Making Headlines in Hindi: Automatic English to Hindi News Headline Translation

Aditya Joshi[1,2] Kashyap Popat[2] Shubham Gautam[2] Pushpak Bhattacharyya[2]
[1]IITB-Monash Research Academy, IIT Bombay
[2]Dept. of Computer Science and Engineering, IIT Bombay
{adityaj,kashyap,shubhamg,pb}@cse.iitb.ac.in

Abstract

News headlines exhibit stylistic peculiarities. The goal of our translation engine '*Making Headlines in Hindi*' is to achieve automatic translation of English news headlines to Hindi while retaining the Hindi news headline styles. There are two central modules of our engine: the modified translation unit based on Moses and a co-occurrence-based post-processing unit. The modified translation unit provides two machine translation (MT) models: phrase-based and factor-based (both using in-domain data). In addition, a co-occurrence-based post-processing option may be turned on by a user. Our evaluation shows that this engine handles some linguistic phenomena observed in Hindi news headlines.

1 Introduction

'*Making Headlines in Hindi*' is a web-based translation engine for English to Hindi news headline translation. Hindi[1] is a widely spoken Indian language and has several news publications. The aim of our translation engine is *to translate English news headlines to Hindi preserving the content as well as Hindi news headline structure to the extent possible*. The engine is based on Moses[2] and has two central parts: modified translation unit and a co-occurrence based post-processing unit. The modified translation unit consists of phrase-based MT (Koehn et al., 2003)) and factor-based MT (Koehn et al., 2007). The automatic post-processing module performs co-occurrence-based replacement for correct sense translation of words by replacing translation of a word with the most frequently co-occurring translation candidate. This paper is organized as follows. Section 2 presents challenges of translating news headlines. Section 3 describes the UI layout. Section 4 discusses technical details of the modified translation unit while section 5 describes the post-processing module that uses co-occurrence-based replacement of words. Finally, Section 6 presents an evaluation of the engine while section 7 concludes our work.

2 Challenges of News Headline Translation

Hindi news headlines have stylistic features that pose challenges to translation as follows:

1. **S-V-O order**: Hindi news headlines often follow the S-V-O order as opposed to S-O-V as commonly seen in Hindi sentences. A common news headline is '*अब तिहाड़ जेल में बिस्कुट बनाएंगे चौटाला (ab tihaaD jel mein biskooT banayenge chauTala; Now Chautala will make biscuits in Tihar jail)*' where the verb '*बनाएंगे (banayenge; will make)*' preceeds the object '*चौटाला (chauTala; Chautala)*'.

2. **Numbers for people**: Use of numbers to indicate a group of people, like in the case of English news headlines, is also common in Hindi news headlines. For example, the word '*Five*' in '*Five held for molesting woman*' stands for five people.

3. **Preferred choice of words**: Words that are commonly used in news headlines are often different from accurate translations. For example, '*RBI*' (abbreviation for 'Reserve Bank of India') is common in English news headlines - however, instead of using its transliterated form, news headlines tend to

[1]https://en.wikipedia.org/wiki/Hindi
[2]http://www.statmt.org/moses/

translate it to '*रिज़र्व बैंक (rizarv bank; Reserve Bank)*' in Hindi news headlines.

4. **Missing verbs**: Often, verbs are also dropped as in the case of '*महाकुंभ में अजब-गजब संतो की भीड़ (mahakumbh mein ajab-gajab santon kii bheeD; Herds of fascinating saints in Mahakumbh (fair))*' where a form of the word 'be' has been dropped.

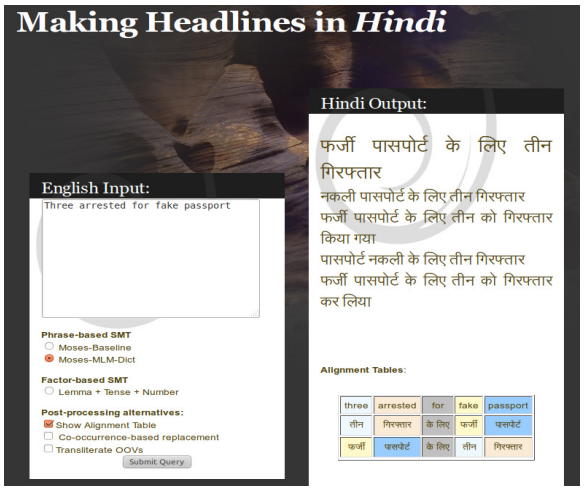

Figure 1: Making Headlines in Hindi: Snapshot of Output

3 UI Layout

The interface of the engine is divided into two vertical blocks for clarity: one for input and another for output. The input to the translation engine consists of:

(a) Text area for English news headline(s),

(b) Option to select Phrase-based v/s Factor-based model,

(c) Checkboxes for co-occurrence based replacement, transliteration for OOVs and displaying alignment table for the output: Each of these options can be turned on/off.

