Preliminary version of work accepted to ACL 2006.

Named Entity Discovery in Multilingual Comparable Corpora

Alexandre Klementiev **Dan Roth**
Dept. of Computer Science
University of Illinois
Urbana, IL 61801
{klementi,danr}@uiuc.edu

Abstract

Named Entity recognition (NER) is an important part of many natural language processing tasks. Current approaches often employ machine learning techniques and require supervised data. However, many languages lack such resources. This paper presents an (almost) unsupervised learning algorithm for automatic discovery of Named Entities (NEs) in a resource free language, given a bilingual corpora in which it is weakly temporally aligned with a resource rich language. NEs have similar time distributions across such corpora, and often some of the tokens in a multi-word NE are transliterated. We develop an algorithm that exploits both observations iteratively. The algorithm makes use of a new, frequency based, metric for time distributions and a resource free discriminative approach to transliteration. Seeded with a small number of transliteration pairs, our approach discovers multi-word NEs, and takes advantage of a dictionary (if one exists) to account for translated or partially translated NEs. We evaluate the algorithm on an English-Russian corpus, and show high level of NEs discovery in Russian.

1 Introduction

Named Entity recognition has been getting much attention in NLP research in recent years, since it is seen as significant component of higher level NLP tasks such as information distillation and question answering. Many successful approaches to NER employ machine learning techniques, which require supervised training data. However, for many languages, these resources do not exist. Moreover, it is often difficult to find experts in these languages both for the expensive annotation effort and even for language specific clues. On the other hand, comparable multilingual data (such as multilingual news streams) are becoming increasingly available.

In this work, we make two independent observations about Named Entities encountered in such corpora, and use them to develop an algorithm that extracts pairs of NEs across languages. Specifically, given a bilingual corpora that is weakly temporally aligned, and a capability to annotate the text in one of the languages with NEs, our algorithm identifies the corresponding NEs in the second language text, and annotates them with the appropriate type, as in the source text.

The first observation is that NEs in one language in such corpora tend to co-occur with their counterparts in the other. E.g., Figure 1 shows a histogram of the number of occurrences of the word *Hussein* and its Russian transliteration in our bilingual news corpus spanning years 2001 through late 2005. One can see several common peaks in the two histograms, largest one being around the time of the beginning of the war in Iraq. A histogram of the word *Russia*, on the other hand, appears to be different. We can exploit such weak synchronicity of NEs across languages to associate them. In order to score a pair of entities across languages, we compute the similarity of their time distributions.

The second observation is that NEs often contain or are entirely made up of words that are phonetically transliterated or have a common etymological origin across languages (e.g. *parliament* in English and *парламент*, its Russian translation), and thus are phonetically similar. Figure 2 shows

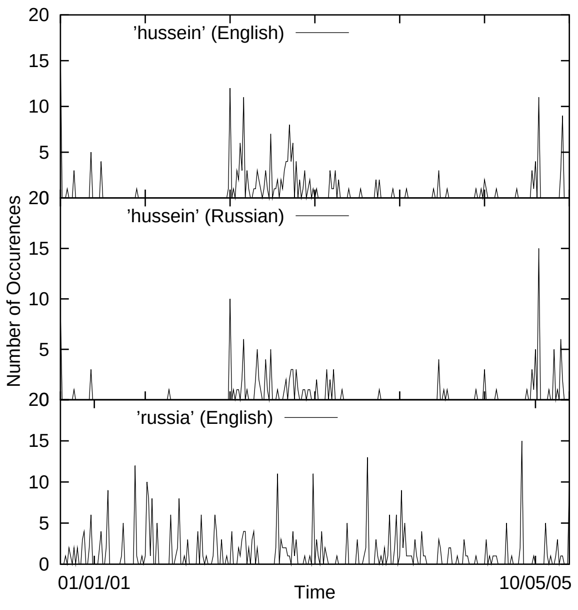


Figure 1: Temporal histograms for *Hussein* (top), its Russian transliteration (middle), and of the word *Russia* (bottom).

English NE	Russian NE
lilic	лилич
fletcher	флетчер
bradford	брэдфорд
isabel	изабель
hoffmann	гофман
kathmandu	катманду

Figure 2: Example English NEs and their transliterated Russian counterparts.

an example list of NEs and their possible Russian transliterations.

