

Data Issues in English-to-Hindi Machine Translation



Ondřej Bojar, Pavel Straňák, Daniel Zeman
 Univerzita Karlova v Praze, Ústav formální a aplikované lingvistiky
 Malostranské náměstí 25, CZ-11800 Praha
 {bojar|stranak|zeman}@ufal.mff.cuni.cz



<http://ufal.mff.cuni.cz/umc/>

Abstract

Statistical machine translation to morphologically richer languages is a challenging task and more so if the source and target languages differ in word order. Current state-of-the-art MT systems thus deliver mediocre results. Adding more parallel data often helps improve the results; if it does not, it may be caused by various problems such as different domains, bad alignment or noise in the new data. We evaluate several available parallel data sources and provide cross-evaluation results on their combinations using two freely available statistical MT systems. We demonstrate various problems encountered in the data and describe automatic methods of data cleaning and normalization. We also show that the contents of two independently distributed data sets can unexpectedly overlap, which negatively affects translation quality. Together with the error analysis, we also present a new tool for viewing aligned corpora, which makes it easier to detect difficult parts in the data even for a developer not speaking the target language.

String	WX Equivalent	Should be
इन्डोमेटिऑन चोमिस्सिओनेर भाषा; ऋ-ऊण्डलुष-	Information Commissioner bhaṣa; r-ūṅṣh-	इन्फोमेशन कोमिशनर Q-UNSCR-
		informešana komišanera danda (sent. end) ???

Table 2: WX encoding is a 1-1 mapping to Latin letters, frequently used to encode Indian languages. If the original text was in one known Indian script (Devanagari in the case of Hindi texts), the original text can be completely reconstructed from the WX-encoded text. However, if the original text contained more than one script (most notably, embedded English words in Latin script), all script changes must be encoded using escape sequences. The Tides corpus apparently has been encoded in WX at some stage but without the embedded Latin letters being taken care of. During reverse conversion to Devanagari, embedded Latin was unfortunately treated as WX-encoded Hindi; so were some control sequences, too.

प्रादेशिक - जनसंख्या बंगाली बंगलादेश ह्यपूर्वी बंगालह से आए अधिकांश विस्थापित दक्षिण अंडमान, नेल, हैवलाक, मध्य अंडमान, उंरI AMDmaana tqaa ilaiTla AMDmaana maoM basaae gae .
 pradesika - janasankhya bangali bangaladesha (hya)purvī bangāla(h) se āe adhikāṁśa viśthāpita dakṣiṇa aṅḍamāna, nel, haivalāka, madhya aṅḍamāna, uṅarI AMDmaana tqaa ilaiTla AMDmaana maoM basaae gae .
 Regional Groups Most of the refugees from Bangladesh (East Bengal) were settled in South Andaman , Neil , Havelock , North Andaman , Middle Andaman and Little Andaman .

Table 3: Broken transcription in the middle of a sentence occurs in more than 200 instances. The Latin characters are not in the WX encoding. Highlighted in red are nonsense Devanagari characters corresponding to parentheses, also lost in conversion process.

English	Hindi
standards	स्टैंडर्ड्स स्टैंडर्डस स्टैंडर्ड्स

Table 4: Nonuniform transcription of English loanwords makes data sparser and complicates experiment evaluation.

English	Hindi/Persian	Hindi/Sanskrit
language	ज़बान	भाषा
book	किताब	पुस्तक
newspaper	अख़बार	समाचार-पत्र
beautiful	खूबसूरत	सुन्दर
meat	गोश्त	मांस
thank you	शुक्रिया	धन्यवाद

Table 5: Sets of synonyms from two sources: original Sanskrit vocabulary vs. Perso-Arabic vocabulary. This complicates translation selection, too.

Text Normalization

- Convert the text to fully decomposed Unicode. For instance, any DEVANAGARI LETTER FA will be replaced by the sequence DEVANAGARI LETTER PHA, DEVANAGARI SIGN NUKTA. Note that both strings are identical in appearance.
- Replace Devanagari digits ("० १ २ ३ ४ ५ ६ ७ ८ ९") by European digits ("0123456789").
- Replace *danda* ("।"), *double danda* ("।।") and the Devanagari abbreviation sign ("•") by period (".").
- Replace *candrabindu* by *anusvar*, e.g. "पॉच" by "पांच".
- Remove all occurrences of *nuksa*, effectively replacing "कखगजडढफ" by "कखगजडढफ".
- Remove various control characters (yes, they occur in the data!), zero-width joiners and the like.
- Replace non-ASCII punctuation by their ASCII counterparts, i.e. "—" by "-".

