

---

# Named Entity Discovery from Multilingual Corpora

---

**Alexandre Klementiev**

Department of Computer Science  
University of Illinois at Urbana-Champaign  
Urbana, IL 61801  
klementi@uiuc.edu

**Dan Roth**

Department of Computer Science  
University of Illinois at Urbana-Champaign  
Urbana, IL 61801  
danr@uiuc.edu

## Abstract

Named Entity recognition (NER) is an important subtask of many natural language processing problems. Most successful approaches to NER employ machine learning techniques, which require supervised data. However, for many languages these resources are scarce.

On the other hand, comparable multilingual data (such as multilingual news streams) are becoming increasingly available and can be cheaply harvested from the Web. In this work, we make several observations about Named Entities encountered in such corpora, and use them to develop an algorithm, which learns to extract pairs of NEs across languages with almost no supervision or language specific knowledge. Specifically, given a bilingual corpus, which is weakly temporally aligned with one side annotated with Named Entities, the algorithm discovers the corresponding NEs in the second language text.

We first observe that NEs often contain or are entirely made up of words that are phonetically transliterated or have a common etymological origin across languages, and thus are phonetically similar.

Secondly, we observe that NEs in one language in such corpora tend to co-occur with their counterparts in the other. We can exploit such weak synchronicity of NEs across languages to associate them. In order to score a pair of entities, we can compute similarity of their time distributions.

Thirdly, an NE and its second language counterpart tend to appear in similar context across languages. If available, a dictionary could thus be used to score their contextual similarity.

Lastly, we expect mentions of a Named Entity to appear in documents from a particular set of topics. Thus, we can measure topic similarity of a pair of NEs by the overlap of topics of the documents in which they appear. In our setting, multilingual document clustering approaches can be used to first cluster documents in our bilingual corpus.

We present an algorithm, which iteratively exploits these observations to train a discriminative transliteration model using the temporal, contextual, and topic similarity scores as supervision signals. During discovery, the trained transliteration model is used to produce a list of highest scoring second language candidates for a given NE. If dictionary is available, the list is also augmented with translations (if any). The list is then re-ranked by a combination of the three similarity scores between the NE and each candidate. Initial experiments with temporal similarity score alone used during training show promising results on a bilingual English-Russian news corpus. They were presented at ACL in 2006.