English-Hindi Translation – Obtaining Mediocre Results with Bad Data and Fancy Models*

Ondřej Bojar, Pavel Straňák, Daniel Zeman, Gaurav Jain, Michal Hrušecký, Michal Richter, Jan Hajič
Univerzita Karlova v Praze, Ústav formální a aplikované lingvistiky,
Malostranské náměstí 25, 118 00, Praha, Czech Republic
{bojar,stranak,zeman,hajic}@ufal.mff.cuni.cz
gauravjain@cse.iitb.ac.in, michal@hrusecky.net, michalisek@gmail.com

Abstract

We describe our attempt to improve on previous English to Hindi machine translation results, using two open-source phrase-based MT systems: Moses and Joshua. We use several approaches to morphological tagging: from automatic word classes, through stem-suffix segmentation, to a POS tagger. We evaluate various combinations of training data sets and other existing English-Hindi resources. To our knowledge, the BLEU score we obtained is currently the best published result for the IIIT-TIDES dataset.

1 Introduction

Machine translation is a challenging task and more so with significant differences in word order of the languages in question and with the target language explicitly marking more details in word forms than the source language does. Precisely this holds for the English-Hindi pair we study.

We try to explore the problems on several fronts: Section 2 describes our careful cleanup and a few additions to the training data. Section 3 is devoted to several additional variants of morphological representation of the target side. Section 4 evaluates the impact of using a hiearchical instead of phrase-based model. Section 7 concludes our experiments by providing a preliminary human evaluation of translation quality.

2 Data

Tides. 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 (Venkatapathy, 2008): 50K sentence pairs for training, 1K development and 1K test data (1 reference translation per sentence).

The corpus is a general domain dataset with news articles forming the greatest proportion. It is aligned on sentence level, and tokenized to some extent. We found the tokenization insufficient and ran our own tokenizer on top of it. Cleaning was also necessary due to significant noise in the data, often caused by misconversion of Latin script.

We used Tides as our primary dataset and all reported scores are measured on its test data. However, note that due to the processing we applied to both training and test data, our results are not directly comparable to the results of the 2008 NLP Tools Contest.[1] We are happy to share our cleaning tools to make the experiments reproducible.[2]

Daniel Pipes. A journalist Daniel Pipes' website:[3] limited-domain articles about the Middle East. The articles are originally written in English, many of them are translated to up to 25 other languages, including Hindi (322 articles, 6,761 sentence pairs).

Emille. EMILLE corpus (Baker et al., 2002) consists of three components: monolingual, parallel and annotated corpora. The parallel corpus consists of 200,000 words of text in English and its accompanying translations in Hindi and other languages. Whenever we mention Emille, we mean the parallel English-Hindi section.

The original Emille turned out to be very badly aligned and had spelling errors, so we worked with a manually cleaned and aligned subset of 3,501

* The research has been supported by the grants GAAV ČR 1ET201120505 (Grant Agency of the Academy of Sciences of the Czech Republic), MSM0021620838 (Czech Ministry of Education), FP7-ICT-2007-3-231720 (EuroMatrix Plus), and GAUK 4307/2009 (Grant agency of the Charles University).

[1] For instance, the Joshua BLEU score of a model trained on Tides only, as we present it later in Table 2, is 12.27 on the cleaned data, and 11.10 on the raw data; see also Table 3.

[2] https://wiki.ufal.ms.mff.cuni.cz/pub-company:icon2009; accept the certificate.

[3] http://www.danielpipes.org/

Proceedings of ICON-2009: 7th International Conference on Natural Language Processing, Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2009

sentence pairs.[4]

We also tried various other small datasets:

ACL 2005. A subset of Emille, used in the shared task of the ACL 2005 workshop on "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond".[5]

Wiki NEs. We have extracted English titles of Wikipedia entries, that are translations (or rather transcriptions) of Hindi, Marathi and Sanskrit named entities (names of persons, artifacts, places, etc.), and their translations in Devanagari.