Approaches that attempt to use these two characteristics separately to identify NEs across languages would have significant shortcomings. Transliteration based approaches require a good model, typically handcrafted or trained on a clean set of transliteration pairs. On the other hand, time sequence similarity based approaches would incorrectly match words which happen to have similar time signatures (e.g., *Taliban* and *Afghanistan* in recent news).

We introduce an algorithm which exploits these observations simultaneously to match NEs on one side of the bilingual corpus to their counterparts on the other. We use a Discrete Fourier Transform (Arfken, 1985) based metric for computing similarity of time distributions, and we score NEs similarity with a linear transliteration model.

We first train a transliteration model on single-word NEs. During training, the current model chooses a transliteration candidate set in another language for a given NE. Time sequence scoring is then used to choose the candidate best temporally aligned with the NE. Pairs of NEs and the best candidates are then used to iteratively train the transliteration model.

Once the model is trained, NE discovery proceeds as follows. For a given NE, transliteration model selects a candidate set for each constituent word. If a dictionary is available, each candidate set is augmented with translations (if they exist). Translations will be the correct choice for some NE words (e.g. for *queen* in *Queen Victoria*), and transliterations for others (e.g. *Bush* in *Steven Bush*). We expect temporal sequence alignment to resolve many of such such ambiguities. It is used to select the best translation/transliteration candidate from each word's candidate set, which are then merged into a possible NE in the other language. Finally, we verify that the NE is actually contained in the target corpus.

A major challenge inherent in discovering transliterated NEs is the fact that a single entity may be represented by multiple transliteration strings. One reason is language morphology. For example, in Russian, depending on a case being used, the same noun may appear with various endings. Another reason is the lack of transliteration standards. Again, in Russian, several possible transliterations of an English entity may be acceptable, as long as they are phonetically similar to the source.

Thus, in order to rely on the time sequences we obtain, we need to be able to group variants of the same NE into an equivalence class, and collect their aggregate mention counts. We would then score time sequences of these equivalence classes. For instance, we would like to count the aggregate number of occurrences of {*Herzegovina, Hercegovina*} on the English side in order to map it accurately to the equivalence class of that NE's variants we may see on the Russian side of our corpus (e.g. {*Герцеговина, Герцеговину, Герцеговины, Герцеговиной*}). One of the objectives for this work was to use as little of the knowledge of both languages as possible. However, to effectively rely on the quality of time sequence scoring, we used a simple, knowledge free approach to group NE variants for Russian. In the rest of the paper, whenever we refer to a Named Entity, we imply an NE equivalence class. Note that better use of language specific knowledge should clearly improve the results, but this would defeat one of the goals of this work.

2 Previous work

There has been other work to automatically discover NE with minimal supervision. Both (Cucerzan and Yarowsky, 1999) and (Collins and Singer, 1999) present algorithms to obtain NEs from untagged corpora. However, they focus on the *classification* stage of *already segmented entities*, and make use of contextual and morphological clues that require knowledge of the language beyond the level we want to assume with respect to the target language.

The use of similarity of time distributions for information extraction, in general, and NE extraction, in particular, is not new. (Hetland, 2004) surveys recent methods for scoring time sequences for similarity. (Shinyama and Sekine, 2004) used the idea to discover NEs, but in a single language, English, across two news sources.

A large amount of previous work exists on transliteration models. Most are *generative* and consider the task of *producing* an appropriate transliteration for a given word, and thus require considerable knowledge of the languages. For example, (AbdulJaleel and Larkey, 2003; Jung et al., 2000) train English-Arabic and English-Korean generative transliteration models, respectively. (Knight and Graehl, 1997) build a generative model for backward transliteration from Japanese to English.

While generative models are often robust, they tend to make independence assumptions that do not hold in data. The discriminative learning framework argued for in (Roth, 1998; Roth, 1999) as an alternative to generative models is now used widely in NLP, even in the context of word alignment (Taskar et al., 2005; Moore, 2005). We make use of it here too, to learn a discriminative transliteration model that requires little knowledge of the target language.

3 An Algorithm for NE Discovery

3.1 The algorithm

In essence, the algorithm we present uses temporal alignment as a supervision signal to iteratively train a transliteration model. On each iteration, it selects a set of transliteration candidates for each NE according to the current model (line 6). It then uses temporal alignment (with thresholding) to select the best transliteration candidate for the next round of training (lines 8, and 9).