While one out of the two options in (b) must be selected, check-boxes in (c) are optional. Each of the components stated above are described in Section 4.

The output consists of:

(a) The *best five translations* obtained in Hindi

(b) A *color-coded alignment table* in case the option to display the alignment table : This helps to understand how each word got translated and then reordered.

(c) *Time taken* for translation

Figure 1 shows a snapshot of the UI. Moses-Baseline indicates the naive translation engine while Moses-MLM-Dict is the modified phrase model.

4 Modified Translation Unit

We implemented two translation models: phrase-based and factor-based. The training corpus consisted of parallel corpus obtained from (a) Gyan-nidhi[3] consisting of 2,27,123 sentences and (b) Mahashabdkosh[4] consisting of 46,825 judicial sentences. To transliterate out-of-vocabulary words, we modified transliteration engine provided by Chinnakotla et al. (2010). The original transliteration was trained for Hindi to English transliteration. For the purpose of our engine, we re-trained this model for English to Hindi transliteration. This section describes each of these components.

4.1 Phrase-based Model

The **Phrase-based MT** model was trained using Moses by (Koehn et al., 2007). In order to improve the quality of translation, we modify different components of the model in two ways. To preserve sentence order, we use a **modified language model** - a language model trained using in-domain data consisting of 20,220 news headlines from BBC Hindi website[5] and 2,02,335 news headlines from Dainik Bhaskar[6] archives of 2010 and 2011. The fact that this modified language model is a better fit to the target data is highlighted by the perplexity value obtained using SRILM toolkit by (Stolcke, 2002). For bi-grams, the perplexity of the Dainik Bhaskar corpus with a test news headline corpus was 434.06 while the perplexity of corpus consisting of tourism documents was 1205.58. Similar trend was observed in case of tri-grams. To enrich the translation mapping table available, we added a **bilingual dictionary** to the parallel corpus used for training the translation

[3]http://www.cdacnoida.in/snlp/digital_library/gyan_nidhi.asp
[4]http://www.e-mahashabdkosh.cdac.in/
[5]http://www.bbc.co.uk/hindi/
[6]http://www.bhaskar.com/

model. This bilingual dictionary was downloaded from CFILT, IIT Bombay[7]. This dictionary contains a total of 1,28,240 mappings and includes words as well as phrases. The fact that this dictionary enriches translations is observed in the case of a news headline containing the word '*catch-22*'. This word does not occur in the parallel news headlines. However, it gets correctly translated to '*जटिल (jaTil)*' according to the entry in the dictionary.

4.2 Factor-based Model

Our **Factor-based MT** model uses a set of factors along with words for translation. The factors used on source and target side are as follows.

1) On the source side, we use POS, lemma, tense and number. The POS tags are obtained from Stanford POS tagger[8] while the lemma are obtained from MIT Wordnet stemmer[9]. Tense and number are derived from POS tags.

2) On the target side, we use CFILT hybrid POS tagger[10] to obtain POS tags.

The factors are combined using options available in Moses. The lemma, tense and number on the source side generate the translated word on the target side. On the target side, words generate POS features. By generating best possible translations using a POS-based target language model, we hope to obtain translations in a POS order best suited to the news headline domain.

5 Post-processing: Co-occurrence-based Replacement

The engine provides an optional **co-occurrence based replacement** strategy to post-process the output. A manual evaluation showed that 14 out of 50 headlines were incorrect because of incorrect sense of one or more words. To overcome this problem, we implemented a post-processing strategy that automatically edits output obtained from the MT model using co-occurrence statistics as found in the in-domain news headline corpus. To elaborate how this works, consider the English news headline '*crpf jawan held on molestation charge*'. The translation obtained was '*सीआरपीएफ़ जवान पर आयोजित उत्पीड़न चार्ज (crpf jawaan par aayojit utpiDan chaarj; molestation charge organized on crpf jawan)*'. The word '*held*' gets translated to '*आयोजित (aayojit; organized/conducted)*' as opposed to '*गिरफ़्तार (giraftar; arrested)*'. The language model relies on n-grams and hence, does not take into account the correct sense of words in cases where the words do not occur together. For this purpose, we implemented a post-processing strategy that considers co-occurrence statistics of a target word with all other words in the sentence to find the best sense translation. In case of the above example, using the co-occurrences in a news headline corpus, we select the sense of '*held*' in Hindi which occurs most frequently with other words and replace the word with this translation. We do not consider co-occurrence statistics for function words. We understand that the above strategy does not work in the case of inflected forms of words in Hindi.