Shabdanjali. A GPL-licensed collaborative English-Hindi dictionary, containing about 26,000 entries, available on the web.

Agriculture domain parallel corpus. English-Hindi-Marathi-UNL parallel corpus from Resource Center for Indian Language Technology Solutions.[6] It contains 17,105 English and 13,248 Hindi words.

The impact of adding additional data to the Tides training data is partially illustrated by Table 2 and Table 6. Adding other data has similar results.

2.1 Normalization

As always with statistical approaches, we want our training data to match the test data as closely as possible. There are style variations throughout our data that can and should be normalized automatically. For instance, the Tides corpus usually (except for some conversion errors) terminates a sentence with the period ("."). However, some of our additional data sets use the traditional Devanagari sign called *danda* ("।") instead. Our normalization procedure replaces all dandas by periods, converts Devanagari digits to ASCII digits and performs other minor reductions of the character set (e.g. normalizing non-ASCII punctuation). Some changes affect the English data as well.

2.2 Lessons Learnt

Hindi as the target language possesses some features that negatively influence MT performance: richer morphology than English, greater noise in training data and harder sparse-data problem due to vocabulary that combines words from various etymological sources.

One common cause of data sparseness is unstable orthography of English and other loanwords (or even transcribed citations), cf. the following counterparts of the English word "standards", all present in the data:

- स्टैंडर्डज *(sṭaiṁḍarḍaja)*
- स्टैंडर्डस *(sṭaiṁḍarḍasa)*
- स्टैंडर्ड्स *(sṭaiṁḍarḍsa)*

Also common is the case where genuine English text has been (during some processing at the site of the data provider) interpreted as Devanagari encoded using Latin letters. Thus, ईन्ङोर्मटिओन् छोम्मिसिओनेर् *(īnṅormaṭion chommisioner)* should actually read (even in the Hindi text) *Information Commis(s)ioner*. If transcribed to Devanagari, it would be इन्फ़ोर्मेशन कोमिशनेर *(informeśana komiśanera)*.

3 Target-Side Morphology

Richer morphology on the target side means that besides selecting the *lexically correct* Hindi equivalent of an English word (such as *book* → किताब *(kitāba)*), the system also must correctly guess the *grammatical features* (such as the direct vs. oblique case in *books* → किताबें *(kitābeṁ)* vs. किताबों *(kitāboṁ)*). Moreover, grammatical agreement (such as मेरी किताब *(merī kitāba)* "my book" vs. मेरा कमरा *(merā kamarā)* "my room") further multiplies possible translations of an English word. A separate model of target-language morphology may help the system select the correct translation.

We have tried two supervised and four unsupervised approaches to Hindi morphology, see below. The results in Table 1 and Table 4 show that we were not able to achieve any improvements by using real POS tags compared to automatic word classes produced by `mkcls` or classes created by hand.

3.1 Supervised Morphology

POS tagger. The only Hindi POS tagger we were able to find on the web, download and use is a part of the GATE framework (Cunningham et al., 2002). That is however more of a demonstration, than a real tagger. Fortunately, we also had the opportunity to work with a CRF-based tagger from IIT Mumbai (Gupta et al., 2006).[7]

[4] We are very grateful to Om Dammani and his colleagues from IIT Mumbai for making their corrected subset of Emille available to us.

[5] Downloaded from the workshop website at `http://www.cse.unt.edu/~rada/wpt05/`. We used this dataset on the assumption that it might be better aligned, compared to the original Emille parallel corpus.

[6] `http://www.cfilt.iitb.ac.in/download/corpus/parallel/agriculture_domain_parallel_corpus.zip`

[7] Many thanks to Pushpak Bhattacharyya for making this tool available to us.

