Name Entity Recognition in Machine Translation

R.C. Balabantaray
International Institute of Information Technology, Bhubaneswar, India
Email: rakeshbray@gmail.com

Abstract- We propose a new name entity recognition method and its translation based on association rules and assumptions. We evaluate the performance of our method with various kinds of texts. The work presented in this paper is part of a larger effort to develop Machine Translation (MT) system which can take care of name entities.

Keywords: MT, NER, ENAMEX, NUMEX, TIMEX.

1. Introduction

Name Entity Recognition (NER) is an information extraction task that is concerned with the recognition and classification of name entity from free text [6]. Name entities classes are, for instance, locations, person named, organization named, dates, times and money amounts. To terms and expressions in the text correspond the entities they represent. For example, in the sentence: *"Indian Prime Minister Mannmohan Singh and President Dr. Abdul Kalam pose for photographers at the Palace",* a name entity recognition process looking for named persons and locations would identify the two persons *Manmohan Singh* and *Abdul Kalam* and the location *Palace*. This recognition can be based on a variety of features of the terms, the sentence, the text and its syntax and could leverage external sources of information such as thesauri and dictionaries, for instance. In the example, a system may have applied a simple rule guessing that the capitalize words directly following the terms '*President*' or '*Minister*' are names of persons. But the most important question is how to convert these name entities to appropriate words in the target language so, that the Machine translation will be fruitful.

Named Entity Recognition (NER), the identification of entity names in free text, is a well-studied problem. In most previous work, NER has been applied to news articles (e.g., (Bikel et al., 1999; McCallum and Li, 2003)), scientific articles (e.g., (Craven and Kumlien, 1999; Bunescu and Mooney, 2004)), or web pages (e.g., (Freitag, 1998)).

India is the second largest in population in the world with more than one billion populations. There are 19 constitutional languages with 10 scripts and over 1650 dialects. Orissa is a state of India situated in the eastern region, with a population of 36.7 million according to 2001 census. Orissa is the first state in India to have been formed on linguistic basis. Oriya is the official as well as spoken language of Orissa, and is one of the constitutional languages of India. We are working to develop a system (OMTrans) which will translate the source language English to target language Oriya.

In this paper, we propose a method for name entity class recognition based on such rules and some assumptions. The rest of the paper is organized as follow. We present and discuss some background and related work on name entity class recognition in the next section. The method is presented in section 3.Finally we conclude and identify the next steps of our research.

2. Related Work

Name Entity Recognition (NER) has been a well-studied problem. This problem is applicable to both formal as well as informal texts. In case of informal text just like emails it is rather easy to identify the name entities as in this case one can find some labels attached with the name entities. But in the formal texts it will be difficult to find it out. There are several classification methods which are successful to be applied on this task. Chieu and Ng[7] and Bender et al.[10] used Maximum Entropy approach as the classifier. Conditional Random Filed (CRF) was explored by McCallum and Li [1] to NER. Mayfield et al.[8] applied Support Vector Machine (SVM) to classify each name entity. Florian et al. [12] even combined Maximum Entropy and hidden Markov Model (HMM) under different conditions. Some other researches are focused more on extracting some efficient and effective features for NER. Chieu and Ng[7] successful used local features, which are near

the word, and global features, which are in the whole document together. Klein et al.[5] and Whitelaw et al.[2] reports that character-based features are useful for recognizing some special structure for the name entity.

3. Methodology

Our algorithm described below is a unique one which is identifying the name entities expressions (ENAMEX), numerical entity expressions (NUMEX), and temporal entity expressions (TIMEX). The most important thing with this technique is that the dictionary should be exhaustive and the lists corresponding to the name entity varieties (like ENAMEX_TYPES, NUMEX_TYPES and TIMEX_TYPES) should be proper. In the name entity ENAMEX there are almost eleven types such as person, organizations, location, facilities, locomotives, artifacts, entertainment, cuisine's, organisms, plants and diseases. Similarly in NUMEX there are four varieties like distance, money quantity, count and in TIMEX there are again four varieties like time, date, day, period. For the above described TYPES we are maintaining different lists consisting of words relevant to the corresponding TYPES arranged according to the ASCII collating sequence so, that the process of searching will be faster. These lists have been prepared with the help of various sources like Corpora and from the texts being used in our day to day life. So, far basically person names are concerned new names are getting created day by day and that might not appeared in the list. In this scenario our system will simply identify it as name entity and it will not be able to categorize it.

Algorithm

Step1: The root words (after removing the tag features from the corresponding words) starting with Uppercase letters which is not present in our Bilingual dictionary or the root words not present in the bilingual dictionary is treated as Name Entity (it may be ENAMEX, NUMEX or TIMEX) with some exception.

Step 2: For categorizing whether it is ENAMEX, NUMEX or TIMEX certain rules are being followed.

Step 3: After that for finding the types we have to refer to the list defined under the corresponding broad categories like ENAMEX, NUMEX and TIMEX.

