationalities were tagged as MISC in CoNLL,[6] but PER in BSNLP 2017.

	trump			ec		
	P	R	F_1	P	R	F_1
ces	56.4	11.7	19.4	45.8	19.5	27.3
hrv	46.8	10.9	17.7	43.7	14.8	22.1
pol	62.4	10.7	18.2	43.9	11.0	17.5
rus	50.3	11.6	18.9	51.4	16.5	25.0
slk	58.0	14.0	22.6	46.2	22.9	30.6
slv	58.8	19.1	28.8	48.4	24.2	32.2
ukr	48.7	6.0	10.7	36.0	2.6	4.9
all	54.8	12.1	19.8	45.7	14.0	21.4

Table 7: Per-language entity coreference.

Within-language entity coreference resolution was similar across the two test sets (see Table 7). Precision was higher than recall, as we expected. Performance merging across the seven languages was lower than for single-language clustering.

6 Conclusions

Using a parallel collection to project named entity tags, and training a named entity recognizer on the resulting collection, is a feasible approach to developing named entity recognition in a variety of languages. Performance of such NER systems is clearly below that achievable with ground truth labels for training data. However, for a variety of downstream tasks, performance such as we see for the Balto-Slavic languages is acceptable.

Acknowledgment

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-16-C-0102. The views, opinions and/or findings expressed are those of the authors and should not

[6]http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt

be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

Peter Christen. 2006. A comparison of personal name matching: Techniques and practical issues. Technical Report TR-CS-06-02, Australian National University.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

James Mayfield, Paul McNamee, Christine Piatko, and Claudia Pearce. 2003. Lattice-based tagging using support vector machines. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 303–308, New York, NY, USA. ACM.

James Mayfield, Paul McNamee, Craig Harmon, Tim Finin, and Dawn Lawrie. 2014. KELVIN: Extracting Knowledge from Large Text Collections. In *AAAI Fall Symposium on Natural Language Access to Big Data*. AAAI Press, November.

Andrew McCallum, Kamal Nigam, and Lyle Ungar. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. In *Knowledge Discovery and Data Mining (KDD)*.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January. Publisher: John Benjamins Publishing Company.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1246.

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*, CoNLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, and Hector Garcia-Molina. 2009. Entity resolution with iterative blocking. In *SIGMOD 2009*, pages 219–232. ACM.

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*, HLT '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.