IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-3, Issue-9 E-ISSN: 2347-2693

Named Entity Recognition for Kashmiri Language using Noun Identification and NER Identification Algorithm

Amir Bashir Malik[1*] and Khushboo Bansal[2]

[1*,2]Department of Computer Science, Deshbhagat University mandigobindgarh - India

www.ijcseonline.org

Received: Aug/23/2015 Revised: Aug/30/2015 Accepted: Sep/24/2015 Published: Sep/30/2015

Abstract- In this study, we present a brief overview of Named Entity Recognition (NER) system, various approaches followed for NER systems and finally NER systems for Kashmiri language. Kashmiri language raises several challenges to Natural Language Processing (NLP) largely due to its rich morphology. Named entity recognition (NER) (also known as entity identification and entity extraction) is one of the important subtask of information extraction that seeks to locate and classify atomic text into predefined categories such as the names of persons, organizations, locations, monetary values, percentages, expressions of times, etc. This paper describes the problems of NER in the context of Kashmiri Language and provides relevant solutions by using noun identification algorithm and named entity recognition identification algorithm Building a named entity recognition system for Kashmiri languages that can understand Kashmiri language has been one of the long-standing goals of (NER) system**.**

Keywords: Named Entity Recognition, Natural Language Processing, Noun Identification, NER Identification, Kashmir language

I. INTRODUCTION

The term Named Entity (NE) was evolved during the sixth Message Understanding Conference (MUC-6,1995). Named Entity Recognition (NER) is also knows as entity identification is a subtask of information extraction (IE). [05][15] NER extracts and classifies the true Named Entities in text. NER system is widely used in different tasks of Natural Language Processing (NLP) and in many commercial applications on internet like Search Engine. Named Entity Recognition (NER) is a process of searching the text to detect entities in a text and to classify them into predefined classes such as the names of persons, organizations, locations, date, time, Designations, Measures, abbreviations and brand etc. Construction of a Named Entity Recognition (NER) system becomes challenging if proper resources are not available. Gazetteer lists are often used for the development of NER systems In many resource-poor languages like Kashmiri gazetteer lists of proper size are not available, but sometimes relevant lists are available in English.

Major tasks in (NER) Given a stream of text, determine which items in the text map to proper names, such as people or places, and what the type of each such name is (e.g. person, location, organization). Note that, although capitalization can aid in recognizing named entities in languages such as English, this information cannot aid in determining the type of named entity, and in any case is often inaccurate or insufficient. For example, the first word of a sentence is also capitalized, and named entities often span several words, only some of which are capitalized. Furthermore, many other languages in non-Western scripts (e.g. Kashmiri or Arabic) do not have any capitalization at all, and even languages with capitalization may not consistently use it to distinguish names. For example, German capitalizes all nouns, regardless of whether they refer to names, and French and Spanish do not capitalize names that serve as adjectives.[12]

Normally, the NER has solved problems in the application of a rule-system of functions. For example, a system has two rules, a recognition rule: enabled, are words candidate organizations "and a classification rule," the kind of candidate units of length more than three words organization. "Those rules are good for the prototype set before.[09] However, real systems tend to be much more complex and their rules are often created by automated learning. Usually, there are three different features to recognize NE: Word-level features, List lookup features and Document and corpus features.

1.1 Word-level features

Word-level features are related to the character makeup of words. They specifically describe word case, punctuation, numerical value and special characters.[13] It contains several features below.

Digit Pattern

Digits can express a wide range of useful information such as dates, percentages, intervals, identifiers, etc.

Common word ending

Morphological features are essentially related to words affixes and roots. For instance, a system may learn that a human profession often ends in ‘ist’ (journalist, cyclist) or that nationality and languages often ends in 'ish' and 'an'(Spanish, Danish, Romanian).[13]

Functions over word

Features can be extracted by applying functions over words

Patterns and summarized patterns

The role of Pattern features is to map words onto a small set of patterns over character types.

1.2 List Look up Features

Lists are the privileged features in NER. The terms gazetteer, lexicon and dictionary are often used interchangeably with the term list. List inclusion is a way to express the relation is a (e.g., Trondheim is a city). It may appear obvious that if a word (Trondheim) is an element of a list of cities, then the probability of this word to be city, in a given text, is high.[13] However, because of word polysemy, the probability is almost never (e.g., the probability of Fast to represent a company is low because of the common adjective fast that is much more frequent).

