યુવકના મોત બાદ કાશ્મીરમાં ફરી હિંસા : ૧૨ ઘાયલ



सोनिया और मोदी भी भर चुके हैं राहुल जैसा फॉर्म

टाइम्स न्यूज नेटवर्क | Mar 16, 2015, 07.48AM IST

Lightweight System for NE-tagged News Headlines corpus creation

Avinash Kumar, Dhaval Patel, Nikita Jain
(Indian Institute of Technology Roorkee)
avinash0161@gmail.com, pateldha@us.ibm.com, nk27jain@gmail.com

Outline

- Motivation
- Research Problem
- Proposed solution
- Experiments
- Future Work

News headlines- small but rich

- Report emerging entities. A new singer finds his place in news headlines before Wikipedia.
- Keep track of trending entities. Use knowledge of headlines to enrich knowledge bases
- Level of association of two entities can be tracked through time by detecting their frequency of co-occurrence

Gandhi image on beer cans: US company apologizes

English headline

भूटान: शाही शादी में शामिल हुए राहुल गांधी

Hindi headline

બીગ બી સહિતના બોલીવૂડ સ્ટાર્સે મુંબઇ કોર્પોરેશનમાં ફરીયાદ કરી

Gujarati headline

We focus on .....

Gandhi image on beer cans : US company apologizes

भूटान : शाही शादी में शामलि हुए राहुल गांधी

Detect named entities (mentions)

Gandhi image on beer cans : US company apologizes

भूटान : शाही शादी में शामलि हुए राहुल गांधी

Research Problem

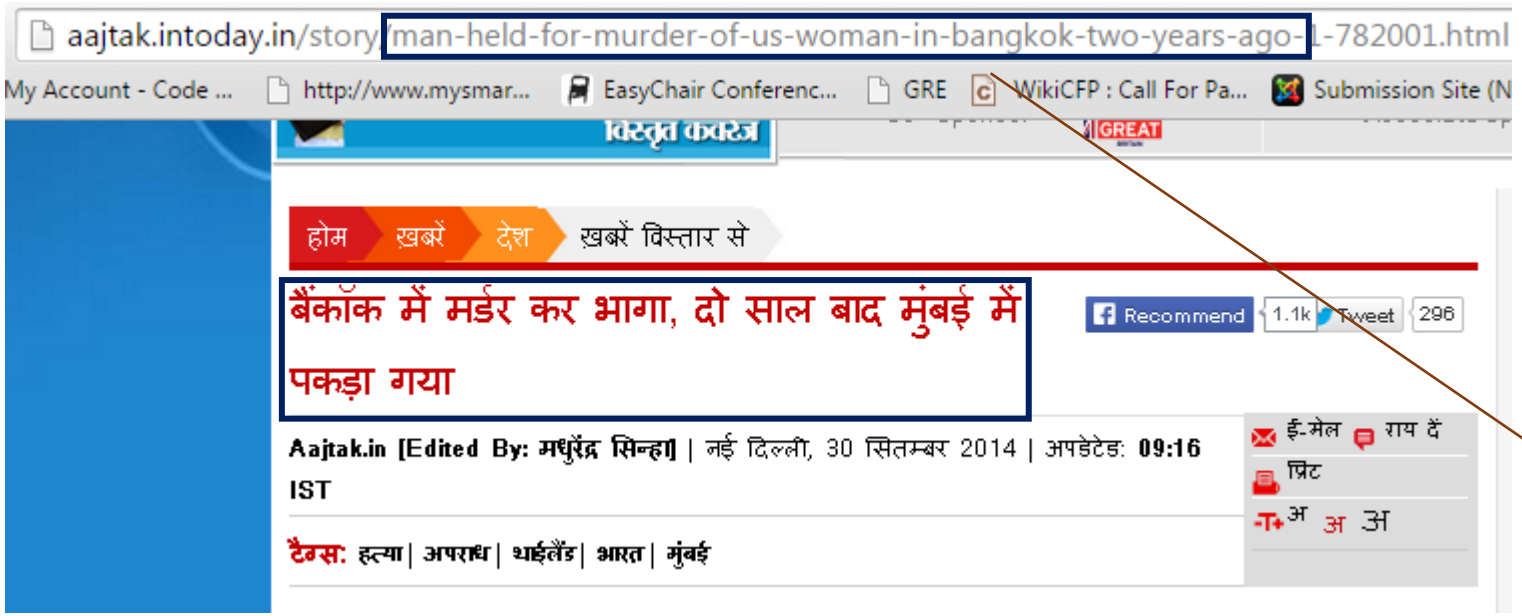
Mention detection in the case of Hindi and other Indian language news headlines

