Toward a Name Entity Aligned Bilingual Corpus

Xiaoyi Ma
Linguistic Data Consortium
3600 Market St. Suite 810
Philadelphia, PA 19104
E-mail: xma@ldc.upenn.edu

Abstract

This paper describes a co-training framework in which, through named entity aligned bilingual text, named entity taggers can complement and improve each other via an iterative process. This co-training approach allows us to 1) apply our method to not only parallel but also comparable text, greatly extending the applicability of the approach; and to 2) adapt named entity taggers to new domains; 3) create a named entity aligned bilingual corpus. Experiment results on Chinese-English data are shown and discussed.

1. Introduction

Named entity aligned bilingual corpora are valuable resources for many NLP applications, including machine translation, cross-lingual information retrieval. Manually annotating such corpora is extremely expensive, time consuming, and it cannot be scaled up easily, which makes automatic creation of these corpora a very attractive approach, given the amount of bilingual text that becomes available everyday.

Automatic bilingual named entity alignment usually involves two steps: 1) identification of names in both halves of the bitext; 2) alignment of names across two languages.

Automatic bilingual named entity alignment faces a couple of difficulties. First and foremost, current state-of-art named entity taggers don't adapt well to new domain and time epochs. Rule-based (Grishman, 1995) and statistical named entity tagging methods, such as hidden markov models (Bikel et al., 1999), maximum entropy models (Borthwick, 1999), and conditional random fields (Li and McCallum, 2003), performs well in the targeting domain, but there performance decreases significantly on data from other domains or time epochs.

Secondly, alignment of names across languages can be tricky due to a number of reasons: 1) name translation and transliteration variations; 2) named entities can be ambiguous, that is, the same "name" can refer to different entities.

This paper is part of our ongoing research on named entity alignment on unlabeled bilingual text, which has the following major goals:

a) improve current state-of-the-art taggers;
b) adapt existing taggers to new domains;
c) automatic alignment of named entities in bilingual texts;

This paper focuses on improving named entity taggers using unlabeled bilingual text and to adapt these taggers to new domains within a co-training framework. Work on aligning the named entities is yet to be completed. However, preliminary experiment result demonstrates that the taggers have good coverage and accuracy on the bitext we're about to conduct entity alignment on, which lays a solid ground for future work on named entity alignment.

This paper is laid out as follows. Section 2 provides the background for this paper. Section 3 describes the entity alignment-based co-training algorithm for enhancing NE taggers, as well as the general approach of entity alignment in parallel and comparable text. Section 4 describes the experiments done on English-Chinese parallel text and comparable text. Section 5 shows the experiment results. Section 6 concludes this paper.

2. Previous Works

Previous works on inducing or enhancing text analysis tools using bilingual text include (Yarowsky et al., 2001) and (Hwa et al., 2005).

(Yarowsky et al., 2001) describes a set of algorithms for automatically inducing text analysis tools – POS taggers, base noun-phrase bracketers, named entity taggers, and morphological analyzers – for an arbitrary foreign language from English, using aligned parallel text corpora. Parallel text corpora were first word/character aligned using the EGYPT system (Al-Onaizan et al., 1999). The English side of the corpus is tagged or bracketed using the state-of-the-art taggers or bracketers, and the English tags/brackets are then projected to the foreign language. Since the direct annotation projection is noisy, the paper presents training procedures and algorithms to bootstrap taggers from noisy and incomplete initial projection.

To induce named entity taggers from aligned parallel text corpora, (Yarowsky et al., 2001) did the initial classification on a per-word basis, using an aggressively smoothed transitive projection model. The co-training-based algorithm given in (Cucerzan and Yarowsky, 1999) was then used to train a named entity tagger from the projected data. To evaluate the performance of the induction algorithm on named entities, (Yarowsky et al., 2001) used the Canadian Hansard corpus with about 2.8M sentence pairs, the English side was first tagged by a tagger trained on MUC-6 training data, then the tags were projected to the French side and the projected data were used to train a French named entity tagger. The named entity tagger achieved 85% classification accuracy measured in terms of per-word entity-type classification accuracy on 4 entity types: FNAME, LNAME, PLACE, and OTHER. The paper claims the induced French tagger is near perfect since the original English tagger achieved only 86% accuracy.