We could enumerate many more list examples but we decided to concentrate on those aimed at recognizing name types.

General Dictionary

Common nouns listed in a dictionary are useful, for instance, in the disambiguation of capitalized words in ambiguous positions

Words that is typical of organization names

Many authors propose to recognize organizations by identifying words that are frequently used in their names.

On the list lookup techniques

Most approaches implicitly require candidate words to exactly match at least one element of a pre-existing list. However we may want to allow some flexibility in the match conditions. At least three alternate lookup strategies are used in the NER field: word can be stemmed, fuzzy matched and accessed using the Sound ex algorithm.

1.3 Document and corpus features

Document features are defined over both document content and document structure. Large collections of documents (corpora) are also excellent sources of features.[14] We list in this section features that go beyond the single word and multi-word expression and include meta-information about documents and corpus statistics.

- Multiple occurrences and multiple casing
- Entity reference and alias
- Document meta-information
- Statistics for Multiword units

II. RELATED WORK

Vishal Gupta and Gurpreet Singh Lehal, 2011, [01] this paper explains the Named Entity Recognition System for Punjabi language text summarization. A Condition based approach has been used for developing NER system for Punjabi language. Various rules have been developed like prefix rule, suffix rule, proper name rule, middle name rule and last name rule. For implementing NER, various resources in Punjabi, have been developed like a list of prefix names, a list of suffix names, a list of proper names, middle names and last names.

Kamal deep Kaur and Vishal Gupta, 2012, [03] built a „NER for Punjabi‟ using rule based and list look up approaches. As mentioned earlier, Punjabi is also a language with high clung and inflections, which leads to linguistic problems. The rule based approach trained the system to identify NEs by writing rules manually for all NE features. The most common words are removed from the database, and then a list look up approach is used with the Gazetteers lists to classify the identified NEs. Their system resulted with 85.88% f-measure.

YassineBenajiba, Paolo Rosso, and Jos´e Miguel Bened´ıRuiz, 2013.[06] In this paper, we present ANER sys: a NER system built exclusively for Arabic texts based-on n-grams and maximum entropy. Furthermore, we present both the specific Arabic language de-pendent heuristic and the gazetteers we used to boost our system. We developed our own training and test corpora and gazetteers to train, evaluate and boost the implemented technique. A major effort was conducted to make sure all the experiments are carried out in the same framework of the CONLL 2002 conference. We carried out several experiments and the preliminary results showed that this approach allows tackling successfully the problem of NER for the Arabic.

NavneetKaurAulakh and Er.YadwinderKaur, 2014, [02] Name entity recognition (NER) techniques are explained and how they find name entity from the text. Translation model calculate the probability of target sentences given the source sentence and decoder maximizes the probability of translated text of target language.

SudhaMorwal, and NusratJahan, 2014.[07]Named Entity Recognition is the process to detect Named Entities (NEs) in a file , document or from a corpus and to categorize them into certain Named entity classes like name of city, State, Country, organization, person, location, sport, river, quantity etc. In this paper our main objective is to perform Named Entity Recognition in Natural languages using Hidden Markov Model (HMM) and provide ways to increase accuracy and the Performance Metrics (Precision, Recall, F-Measure). Named entity recognition (NER) is one of the applications of Natural Language Processing and is considered as the subtask of information retrieval.

III. PROPOSED WORK

Objectives

- Too understand the language
- To generate the Kashmiri dictionary.
- To make algorithm of Kashmiri language system.
- To implement and test Kashmiri language system

3.1 Noun Identification

It is useful to recognize nouns and eliminate non-nouns. The Kashmiri morphological analyzer developed here has been used to obtain the categories. Structure of Kashmiri nouns is root stem along with number marker along with case markers. Nouns in Kashmiri follow the traditional classification scheme of (i) Proper (human animate, non-human animate, and inanimate) nouns, and (ii) Common (count, mass) nouns. Nouns are not formally distinguished for being definite or indefinite.

Noun Inflection

Nouns are inflected for gender, number and case.

