Surprise! What's in a Cebuano or Hindi Name?

JONATHAN MAY, ADA BRUNSTEIN, PREM NATARAJAN,
AND RALPH WEISCHEDEL
BBN Technologies

Empirical results are presented for creating training data and training a statistical name learning algorithm on Cebuano and Hindi in roughly three weeks time. The empirical study compares performance in a compressed time frame against performance of the same statistical language model in English (where there was no compressed time frame). Rapid development of several co-reference heuristics in Hindi are also described, and co-reference performance in Hindi is compared to previously developed English techniques.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Language parsing and understanding; Language models; Text analysis*; G.3 [**Mathematics of Computing**]: Probability and Statistics – *Markov processes; Probabilistic algorithms*

General Terms: Algorithms; Experimentation; Languages

Additional Key Words and Phrases: Extraction, Cebuano; Hindi

1. INTRODUCTION

Named entity extraction, the automatic bracketing and labeling of names in running text, has been studied for several years. By itself, name extraction provides a useful indexing function for retrieval of documents from large collections, and for visualizing those documents by highlighting particular classes of entities. Name extraction is also a vital component of higher-level information extraction, as the identification of names is a crucial first step toward automatic identification of real-world entities in text, relationships between them, and ultimately for derived structural semantic networks. These networks are of great interest to the intelligence community and other bodies that would benefit from efficient automated analysis of large amounts of human-generated data.

Formal evaluations of named entity extraction (NE) in the past have focused on several languages, including Chinese, English, Japanese, and Spanish. In the first half of 2003, a group of researchers and the U.S. National Institute of Standards and Technology (NIST) collaborated in a unique porting experiment: trying to obtain data and bring up algorithms for a new language in 30 calendar days. The chosen languages were Cebuano

This research is sponsored by the Defense Advanced Research Projects Agency and managed by SPAWAR under contract N66001-00-C-8008. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the Defense Advanced Research Projects Agency, the Air Force Research Laboratory, or the United States Government.

Authors' addresses: BBN Technologies, 10 Moulton St., Cambridge, MA 02138

and Hindi. Cebuano is one of the eight dialects of the Philippines and is spoken by approximately 15,000,000, mostly in the South Philippines. The printed form uses a Latin character set and has punctuation and word separation similar to English. Hindi is the lingua franca of India and is the primary language of approximately 315,000,000. The printed form is a non-Latin script called Devanagari. Punctuation is present as is word separation.

We overview the task and scoring procedure in section 2. The statistical learning algorithm used in this 30-day experiment is based on a hidden Markov model (HMM) and has been published before [Bikel et al. 1999]. It has been evaluated formally in English and Spanish, and informally benchmarked in Arabic and Chinese. It is described briefly in Section 3.

The most challenging aspect of the task was the time frame. In previous evaluations, such as the Message Understanding Conferences (MUC), a significant sample of manually annotated data was provided at the outset of the experiment, and the language was known well in advance. In the experiments reported here, from the announcement of the language, one had to find documents in the language, account for challenges unique to the language, manually annotate the data, and train the algorithms in no more than 30 days. BBN's role in the data acquisition and processing path is described in section 4.

Our results are in section 5, including

- An analysis of algorithm performance in Cebuano, English, and Hindi, given comparably sized training sets.
- Comparison of performance in Hindi on material from sources (news publishers) in the training versus sources (news publishers) not in the training set (of course, the test set was blind -- no document in the test set was in the training).
-

In section 6, we report related work. In section 7, we describe a rule-based, name co-reference system developed in fewer than two weeks; section 7 also reports its performance and our suggestions for future improvements appropriate in the rapid-development context. Section 8 concludes.

2. TASK & SCORING

The named entity task used for this evaluation requires the system to identify all named locations, named persons, and named organizations. The task definition [Strassel 2003] is a simplified form based on that given in [Chinchor et al. 1998]. Examples of the task are shown in Figure 1 for Cebuano and in Figure 2 for Hindi. The tool for human annotation is called IdentiTagger, a Unicode-based GUI that supports annotation. The tool tries to give the look and feel of having one highlighting pen for every type of name to be recognized.