(Hwa et al., 2005) adopted a similar approach to bootstrap non-English syntactic parsers from English by using a state-of-the-art English parser and parallel text. The English side of the parallel text is first analyzed using the state-of-the-art parser, the parse trees are then converted to dependency structures, which are projected across the word alignment to the non-English side using a direct project algorithm. To address the structural differences between English and non-English languages, (Hwa et al., 2005) apply a small set of manually compiled, language-specific post-projection transformation rules on the projected trees. Finally, (Hwa et al., 2005) uses aggressive filtering strategy to automatically prune out projected trees that are believed to be of poor quality. The resulting trees are then used to train a new dependency parser.

Co-training (Blum and Mitchell, 1998) assumes features can be partitioned into two different sets to represent different views of the same data, and in addition, it assumes each view by itself would be sufficient for learning if there were enough labeled data. Initially two separate classifiers are trained with labeled data. Each classifier was then used to classify the unlabeled data and each classifier's prediction on the unlabeled data is used to augment the training set of the other. Each classifier is retrained with the additional training data provided by the other classifier, and the process repeats.

(Blum and Mitchell, 1998) applied the co-training algorithm to web page classifiers which are trained to identify course web pages from a set of web pages collected from Computer Science department websites at four universities. Three naive Bayes based classifiers were trained on the labeled data, one page based, one hyperlink based, and the third page-hyperlink combined. Experiment results show the co-training algorithm improves all three classifiers significantly, and in the case of combined classifier, the co-training algorithm was able to reduce the error rate by more than 50%.

3. The Co-training Algorithm

Named entity tagging in the context of bilingual text fits the co-training framework nicely. Bilingual texts of the same content (news event, biomedical paper, etc) are naturally two views of the same data. Each view is sufficient for learning of named entity tagging, given enough labeled data.

Figure 1 illustrates the co-training algorithm, which utilizes parallel text to improve NE taggers, using English and Chinese as an example.

In essence, the algorithm iteratively selects new training instances from unlabeled text to augment labeled training data. During the initialization stage, both sides of the parallel text are labeled by the baseline taggers trained on labeled training data, $E_{labeled}$ and $C_{labeled}$ (lines 2 to 5). On each iteration, using labeled English data for supervision, the algorithm selects Chinese data that the current Chinese NE tagger fails to label correctly, and these data (with their labels corrected) are used to augment the training data for Chinese (line 13). The augmented training data is used to train a new and better Chinese named entity tagger (line 14). The new Chinese tagger is then used to re-tag the Chinese text (line 16). Using the newly tagged Chinese text for supervision, English training data is augmented (line 18) and used to train a new English tagger (line 19). And the process repeats for *N* iterations.

```
1  Initialization
2      train English NER model $Etagger_{baseline}$ on $E_{labeled}$
3      train Chinese NER model $Ctagger_{baseline}$ on $C_{labeled}$
4      $ETagged_{baseline} \leftarrow$ tag English side of the parallel
       text using $Etagger_{baseline}$
5      $CTagged_{baseline} \leftarrow$ tag Chinese side of the
       parallel text using $Ctagger_{baseline}$
6      $ETagged_{latest} = ETagged_{baseline}$
7      $CTagged_{latest} = CTagged_{baseline}$
8
9  For i in 1 to N
10     $Ctrain_{add} \leftarrow \phi$
11     $Etrain_{add} \leftarrow \phi$
12
13     $CTrain_{add} \leftarrow augE2C(ETagged_{latest}, CTagged_{baseline})$
14     train Chinese NER model $Ctagger_i$ on
       combine($C_{labeled}$, $CTrain_{add}$)
15
16     $CTagged_{latest} \leftarrow$ tag Chinese side of the parallel
       text using $Ctagger_i$
17
18     $ETrain_{add} \leftarrow augC2E(ETagged_{baseline}, CTagged_{latest})$
19     train English NER model $Etagger_i$ on
       combine($E_{labeled}$, $ETrain_{add}$)
20
21     $ETagged_{latest} \leftarrow$ tag English side of the parallel
       text using $Etagger_i$
22 done
```

Figure 1 Co-training algorithm for English and Chinese named entity taggers

Given both sides of the parallel text with automatic labels, functions augE2C and augC2E augment Chinese and English training data respectively by projecting tags from English to Chinese and Chinese to English.

3.1. Filtering Noises

Noises may be added to the training data and propagates, leading to the deterioration of the NE taggers' performance. The noises come from two sources: 1) incorrectly labeled tokens on both sides of parallel text; 2) the name projection process, which can project correctly labeled tokens incorrectly across languages.