Suffix features Every language uses some specific patterns which may act as ending words in proper names and the list of this type of words is called as suffix list.[10] The following suffixes added to nouns indicate their masculine formation: - As a result of adding of these suffixes certain morphophonemic changes take place.

duka:n/(دُکان،وان)	'shop'
duka:nda:r/(دُکان داوانہٕ وول)	'shopkeeper'
The:kI / (ٹھیکہٕ،مُہاد)	'contract'
The:kIdar ٹھیک دار)	'contractor' .

The following suffixes added to nouns indicate their feminine formation: -en',-In', -A:n', -ba:y, -Ir.

ma:sTarba:y(اُستاد)	' teacher'
ta:leem(تعلیٖم)	'education'

Morphological features

Indian languages are morphologically rich. Words are inflected in various forms depending on its number, tense, person, case, etc. Identification of root word is very difficult in Indian languages like Kashmiri [4] [8].

The 3.3 algorithm shows noun identification various test are to be performed to achieve a good performance by using different techniques.

Gender Nouns are divided into two classes : Masculine and feminine. Animates follow the natural gender system. Gender of a large number of inanimate nouns can be predicted by their endings. Gender formation processes from masculine to feminine or vice versa are irregular. Main gender formation processes involve

- suffixation
- changes in vowels and consonants, and
- Most of the phonological and morphological changes are regular.

Number

There are two numbers: singular and plural. Most count nouns form their plurals from singular form. Some count nouns have the same form for both numbers. Mass nouns do not show number distinction. Plurals are formed from singulars by suffixation, palatalization and vowel changes.

Case. Case suffixes added to nouns/noun phrases occur as bound morphemes. Following table gives the case suffixes added to the nouns agreeing in number and gender.

3.2 NER Identification

Features of Kashmiri Language can be exploited for development of a good Named Entity Recognizer. Some features considered are as:
Kashmiri borrows words from Perso-Arabic, Sanskrit, Hindi Urdu and English. Nativized loans from these languages fall in two genders: masculine and feminine. It is interesting to note that a large number of words borrowed from Hindi-Urdu have different genders from their sources (see for details Koul 1983). A few examples are given below in table 3.2

Hindi-Urdu word	Kashmiri word	English word	NER
Adat	a:dat /(عادَت)	Habit	Noun
Kimat	Kl:mat/(قِمَت طَے کَرُن)	Price	Verb
Dava	Dava/(دَوا)	Medicine	noun

3.3 Algorithm for Noun identification.

```
Read file and divide into sentences.
Read sentences and divide into tokens
    Read each token
For each (word) Loop
   Match with Kashmiri dictionary
   If direct match with dictionary then
assign category
else if no match with dictionary then
```

```
apply noun Morphological suffixes
if suffixes are found and root is found in
Kashmiri dictionary then
assign category
else if suffix matches and root     is not found
then
token may be noun
else if token ending with consonant  then
the word may be loan word
assign noun
else
assign the category "unknown"
End loop.
```

3.4 Algorithm for NER identification.

```
Read list of nouns identified by noun identification
Check gazetteers lists for NER features
For each (noun) Loop
if suffix features found  then
   Assign NER tag
else if  context features found then
   Assign NER tag
else if found in NER dictionary then
   Assign NER tag
else
   Assign "Miscellaneous word"
If a
ambiguity is found then
call disambiguation technique
remove disambiguation
End loop
```

IV. RESULTS

Kashmiri Language NER system for words or sentence has been implemented in JAVA NETBEANS at front end and MS Access at back end. Regarding condition based NER system, an in depth analysis of output has been done over 1000 Kashmiri words as input. It is producing result 93.32%, An in depth error analysis of condition based system has been done over 1000 words and it is giving 07.75% errors. In the First phase we have conducted various tests for noun identification and got very good performance by using dictionary gazetteers lists, morphological suffix mapping techniques and other features. A good number of nouns are identified in the first phase. These nouns may be common nouns or Named Entities or loan words. The identified nouns are given as input to the second phase. In the second phase we are checking each noun with gazetteers lists which contains beginnings, endings, contexts and suffixes of various tags. According to the category NE tags are assigned ambiguity is also resolved by using gazetteers lists and features. After conducting NER identification it is observed that good performance is achieved by the system. For Kashmiri language it is first time that these resources have developed and these may be helpful for future NLP applications in Kashmiri language.