The task was evaluated using the MUC scorer [Chinchor et al. 1993] in terms of precision (P) and recall (R),

$$\text{where } P = \frac{\text{number of correct responses}}{\text{number of hypothesized responses}}$$

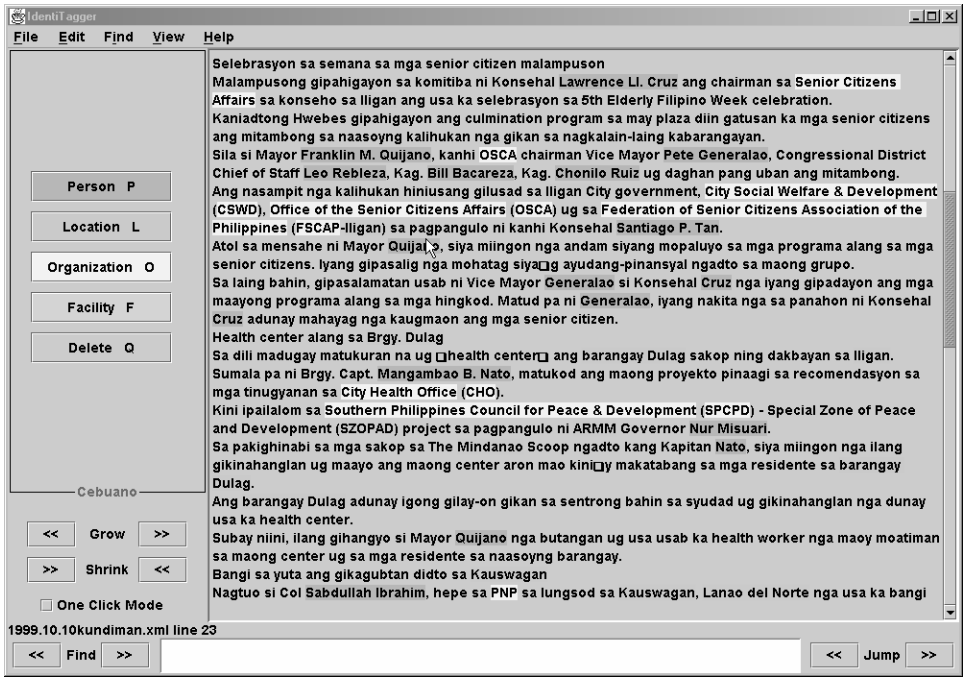

Figure 1: Screen shot of name annotation of Cebuano text.

and

$$R = \frac{\text{number of correct responses}}{\text{number of tags in reference}}.$$

The F-measure is the uniformly weighted harmonic mean of precision and recall:

$$F = \frac{RP}{(R+P)/2}.$$

In NE, a correct response is one where the label and both boundaries are correct. A response is half correct if the label is correct and the response string overlaps with the reference string.

3. ALGORITHM

The information extraction system tested is IdentiFinder(TM), which has previously been reported in [Bikel et al. 1997]. In that system, an HMM labels each word with one of the desired classes (e.g., person, organization, etc.) or with the label NOT-A-NAME (to represent none of the desired classes). The states of the HMM fall into regions, one region for each desired class plus one for NOT-A-NAME (See Figure 3). The HMM thus has a model of each desired class and of the other text. Note that the implementation is