Textbook suffixes. Since the main morphological device of Hindi is alternating word suffixes, one can model suffixes instead of morphological tags. Our unsupervised approaches (see below) automatically acquire long but noisy lists of Hindi suffixes, some of which may bear grammatical meaning. As an alternative, we compiled a short list of 31 suffixes (including duplicates) for regular declension of nouns, verb formation etc. The cost of such a list was negligible: one of the authors (who does not speak Hindi) spent less then two hours with a Hindi textbook (Snell and Weightman, 2003), looking for grammar-describing sections.

3.2 Unsupervised Morphology

We have tried two simple unsupervised methods:

- Automatic classes created by `mkcls`[8] tool contained in GIZA++ installation.

- simple n-character long suffixes

and two more elaborate methods:

Affisix is an open source tool for unsupervised recognition of affixes in a given language. It takes a large list of words and generates a list of possible suffixes or prefixes scored according to a selected method or a combination of methods. In this case, we used it to obtain the 100 top scoring suffixes using backward entropy and backward difference entropy methods (see Hlaváčová and Hrušecký (2008) for details). The longest possible suffix seen in a word was then treated as the morphological tag of the word, leaving some words with a blank tag.

Hindomor is another tool for unsupervised segmentation of words into morphemes. It was originally published in the context of information retrieval (Zeman, 2008).

The tool has been trained on the word types of the Hindi side of the Tides corpus. For every word the algorithm searches for positions where it can be cut in two parts: the stem and the suffix. Then it tries to filter the stem and suffix candidates so that real stems and suffixes remain. The core idea is that real stems occur with multiple suffixes and real suffixes occur with multiple stems.

Given the lists of stems and suffixes obtained during training, we want to find the stem-suffix boundary in a word of the same language. Theoretically, we could use the learned stem-suffix combinations to require that both stem and suffix be known. However, this approach proved too restrictive, so we ended up in using just the list of suffixes. If a word ends in a string equal to a known suffix, the morpheme boundary is placed at the beginning of that substring.

factor	BLEU	factor	BLEU
tag	12.03±0.75	hitbsuf	11.58±0.74
wc50	11.97±0.73	hindomor2	11.55±0.74
wc10	11.76±0.74	hindomor1	11.54±0.71
lcsuf3	11.66±0.75	affddf	11.50±0.7
lcsuf1	11.63±0.72	affbdf	11.33±0.72
hindomor3	11.60±0.73	lcsuf2	11.14±0.74

Table 1: Target side morphology: Using different additional factors for second language model of MT system and its effect on BLEU score. Trained on IIIT-TIDES only. tag – POS tags; wc*n* – *n* word classes from mkcls; hitbsuf – word classes created by hand; lcsuf*n* – simple n-character suffixes; hindomor*n*; aff*xxx* – Affisix

The best BLEU score we achieved with Moses while making experiments with different factors for target side morphology was 12.22±0.78: trained on Tides and Daniel Pipes data with `hitbsuf` as the additional factor.

4 Hierarchical Phrase-Based Models

One of the major differences between English and Hindi is the word order. Massive reordering must take place during translation. That is why we conducted several experiments with hierarchical translation models (Chiang, 2007), namely with Joshua (Li et al., 2009), and compared the results with classical phrase-based models (Moses, Koehn et al., 2007).

Hierarchical phrase-based translation is based on synchronous context-free grammars (SCFG). Like phrase-based translation, pairs of corresponding source- and target-language phrases (sequences of tokens) are learnt from training data. The difference is that in hierarchical models, phrases may contain “gaps”, represented by nonterminal symbols of the SCFG. If a source phrase f contains a nonterminal X_1, then its translation e also contains that nonterminal, and the decoder can replace the nonterminal by any phrase f_1 and its translation e_1, respectively. An illustrative rule for English-to-Hindi translation is

$$X \rightarrow \langle X_1 \cdot of \cdot X_2 \rangle, \langle X_2 \cdot \text{का} \cdot X_1 \rangle$$

[8] `http://www.fjoch.com/mkcls.html`

where the Hindi word का *(kā)* is one of several grammatical forms of the Hindi equivalent of "of", and the subscripts on the nonterminals cause the two (noun) phrases around *of* to be reordered around का in the translation.

