Dealing with Hinglish Named Entities in English Corpora

Renu Balyan
Speech & Natural Language Processing Lab.
Centre for Development of Advanced Computing, Noida,
India
renubalyan.iitd@gmail.com

Abstract

Mixing words of one language into another is a frequently encountered phenomenon in day-to-day natural language communication. This phenomenon of mixing words of different languages has almost become a trend and is very prevalent these days. For different communicative purposes, a language uses words from other languages that may be very common. This gives rise to a mixed language which is neither totally the host language nor the foreign language. This kind of mixed language poses various challenges in natural language processing areas such as machine translation, named entity recognition etc. Thus, it is necessary to identify the "foreign" elements in the source language and process them accordingly. One such problem that has been analyzed and discussed in detail in this paper is incorrect classification of named entities due to Hindi-English (Hinglish) words occurring together as a named entity. In this paper foreign elements are identified and used as cues which may help in proper classification of the so called Hinglish named entities. The paper also suggests finer-grained classification for some of the named entity categories related to location class. This paper is a position paper and hence the aim of the paper is to bring up an issue that needs consideration and not solving it. However, the author has suggested some possible measures for solving the issue.

1 Introduction

Most of the early work formulates the named entity recognition (NER) problem as recognizing "proper names" in general. The entity types that are most studied are three specializations of "proper names": names of "persons," "locations," and "organizations." These types are collectively known as "enamex" since the MUC-6 competition. However, these can be further sub divided into finer categories. For example, the named entity (NE) "person" can have subcategories, as "politician", "entertainer," "sports person" etc. and this work has been discussed in Fleischman and Hovy (2002). The "miscellaneous" NE type is used in the CONLL-2002 and 2003 conferences, and includes proper names falling outside the "enamex." The "timex": "date" and "time" types, and the "numex": "money" and "percent" types are also quite predominant in the literature. Some other NEs may also include entities like, "email address" and "phone number" (Witten et al., 1999; Maynard et al., 2001); "titles" (Cohen and Sarawagi, 2004); and "brand" (Bick, 2004). Some work also exist that does not limit the possible types to extract and is referred to as "open domain" NER (Alfonseca and Manandhar, 2002; Evans, 2003). On the similar lines Sekine and Nobata (2004) defined a named entity hierarchy of nearly 200 categories, which includes many fine grained subcategories, such as international organization, river, or airport, and adds a wide range of categories, such as event, animal, religion, color etc. All these techniques work well for English as lot of research has already taken place in this field. But all these techniques perform poorly when the English text used for identification and classification of NEs contains words not only from English but also Hindi. It was observed that such mixed word NEs are though identified correctly but are classified incorrectly. This paper presents a simple strategy based on some heuristics for dealing with such NEs that contain mixed words of English and Hindi so that these can be classified correctly.

The rest of the paper is organized as follows: in Section 2 NE identification and classification is discussed. Two methods have been studied here. The experimental results and the observations are discussed in Section 3. The approach for proper classification of mixed word NEs is

discussed in Section 4. Future work and the conclusion are presented in Section 5.

2 NE identification & Classification

NEs play a major role in number of natural language processing (NLP) applications. Some of these could be machine translation, Cross Lingual information retrieval, question answering etc. NEs can also be used for event detection as discussed by Smith (2002). For instance, conferences are usually made up of four parts: conference name, location, and dates. Let us illustrate this with an example The NLP conference ICON (conference name) is to be held at Noida (Location) and its start date is "December 18th 2013," and end date is: "December 20th 2013". Thus identifying this event correctly would require identification and classification of all these four NEs from the text. Similarly, a person's birth may be a combination of person name, place, date and Time (e.g., name: "Aryan Kumar," Place: "United States", date: "May 3rd, 2002", Time: "6:15 PM"). Question answering (Srihari and Li, 1999) also often involves NER at the core of the answering capabilities. Local search (Wang et al. 2005) is the task of using location information expressed in a query (e.g., Hyderabad restaurants) to return locally relevant results, such as a list of nearby restaurants.

The problem of identifying Acronyms is also related to NER because many organization names are acronyms (GE, IISC, IIT, IIIT, NRC, etc.). Acronym identification (Nadeau and Turney, 2005) is described as the identification of an acronym's definition (e.g., "IBM" stands for "International Business Machines") in a given document. Thus in this paper we do not restrict ourselves to some specific categories and classify them as NEs. Infact the author considers some finer-grained categories for NEs as also discussed by Nadeau (2007).