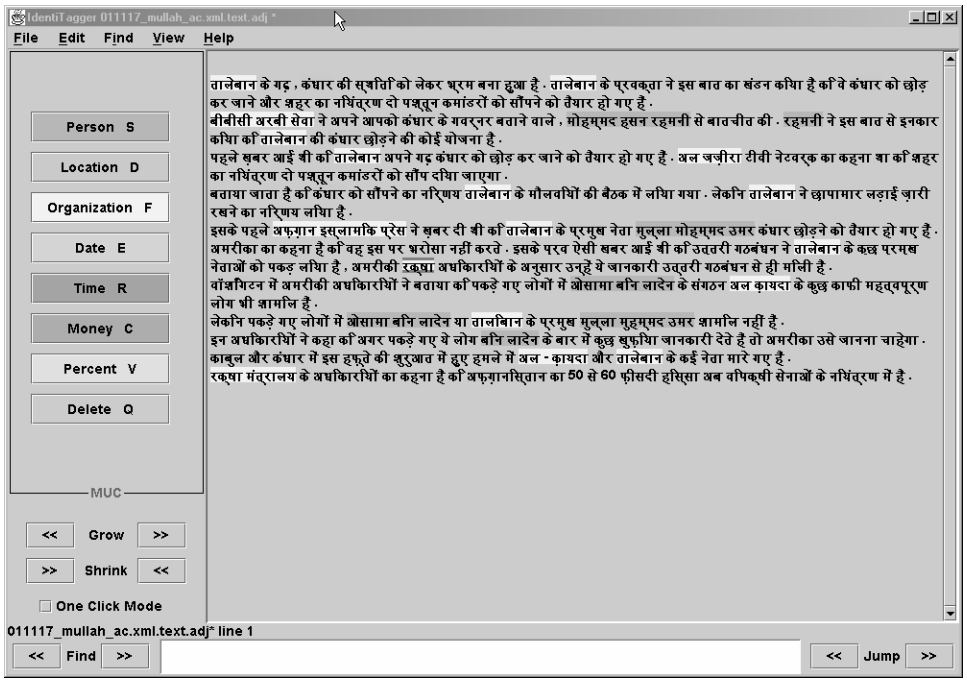

Figure 2: Screen shot of name annotation of Hindi text.

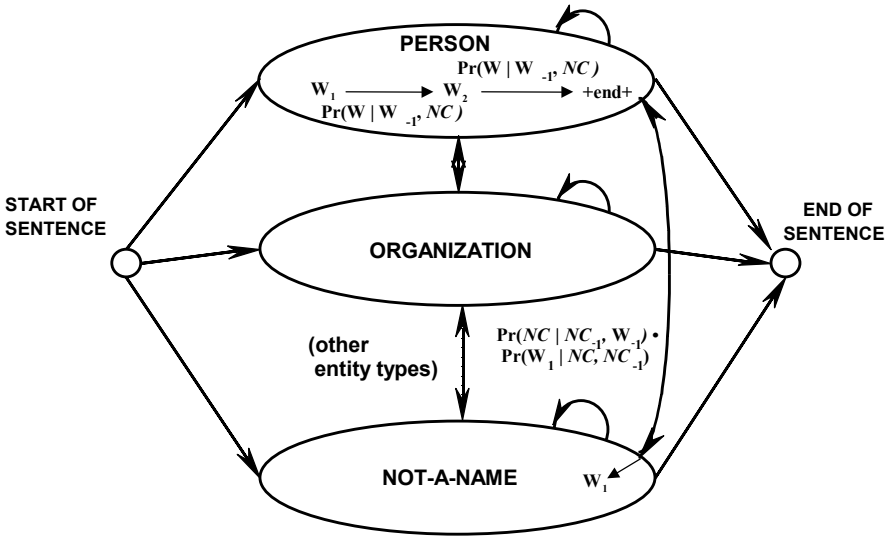

Figure 3: Structure of the hidden Markov model (HMM) for named entity recognition.

not confined to the three name classes used in the NE task; the particular classes to be recognized can be easily changed via a parameter.

Within each of the regions, we use a statistical bigram language model, and emit exactly one word upon entering each state. Therefore, the number of states in each of the name-class regions is equal to the vocabulary size. Additionally, there are two special states, the START-OF-SENTENCE and END-OF-SENTENCE states. The model estimates the probability of traversing the finite state machine from start of sentence to end of sentence, generating a word and label (person, organization, etc.). See Figure 3.

It is important to note that there are virtually no differences in the versions of Identifinder used from language to language throughout all the experiments presented. The Identifinder system does recognize certain word features such as "is capitalized" or "contains Latin characters" in order to more effectively handle unknown words. Some of these features are irrelevant in some languages – there is no capitalization in Hindi or Arabic – and the presence of Latin characters does not affect performance in English. However all systems recognize all features, and it is only the training data that differentiates between language versions of Identifinder.