For example in the sentence like "*Mannmohan Singh is the prime minister of our country.*" The word *Monnmohan* and *Singh* both are starting with uppercase letters and not present in the dictionary so both will be identified as name entity and as ENAMEX category as per our rules. Now binary search will be performed in the list defined under ENAMEX category to get the subcategory and both the word will be found as person name (Individual).

<ENAMEX TYPE="PERSON"> Mannmohan Singh</ENAMEX>

In the sentence like "*I am a citizen of India."* the word *India* is starting with uppercase letter but it is present in the dictionary so, it will not be treated as name entity but here as per our exception rules and the information retrieved from dictionary this is name of a place so, it will be treated as ENAMEX and type will be location.

<ENAMEX TYPE="LOCATION"> India</ENAMEX>

Similarly in the sentence like "*it is 5:30 AM in the morning.*" The token *5:30* will be treated as name entity but since it contains numeric figure it can be NUMEX or TIMEX but as per our rules it is followed by AM so, it will be TIMEX. Then after consulting the lists the type will be finalized as *Time*.

<TIMEX TYPE="TIME"> 5:30 AM</TIMEX>

The table given below shows the names of the lists we have maintained for identifying the types of the name entities under the three broad categories ENAMEX, NUMEX and TIMEX.

Name of broad categories	Name of Lists
ENAMEX	ENAMEX_PERSON
	ENAMEX_ORGANIZATION
	ENAMEX_LOCATION
	ENAMEX_FACILITIES
	ENAMEX_LOCOMOTIVES
	ENAMEX_ARTIFACTS ENAMEX_ENTERTAINMENT
	ENAMEX_CUISINE
	ENAMEX_ORGANISMS
	ENAMEX_PLANTS
	ENAMEX_DISEASE
NUMEX	NUMEX_DISTANCE
	NUMEX_MONEY
	NUMEX_QUANTITY
	NUMEX_COUNT
TIMEX	TIMEX_TIME
	TIMEX_DATE

	TIMEX_DAY
	TIMEX_PERIOD

4. Conclusion

We have presented a new name entity class recognition method based on association rules and assumptions. The dictionary which we are following for this task is exhaustive but the contents of the lists are growing day by day. Once the list will be exhaustive it is guaranteed that our method will show a higher degree of precession than the other methods. The technique used over here to identify the name entities can also be applicable to the translation system of English to other Indian languages.

We are in the process of transliterating these name entities to the target language Oriya so, that this will be a standard for the translation system of English to other Indian Languages as all Indian languages are phonetically linear.

References

[1]. A. McCallum and W. Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *CoNLL- 2003*.

[2]. Casey Whitelaw and Jon Patrick *2003.* Named Entity Recognition Using a Character-based Probabilistic Approach. In: *Proceedings of CoNLL-2003*, Edmonton, Canada, 2003, pp. 196-199.

[3]. D. M. Bikel, R. L. Schwartz, and R. M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34:211–231.

[4]. D. Freitag. 1998. Information extraction from html: application of a general machine learning approach. In
AAAI-98.

[5]. Dan Klein, Joseph Smarr, Huy Nguyen and Christopher D. Manning, *2003.* Named Entity Recognition with Character-Level Models. In: *Proceedings of CoNLL-2003*, Edmonton, Canada, pp. 180-183.

[6]. Grishman, R., 1997. "Information Extraction: Techniques and Challenges", Lecture Notes in Computer Science, Vol. 1299,Springer-Verlag.

[7]. Hai Leong Chieu and Hwee Tou Ng, 2003. Named Entity Recognition with a Maximum Entropy Approach. In: *Proceedings of CoNLL-2003*, Edmonton, Canada, pp. 160-163.

[8]. James Mayfield, Paul McNamee and Christine Piatko, 2003. Named Entity Recognition using Hundreds of Thousands of Features. In: *Proceedings of CoNLL-2003*, Edmonton, Canada, pp. 184-187.

[9]. M. Craven and J. Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB-99*.

[10]. Oliver Bender, Franz Josef Och and Hermann Ney, 2003. Maximum Entropy Models for Named Entity Recognition In: *Proceedings of CoNLL-2003*, Edmonton, Canada, pp. 148-151.

[11]. R. Bunescu and R. J. Mooney. 2004. Relational markov networks for collective information extraction. In *ICML-2004 Workshop on Statistical Relational Learning*.

[12]. Radu Florian, Abe Ittycheriah, Hongyan Jing and Tong Zhang, 2003. Named Entity Recognition through Classifier Combi nation. In: *Proceedings of CoNLL-2003*, Edmonton, Canada, pp. 168-171.

Dr. Rakesh Chandra Balabantaray is currently working as Assistant Professor in the Department of Computer Science & Enginerring at IIIT, BHUBANESWAR, Orissa, India. He did his Masters in Computer Science in the year 2001 and Ph.D. in Computer Science in the year 2008 from Utkal University, Orissa, India. He was born in the year 1978. He has more than twenty publications in various journals and conferences. His major area of research is Artificial Intelligence, Natural Language Processing & Information Retrieval.