6 Evaluation

We evaluated the engine using a test set of 787 headlines downloaded from the website of a popular English daily, The Hindu[11] and manually translated into Hindi by native speakers. A BLEU score of 13.40 is obtained for phrase-based MT and 5.73 for factor-based MT. In order to understand how the engine performs for different kinds of linguistic phenomena, we also performed a qualititative evaluation of the output. The following are examples of output from our engine. They handle different linguistic phenomena as follows:

1. **Ambiguity**:
   Input: Industrialist remembered
   Output: उद्योगपति को याद किया
   (udyogpati ko yaad kiyaa;
   Industrialist remembered)
   The input sentence in domains other than news headlines is ambiguous as it could mean that either an industrialist was remembered or an industrialist remembered something. In case of news headline, however, the former holds true. This is correctly reflected in the Hindi translation.

2. **S-V-O order**:
   Input: Now Jaganmohan will make biscuits in jail
   Output: अब जगमोहन करेंगे बिस्कुट जेल में
   (ab jaganmohan karenge biskoot jel mein;
   Now Jaganmohan will do biscuits in jail)

[7] http://www.cfilt.iitb.ac.in

[8] http://www-nlp.stanford.edu/software/tagger.shtml

[9] http://projects.csail.mit.edu/jwi/api/edu/mit/jwi/morph/WordnetStemmer.html

[10] http://www.cfilt.iitb.ac.in/Tools.html

[11] www.thehindu.com

The verb '*will do*' gets placed correctly in the target sentence thus preserving the verb order. However, the translation '*karenge (will do)*' is incorrect and must be '*banaange (will make)*'.

3. **Numbers for people**:
   Input: Five killed in bomb blast
   Output: पांच बम विस्फोट में मारे गये
   (paanch bum visfot mein maare gaye;
   Five killed in bomb blast)
   The output sentence is a perfect translation and correctly translates 'five' as 'paanch'. However, the news headline order is not retained in this case.

4. **Missing verbs**:
   Input: Veteran journalist dead
   Output: दिग्गज पत्रकार मृत
   (diggaj patrakaar mrut;
   Veteran journalist dead)
   The output sentence is a perfect translation although a form of '*be*' is absent in the source sentence.

5. **Translation of idioms**:
   Input: Croatia and Serbia bury the hatchet
   Output: क्रोटिया और सर्बिया जगड़ा ख़तम करना
   (kroatia aur serbia jhagDa khatam karna;
   Croatia and Serbia do-end-quarrel)
   The idiom '*bury the hatchet*' gets correctly translated to '*जगड़ा ख़तम करना; jhagDa khatam karna; to end a quarrel*' as a complete entity. This is a direct mapping from the bilingual dictionary and does not have the correct inflection.

6. **Sense correction due to co-occurrence based replacement**:
   Input: No hike in AMU tuition fees
   Moses-MLM-Dict: amu अध्यापन फ़ीस में कोई वृद्धि नहीं
   (amu adhyaapan fees mein koi pad-yaatra;
   hike (trek) in AMU tuition fees)
   Moses-CoOcc: amu शिक्षण क्षेत्र में कोई वृद्धि नहीं
   (amu shikshan fees mein koi vriddhi;
   hike (increase) in AMU tuition fees)
   We observe that our post-processing unit improves the output in some cases. The original output translates '*hike*' as '*पदयात्रा (pad-yaatra ; hike)*'. The co-occurrence-based replacement unit identifies and corrects the sense to '*वृद्धि (vriddhi; increase)*'. We understand that the '*no*' gets missed out in the translation.

7 Conclusion & Future Work

We presented '*Making headlines in Hindi*', a translation engine that aims to translate English news headlines to Hindi while preserving news headline styles in the target language. Our engine includes a phrase-based model and a factor-based model. The phrase-based model uses an in-domain language model and a bilingual dictionary. The factor-based model uses factors like POS, lemma, tense and number. In addition, we also described our post-processing strategy that performs co-occurrence-based replacement of words to obtain correct sense of target language words. An evaluation of the output of our translation engine shows that it performs well for many linguistic styles used in Hindi news headlines.

The co-occurrence-based strategy is naive. As a future work, co-occurrence-based strategy can be improved to incorporate inflections of words. Also, other approaches to improve translation quality may be considered.

References

Manoj Kumar Chinnakotla, Om P. Damani and Avijit Satoskar. 2010. Transliteration for Resource-Scarce Languages. *Proc. ACM Trans. Asian Lang. Inf. Process.*,

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation *Proc. of ACL 2007, demonstration session*, Prague, Czech Republic

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. *Proc. of EMNLP-CoNLL 2007*, Prague, Czech Republic

Philipp Koehn and Franz Josef Och and Daniel Marcu,. 2003. Statistical phrase-based translation *Proc. of NAACL 2003*, Edmonton, Canada

A. Stolcke. 2002. SRILM - An extensible language modeling toolkit. *Proc. International Conference on Spoken Language Processing, vol. 2*