Hierarchical models have been known to produce better results than classical phrase-based models (Chiang et al., 2005).

In our experiments we used the Joshua implementation of the hierarchical phrase-based algorithms.[9] We set the maximum phrase length to 5, MERT worked with 300 best translation hypotheses per iteration, and the number of iterations was limited to 5.

Unless stated otherwise, our experiments use a trigram language model trained on the target side of the Tides training data. In accord with Bojar et al. (2008), we found additional out-of-domain language models damaging to BLEU score.

Symmetrical word alignments for grammar extraction were obtained using Giza++ and the scripts accompanying Moses. Alignments were computed on first four (lowercased) characters of each training token, and the *grow-diag-final-and* symmetrization heuristic was used.

The results of the hierarchical models trained on various datasets, compared with classical phrase-based models are shown in Table 2.

Parallel data	**Joshua**	**Moses**
Tides	12.27±0.83	11.46±0.72
Tides+DP	**12.58±0.77**	11.93±0.75
Tides+DP+Emille	11.32±0.74	10.06±0.72
Tides+DP+Dict	12.43±0.79	11.90±0.78

Table 2: Results of Joshua compared with Moses

5 Results

As far as automatic evaluation is concerned, the best result reported on in this paper is the 12.58 BLEU of Joshua trained on Tides and Daniel Pipes (Table 2). Moses was not able to outperform this score despite its ability to learn factored models. The best Moses score is 12.22 (morphology).

The greatest mystery is the fact that adding Emille to the training data does not improve results with either system. Although we are not able to explain this in full, we made a few observations:

1. When asked to extract its SFCG, Joshua needs to see the source (English) side of the data to be decoded, and extracts only the rules relevant to the dataset. Thus we extract one grammar for the development data (used for minimum error rate training) and another for the test data. While the development grammar changes when Emille is added, the test grammar remains the same. As if Emille was unable to approximate the test data in any way.

2. MERT optimizes 5 feature weights of Joshua: the language model probability, the word penalty (preference for shorter translations), and three translation features: $P(e|f)$, $P_{lex}(f|e)$ and $P_{lex}(e|f)$. When Emille is involved, MERT always pushes the non-lexical translation probability extraordinarily high, and causes overfitting. While for other experiments we usually saw better BLEU scores on test data than on development data, the opposite was the case with Emille.

6 Related Research

Ramanathan et al. (2009) improve Hindi morphology by transfering additional information from English parse tree and semantic roles, in addition to some hard-coded syntax-based reordering. However, they use an unspecified corpus making their results incomparable.

The ICON 2008 NLP Tools Contest (Venkatapathy, 2008) included translation of Tides test data. Although our retokenized and cleaned data make a direct comparison impossible, our preliminary experiments on the unclean data are comparable. In Table 3 we compare four results presented at ICON 2008 with our hierarchical model trained on (unclean) Tides, with a Tides trigram language model.

System	**BLEU**
Mumbai (Damani et al., 2008)	8.53
Kharagpur (Goswami et al., 2008)	9.76
Prague (Bojar et al., 2008)	10.17
Dublin (Srivastava et al., 2008)	10.49
present Joshua	11.10

Table 3: Previous work compared to our hierarchical model (Joshua) on unclean data

7 Human Evaluation

The richer morphology and a high degree of reordering are known to render BLEU unreliable. While we still optimize for BLEU, our manual

[9] Many thanks to the team at Johns Hopkins University for creating Joshua and making it publicly available.

analysis reveals relatively low correlation. An extensive study similar to Callison-Burch et al. (2009) would be very valuable.

We were able to conduct three small-scale manual evaluations. Our annotator was given the source English text and four or five Hindi translations in a randomized order of 100 sentences (each time a different subset of the full test set). He assigned a mark to each of the four Hindi translations: giving no mark ("0") indicates a completely incomprehensible translation. A single star ("*") denotes sentences with severe errors and hard to understand but still related to the input. Two stars ("**") are assigned to more or less acceptable translations, possibly with many errors but understandable and conveying most of the original meaning.