The noise filtering approaches we adopted include local and global validation, and orthography-based filtering. Local and global validation validates strings that are identified as names. Orthography-based filtering makes sure all names in a sentence have been identified.

Local and global validation – For any given name label pair (*n, t*) where *n* is a name and *t* is the type of the name, local and global validation seeks supporting evidence that *t* is the correct label for *n*. If (*n, t*) fails both local and global validation, the word label pair would be deemed unreliable, it wouldn't be used for tag projection, and sentences containing the word would be disqualified as new training examples.

Local validation of (*n, t*) passes if there is at least another instance of string *n* within the same document and if all instances of name *n* bear the same label *t*. Local validation is based on the hypothesis that within a document the name type of the same name should be highly consistent.

If a name fails local validation (either because there aren't other instances in the same document, or the instances of the name aren't labeled consistently), global validation would decide if the name and type are valid. Global validation considers how a name is labeled in the entire corpus. If the label consistency of a name exceeds a preset threshold (85% in all experiments in this paper), the name and type would be considered valid.

If we look at each iteration of the co-training algorithm as a two-step process, in the first step to project names from one language to another, and in the second step to select sentences to augment the training data, then the name validation can be applied to both steps. Before a name is projected to the other language, the name has to be validated either locally or globally so that we're fairly confident with the label. Also, before a sentence is added to the training data, we validate all the names in the sentence. A sentence would be disqualified as new training data if any name in the sentence fails both local validation and global validation.

Orthography-based filtering – local and global validation filters out names that are identified but incorrectly labeled (for example, *George Bush* labeled as an organization). Another type of noise is those names that aren't identified at all (those labeled as *O* in BIO scheme). Sentences containing unidentified names should not be used as new training examples. In languages that exhibit orthographic differences between names and non-names, such as English, exploring the orthographic differences can effectively filter out sentences containing unidentified names. For example, in case of English, person names, location names and organization names are written with the initial letter of each word capitalized, while most non-names are not. So an aggressive and simple heuristic for filtering out sentences with unidentified names is to discard all sentences containing words (except the first word in a sentence) with the initial letter capitalized but not identified as a name.

We compiled a capital_non_name list from ACE 2007 English data. The list consists non-names that are usually written with the first letter capitalized, including job titles, days of a week, months of a year, and names of other types, for example books, movies, drugs.

The orthography-based filter disqualifies English sentences containing word(s) that satisfy all three conditions as follows:

1) the initial letter of the word is capitalized (except when the word is the initial word of the sentence);
2) the word is labeled as a non-name;
3) the word is not on the capital_non_name list;

This procedure inevitably filters out some good sentences as well, which isn't a big concern to us because we have large quantities of unlabeled data.

Orthography-based filtering cannot be applied to languages such as Chinese and Arabic, which don't distinguish names from non-names orthographically. It is still effective if the language pair involves one language that does have orthographical differences between names and non-names.

3.2. Maintaining Data Distribution

A caveat of applying statistical semi-supervised methods, co-training included, is that the new training data extracted from unlabeled data should conform to the underlying data distribution, otherwise the additional training data may skew the statistics and end up hurting the retrained classifier. We choose to use the data distribution in the manually labeled data as the underlying data distribution. The new training data is selected in a way so that it matches the ratio of person names, location names, and organization names in the labeled data.

3.3. Weighting training data

The new training examples extracted from parallel text will undoubtedly contain incorrectly labeled tokens. Naturally the manually labeled data and the extracted sentences should be weighted differently to favor manually labeled data. Between count merging and creating multiple models and calculating weights for each model (model interpolation), (Bacchiani et al., 2006) shows that count merging is more effective, which is what we employed in our system. We implement count merging by concatenating the training sets, possibly with multiple copies of each to account for weighting.

3.4. Entity Alignment for Chinese-English Bilingual Text

For the co-training algorithm to work, names need to be aligned correctly across bilingual text. If the text is parallel text, alignment can be acquired via automatic word alignment of the parallel text, which is a topic well studied in the context of Machine Translation. The problem with the word alignment approach is two fold. First, it only works on parallel text. Second, it requires large quantities of parallel text to work well.