4. DATA

At the start of the effort, the first task was to find any online news text. For Cebuano, the Linguistic Data Consortium (LDC) identified 246k words from the news site http://www.iliganon.com. For Hindi, LDC made 600k words of online news available from the BBC, from the EMILLE corpus at Sheffield University [McEnery et al. 2000], from India News Daily, from Naidunia, from Rediff, from United News of India, and from Voice of America. While Cebuano uses the Roman character set with some extensions, Hindi uses Devanagari, a script for which there are several (often proprietary) character encodings. In order to be useable, the data had to be converted to a consistent Unicode format. This conversion process made the data acquisition process more difficult than had been originally anticipated.

Given the online documents, the next step was to annotate the data. For Cebuano, LDC annotated 28k words of data, and BBN 218k. For Hindi, LDC, NYU, and BBN all annotated data, as shown below in Figure 4. Annotator difficulties in Hindi were similar to those in English text with unreliable capitalization. For example, in the span *Stolitsa magazine* it is difficult to determine if *magazine* is part of the organization name. Determining what is a name when presented with difficult words like *congress* can be more of a problem in a mono-case language like Hindi as well. These ambiguities were largely not a concern in Cebuano, which uses case to determine names almost unambiguously. Other common problematic cases such as type ambiguity in difficult spans like *White House* were present in all the languages.

5. RESULTS

For Cebuano, there was no official test set. Therefore, we chose 20% of the annotated material, approx. 50k words, as blind test and used the remainder as training.

For Hindi, a test set of 25 previously unseen documents, consisting of 50k total words, was common to all participating sites. Not only were the documents previously unseen, but their distribution among sources was drastically different than in training, as is evident in Table 1 below. Therefore, though the genre (news) was the same as in training, over 90% of the test was from sources covered by only 9% of the training data.

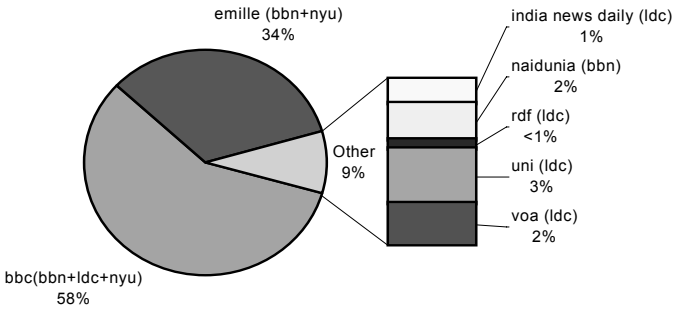

Figure 4: Distribution of training data among seven Hindi news sources.

Table 1: Distribution of Blind Test Material Among Sources

Source	# of Documents in Test
BBC	2
India Today	5
VOA	2
Naidunia	5
Rediff	6
United News of India	5

This is in stark contrast to all previous named entity evaluations. In the results reported for Cebuano and English, all training and test data comes from a single news source.

In Figure 5, we graph named entity performance in English, Cebuano, and Hindi as a function of training set size, starting with as little as several thousand words of text through as much as 650k words of text. Several observations stand out:

- Comparing Cebuano performance to English, Cebuano performance is very good, comparable to the performance of English.
- Comparing Hindi performance to English, there is a large gap.
- Comparing Hindi performance to historical benchmarking results in Arabic and Chinese (figure 6) the results are comparable to Hindi. Chinese and Arabic tests were run over mono-source data and do not use capitalization.
- Comparing Hindi performance to mono-case English (since Hindi does not use capitalization to signal names), the gap in performance from English to Hindi is less pronounced, but still noticeable. However, there are other factors to explain the remaining gap.
- In particular, the dominance in the test set of sources not covered well in training is part of the explanation. Comparing performance on BBC-only (the source of 60% of the training data) with mono-case, mono-source English, the algorithm performs comparably on both cases.