Some of the common heuristics used for the identification of NEs for English are proper names start with upper-case and acronyms are all in upper-case. NE identification involves the detection of their boundaries, i.e. the start and the end of all the possible spans of tokens that are likely to belong to a named entity. Identification involves three stages: initial delimitation, separation and exclusion. Initial delimitation involves application of general patterns. These patterns are combinations of a limited number of words, selected types of tokens (e.g. tokens consisting of capital characters), special symbols and punctuation marks. At the separation stage, possible named entities that are likely to contain more than one named entity or a named entity attached to a non named entity, are detected and attachment problems are resolved. Finally all those entities are excluded that may fail some of the rules.

Usually, the NER problem is resolved by applying a rule system over the features. For instance, a system might have two rules, a recognition rule ("capitalized words are entity candidates") and a classification rule ("the type of entity candidates of length greater than 3 words is organization"). These rules work well for some cases but real systems tend to be much more complex.

Here we discuss some features at word-level that are most often used for the recognition and classification of named entities. The word-level features could specifically describe word case, punctuation, numerical value, and special characters. Digits can express a wide range of useful information such as dates, percentages, intervals, etc. Some particular patterns of digits may be, For example, two-digit and four-digit numbers can stand for years (Bikel et al., 1997), and when followed by an "s," they can stand for a decade; one and two digits may stand for a day or a month (Yu et al., 1998). Also common word endings may also provide a big indication towards the type of entity. For instance, the words that end in "ist" (e.g., journalist, cyclist) may often indicate human profession. Nationality and languages often ends in "ish" and "an" (e.g., Indian, Canadian, Irish, British, Spanish). Other examples of common word endings are religion names that end in "ism" (Hinduism, Jainism, Buddhism).

3 Experimental Results & Observations

The identification and classification results for the named entities for tourism domain corpus for fifteen thousand sentences (1,50,000 words approx.) have been analyzed. The corpus used as the test set is in raw form. The seven class model of Stanford NER has been used for the analysis. This identifies seven NEs which are location, person, organization, date, time, money and percent. The same corpus has also been preprocessed through the pre-processor module of a rule-based machine translation system. This module also pre-processes the English input and tags entities such as date, time, currency, acro-

nyms, numbers, MWEs to simplify the translation process. This approach does not explicitly mention the location, person or organization categories but marks the entities as an MWE as these are not required by the translation engine. However these can be extracted and also classified in these categories as the system maintains a list of places, names and organizations. Classifying the identified entities into these 3 categories would involve lexicon look-up. However the entities that are not found in the lexicon will not be classified. The results of the pre-processor module of the rule-based system are shown below in Table 1.

Named Entities	Total Tagged Entities	Incorrectly tagged Entities
Acronyms	347	87
Numbers	3805	77
Time	413	43
Currency	72	12
Date	258	26
MWE	24508	209

Table 1. Pre-processed output

The approach used here is completely rule-based so the results depend on how well-defined and fine-grained rules we use. The MWE further classification will depend on static database and it is neither possible nor suggested to enter all the possible NEs related to person, location and organization. Thus we use a different strategy for their classification that we discuss in the next section.

The Stanford NER has identified 15023 named entities from the same corpus used above. Total unique entities identified were 6549. The results are shown in Table 2. It was observed that the NER performs well for the time, money and percent entities classification. But, it performs poorly for identification of these entities as it fails to identify a large number of 'timex' and 'numex' entities as compared to the rule-based approach. It has also been observed that the approaches are complementing each other and we need to use the hybrid approach instead of focusing on any single approach. Thus the author feels that the rule-based approach should be preferred for classification of 'timex' and 'numex' entities whereas the NER can be used for an initial classification of 'enamex' entities, which can be further refined using the strategy discussed in the next Section.

Named Entities	Total Tagged Entities	Incorrectly tagged Entities
Location	1817	104
Person	1515	261
Time	24	0
Money	7	2
Percent	31	0
Date	411	40
Organization	2744	--

Table 2. 7-class model Stanford NER output

The results show that identification of 'enamex' entities is performed well by the NER but the classification of these entities is full of errors. It was observed that large number of entities like "Aga Khan Palace, Anu Sagar Lake, Bharat Museum" were incorrectly classified as person. It was also observed that number of persons like "Gandhiji, Aurangzeb, Christ, Guru Nanak" was classified as location by the Stanford NER. It was also seen that other entities such as names of languages, religions, festivals and various forms of dances were also incorrectly classified as either location or person.