To achieve named entity alignment, we probably don't need a full-fledged word alignment. There are certain properties of named entity translation that we can take advantage of to achieve high accuracy without aligning every word in the sentence. One observation is that names and non-names are translated differently: names are usually transliterated – with the exception of organization names – while non-names are mostly translated. In addition, large percentage of transliterated names is proper names in the target language that don't overlap with other word categories. These two properties are very effective and quite enough to remove most false positives, as shown in the experiments described in the following sections.

These properties can be explored to align named entities in bilingual document pairs. For parallel text, these pairs are the source text and the translation. In case of comparable text, we use a lexicon based content matching tool to identify document pairs that have similar content.

We employ four approaches to align names, in order of accuracy:

1) Pinyin mapping – a deterministic process to transliterate Chinese into English;
2) Dictionary lookup – looking up possible translation/transliterations from existing bilingual name lists;
3) Transliteration model – use transliteration model trained on transliterated Chinese English name pairs to generate and search for possible transliterations of a name. Models were trained using Moses (Koehn et al., 2007).
4) Google translation – use the Google online translation tool[1] to translate a name.

Some of these methods can be applied to certain entity types only. For example, we don't use transliteration model on organization names, because organization names are usually translated.

4. Experiments

We first trained baseline Chinese English named entity taggers, then applied the co-training algorithm using Chinese English parallel and comparable text.

4.1. The Data

The baseline taggers were trained on the Chinese and English data from ACE 2005 Multilingual Training Corpus(Doddington et al., 2004). The Chinese training data contains about 308K characters, the English about 190K words. The ACE test data contains about 74K Chinese characters and 58K English words.

The CRF based taggers which identify person, location and organization names uses features such as unigram, bigram, trigram, and pre-defined lexicons.

The parallel text used in this experiment is the FBIS data[2], which consists of the Chinese and the English translation of news stories and editorials from major news agencies and newspapers in mainland China. The parallel text that was used for training contains 12.0M Chinese characters and 9.7M English words in total. A small portion of parallel text(132K Chinese characters, 106K English words) from the same corpus was manually annotated to be used as the test data, hereafter referred to as PTtest_CN and PTtest_EN.

The comparable text were extracted from the 1995 – 2001 Xinhua sections of Chinese Gigaword Third Edition[3] and English Gigaword Third Edition[4], using the lexicon based content matching algorithm. A total of 15,133 document pairs, or 5.8M Chinese characters, 2.9M English words were extracted by this method.

4.2. The Experiments

The co-training algorithm was run on the parallel text for six iterations. The Chinese tagger and the English tagger at the end of each iteration were then tested on the ACE test data and the PTtest data.

The same experiments were also run on the comparable text, and the Chinese tagger and the English tagger at the end of each iteration were tested on the ACE test data.

5. Results

Figure 2 to 7 illustrate the precision, recall and f-measure of the co-trained taggers on different test sets, where *PT* stands for parallel text, and *CT* for comparable text. Note all f-measures improve significantly before deteriorating after the third iteration. The deterioration is caused by propagating noises coming from labeling and projection errors during the co-training. The F-measures with comparable text degraded much faster than with

[1] http://www.google.com/language_tools

[2] FBIS data were made available to TIDES and GALE researchers by LDC, but not to general public.

[3] LDC catalogue number LDC2007T38

[4] LDC catalogue number LDC2007T07

parallel text, because name projection with comparable text is more difficult and noises are easier to find their way into the training data.

Table 1 shows co-trained Chinese/English named entity taggers' performance on the ACE test data and PTtest data. Because co-training using parallel text was run six iterations, there are six co-trained Chinese taggers and six English taggers. Due to the space limitation, the table only shows the best f-measure (column *BestF*) achieved by the six taggers.

The table clearly shows that co-trained taggers have great improvement over the baseline taggers. In addition, in this experiment, using comparable text achieved about the same result as using parallel text. Note that using comparable text showed significantly better result on ACE Chinese test data than using parallel text.

The test results on ACE test data show that co-training with parallel and comparable test can effectively enhance a name tagger's performance in the domain the tagger was originally trained on.

The test results on PTtest demonstrate that co-training with bilingual text can be used to adapt existing taggers to new domains.