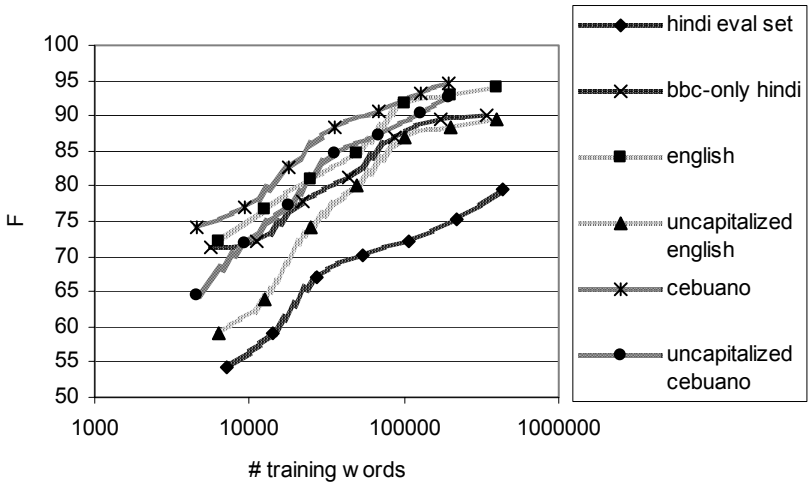

Figure 5: Named entity performance in English, Cebuano, and Hindi as a function of training set size.

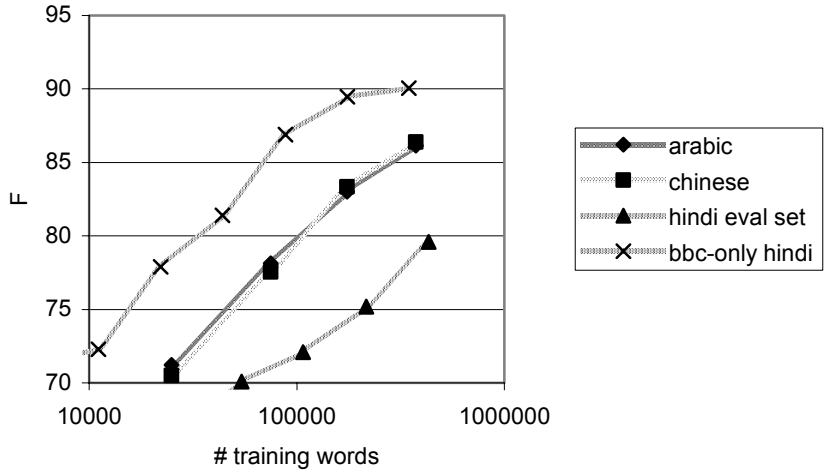

Figure 6: historical benchmarking results on Chinese and Arabic vs. Hindi

This suggests that we look further at test performance on the sources with little training data. In Figure 7, we see that the performance on all sources combined is quite comparable to that covered by 60% of the training data. Of the five sources that comprised 90% of the test material, but only 9% of the training – i.e. the sources that are not BBC – performance was slightly better than on BBC for India Today, Naidunia, and