Once the training is complete, lines 4 through 10 are executed without thresholding for each constituent NE word. If a dictionary is available, transliteration candidate sets $\mathcal{NE}_\mathcal{T}$ on line 6 are augmented with translations. We then combine the best candidates (as chosen on line 8, without thresholding) into complete target language NE. Finally, we discard transliterations which do not actually appear in the target corpus.

```
   Input: Bilingual, comparable corpus ($\mathcal{S}, \mathcal{T}$), set of
          named entities $\mathcal{NE}_\mathcal{S}$ from $\mathcal{S}$, threshold $\theta$
   Output: Transliteration model $\mathcal{M}$
1  Initialize $\mathcal{M}$;
2  $\forall \mathcal{E} \in \mathcal{NE}_\mathcal{S}$, collect time distribution $\mathcal{Q}_{\mathcal{E}\mathcal{S}}$;
3  repeat
4      $\mathcal{D} \leftarrow \emptyset$;
5      for each $\mathcal{E}_\mathcal{S} \in \mathcal{NE}_\mathcal{S}$ do
6          Use $\mathcal{M}$ to collect a set of candidates $\mathcal{NE}_\mathcal{T} \in \mathcal{T}$
           with high transliteration scores;
7          $\forall \mathcal{E} \in \mathcal{NE}_\mathcal{T}$ collect time distribution $\mathcal{Q}_{\mathcal{E}\mathcal{T}}$;
8          Select candidate $\mathcal{E}_\mathcal{T} \in \mathcal{NE}_\mathcal{T}$ with the best
           $\omega = score(\mathcal{Q}_{\mathcal{E}\mathcal{S}}, \mathcal{Q}_{\mathcal{E}\mathcal{T}})$;
9          if $\omega$ exceeds $\theta$, add tuple $(\mathcal{E}_\mathcal{S}, \mathcal{E}_\mathcal{T})$ to $\mathcal{D}$ ;
10     end
11     Use $\mathcal{D}$ to train $\mathcal{M}$;
12 until D stops changing between iterations ;
```

Algorithm 1: Iterative transliteration model training.

3.2 Time sequence generation and matching

In order to generate time sequence for a word, we divide the corpus into a sequence of temporal bins, and count the number of occurrences of the word in each bin. We then normalize the sequence.

We use a method called the F-index (Hetland, 2004) to implement the $score$ similarity function on line 8 of the algorithm. We first run a Discrete Fourier Transform on a time sequence to extract its Fourier expansion coefficients. The score of a pair of time sequences is then computed as a Euclidean distance between their expansion coefficient vectors.

As we mentioned in the introduction, an NE may map to more than one transliteration in another language. Identification of the entity's equivalence class of transliterations is important for obtaining its accurate time sequence. In order to keep to our objective of requiring as little language knowledge as possible, we took a rather simplistic approach to take into account morphological ambiguities of NEs in Russian. Two words were considered variants of the same NE if they share a prefix of size five or longer. A cumulative distribution was then collected for such equivalence classes.

3.3 Transliteration model

Unlike most of the previous work considering *generative* transliteration models, we take the *discriminative* approach. We train a linear model to decide whether a word $\mathcal{E}_\mathcal{T} \in \mathcal{T}$ is a transliteration of an NE $\mathcal{E}_\mathcal{S} \in \mathcal{S}$. The words in the pair are partitioned into a set of substrings $s_\mathcal{S}$ and $s_\mathcal{T}$ up to a particular length (including the empty string _). Couplings of the substrings $(s_\mathcal{S}, s_\mathcal{T})$ from both sets produce features we use for training. Note that couplings with the empty string represent insertions/omissions.

Consider the following example: $(\mathcal{E}_\mathcal{S}, \mathcal{E}_\mathcal{T})$ = (powell, pauel). We build a feature vector from this example in the following manner:

- First, we split both words into all possible substrings of up to size two:

  $\mathcal{E}_\mathcal{S} \rightarrow \{\_, p, o, w, e, l, l, po, ow, we, el, ll\}$

  $\mathcal{E}_\mathcal{T} \rightarrow \{\_, p, a, u, e, l, pa, au, ue, el\}$

- We build a feature vector by coupling substrings from the two sets:

  $((p, \_), (p, a), ...(w, au), ...(el, el), ...(ll, el))$

We use the observation that transliteration tends to preserve phonetic sequence to limit the number of couplings. For example, we can disallow the coupling of substrings whose starting positions are too far apart. For instance, we might not consider a pairing (po, ue) in the above example.