The author also observed that large number of named entities that were incorrectly classified were actually mixed word entities that were using a mix of English and Hindi words. Some of the examples are "Shiva Mandir, Mumtaz Mahal, Vyas Gufa, Ram Nagar, Zaveri Bazaar". These types of examples are large in number and if handled correctly will make a significant improvement in classifying the NEs correctly. Thus in the next section we discuss how to handle these mixed word entities to improve the classification process.

4 Classification for Mixed words NEs

It was observed that large number of NEs that were incorrectly classified was due to the presence of both English and Hindi words as a single entity. The NEs were correctly identified but could not be put in the correct class. The author however noticed that the Hindi words that were part of the entity can be used as a clue to indicate the class of the entity to which it may or may not belong. Let us illustrate this with the help of a

few examples that have been classified as person by the Stanford NER.

- Alsisar Haveli
- Apsara Vihar
- Arunachal Pradesh
- Bara Bazaar
- Bharat Kala Bhavan
- Ganesh Murti
- Dungar Darwaza
- Chandni Chowk
- Char Minar
- Chang Gali
- Gita Mandir
- Bibi Ka Maqbara
- Char Dham Yatra
- Jama Masjid
- Jawahar Sagar
- Kamala Basti
- Rishi Mandi
- Ardh Kumbh Mela

These are only some of the entities that have been incorrectly classified. The words within the identified entities are strong clues to clearly indicate that these are not persons. So, we can use these Hindi words as cues for correcting the classification. There are also examples where the NE is not a mixed word NE but the classification is still incorrect and some of these NEs include "Rai Garh Museum, Zari Centre, Sree Sankara Gardens, Kanaria Lake, Kalina Campus, Brij Raj Bhawan Palace, Dabolimari Port, Glen View Hotel, Albert Hall, Badra Fort, Ravi River". These entities are also classified as person.

Similarly for organization and location such mis-classification for NEs is shown by the NER. The author suggests a set of cue words that can be used to straightaway select or reject an entity to be classified to a particular class.

4.1 Cue Words for Classification

For the person entity we shall prefer the exclusion strategy as we have very few cue words for this category. However for the location and organization entity, first we look for the entity or part of entity to be one of the cue words and allocate that class to the whole NE. For example if an NE is classified as a person, but it contains a cue word that belongs to the location or the organization cue word list, then it is for sure that the NE is not a person. However, now we can either allocate the NE to the class for which the cue words are defined or allocate a finer-grained category. This problem was faced by the author while allocating classes to names of dances, languages, religion, festival etc. As these NEs is neither an organization nor location. So the author feels that there is a need for dividing the NE categories further so that they can be allocated a proper class. The cue words for organization class are shown in Table 3 and first level of classification cue words for location class are shown in table 4.

Organization NE
University, Union, Association, Council, & Co., Club, College, School, Bank, Ltd. Institute, Organization, Committee, Group, Force, Corporation, Municipal, Bureau, Board, Department, Memorial, Trust, Station, Railways, Terminus, Terminal, Sangha, Banquet , Vedic Shala, Yogshala

Table 3. Cue words for an organization NE

The lists of cue words for various classes (organization and location) are not exhaustive and can be extended with more corpus as currently this list is based on the analysis of corpus which consists of only 15,000 sentences from tourism domain. The cue-words list for location can be broken down further to finer sub categories such as rivers, lakes, mountains, buildings, hills, valleys, villages, hotels, ghats and so on. For this study we are considering only the top level class and consider all these entities as belonging as one class- which is location only.

Location NE
Palace, Hall, Haveli, Lake, Pradesh, Gaon, Vihar, Point, Math, Bazaar, Hill, Valley, Haat, Bhawan, Museum, Niwas, Chowk, Minar, Fort, Darwaza, Gali, Port, Bagh, Church, Village, Town, City, Teerthsthal, Mahal, Nagar, Niketan, Hotel, Temple, Mandir, Road, Regency, Dham, Garden, Shrine, Campus, Masjid, Mosque, Mohallah, Maqbara, killa, Beach, River, Airport, Sanctuary, Vatika, Junction, Caves, Gufa, Café, Resort, Basin, Lodge, Harbour, Coast

Table 4. Cue words for a location NE

These are some cue words that can be searched for within the NE. However, we can also look for some clues with places names that end with "pur" (e.g. Kanpur, Jaipur, Raipur etc.) and also words that end with "garh" (e.g. Chandigarh, Raigarh, Chattisgarh etc.). Similar other kind of clues can be formed for location NE.