Test Set	Parallel Text		Comparable Text	
	Baseline	BestF	Baseline	BestF
ACE Chinese	80.45%	83.43%	80.45%	84.12%
ACE English	86.79%	88.55%	86.79%	88.29%
PTtest Chinese	84.89%	88.65%	NA	NA
PTtest English	81.55%	85.43%	NA	NA

Table 1 F-measures of co-trained taggers on test sets; BestF indicates the best F-measure co-trained taggers achieved

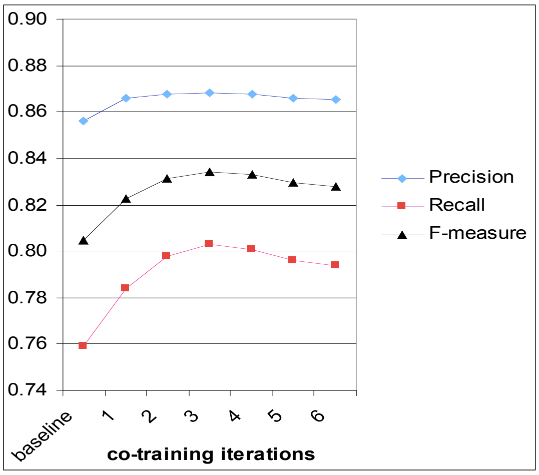

Figure 2 PT models on ACE Chinese test data

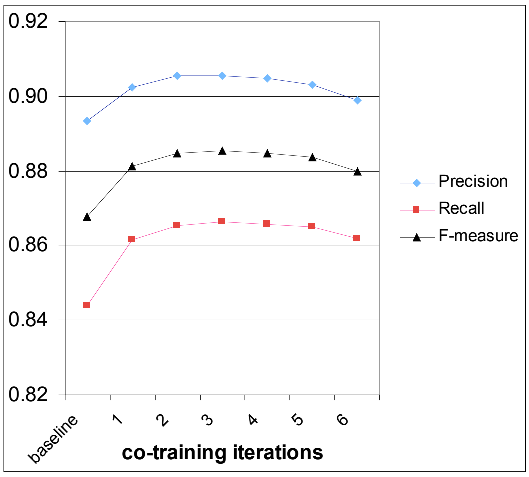

Figure 3 PT models on ACE English test data

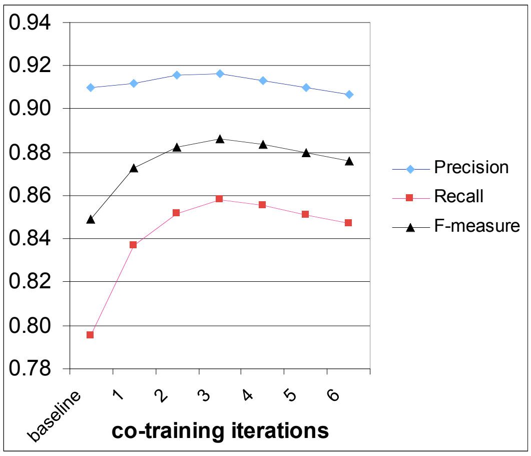

Figure 4 PT models on Chinese PTtest data

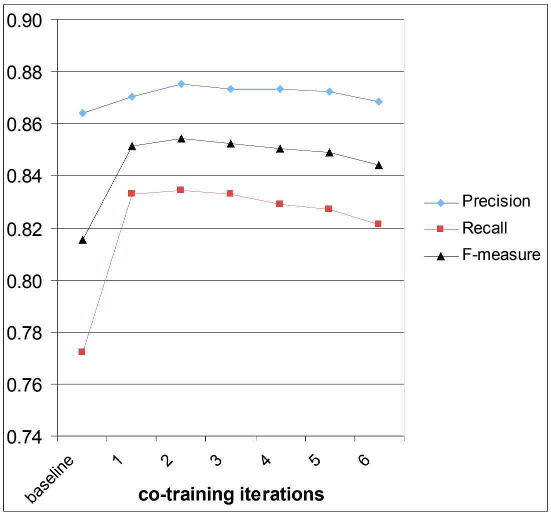

Figure 5 PT models on English PTtest data

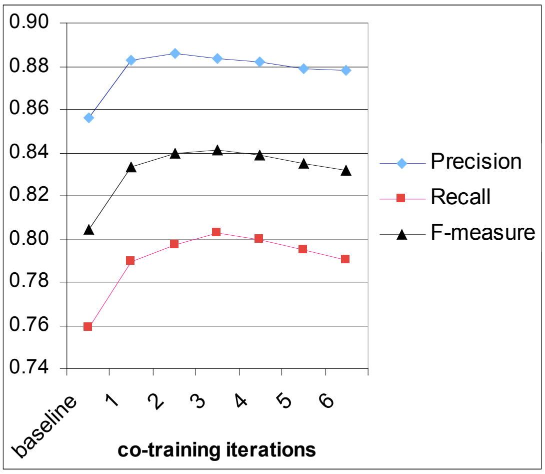

Figure 6 CT models on ACE Chinese test data

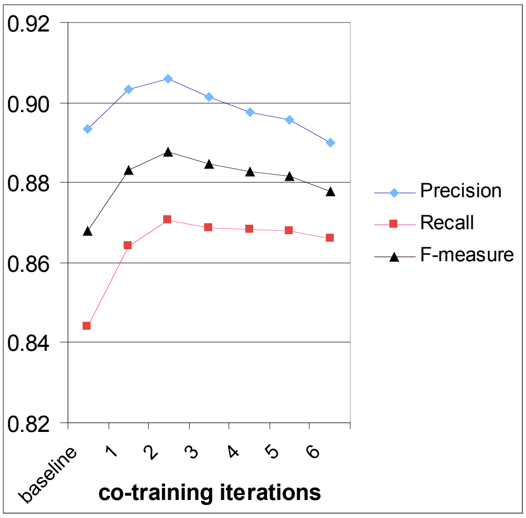

Figure 7 CT models on ACE English test data

6. Conclusion

We have demonstrated that applying co-training on unlabeled bilingual data can improve current state-of-the-art NE taggers, and adapt existing taggers to new domains. Together with entity alignment, we can extend our method from parallel text to comparable text, which has a much greater availability in many domains.

The co-training and entity alignment algorithm we presented have several advantages over previous approaches – the same algorithm can be applied on comparable text; the amount of data required to make the algorithm work is less than word alignment-based approaches; the algorithm can improve NE taggers of both sides of the bilingual text.

The improved tagger performance lays a solid foundation for future works on named entity alignment.

7. References

AL-ONAIZAN, Y., CURIN, J., JAHR, M., KNIGHT, K., LAFFERTY, J., MELAMED, D., OCH, F., PURDY, D., SMITH, N. & YAROWSKY, D. (1999) Statistical Machine Translation. Final Report, JHU Summer Workshop.