We use the perceptron (Rosenblatt, 1958) algorithm to train the model. The model activation provides the score we use to select best transliterations on line 6. Perceptron takes variable number of features in its examples; each example is a subset of all features seen so far that are active in the input. As the iterative algorithm observes more data, it discovers and makes use of more features. This is the so-called infinite attribute model and it follows the perceptron version of SNoW (Roth, 1998).

Positive examples used for iterative training are pairs of NEs and their best temporally aligned (thresholded) transliteration candidates. Negative examples are English non-NEs paired with random Russian words.

4 Experimental Study

We ran experiments using a bilingual comparable English-Russian news corpus we obtained by crawling a Russian news web site. We collected English articles together with their loose Russian translations spanning from 1/1/2001 through 10/05/2005. The corpus consists of 2,327 documents, with 0-8 documents per day. English side was tagged with a publicly available NER system based on the SNoW learning architecture (Roth, 1998). This set of English NEs was hand-pruned to remove incorrectly classified words to obtain 978 single word NEs.

In order to reduce running time, some limited pre-processing was done on the Russian side. All classes, whose temporal distributions were close to uniform (i.e. words with a similar likelihood of occurrence throughout the corpus) were deemed common and not considered as NE candidates. Unique words were thus grouped into 14,781 equivalence classes.

Unless mentioned otherwise, the transliteration model was initialized with a set of 20 pairs of NEs with their transliteration equivalence classes initially discovered based only on high temporal distribution similarity and hand pruned. Negative examples here and during the rest of the training were pairs of randomly selected non-NE English and Russian words. In our previous experience, increasing the initial transliteration example set size past 20 did not offer significant performance improvements, although it did shorten the algorithm convergence times.

New features were discovered throughout training; all but top 3000 features from positive and 3000 from negative examples were pruned based on the number of their occurrences so far. Features remaining at the end of training were used for NE discovery.

Insertions/omissions features were not used in the experiments as they provided no tangible benefit for the languages of our corpus.

In each iteration, we used the current transliteration model to find a list of 30 best transliteration equivalence classes for each NE. We then computed time sequence similarity score between NE and each class from its list to find the one with the best matching time sequence. If its similarity score surpassed a set threshold, it was added to the list of positive examples for the next round of training. Positive examples were constructed by pairing an NE with the common stem of its transliteration equivalence class.

We used the Mueller English-Russian dictionary to obtain translations in our multi-word NE experiments. We only considered the first dictionary definition as a candidate.

For evaluation, random 727 of the total of 978 NEs were matched to correct transliterations by a language expert. Accuracy was computed as the percentage of NE occurrences correctly transliterated by the algorithm.

In the multi-word NE experiment, 282 random multi-word (2 or more) NEs and their transliterations/translations discovered by the algorithm were verified by a language expert.

4.1 NE discovery

Figure 3 shows the proportion of correctly discovered NE transliteration equivalence classes throughout the training stage. The figure also shows the accuracy if transliterations are selected according to the current transliteration model (top scoring candidate) and sequence matching alone. The transliteration model alone achieves an accuracy of about 38%, while the time sequence alone gets about 41%. The combined algorithm achieves about 63%, giving a significant improvement.

In order to understand what happens to the transliteration model as the training proceeds, let us consider the following example. Figure 4 shows parts of transliteration lists for NE *forsyth* for two iterations of the algorithm. The weak transliteration model selects the correct transliteration (italicized) as the 24th best transliteration in the first iteration. Time sequence scoring function chooses it to be one of the training examples for the next round of training of the model. By the eighth iteration, the model has improved to select it as a best transliteration.

Not all correct transliterations make it to the top of the candidates list (transliteration model by itself is never as accurate as the complete algorithm on Figure 3). That is not required, however, as the