Two step process :

1. Parallel Corpus Creation of the headlines and their approximate English translation

2. NE-tagging the headlines

1. Parallel corpus creation



man-held-for-murder-of-us-woman-in-bangkok-two-years-ago

Courtesy : aajtak.indiatoday.in

1. Parallel corpus creation

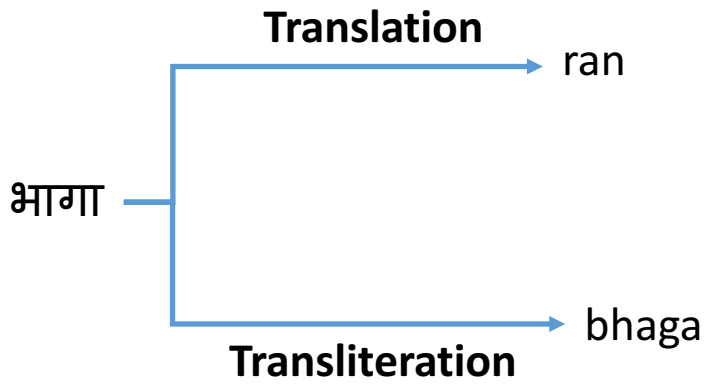
Hindi headlines	English translation from URL
बैंकॉक में मर्डर कर भागा, दो साल बाद मुंबई में पकड़ा गया	**man held for murder of us woman in bangkok two years ago**
मुंबई हमले की तरह सडिनी में आतंकी हमला	**sydney terror hostage at chocolate café similar to mumbai terror attack**
आईपीएल-7 के यूएई चरण में मैक्सवेल , नरीन का जलवा	**maxwell narine make it large at uae leg of ipl 7**

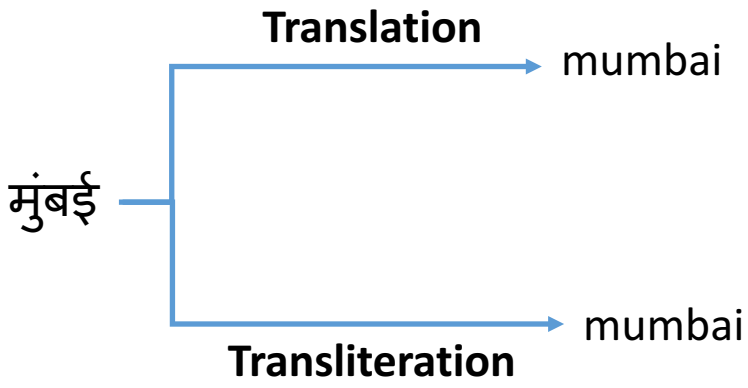
Table: A section of the parallel corpus for Hindi headlines*

*The entire corpus of 600,000 Hindi headlines and their translation obtained from the URL is shared at https://goo.gl/OUwMc3

2. NE tagging module

We use the feature of transliteration to tag the named entities in the headlines. The intuition used is that if the translation of a word is in fact a transliteration then the word is a named entity.