BIKEL, D. M., SCHWARTZ, R. L. & WEISCHEDEL, R. M. (1999) An Algorithm that Learns What's in a Name. *Machine Learning,* vol. 34**,** issue 1-3, pp. 211-231.

BLUM, A. & MITCHELL, T. (1998) Combining Labeled and Unlabeled Data with Co-training. *In Proceedings of the Workshop on Computational Learning Theory,* pp. 92-100. Morgan Kaufmann Publishers.

BORTHWICK, A. (1999) A Maximum Entropy Approach to Named Entity Recognition. Ph.D dissertation, New York University.

CUCERZAN, S. & YAROWSKY, D. (1999) Language independent named entity recognition combining morphological and contextual evidence. *In Proceedings of 1999 Joint SIGDAT Conference on EMNLP and VLC,* pp. 90-99.

DODDINGTON, G., MITCHELL, A., PRZYBOCKI, M., RAMSHAW, L., STRASSEL, S. & WEISCHEDEL, R. (2004) Automatic Content Extraction (ACE) program - task definitions and performance measures. *In Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation,* pp. 837-840.

GRISHMAN, R. (1995) The NYU System for MUC-6 or Where's the Syntax? *In Proceedings of the MUC-6 workshop,* pp. 167-175. Washington.

HWA, R., RESNIK, P., WEINBERG, A., CABEZAS, C. & KOLAK, O. (2005) Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Journal of Natural Language Engineering, special issue on parallel text,* 11:3**,** pp. 311-325.

LI, W. & MCCALLUM, A. (2003) Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction. *In Proceedings of ACM Transactions on Asian Language Information Processing, 2003,* pp. 290-294.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. & HERBST, E. (2007) Moses: Open Source Toolkit for Statistical Machine Translation. *In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session,* pp. 177-180. Prague, Czech Republic.

YAROWSKY, D., NGAI, G. & WICENTOWSKI, R. (2001) Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. *In Proceedings of HLT 2001, First International Conference on Human Language Technology Research,* pp. 161-168.