Table 4 evaluates some basic configurations of Moses. For a comparison, we include the reference translations (REF) among the hypotheses (due to the randomization, the annotator cannot be entirely sure which of the hypotheses is the reference but often it can be guessed from the striking difference in quality). We see that even 6 reference translations were marked as inadequate and 11 as very bad. Because the test set is a part of the Tides corpus, these figures give a rough estimate of the overall corpus quality.

The remaining figures in Table 4 document that (while somewhat suspicious by itself), the Tides training data are most informative with respect to the (Tides) test set: a system trained completely out-of-domain on everything except Tides was able to deliver only 3 acceptable translations (OOD). The TIDP and WC10 systems compare the effect of adding more parallel data (TIDP for Tides+DanielPipes) vs. adding the unsupervised morphological factor (WC10, 10 word classes). We observe that the difference indicated by human evaluation is rather convincing: nearly twice as many acceptable sentences in TIDP, but negligible in BLEU. This confirms our doubts about the utility of BLEU scores for languages with rich morphology and the neccessity to regularly run manual evaluations.

In Table 5, we evaluate the difference between phrase-based (Moses) and hierarchical (Joshua) translation model. Both systems are trained in identical conditions: both the translation and the language model are trained on Tides and DanielPipes. For Moses, we wanted to confirm the observation that more data are more important than ensuring morphological coherence on another 100 manually judged sentences. Therefore, we report also Moses-DPipes+POStags, a setup trained on Tides only but including the supervised morphology of Mumbai tagger.

System	0	*	**	BLEU
REF	6	11	83	–
OOD	80	17	3	1.85±0.24
TIDP	26	44	30	11.93±0.75
WC10	38	46	16	11.76±0.74

Table 4: Manual evaluation of some basic Moses setups: training out of domain (OOD) and adding either more parallel data (TIDP) or a factor for morphological coherence (WC10) compared to the quality of the reference translation (REF).

We see that while the BLEU scores indicate the superiority of the hierarchical model over the phrase-based model with lexicalized reordering, the difference in manual judgments seems less convincing. Only 3 sentences were ranked better. On the other hand, we confirm that even the supervised morphology is less important than a good parallel data source.

System	0	*	**	BLEU
REF	6	10	84	–
Joshua	32	37	31	12.58±0.77
Moses	35	35	30	11.93±0.75
Moses-DPipes+POStags	32	42	26	12.03±0.75

Table 5: Manual judgements of hierarchical (Joshua) vs. phrase-based (Moses) translation.

Table 6 provides manual analysis of our baseline Moses setup (simple phrase-based translation, lexicalized reordering, language model trained on the full target side of the parallel data) trained on various subsets of our parallel data. We start with the combination of Tides and DanielPipes (TIDP), add Emille (EM), all other corpus sources (oth) and finally the dictionary Shabdanjali in two variants: full (DICTFull) and filtered to contain only Hindi words confirmed in a big monolingual corpus (DICTFilt). The BLEU scores indicate that Emille hurts the performance when tested on Tides test set. This surprising result was indeed confirmed by the manual analysis: only 12 instead of 19 sentences were translated acceptably. Adding further data reduces the detrimental effect of Emille (not observed in BLEU scores) but the best performance is achieved by Tides+DanielPipes only.

Note that this final manual evaluation was based on a smaller dataset of 53 sentences only.

System	0	*	**	BLEU
REF	0	8	45	–
TIDP	20	14	19	11.89±0.76
TIDPEM	22	19	12	9.61±0.75
TIDPEMoth	17	25	11	10.97±0.79
TIDPEMothDICTFilt	23	17	13	10.96±0.75
TIDPEMothDICTFull	22	16	15	10.89±0.69

Table 6: The effect of additional (out-of-domain) parallel data in phrase-based translation.