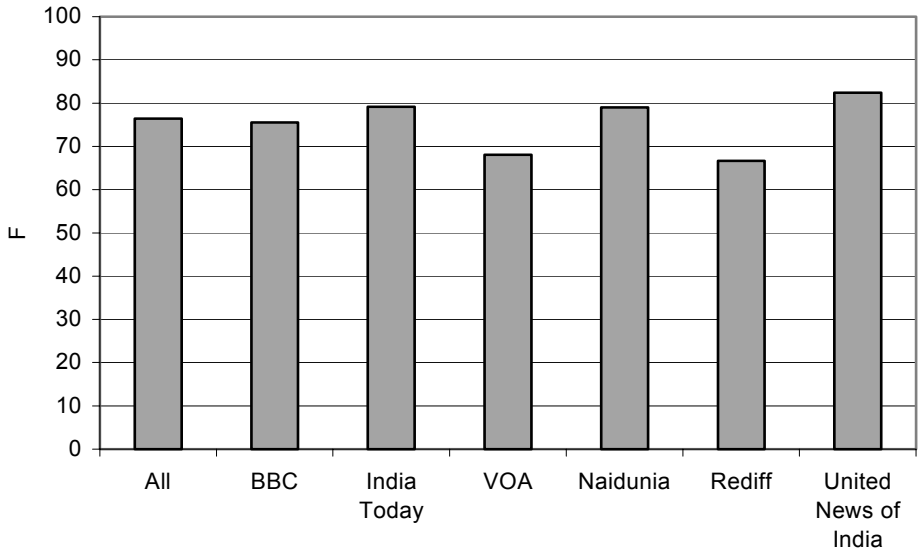

Figure 7. Performance by Source

United News of India, but for the remaining two sources performance was noticeably worse than on BBC. Not too much should be read into these numbers due to the small number of entities in the test set for any of the six sources; nevertheless, it is an encouraging sign that performance does not fall apart on unseen sources of the same genre.

One other interesting experiment was to try to supplement the training data with lists of names. Intuitively, one would hope this would improve performance, particularly due to the dominance of lightly trained sources. Based on this intuition, we added a list of locations into the system. However, empirically this actually reduced scores by about 1%. We analyzed the differences and found that there were only two: in both cases, the inclusion of the location list caused entities of other types (one organization and one person) to be incorrectly tagged as locations.

6. BUILDING A NAME CO-REFERENCE ALGORITHM FOR HINDI

Name co-reference is the problem of determining whether two names refer to the same entity. The problem is sometimes no more difficult than substring matching. (e.g. "George Bush" and "Bush"), but more complex cases often arise. For example, "NASA" and "National Aeronautics and Space Administration" refer to the same organization, while "Bush" and "Bush" may refer to different people in a document discussing the current and former presidents. In previous work, BBN developed statistical models of co-reference, and estimated the parameters of those models from large annotated corpora.

During the 30 day period of the surprise language experiment, however, it was determined that this approach would be too time consuming. Rather, an experiment was conducted to determine the effectiveness of name co-reference cases based only on a few simple rules.

In the great majority of cases seen, co-referring names share at least one word. However, the reverse is not necessarily true; names that share at least one word do not always co-refer. Certain words – those in a particular position of a name or words that are commonly found in lots of names, for instance – are in fact more likely *not* to signify a link than other words. Given the name "John Quincy Anderson," a subsequent mention of "Anderson" and even possibly of "John" could sensibly link, but a mention of "Quincy" would not, due to the word's position in the original mention. Additionally, a mention of "XYZ" should link to an earlier mention of "XYZ, Ltd.," but "Ltd." should not link, since that word is commonly seen in organization names. Similar patterns occur in Hindi as well as in English.

After consultation with a native Hindi speaker, rules for constructing alternate matching names were derived for each entity type. For example, given the name of a person "Aaa Bbb Ccc," valid matches would be "Aaa Ccc," "Ccc," and "Bbb Ccc," but not "Aaa" or "Aaa Bbb" since these are uncommon constructions for coreference for this language. In Hindi certain endings are often added to person names. First and last names may receive "shri," "shrimati," or "kumari" as suffixes. Thus, additional valid matches would also be "Aaashri," "Cccshrimati," and so on. Subsequent names that are in the set of valid alternate matching names are linked to the original name.

Many of the cases of name co-reference that do not share at least one word depend on external knowledge, and thus require the construction of alias lists. One case that does not is that of acronyms, which can be constructed heuristically. In English this is done by joining initial letters of a name, possibly after filtering for stop words. Acronyms in Hindi, however, are often expressed as the pronunciation of the initial English letters of either the English transliteration or translation of the original Hindi. Thus, in order to predict an appropriate acronym given a Hindi name, it is necessary to translate or transliterate each word, then transcribe the pronunciation of each initial English letter into Hindi characters. In order to accomplish this task, a statistical lexicon was used, containing 438k entries and provided by IBM, as was a mapping from English letters to their Hindi pronunciation, provided by a native Hindi speaker. Figure 8 shows an example of this construction – The name "Bharat Tibet Sima Police" is made into an acronym by first translating each word into English, then mapping English letters to appropriate Hindi phonemes.

