Named Entity System for Tweets in Hindi Language

Arti Jain, Jaypee Institute of Information Technology, Noida, India
Anuja Arora, Jaypee Institute of Information Technology, Noida, India

ABSTRACT
Due to the growing need of smart-health applications in Hindi language, there is a rapid demand for health-related Named Entity Recognition (NER) system for Hindi. For the purpose of the same, this research considers Twitter social network to extract tweets dated 1st October 2016 to 15th October 2017 from Patanjali, Dabur and other Hindi language-oriented Twitter based health sites; while considering four NE types—Person, Disease, Consumable and Organization. To the best of its knowledge, the considered Twitter dataset and NE types for Hindi language is one of the first resources that is being taken care. This article introduces three stage NER system for Tweets in Hindi language (HinTwtNER system)- pre-processing stage; machine learning stage (Hyperspace Analogue to Language (HAL) and Conditional Random Field (CRF)); and post-processing stage. HinTwtNER looks into binary features and achieves an overall F-score of 49.87% which is comparable to the Twitter based NER systems for English and other languages.

KEYWORDS
Analogue to Language, Conditional Random Field, Hindi, Hyperspace Machine Learning, Named Entity Recognition, Online Social Network, Tweets, Twitter

1. INTRODUCTION
Online Social-media Networks (OSNs) (Kwak et al., 2010; Lin & Huang, 2013; He et al., 2014; Baldwin et al., 2015; Zhu et al., 2016; Cresci et al., 2017; Egele et al., 2017) such as Facebook, Twitter etc. are becoming an everyday part of many peoples’ lives, and they play a major role in the modern society. As these OSNs act as key elements to transform the cultural, social, technological and other diverse aspects of modern civilization. This in turn impacts various sectors, namely-business, education, health, psychology etc. Statistics (Facebook, 2017) reveal that currently 2.01 billion monthly active users are Facebook users. And, on Twitter (Twitter, 2017) on an average, every second around 6,000 tweets are tweeted, corresponding to over 350,000 tweets per minute, 500 million tweets per day and around 200 billion tweets per year. Tweets are short messages with restriction of maximum length of 140 characters. These tweets are often noisy having spelling and grammatical mistakes (because of informal, mix and gibberish language); short-forms of words (because of slang language); multi-words merged together; special symbols and characters (such as emoticons (:_)) that are embedded within words. Still now-a-days, users prefer to tweet due to the following reasons:

DOI: 10.4018/IJIIT.2018100104

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
• Users aren’t getting preferable posts on their newsfeed i.e. system doesn’t analyze and display posts according to users’ interest perfectly;
• Users don’t prefer to read long posts even on topics of their interests and prefer short posts most of the time;
• Users prefer posts with images which have greater understanding than only with facts.

So, Twitter is used by the large number of users to share their posts, incorporate follow-ups, re-tweets etc. on variety of trending topics as tweets. Although it generates an idea of what is current, important and popular to twitter users, it becomes tedious to sift through the vast pool of tweets. In order to filter out certain specific tweets from millions of tweets, researchers have applied numerous Natural Language Processing (NLP) utilities such as Named Entity Recognition (NER) (Liu et al., 2011; Li et al., 2012; Cano et al., 2014; Derczynski et al., 2014; Godin et al., 2015; Rizzo et al., 2015; Belainine et al., 2016; Sikdar & Gambäck, 2016; Baksa et al., 2017; Lopez et al., 2017; Tran et al., 2017). Usually in researchers work NER based tweet topic extraction plays a vital role and seems to provide effective results as compared to any other approach. So, while taking advantage of NER, this research work filters out specific theme relevant tweets.

1.1. Brief About Named Entity Recognition

Named Entity Recognition (NER) (Nadeau & Sekine 2007; Ekbal & Bandyopadhyay 2008; Leaman & Gonzalez 2008; Srivastava et al., 2011; Rodríguez et al., 2012; Marrero et al., 2013; Fromreide et al., 2014; van Keulen & Habib 2014; Küçük 2015; Bhoi et al. 2016; Ekbal et al. 2016; Baksa et al. 2017) is defined as a non-trivial, automated sequence labelling task which comprises of identification, and classification of Named Entities (NEs). Identification of NE means marking the presence of a word/term/phrase i.e. names (proper noun/noun phrase) as NE in a given text. And, classification of NE means denoting the role of an identified NE into certain well-defined categories such as Person name, Location name, Organization name etc. NER is proved to be successful in vivid areas such as Question Answering (Lee et al., 2006), Machine Translation (Hermjakob et al., 2008), Automatic Text Summarization (Schifman et al., 2002), Information Retrieval (Khalid et al., 2008; Katta & Arora, 2015), Word Sense Disambiguation (Moro et al., 2014), Coreference Resolution (Bontcheva et al., 2002) and so on.

The research work which is discussed in this paper focuses towards tweets posted in Hindi Language. So, literature has been studied with respect to the work done to recognize NEs for an Indian Language. Various computational techniques, models and approaches are used by researchers to identify NEs in Hindi, Bengali, Tamil, Oriya and other Indian languages.

NER for Indian languages (Cucerzan & Yarowsky, 1999; Li & McCallum, 2003; Kumar & Bhattacharyya, 2006; Ekbal, & Bandyopadhyay, 2007; Ekbal & Bandyopadhyay, 2009; Patel et al., 2009; Saha et al., 2010; Ekbal & Bandyopadhyay, 2011; Ekbal et al., 2012; Saha et al., 2012; Sikdar et al., 2012; Saha & Ekbal, 2013; Jain et al., 2014; Nanda, 2014; Rao et al., 2015; Athavale et al., 2016; Ekbal et al., 2016), especially Hindi is thriving as a budding research topic for more than a decade. This situation has probably contributed to the fact that Hindi has wider perspective; encompasses a wider range of content, and so deserves more attention towards language-based research. Henceforth, Department of Electronics and Information Technology (DeITY)\(^1\), Government of India has initiated Information Retrieval Society of India (IRSI)\(^2\) to promote research & development in this regard. Hindi language is written in the Devanagari (Gupta et al., 2011) script and is considered as an official language of the Government of India, in addition to English. Hindi is also one of the 22\(^3\) scheduled languages of the State of India. Hindi is the lingua franca\(^4\) of the Hindi belt\(^5\) in India. And outside India, it is an official language in Fiji, and regional language in Mauritius, Trinidad and Tobago, Guyana, and Suriname. Hindi is the fourth most-spoken first language in the world, after Mandarin, Spanish and English. Hindi is highly inflectional, morphologically rich and primarily suffixing
Related Content

Imprecise Knowledge and Fuzzy Modeling in Materials Domain
www.igi-global.com/chapter/imprecise-knowledge-and-fuzzy-modeling-in-materials-domain/178393?camid=4v1a

Using Fuzzy Goal Programming Technique to Solve Multiobjective Chance Constrained Programming Problems in a Fuzzy Environment
www.igi-global.com/article/using-fuzzy-goal-programming-technique/63356?camid=4v1a

Rough Set-Based Neuro-Fuzzy System
Kai Keng Ang and Chai Quek (2009). Encyclopedia of Artificial Intelligence (pp. 1396-1403).
www.igi-global.com/chapter/rough-set-based-neuro-fuzzy/10422?camid=4v1a
Cyber Security Patterns Students Behavior and Their Participation in Loyalty Programs


[www.igi-global.com/article/cyber-security-patterns-students-behavior-and-their-participation-in-loyalty-programs/205573?camid=4v1a](http://www.igi-global.com/article/cyber-security-patterns-students-behavior-and-their-participation-in-loyalty-programs/205573?camid=4v1a)