8 Conclusions and Future Work

We have tried several ways to improve state-of-the-art in English-Hindi SMT, however results are mixed:

There are quite a few strange results:

- POS tagging does not give better results than automatic word classes. Hindi textbook-based manual word classes were even better.
- Adding more training data to Tides either helps insignificantly, or (more often) hurts BLEU score.
- BLEU may not correlate with human judgement. Our limited experiments show that adding training data may hurt BLEU but improve quality by human judgement.

In our future work we want to further explore problems with existing datasets, the use of morphology, and the relation of output quality measured in terms of BLEU vs. human judgement. We also believe that there is room for improvement in the quality and amount of available parallel data.

On the other hand, we have shown that our hand-crafted word classes and some additional data help Moses achieve significantly better results than reported previously. Hierarchical decoder Joshua can capture word order even better than Moses. Its results are always slightly better. And as far as we know, our current results are the best that have been reported on this dataset.

References

Paul Baker, Andrew Hardie, Tony McEnery, Hamish Cunningham, and Rob Gaizauskas. EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation. In *Proc. of LREC 2002*, pages 819–827, 2002. 2

Ondřej Bojar, Pavel Straňák, and Daniel Zeman. English-Hindi Translation in 21 Days. In *Proc. of ICON-2008 NLP Tools Contest*, 2008. 4, 6

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 workshop on statistical machine translation. In *Proc. of the Fourth Workshop on Statistical Machine Translation, EACL 2009*, 2009. 7

David Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, 2007. 4

David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, Philip Resnik, and Michael Subotin. The Hiero Machine Translation System: Extensions, Evaluation, and Analysis. In *Proc. of HLT/EMNLP*, pages 779–786, 2005. 4

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proc. of ACL 2002*, 2002. 3.1

Om P. Damani, Vasudevan N., and Amit Sangodkar. Statistical machine translation with rule based re-ordering of source sentences. In *Proc. of ICON-2008 NLP Tools Contest*, 2008. 6

Sumit Goswami, Nirav Shah, Devshri Roy, and Sudeshna Sarkar. NLP Tools Contest: Statistical Machine Translation (English to Hindi). In *Proc. of ICON-2008 NLP Tools Contest*, 2008. 6

Kuhoo Gupta, Manish Shrivastava, Smriti Singh, and Pushpak Bhattacharyya. Morphological richness offsets resource poverty- an experience in building a pos tagger for hindi. In *Proc. of COLING/ACL-2006*, 2006. 3.1

Jaroslava Hlaváčová and Michal Hrušecký. Affisix: Tool for Prefix Recognition. In *Proc. of Text, Speech and Dialogue, LNAI 5246*, pages 85–92. Springer, 2008. 3.2

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL 2007 Companion Volume Proc. of the Demo and Poster Sessions*, pages 177–180, 2007. 4

Zhifei Li, Chris Callison-Burch, Sanjeev Khudanpur, and Wren Thornton. Decoding in Joshua: Open Source, Parsing-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 91:47–56, 2009. 4

Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. Case markers and morphology: Addressing the crux of the fluency problem in english-hindi smt. In *Proc. of ACL/IJCNLP*, 2009. 6

Rupert Snell and Simon Weightman. *Teach Yourself Hindi*. Hodder Education, London, UK, 2003. 3.1

Ankit Kumar Srivastava, Rejwanul Haque, Sudip Kumar Naskar, and Andy Way. MaTrEx: The DCU Machine Translation System for ICON 2008. In *Proc. of ICON-2008 NLP Tools Contest*, 2008. 6

Sriram Venkatapathy. Nlp tools contest – 2008: Summary. In *Proc. of ICON-2008 NLP Tools Contest*. NLP Association of India, 2008. 2, 6

Daniel Zeman. Unsupervised acquiring of morphological paradigms from tokenized text. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007. LNCS 5152*, pages 892–899. Springer, 2008. 3.2