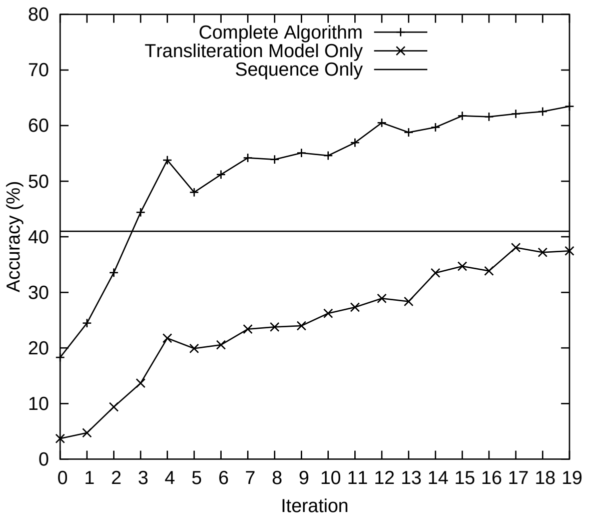


Figure 3: Proportion of correctly discovered NE pairs vs. iteration. Complete algorithm outperforms both transliteration model and temporal sequence matching when used on their own.

model only needs to be good enough to place the correct transliteration anywhere in the candidate list.

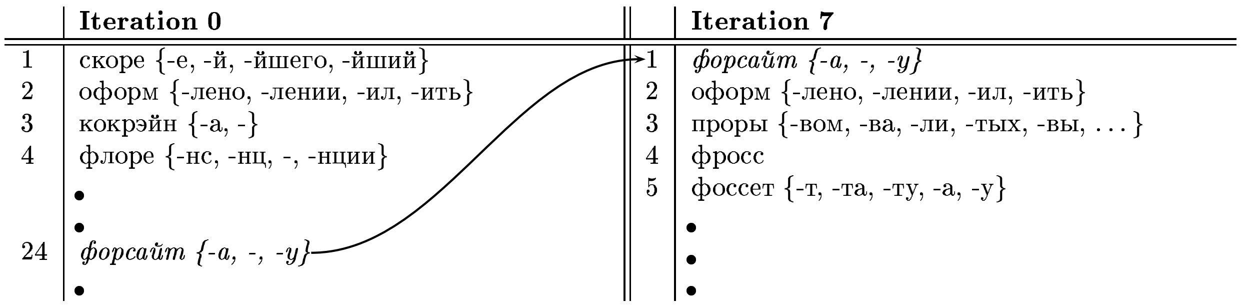
Not surprisingly, some of the top transliteration candidates start sounding like the NE itself, as training progresses. On Figure 4, candidates for *forsyth* on iteration 7 include *fross* and *fossett*.

Once the transliteration model was trained, we ran the algorithm to discover multi-word NEs, augmenting candidate sets of dictionary words with their translations as described in Section 3.1. We achieved the accuracy of about 66%. The correctly discovered Russian NEs included entirely transliterated, partially translated, and entirely translated NEs. Some of them can be seen on Figure 5.

4.2 Comparison of time sequence scoring functions

We compared the performance of the DFT-based time sequence similarity scoring function we use in this paper to the commonly used *cosine* (Salton and McGill, 1986) and *Pearson*'s correlation measures.

We perturbed the Russian side of the corpus in the following way. Articles from each day were randomly moved (with uniform probability) within a k-day window. We ran single word NE temporal sequence matching alone on the perturbed corpora using each of the three measures



	Iteration 0		Iteration 7
1	скоре {-е, -й, -йшего, -йший}	1	*форсайт {-а, -, -у}*
2	оформ {-лено, -лении, -ил, -ить}	2	оформ {-лено, -лении, -ил, -ить}
3	кокрэйн {-а, -}	3	проры {-вом, -ва, -ли, -тых, -вы, . . . }
4	флоре {-нс, -нц, -, -нции}	4	фросс
	•	5	фоссет {-т, -та, -ту, -а, -у}
	•		•
24	*форсайт {-а, -, -у}*		•
	•		•

Figure 4: Transliteration lists for *forsyth* for two iterations of the algorithm. As transliteration model improves, the correct transliteration moves up the list.

	$k = 1$	$k = 3$	$k = 5$
Cosine	41.3	5.8	1.7
Pearson	41.1	5.8	1.7
DFT	41.0	12.4	4.8

Table 1: Proportion of correctly discovered NEs vs. corpus misalignment (k) for each of the three measures. DFT baset measure provides significant advantages over commonly used metrics for weakly aligned corpora.

	$w = 1$	$w = 2$	$w = 3$
Cosine	5.8	13.5	18.4
Pearson	5.8	13.5	18.2
DFT	12.4	20.6	27.9

Table 2: Proportion of correctly discovered NEs vs. sliding window size (w) for each of the three measures.

(Table 1).