हिंदी

Error Analysis

<https://wiki.ufal.ms.mff.cuni.cz/user:zeman:addicter>

For

Examples of the word in the r data: The word 'For' occurs in 10 sentences. This is the sentence number 217 in file S.

For	many	years	it	was	the	single	largest	earner	of	foreign	exchange
लक	अनेक	वर्षों	.	यह	उद्योग	सबसे	ही अकेला	अधिक	अर्जक	विदेशी	मुद्रा
0-2	1-0	2-1	3-3	4-4	5-5	6-8	7-6 7-7	8-9	9-12	11-10	12-11
13-15											

Alignment summary

Table 9: Out-of-vocabulary rate. Number of tokens and types encountered in test / development data that were not known from the training data. English and Hindi evaluated separately. Individual columns represent various sets of training data, rows correspond to test data sets. The figures clearly show that non-Tides data are not able to significantly reduce the OOV rate. The Tides training data is much larger than all the other sets, and its domain is a better match for the Tides test and development data.

OOV	tokens unseen in train					types unseen in train				
	Tides	Tides + DP	Tides + dict	Tides + DP + dict	All - Tides	Tides	Tides + DP	Tides + dict	Tides + DP + dict	All - Tides
Tides-test-en	369	348	363 (1.336%)	343 (1.262%)	2429 (8.940%)	643	343	357 (6.011%)	338 (5.691%)	1901 (32.009%)
Tides-test-hi	839	830	836 (2.926%)	828 (2.898%)	3310 (11.584%)	642	633	639 (10.882%)	631 (10.746%)	2465 (41.979%)
Tides-dev-en	464	421	462 (2.055%)	419 (1.863%)	1873 (8.330%)	459	418	457 (8.167%)	416 (7.434%)	1608 (28.735%)
Tides-dev-hi	619	607	618 (2.537%)	606 (2.487%)	2661 (10.922%)	580	568	579 (10.262%)	567 (10.050%)	2129 (37.735%)

- A dataset originally collected for the DARPA-TIDES surprise-language contest in 2002, later refined at IIIT Hyderabad and provided for the NLP Tools Contest at ICON 2008.
- A journalist Daniel Pipes' website (<http://www.danielpipes.org/>) limited-domain articles about the Middle East. Written in English, many of them translated to up to 25 other languages.
- Monolingual, parallel and annotated corpora for fourteen South Asian languages (including Hindi) and English. (ELDA)
- English-Hindi-Marathi-UNL parallel corpus from Resource Center for Indian Language Technology Solutions (http://www.cfit.iitb.ac.in/download/corpus/parallel/agriculture_domain_parallel_corpus.zip).
- Aligarh District ... अलीगढ़ जिला
- Allahabad University ... इलाहाबाद विश्वविद्यालय
- Amit Vilasrao Deshmukh ... अमित तिलासराव देशमुख
- mountain ... पहाड़
- mountain ... पर्वत
- Mr ... श्री

Corpus	Sentences	En Tokens	Hi Tokens
Tides.train	50,000	1,226,144	1,312,435
Tides.dev	1,000	22,485	24,363
Tides.test	1,000	27,169	28,574
Daniel Pipes	6,761	176,392	122,108
Emille	3,501	55,660	71,010
ACL 2005	3,441	55,967	69,349
Agrocorpus	527	11,977	7,156
Wiki NE 2008	853	1,666	1,394
Wiki NE 2009	774	1,397	1,259
Shabd full	32,159	35,999	44,546
Shabd filt	1,422	1,470	1,422

Table 1: Overview of parallel corpora and dictionaries. Counts reflect clean subsets we used for experiments.