Using these very basic algorithms entity co-reference was performed over a set of development test data. The ACE EDT scorer was used for evaluation [Doddington 2001] – names not linked generate false alarms, names erroneously linked generate misses; the score is 100*(correct entities found – false alarms)/total correct entities. Table 2 shows performance in the rapidly constructed Hindi system vs. BBN's English system, which incorporates the statistical models and knowledge bases described above. Relative performance on "perfect" name finding is a more informative number for this comparison, as it demonstrates the gap in performance between name linking algorithms for the two languages without the compounded error caused by faulty automatic name finding.

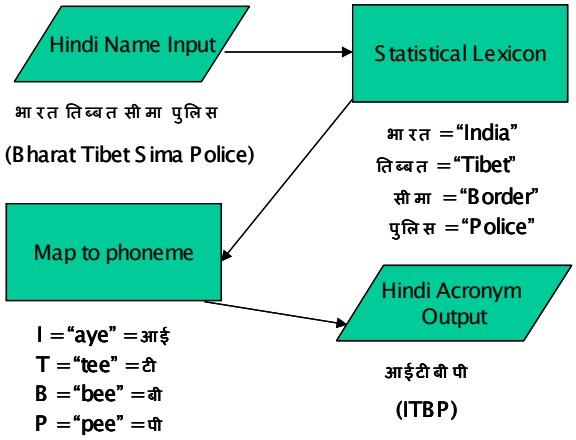

Figure 8: Illustration of the process of hypothesizing acronyms in Hindi.

Table 2: EDT Co-Reference Scores

	Hindi	English
Perfect name finding	84.6	92.0
Automatic name finding	44.1	75.2

A heuristic system is improved by cycles of development, testing, and analysis. After the limited cycles performed, shortcomings in the algorithms used and ways to improve them were immediately obvious. For example, acronyms that are pronounced as words instead of as sequences of letters can be transliterated en masse, rather than letter-by-letter. This would resolve problematic acronyms like "BHEL" for "Bharat Heavy Electricals Limited," which is pronounced (and thus spelled) as a single syllable word, not as four English letters. A more advanced transliteration system would also come in handy here – transliteration knowledge for acronym generation was obtained as a result of the statistically generated bilingual lexicon – thus those transliterations were essentially whole word translations, and not induced from phoneme-based pronunciation models.

Context clues can be used to resolve ambiguous co-reference decisions. In a very simple form, a choice between linking a mention to one of two otherwise equivalent, previously mentioned names can be resolved by choosing the one closer to the mention in question. The most effective improvements to the algorithms, however, will come from the addition of lists based on world knowledge. Abbreviations, particularly those unique to the language in question (e.g. "Pak" for "Pakistan"), nicknames of widely known people and organizations (Harshad Mehta, a figure involved in massive Indian securities scams, was widely known as "Big Bull", much as IBM is widely known as "Big Blue"), and alternate spellings – useful especially for linking foreign names that, particularly in Hindi news sources, can vary within the same document – are easily obtained by working

with language experts and cover the handful of cases that are not easily resolvable through heuristic means.

If the efforts in Hindi are any indication, a name co-reference system that provides a useful level of performance is well within the grasp of limited-time development on a surprise language with only annotation and development support – the vast majority of cases can be covered without the need for language expertise. With only a week of development time a system performing at 92% of English performance was created – a month of collection, annotation, and development efforts dedicated to co-reference would certainly close that distance.

7. RELATED WORK

Other research sites working on this problem have used a variety of techniques for their extraction systems. The University of Sheffield adapted their rule-based extraction system with a non-Hindi rule writer and gazetteer lists induced from training data [Maynard et al. 2003]. The University of Massachusetts employed conditional random fields, with feature induction [McCallum and Li 2003]. MITRE used a hidden Markov model essentially the same as ours [Palmer et al. 1999].

This effort is not the first to note performance degradation from lack of case. [Cucerzan and Yarowsky 1999] point to this as the reason for poorer recall in Hindi than in other languages in their experiments using a language-independent extraction system based on iterative learning and re-estimation of contextual and morphological patterns.