2. NE tagging module

Transliteration pairs

Hindi headlines	English translation from URL
बैंकॉक में मर्डर कर भागा, दो साल बाद मुंबई में पकड़ा गया	**man held for murder of us woman in bangkok two years ago**
मुंबई हमले की तरह सडिनी में आतंकी हमला	**sydney terror hostage at chocolate café similar to mumbai terror attack**
आईपीएल-7 के यूएई चरण में मैक्सवेल , नरीन का जलवा	**maxwell narine make it large at uae leg of ipl 7**

Table: Named entities mostly have their transliteration in the obtained English translation

2. NE tagging module

Hindi headlines	English translation from URL
बैंकॉक में मर्डर कर भागा, दो साल बाद मुंबई में पकड़ा गया	**man-held-for-murder-of-us-woman-in-bangkok-two-years-ago**
मुंबई हमले की तरह सडिनी में आतंकी हमला	**sydney terror hostage at chocolate café similar to mumbai terror attack**
आईपीएल-7 के यूएई चरण में मैक्सवेल , नरीन का जलवा	**maxwell-narine-make-it-large-at-uae-leg-of-ipl-7**

Table: Named entities mostly have their transliteration in the obtained English translation

2. NE tagging module

Hindi headlines	English translation from URL
बैंकॉक में मर्डर कर भागा, दो साल बाद मुंबई में पकड़ा गया	**man-held-for-murder-of-us-woman-in-bangkok-two-years-ago**
मुंबई हमले की तरह **सडिनी** में आतंकी हमला	**sydney terror hostage at chocolate café similar to mumbai terror attack**
आईपीएल-7 के यूएई चरण में मैक्सवेल , नरीन का जलवा	**maxwell-narine-make-it-large-at-uae-leg-of-ipl-7**

Table: Named entities mostly have their transliteration in the obtained English translation

2. NE tagging module

Hindi headlines	English translation from URL
बैंकॉक में मर्डर कर भागा, दो साल बाद मुंबई में पकड़ा गया	**man-held-for-murder-of-us-woman-in-bangkok-two-years-ago**
मुंबई हमले की तरह सडिनी में आतंकी हमला	**sydney terror hostage at chocolate café similar to mumbai terror attack**
आईपीएल-7 के यूएई चरण में मैक्सवेल , नरीन का जलवा	**maxwell-narine-make-it-large-at-uae-leg-of-ipl-7**

Table: Named entities mostly have their transliteration in the obtained English translation

2. NE tagging module

Hindi headlines	English translation from URL
बैंकॉक में मर्डर कर भागा, दो साल बाद मुंबई में पकड़ा गया	**man-held-for-murder-of-us-woman-in-bangkok-two-years-ago**
मुंबई हमले की तरह सडिनी में आतंकी हमला	**sydney terror hostage at chocolate café similar to mumbai terror attack**
आईपीएल-7 के **यूएई** चरण में मैक्सवेल , नरीन का जलवा	**maxwell-narine-make-it-large-at-uae-leg-of-ipl-7**

Table: Named entities mostly have their transliteration in the obtained English translation

2. NE tagging module

Hindi headlines	English translation from URL
बैंकॉक में मर्डर कर भागा, दो साल बाद मुंबई में पकड़ा गया	**man-held-for-murder-of-us-woman-in-bangkok-two-years-ago**
मुंबई हमले की तरह सडिनी में आतंकी हमला	**sydney terror hostage at chocolate café similar to mumbai terror attack**
आईपीएल-7 के यूएई चरण में **मैक्सवेल** , नरीन का जलवा	**maxwell-narine-make-it-large-at-uae-leg-of-ipl-7**

Table: Named entities mostly have their transliteration in the obtained English translation

2. NE tagging module

Hindi headlines	English translation from URL
बैंकॉक में मर्डर कर भागा, दो साल बाद मुंबई में पकड़ा गया	**man-held-for-murder-of-us-woman-in-bangkok-two-years-ago**
मुंबई हमले की तरह सडिनी में आतंकी हमला	**sydney terror hostage at chocolate café similar to mumbai terror attack**
आईपीएल-7 के यूएई चरण में मैक्सवेल , **नरीन** का जलवा	**maxwell-narine-make-it-large-at-uae-leg-of-ipl-7**

Table: Named entities mostly have their transliteration in the obtained English translation

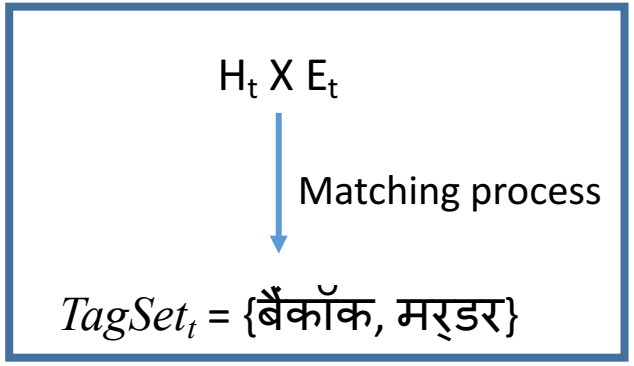
How to identify such words ?

