Microsoft

Data Sciences Meet

28-June-2016 | Hyderabad | India

dSmL studio

NE-tagged News Headlines corpus creation

Avinash Kumar, Dhaval Patel, Nikita Jain (Indian Institute of Tech., Roorkee)

News headlines are short informational texts which often contain named entities. For example, the news headline 'Microsoft to acquire LinkedIn' contains two named entities – *Microsoft* and *LinkedIn*.

Benefits of NE-tagging news headlines:-

1. News headlines are continuous source of textual data. Tagging them will lead to a huge tagged corpora.
2. News headlines encapsulate information about the named entities. When tagged, the headline can be associated with the entity in a named entity repository.

Heuristic used: We use transliteration to identify the named entities. The heuristic we use is that if the translation of a word is in fact a transliteration, then the word is likely to be a named entity.

- साल transliterates to saal, साल translates to year
- मुंबई transliterates to Mumbai, मुंबई translates to Mumbai

The practice of providing an approximate English translation of the headline, in the URL of the webpage carrying the headline and the accompanying news story, is standard in many online news media sources. Table below provides a short list of such news websites.

News website	Language
http://aajtak.intoday.in	Hindi
http://www.patrika.com	Hindi
http://www.gujaratsamachar.com	Gujarati
http://tamil.oneindia.com	Tamil
http://www.sakshi.com	Telugu
http://www.anandabazar.com	Bengali
http://www.bbc.com/urdu	Urdu
http://malayalam.oneindia.com	Malayalam
http://abpmajha.abplive.in	Marathi
http://kannada.oneindia.com	Kannada

Table below shows the approximate translation of headlines in the URL of Hindi headlines. The third column provides the accurate English translation, provided by the authors, of the Hindi headline in the first column.

Hindi Headline	URL of the page carrying the story	Accurate English translation provided by authors
बैंकॉक में मर्डर कर भागा, दो साल बाद मुंबई में पकड़ा गया	http://aajtak.intoday.in/story/**man-held-for-murder-of-us-woman-in-bangkok-two-years-ago**-1-782001.html	Ran after committing murder in Bangkok, caught after two years in Mumbai
मुंबई हमले की तरह सिडनी में आतंकी हमला	http://aajtak.intoday.in/video/**sydney-terror-hostage-at-chocolate-cafe-simillar-to-mumbai-terror-attack**-1-791629.html	Terrorist attack in Sydney similar to Mumbai
आईपीएल-7 के यूएई चरण में मैक्सवेल, नरीन का जलवा	http://www.patrika.com/news/**maxwell-narine-make-it-large-at-uae-leg-of-ipl-7**/1004407	Maxwell and Narine rock at IPL-7 UAE phase
व्हाइट हाउस के बाहर मोदी समर्थकों ने किया गरबा	http://aajtak.intoday.in/video/**modi-supporters-performed-garba-outside-white-house**-1-781921.html	Modi supporters perform Garba outside the White House

The entire process of NE-tagging consists of two parts:

1. Parallel corpus creation
2. NE-tagging of the news headlines.

NE-Tagging the headlines in parallel corpus consists of three parts :

1. *Transliteration heuristic application*
2. *Remove English Lexicon Words*
3. *Identify Non-Transliterated Entity*

Transliteration heuristic application

We attempt to find the transliteration pairs in the parallel corpus obtained. Khapra et. al. 2014 describe a matching process to find transliteration pairs. The matching process is a language independent, four-step process to match a particular word written in language A, to its transliteration in language B. It checks boundary vowels, removes other vowels and then uses the consonant mapping table to determine if words are transliteration pairs.

क (ka)	k,q,c,**ck,kh,ch**	ञ	j,**nj**	द (da)	d	य (ya)	y,**NULL**	श्र (khsa)	s
ख (kha)	k,**kh**	ट (ta)	t,**th**	द् (d)	**j,g,z,s,jh**	र (ra)	r,**rh**	त्र (tra)	t,**tr**
ग (ga)	g,**gh**	ठ (tha)	t,**th**	ध (dha)	d,**dh**	ल (la)	l	ज्ञ (gnya)	j,g,**gy,gny**
घ (gha)	g,**gh**	ड (da)	d,**dh**	न (na)	n	व (va)	v,w	क्ष (shr)	s,**ksh**
ङ	g,**ng**	ढ (dha)	d,**dh**	प (pa)	p	श (sha)	s,**c,sh,t**	क्स	**x**
च (cha)	c,**ch**	ढ (dha)	d,**dh**	फ (fa)	f,p,**ph,gh**	ष (sha)	s,**sh,shh**	ी	**y**
छ (chha)	c,**ch,chh**	ण (na)	n	फ़ (f)	f,p,**ph,gh**	स (sa)	s,**sh**	ि	**y**
ज (jha)	j,g,z,s,**jh**	त (ta)	t,**th**	ब (ba)	b	स् (s)	**c,s,sh,t**	◌	**NULL**
ज़ (jha)	z,x,s	त् (t)	**c,ch**	भ (bha)	b,**bh**	ह (ha)	h,**gh**	◌	**n,m,NULL**
झ (jha)	j,s,z,**jh**	थ (tha)	t,**th**	म (ma)	m	ँ	**n,m**	ृ	**r**

However, transliteration heuristic has some problems :-

- भारत is the Hindi word for the country which is called India in English. (false negative)
- Named entity present in headline but not in translation (false negative)
- English words which have become a commonplace in the Hindi language (like मर्डर) are tagged (false positive)

Remove English Lexicon Words

The Webster's Unabridged English Dictionary, which contains about 118,830 words, is taken and transliterated into Hindi. This transliteration can be done locally through a simple java program, on any machine which contains the transliteration tools like Google Input Tools or Microsoft Indic Language Input Tool. All the named entities identified by the transliteration heuristic are searched in the transliterated lexicon. If found, the named entity can be marked as a false positive.

Identify Non-Transliterated Entity

The first headline in Table on first page contains the named entity मुंबई (Mumbai) while its approximate translation obtained from the URL does not contain Mumbai.

In order to identify such false negatives, we create a Gazetteer list (called *CorpusGazetteer*) which contains the named entities identified till now in all the headlines (after removing the English lexicon words). Then, we use *CorpusGazetteer* search on the headlines to identify the named entities which have been missed.

Intuition behind using CorpusGazetteer :-

- Named entity may be identified in headlines from other news sources.
- Popular named entity may be identified in headlines published after that day.
- Gazetteer lists available on web consists of a few hundred entities for Hindi. The *CorpusGazetteer* for 600,000 Hindi headlines contained about 31,000 named entities.