8. CONCLUSIONS

With a multi-site effort, it is possible to create a name extraction capability and name co-reference capability in under 30 days; this includes selection of language, searching for online text, revision of guidelines, annotation of data, and training of learning algorithms. Name tagging was achieved for Cebuano, where online resources and native speakers were scarce. Both name extraction and name co-reference was achieved for Hindi, where multiple (often proprietary) character encodings were a challenge. In fact, we were able to make Hindi name tagging available as a web service within three weeks of selecting Hindi.

Statistical language models produced good performance in both Cebuano and Hindi in such a short time frame. Performance in Cebuano proved comparable with that of English, in spite of the time constraints. Performance in Hindi proved lower than English due to the following:

- The lack of capitalization in Hindi and
- The fact that several Hindi sources that dominated the test set were only lightly included in training.

Additionally, inter-annotator agreement in Hindi was measured as an F of only 90, whereas inter-annotator agreement in English typically achieves an F of 97, 3 times less disagreement than in Hindi. This could be rectified with more time.

Thus, even under severe time constraints it is feasible to achieve results near state of the art armed only with a limited set of established algorithms and annotation tools, and a few native speakers. Additionally, a few handcrafted rules provided name co-reference for persons, organizations, and locations in the same 30 days, yielding 92% of performance on a knowledge-rich system built for English.

REFERENCES

BIKEL, D., MILLER, S., SCHWARTZ, R., AND WEISCHEDEL, R. 1997. Nymble: a high-performance learning name-finder. In *Fifth Conference on Applied Natural Language Processing*, Washington, DC, USA, April, 1997, 194-201.

BIKEL, D., SSCHWARTZ, R., AND WEISCHEDEL, R. 1999. An algorithm that learns what's in a name. *Machine Learning 34*, 211-231.

CHINCHOR, N., HIRSHMAN, L., AND LEWIS, D. 1993. Evaluating message understanding systems: an analysis of the third Message Understanding Conference (MUC-3). *Computational Linguistics 19:3*, 409-449.

CHINCHOR, N., ROBINSON, P., AND BROWN, E. 1998. HUB-4 Named entity task definition version 4.8. Available at http://www.nist.gov/speech/tests/bnr/hub4_98/h4_iene_task_def.4.8.ps on 9/26/2003.

CUCHERZAN, S., AND YAROWSKY, D. 1999. Language independent named entity recognition combining morphological and contextual evidence. In *Proceedings, 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, College Park, MD, USA, June, 1999, 90-99.

DODDINGTON, G. 2001. Value-based evaluation of EDT. Technical report on the ACE 6-month meeting. Available via ftp at ftp://jaguar.ncsl.nist.gov/ace/phase2/nyu-meeting/nist-2001.05-edt-cost-model-v3.pdf on 8/20/2003.

MAYNARD, D., TABLAN, V., AND CUNNINGHAM, H. 2003. NE recognition without training data on a language you don't speak. 2003. *ACL Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models*, Sapporo, Japan..

MCCALLUM, A. AND LI, W. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Seventh Conference on Natural Language Learning (CoNLL)*, Edmonton, Canada, June, 2003.

MCENERY, A. M., BAKER, P., GAIZAUKAS, R., AND CUNNINGHAM, H. 2000. EMILLE: Building a corpus of South Asian languages. *Vivek, A Quarterly in Artificial Intelligence 13:3*, 23-32.

PALMER, D. D., BURGER, J. D., AND OSTENDORF, M. 1999. Information extraction from broadcast news speech data. *Proceedings Of The DARPA Broadcast News Workshop, February 28-March 3,* Morgan Kaufmann Publishers, 41-46.

PRZYBOCKY, M. A., FISCUS, J. G., GAROFOLO, J. S., AND PALLETT, D. S. 1999. 1998 Hub-4 information extraction evaluation. *Proceedings Of The DARPA Broadcast News Workshop, February 28-March 3,* Morgan Kaufmann Publishers, 13-18

STRASSEL, S. 2003. Simple Named Entity Guidelines. Available at http://www.ldc.upenn.edu/Projects/SurpriseLanguage/Annotation/NE/index.html on 9/ 26/2003.