2. NE tagging module (Phase 1)

- We use a language independent matching process described in Khapra et. al., 2014 to find those words in the headlines whose transliterations occur in the approximate English translation in the headlines.

बैंकॉक में मर्डर कर भागा, दो साल बाद मुंबई में पकड़ा गया → H_t = {**बैंकॉक**, में, **मर्डर**, कर, भागा, दो, साल, बाद, मुंबई, में, पकड़ा, गया}

man held for murder of us woman in bangkok two years ago → E_t ={man, held, for, **murder**, of, us, woman, in, **bangkok**, two, years, ago}



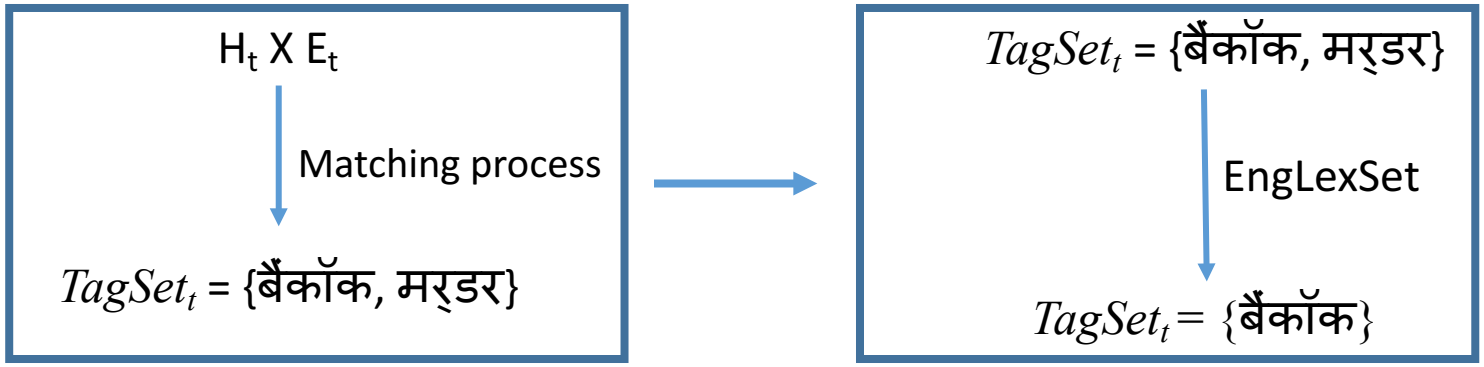
2. NE tagging module (Phase 2)

- Many words of English lexicon have become a commonplace in other languages.

Eg: मर्डर (murder), ट्रेन (train), पेपर (paper), बस (bus), रेल (rail), लेट (late)