The cue words for person are not too many and we also feel that these may not be sufficient to correctly classify the person NE. We suggest that in that case we may also use a window of

size 3 to the left and the right of the NE in the sentence and look for some clues like pronouns and the nouns to which these are associated using dependencies between the words. However, this still needs to be explored further. Some cue words that may be used for classifying person NE are Lord, Emperor, Bhagwan, Mahadev, God, Goddess, Raja, Rani, Guru, Sree, Shri, ji, Mr., Jr., Dr. (other applicable titles). Currently the cue-word list has been prepared manually but an automatic method would be more useful for extraction of these words.

Some other categories that are needed for proper classification are religion, language, dance names, festivals, food_item as none of these belong to location, organization or person class. Currently entities belonging to religion, language, dance names, festivals, food_item are being classified as either location, organization or person, which is not correct. The author could not however decide the class for some NEs like Bahamani Saltanat, Chandela Dynasty, Vijaynagar Empire.

5 Future Plan & Remarks

In this work the NE identification and classification by two systems was studied. The paper also described the issues related to classification of NE independent of the approach. None of the approach seems sufficient or complete and hence the authors have proposed a hybrid approach to improve performance of classification of NEs. The authors also suggest a few NE categories that should be classified separately as they do not fit into either of the existing categories. The work described here is part of an on-going implementation effort and the approach presented in this paper is based on hand crafted rules however, a probabilistic approach may also be thought of which may prove to be more robust.

Acknowledgments

The author would like to thank the anonymous reviewers for their valuable comments and her colleagues in the SNLP laboratory for their support.

References

Alfonseca, Enrique and Manandhar, S. (2002). *An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery.* Proceedings of International Conference on General WordNet.

Bick, Eckhard. (2004). *A Named Entity Recognizer for Danish.* Proceedings of Conference on Language Resources and Evaluation.

Bikel, Daniel M., Miller, S., Schwartz, R. and Weischedel, R. (1997). *Nymble: a High-Performance Learning Name-finder.* Proceedings of Conference on Applied Natural Language Processing.

Cohen, William W. and Sarawagi, S. (2004). *Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods.* Proceedings of Conference on Knowledge Discovery in Data.

Evans, Richard (2003). *A Framework for Named Entity Recognition in the Open Domain.* Proceedings of Recent Advances in Natural Language Processing.

Fleischman, Michael and Hovy. E. (2002). *Fine Grained Classification of Named Entities.* Proceedings of Conference on Computational Linguistics.

Maynard, Diana, Tablan, V., Ursu, C., Cunningham, H. and Wilks, Y. (2001). *Named Entity Recognition from Diverse Text Types.* Proceedings of Recent Advances in Natural Language Processing.

Nadeau, David and Turney, P. (2005). *A Supervised Learning Approach to Acronym Identification.* Proceedings of Canadian Conference on Artificial Intelligence.

Nadeau, David. (2007). *Semi-supervised Named Entity Recognition: Learning to recognize 100 Entity Types with Little Supervision.* University of Ottawa. PhD thesis.

Sekine, Satoshi and Nobata, C. (2004). *Definition, dictionaries and tagger for Extended Named Entity Hierarchy.* Proceedings of Conference on Language Resources and Evaluation.

Smith, David A. (2002). *Detecting and Browsing Events in Unstructured Text.* Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval.

Srihari, Rohini and Li, W. (1999). *Information Extraction Supported Question Answering.* Proceedings of Text Retrieval Conference.

Witten, Ian. H., Bray, Z., Mahoui, M. and Teahan W. J. (1999). *Using Language Models for Generic Entity Extraction.* Proceedings of International Conference on Machine Learning. Text Mining.

Yu, Shihong, Bai S. and Wu, P. (1998). *Description of the Kent Ridge Digital Labs System Used for MUC-7.* Proceedings of Message Understanding Conference.