Some accuracy drop due to misalignment could be accommodated for by using a larger temporal bin for collecting occurrence counts. We tried various (sliding) window size w for a perturbed corpus with $k = 3$ (Table 2).

DFT metric outperforms the other measures significantly in most cases. NEs tend to have distributions with few pronounced peaks. If two such distributions are not well aligned, we expect both Pearson and Cosine measures to produce low scores, whereas the DFT metric should catch their similarities in the frequency domain.

5 Conclusions

We have proposed a novel algorithm for cross lingual NE discovery in a bilingual weakly temporally aligned corpus. We have demonstrated that using two independent sources of information (transliteration and temporal similarity) together to guide NE extraction gives better performance than using either of them alone (see Figure 3).

We developed a linear discriminative transliteration model, and presented a method to automatically generate features. For time sequence matching, we used a scoring metric novel in this domain. We provided experimental evidence that this metric outperforms other scoring metrics traditionally used.

In keeping with our objective to provide as little language knowledge as possible, we introduced a simplistic approach to identifying transliteration equivalence classes, which sometimes produced erroneous groupings (e.g. an equivalence class for NE *congolese* in Russian included both *congo* and *congolese* on Figure 5). We expect that more language specific knowledge used to discover accurate equivalence classes would result in performance improvements.

Other type of supervision was in the form of a very small bootstrapping transliteration set. The intuition for not needing more initian examples is the following: the few examples in the initial training set produce features corresponding to substring pairs characteristic for English-Russian transliterations. Model trained on these (few) examples chooses other transliterations containing these same substring pairs. In turn, the chosen positive examples contain other characteristic substring pairs, which will be used by the model to select more positive examples on the next round, and so on.

6 Future Work

The algorithm can be naturally extended to comparable corpora of more than two languages. Pair-wise time sequence scoring and translitera-

Eng. NE	Rus. NE equiv. class
carla del ponte	карла{-, -йл} дель понте
marc dutroux	марк дютру
pangbourne	пангбурн
supreme council	верхо{-вный} совет{...}
congolese	конго{-, -лезской}
north carolina	север{...} карол{-ина,...}
junichiro koizumi	дзюнитиро коидзуми
rehman	реман{-, -а}

Figure 5: Example of correct transliterations discovered by the algorithm.

tion models should give better confidence in NE matches.

The ultimate goal of this work is to automatically tag NEs so that they can be used for training of an NER system for a new language. To this end, we would like to compare the performance of an NER system trained on a corpus tagged using this approach to one trained on a hand-tagged corpus.

7 Acknowledgments

We thank Richard Sproat, ChengXiang Zhai, and Kevin Small for their useful feedback during this work, and the anonymous referees for their helpful comments. This research is supported by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program and a DOI grant under the Reflex program.

References

Nasreen AbdulJaleel and Leah S. Larkey. 2003. Statistical transliteration for english-arabic cross language information retrieval. In *Proceedings of CIKM*, pages 139–146, New York, NY, USA.

George Arfken. 1985. *Mathematical Methods for Physicists*. Academic Press.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classifi cation. In *Proc. of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*.

Silviu Cucerzan and David Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence. In *Proc. of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*.

Magnus Lie Hetland, 2004. *Data Mining in Time Series Databases*, chapter A Survey of Recent Methods for Effi cient Retrieval of Similar Time Sequences. World Scientifi c.

Sung Young Jung, SungLim Hong, and Eunok Paek. 2000. An english to korean transliteration model of extended markov window. In *Proc. the International Conference on Computational Linguistics (COLING)*, pages 383–389.

Kevin Knight and Jonathan Graehl. 1997. Machine transliteration. In *Proc. of the Meeting of the European Association of Computational Linguistics*, pages 128–135.

Robert C. Moore. 2005. A discriminative framework for bilingual word alignment. In *Proc. of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 81–88.

Frank Rosenblatt. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65.

Dan Roth. 1998. Learning to resolve natural language ambiguities: A unifi ed approach. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 806–813.

Dan Roth. 1999. Learning in natural language. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 898–904.

Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Yusuke Shinyama and Satoshi Sekine. 2004. Named entity discovery using comparable news articles. In *Proc. the International Conference on Computational Linguistics (COLING)*, pages 848–853.

Ben Taskar, Simon Lacoste-Julien, and Michael Jordan. 2005. Structured prediction via the extragradient method. In *The Conference on Advances in Neural Information Processing Systems (NIPS)*. MIT Press.