To remove such words from TagSet :-

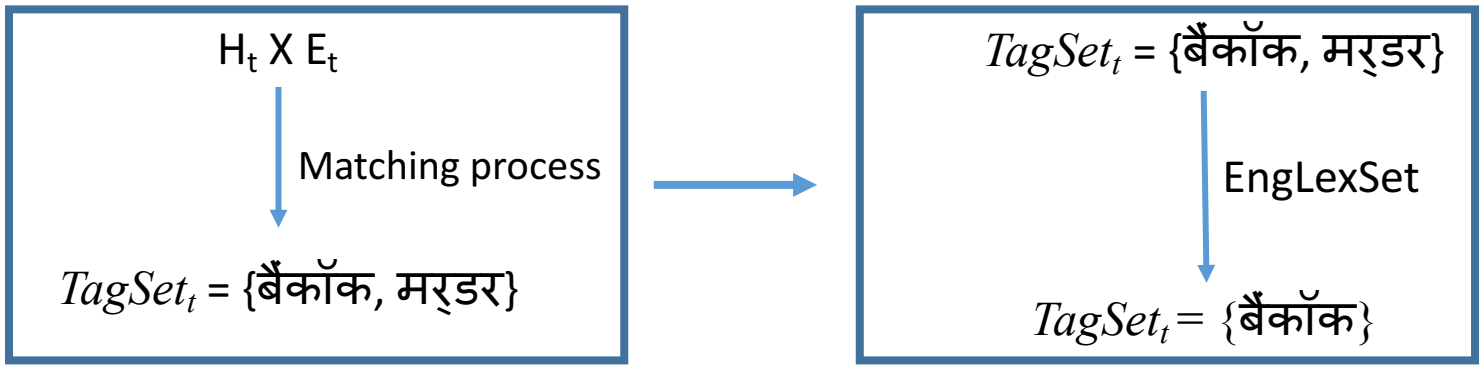
- Use Webster's English lexicon and transliterate it to the language of the headlines (use Google input tools)



2. NE tagging module (Phase 3)

- The named entity मुंबई(mumbai) has still not been tagged in headline.

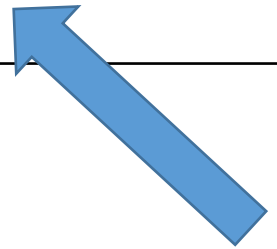
Hindi headlines	English translation from URL
बैंकॉक में मर्डर कर भागा, दो साल बाद मुंबई में पकड़ा गया	man-held-for-murder-of-us-woman-in-bangkok-two-years-ago



2. NE tagging module (Phase 3)

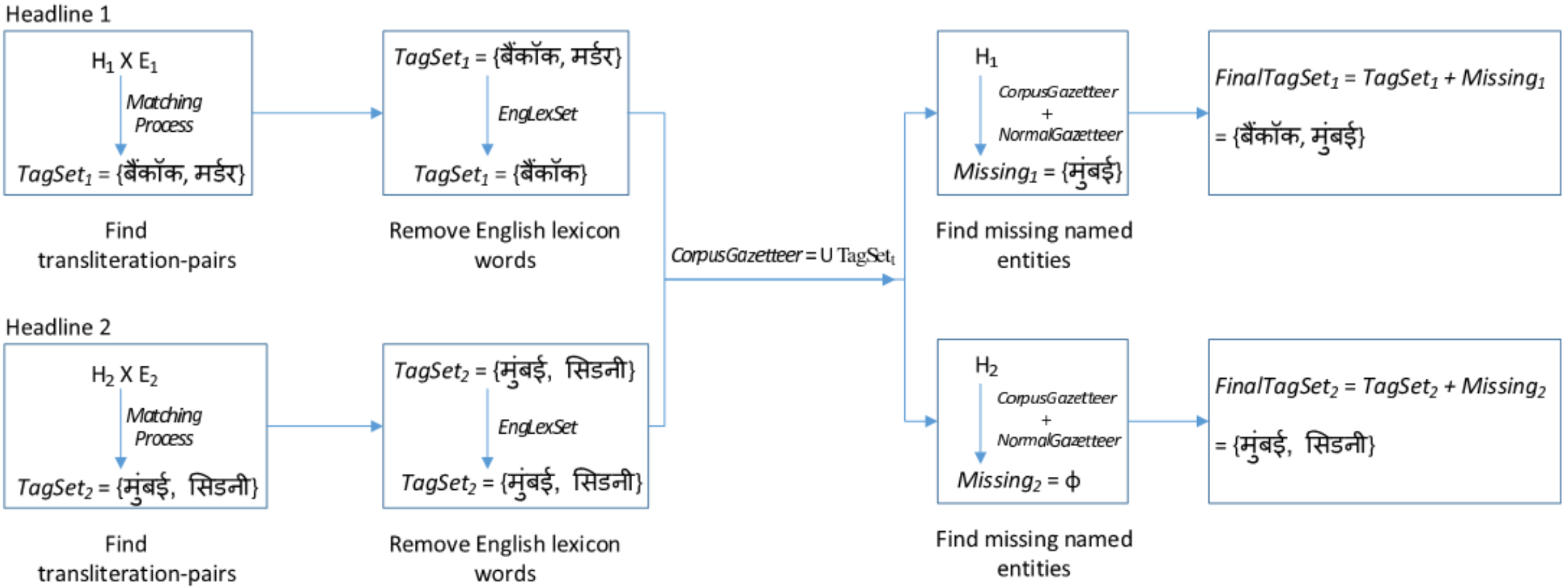
- The named entity मुंबई(mumbai) has still not been tagged in headline.