Line No.	English Sentence	Line No.	Hindi Sentence	Transliteration
1	A Shopper 's Guide	1	शिकायत करने का तरीका मदद कहें	shikāyata karane kā tariqā madada kahē
2	Your legal rights	2	A Shopper 's Guide	A Shopper 's Guide
3	How to complain	3	s Guide	s Guide
...
53	Before you buy	17	खरीदने से पहले	harīdane se pahale
54	Do you know what precautions	18	क्या आप जानते हैं कि कोई इस्तेमाल	kyā āpa jānate hain ki koi istemāla ki
...
64	Buying goods	56	सामान खरीदना	sāmāna harīdanā
65	The law says that goods must	57	चीजों के बारे में कानून कहता है	chīzōn ke bāre mē kānūna kahatā hā
...
342	See page 93 for the address .	871	प ते के लिए पृष्ठ 36 देखिए ।	pa te ke lie pṛṣṭha 36 dekhīe .
343	Useful Organisations		end of file	
344	Listed here are some of the ma			

Table 6: Sentence alignment of Emille cannot be performed automatically because of large amounts of nonparallel text, especially at the beginning and end of documents. We used a manually aligned and cleaned subset of Emille, courtesy of the team of IIT Mumbai.

Why does Emille hurt the BLEU score?

TM	LM	DT	Dbleu	Tbleu
Emille	Emille	Emille	9.33	10.16
Tides	Tides	Tides	11.45	12.08
Tides + DP	Tides	Tides	11.24	12.58
Tides + Emille	Tides	Tides	13.05	11.05
Tides + DP + Emille	Tides	Tides	12.98	11.32
Emille	Emille	Tides	9.03	1.75
Tides	Tides	TideSwap	12.78	10.66
Tides + DP	Tides	TideSwap	12.82	10.75
Tides + Emille	Tides	TideSwap	12.74	11.75
Tides + DP + Emille	Tides	TideSwap	12.64	11.68
Emille	Emille	TideSwap	2.26	7.38

Table 7: Cross-evaluation of various corpora, using Joshua (Li et al. 2009) MT system. TM = training data for translation model; LM = training data for language model; DT = development and test data (Emille: we split the 3501 Emille training pairs to 3151 training, 175 development and 175 test; TideSwap: Tides test data were used as development data, and vice versa). Dbleu: final BLEU score on development data. Tbleu: BLEU score on test data. General observation: training on Emille and tuning on Tides (not TideSwap) causes overfitting.

Sentences (per cent)	of	also found in
2320	5.00%	English Tides training
107	11.00%	English Tides dev
0	0.00%	English Tides test
2320	69.00%	English Emille

Table 8: The answer: Emille significantly overlaps with Tides!

Addicter

Addicter (Automatic Detection and Display of Common Translation Errors) is a tool for analysis of errors of machine translation systems. It indexes aligned parallel corpora and shows examples of words in context. Besides training data, we can also post-compute word alignments of development and test data (both with reference translation and with system hypothesis) and view these, too.

Dynamic HTML is used for the viewing. All words are clickable to quickly navigate to the sequence of examples of every word in the corpus. An alignment summary (right) shows the most frequent counterparts a particular word has got aligned to. For most words this gives a good clue of the actual meaning of the word (i.e. you do not need to understand both languages in the pair). Words in languages that do not use the Latin script (such as Hindi) are also shown transliterated.

Future versions should include browsing the phrase table / extracted grammar. A lemmatized version of the data (provided a lemmatizer is available) will help to identify morphology-related translation errors. Another (already existing) tool to be integrated to Addicter is a system comparator that highlights common and missing N-grams in two system outputs vs. the reference translation.

Alignment summary

The word 'book' got aligned to 49 distinct words/phrases:

- पुस्तक / pustaka (94)
- किताब / kitāba (34)
- (15)
- बैच / baichā (11)
- पुस्तिका / pustikā (8)
- कृति / kṛti (7)
- किताबों / kitābōn (4)
- बुक / buka (4)
- पुस्तक ? / pustaka ? (3)
- चेकबुक / cekabuka (2)
- पुस्तक निबन्ध / pustakāni (2)
- पुस्तकों / pustakōn (2)
- पुस्तक / pustaka (2)
- बुकिंग / bukinga (1)
- प्रकाशित पुस्तक / prakāśita pustaka (1)
- निषिद्ध किताब / nisiddha kitāba (1)
- लिखा / likhā (1)
- ? पुस्तक / ? pustaka (1)