V. CONCLUSION AND FUTURE SCOPE

Through the research the areas of NER, we can learn the basic and detail definition of NER. We explained when/how NER is used in applications. We list the main challenges of NER systems, and also the benefit of using NER as part of other systems. NER is the task of recognizing proper nouns from the given text. Recognition of NEs can be done through framing grammar rules by language experts. Finally, the output of the system is a list of NEs with good precision tagged accordingly and the categories used for tagging throughout the paper are PERSON, LOCATION and ORGANIZATION. The approach specified is simple yet effective and can be extended to any language as none of the language dependent tools or language experts are involved. There are several issues taken care as complete usage of words, overcoming the challenges, etc., which are unique to this paper. Thus, a language independent named entity recognization system for kashmiri language is developed.

Future Scope: As I discussed before in this paper, English has been research very well. Even Chinese, Germany, Japanese and so many other languages has been present as the field of NER, but Kashmiri language has not research very well. How to find the right way to deliberate Named Entity Recognition system for Kashmiri language can be consideration in the future. In the future scope the accuracy of the system in the script of Kashmiri language can be improved. As no work has done in NER for Kashmiri language So, Prediction ability of the research system can be improved. In the prediction ability of the system the System will be able to produce the accurate results to the NER Kashmiri language. In future, quality can be improved to increase the accuracy of ner system.

REFERENCE

[01] Vishal Gupta, Gurpreet Singh Lehal, "Named Entity Recognition for Punjabi Language Text Summarization". International Journal of Computer Applications (**0975 – 8887**) Volume **33**– No.**3**, November **2011.**

[02] NavneetKaurAulakh, Er.YadwinderKaur. "Review Paper on Name Entity Recognition of Machine Translation".International Journal of Advanced Research in Computer Science and Software Engineering ISSN: **2277 128X** Volume 4, April **2014**

[03] Kamal deep Kaur, Vishal Gupta. "Name Entity Recognition for Punjabi Language".International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: **2249-9555** Vol. **2**, No.**3**, June **2012.**

[04] Ganapathiraju, M., et al. OM: "One Tool for Many (Indian) Languages". in ICUDL: International Conference on Universal Digital Library. **2005**. Hang Zhou

[05] Fleischman, Michael. **2001**. Automated Subcategorization of Named Entities. In Proc. Conference of the European Chapter of Association for Computational Linguistic.

[06]YassineBenajiba, Paolo Rosso, and Jos´e Miguel Bened´ıRuiz, **2013**.” NER system built exclusively for Arabic texts based-on n-grams and maximum entropy” april**2013**

[07]SudhaMorwal, and NusratJahan. “Named Entity Recognition Using Hidden Markov Model (HMM): An Experimental Result on Hindi, Urdu and Marathi Languages”. International Journal of Advanced Research in Compu**ter Science** and Software Engineering ISSN**: 2277 128**X Volume **3**, Issue **4**, April **2013**

[8] http://www.iiit.net/ltrc/morph/morph_analyser.html

[09] Pallavi, Dr. Anitha S Pillai. “Named Entity Recognition for Indian Languages: A Survey”. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: **2277 128**X, Volume **3**, November **2013**

[10] McDonald D. 1996. Internal and external evidence in the identification and semantic categorization of proper names. In: B.Boguraev and J. Pustejovsky (eds), Corpus Processing for Lexical

[11**]** http://www.aclweb.org/anthology/C**12-1153**

[12] Pallavi, Dr. Anitha S Pillai. “Named Entity Recognition for Indian Languages: A Survey”. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: **2277 128**X, Volume 3, November **2013** Acquisition, pp. **21-39**.

[13]https://daim.idi.ntnu.no/masteroppgaver/**005/5654/**masteroppgave.

14]GitimoniTalukdar, and PranjalProtim Borah. “A Survey of Named Entity Recognition in Assamese and other Indian Languages”.International Journal on Natural Language Computing (IJNLC) Vol. **3**, No.**3**, June **2014**

[15]Grishman and Sundheim, Message Understanding Conference-6: a brief history, International Conference On Computational Linguistics, Proceedings of the **16**th conference on Computational linguistics,**1996**

[16] S Amarappa, Dr. S V Sathyanarayana. “Named Entity Recognition and Classification in Kannada Language”. International Journal of Electronics and Computer Science Engineering, ISSN- **2277-1956** November **2012.**