Hindi headlines	English translation from URL
बैंकॉक में मर्डर कर भागा, दो साल बाद मुंबई में पकड़ा गया	**man-held-for-murder-of-us-woman-in-bangkok-two-years-ago**
मुंबई हमले की तरह सडिनी में आतंकी हमला	**sydney terror hostage at chocolate café similar to mumbai terror attack**



Use this headline to find **मुंबई** in first headline

2. NE tagging module

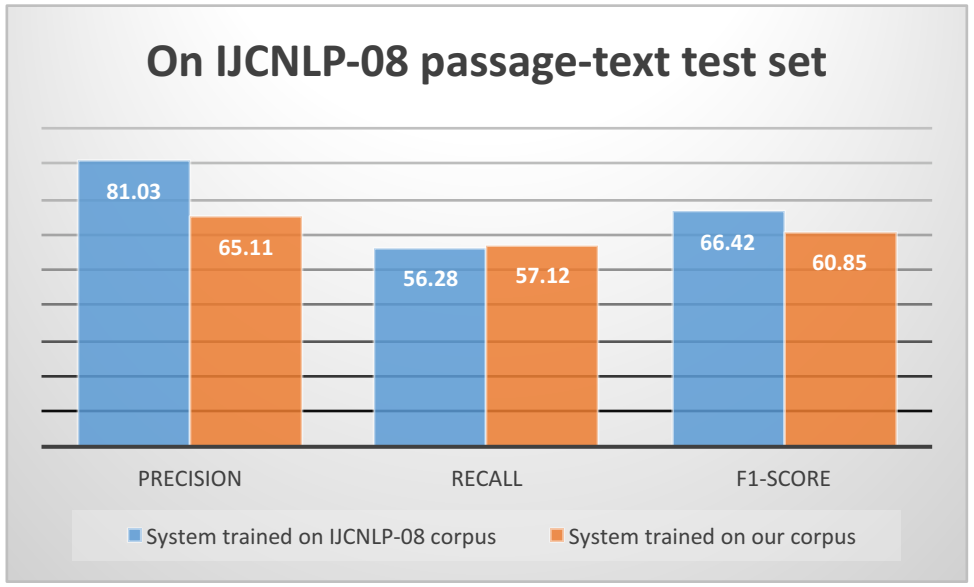


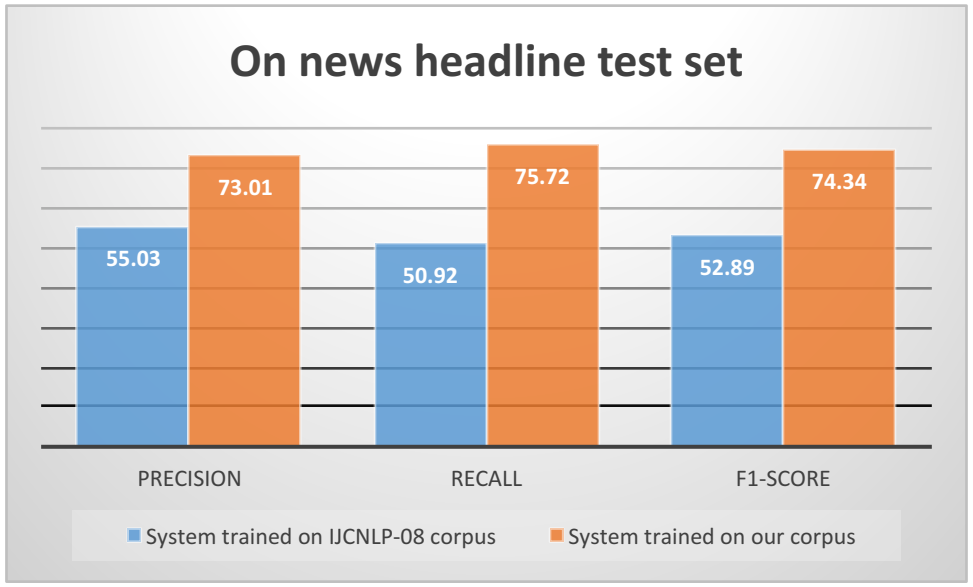
Experiments- Accuracy of NE-tagging process

- We randomly take 10000 headlines from our Hindi NE-tagged corpus of 600,000 headlines and manually annotate the named entities present in them
- Ground truth contained 13344 named entities. Our NE-tagging process annotated 14467 named entities, out of which 10471 were present in the ground truth.
- Precision = 72.38%, Recall = 78.47%, F1-score = 75.30%

Experiments- Comparison with gold-standard corpus for training NEI models

- NEI models were trained on gold standard corpus from IJCNLP-08 and corpus generated by our system. The features used to train the models were the same as used by Saha et al., 2008, who achieved the best results for Hindi in IJCNLP-08 NER-SSEA workshop.
- The two models were tested on test set from IJCNLP-08 and on a test-set of 10k news headlines created by us.





Headlines, URLs and accurate English translations for other Indian languages

Language	Headlines	URL of the page carrying the story	Accurate English translation provided by authors
Gujarati	યુવકના મોત બાદ **કાશ્મીરમાં** ફરી હંસિા : ૧૨ ઘાયલ	http://www.gujaratsamachar.com/index.php/articles/display_article/national/national-after-the-death-of-youth-violence-in-**kashmir**-12-injured	12 people injured in violence following death of youth in Kashmir
Bengali	**রতন-সাইরাস** যুদ্ধে হাতছাড়া হতে পারে সিংহাসন, উদ্বগ্নি পারসিরা	http://www.anandabazar.com/national/**ratan**-tata-vs-**cyrus**-mistry-may-be-missing-out-on-throne-parsi-community-concerned-dgtlx-1.508354	Due to Ratan-Cyrus war, the CEO position is in conflicts, say agitated parsees
Marathi	**दल्लिीत** वषिारी धुरक्याचा कहर, घराबाहेर नघिणंही कठीण	http://abpmajha.abplive.in/india/pollution-situation-in-**delhi**-is-going-worse-310633	Delhi pollution crisis, going out of home becomes difficult
Tamil	**புதிய கல்வி கொள்கையில் அரசியல் இல்லை.. தேசே வளர்ச்சி மட்டும்தான்... பிரகாஷ் ஜவடகேர்**	http://tamil.oneindia.com/news/tamilnadu/new-education-policy-development-**prakash-javadekar**-266498.html	There is no politics in new educational policy. Only growth of the country – Prakash Javadekar
Telugu	**తెలంగాణలో** ఆత్మహత్యలు ఆగిపోవాలి	http://www.sakshi.com/news/telangana/stop-suicides-in-telangana-419770?pfrom=home-**telangana**-news	Suicides should stop in Telangana

Other advantages of our proposed system

- **Corpus not restricted to any field**: Unlike corpora like Health-ILCI, Tourism-EIMLT
- **Automated production of huge number of transliteration pairs**

- **Automated creation of gazetteer lists and inclusion of emerging entities**

- **Corpus for machine translation**: The news headlines and their approximate English translations can be used to learn word-alignment models using softwares like GIZA++. This can be used for statistical machine translation of languages.

Future Work

- Currently, in order to keep the system lightweight, we haven’t considered the POS tag information of the word when removing it in the Remove English Lexicon Words stage of our process. Although this step of removing English lexicon words is necessary, it also ends up removing words like ‘speaker’, which in some cases might refer to ‘Speaker of the Parliament’.
- An implementation of the streaming algorithm version of our method.

THANK YOU

Any Questions ?