Benchmarking of English-Hindi parallel corpora

Jayendra Rakesh Yeka, Prasanth Kolachina, Dipti Misra Sharma
LTRC, IIIT-Hyderabad
Hyderabad, India
rakesh.yeka@research.iiit.ac.in, prasanth_k@research.iiit.ac.in, dipti@iiit.ac.in

Abstract

In this paper we present several parallel corpora for English↔Hindi and talk about their natures and domains. We also discuss briefly a few previous attempts in MT for translation from English to Hindi. The lack of uniformly annotated data makes it difficult to compare these attempts and precisely analyze their strengths and shortcomings. With this in mind, we propose a standard pipeline to provide uniform linguistic annotations to these resources using state-of-art NLP technologies. We conclude the paper by presenting evaluation scores of different statistical MT systems on the corpora detailed in this paper for English→Hindi and present the proposed plans for future work. We hope that both these annotated parallel corpora resources and MT systems will serve as benchmarks for future approaches to MT in English→Hindi. This was and remains the main motivation for the attempts detailed in this paper.

Keywords: Machine Translation, Parallel Corpora, Indian languages, English↔Hindi, Benchmarking

1. Introduction

Machine Translation (MT) has occupied a majority of the spectrum of efforts in NLP in the last couple of decades. Statistical approaches to Machine Translation (SMT) have been gaining more prominence in the recent past. Indian languages are one set for which approaches to SMT have only recently been studied (Ramanathan et al., 2009; Venkatapathy and Bangalore, 2009; Venkatapathy et al., 2010; Arafat et al., 2010). Compared to language pairs for which large amounts of parallel corpora exist, Indian languages fall short in terms of quantity that can be used for SMT. But, parallel corpora resources that may be used by researchers for sake of comparative analysis exist. Hindi being an Indian language spoken by the majority of the country, has managed to find more sizable resources when compared to other Indian languages. More efforts are ongoing into building large collections of parallel corpora for all Indian languages to help create general purpose SMT systems. Most of these efforts are distributed and result in different corpora sets with variations across texts. Corpora resources created in this manner lack normalization across efforts.

Inspite of lack of parallel corpora for statistical methods, MT from English to Hindi has been the focus of research efforts for close to two decades. The first known system for translation to Hindi, *Anusaaraka*[1] is a transfer-based system consisting of hand-crafted grammatical rules and large bilingual dictionaries for translation from English to many Indian languages. Apart from this, other attempts at creating general purpose translation systems for translation to Hindi have been made leading to reasonable success. These translation systems use a customized pipeline for carrying out the task of translation, leading to difficulties in comparing their respective approaches to MT. The same is also the case for MT systems created using statistical methods, making reproducibility of results impossible.

What is lacking in these efforts is lack of a standard linguistic analysis benchmark that can be used when evaluating different translation systems. Different translation systems based on the same paradigm may result in significantly different translations due to variation in quality of linguistic analysis provided to these translation systems.

The main contribution of the current work is a proposal for a standard pipeline for uniform linguistic analysis of parallel corpora to be used across different translation systems. Such a pipeline will provide a framework to create annotated corpora that can be used to compare and analyze different approaches to MT. The goal of this study is to create both annotated corpora resources along with establishing a pipeline for processing parallel texts for English→Hindi MT.

This paper is organized as follows: we present different parallel corpora available available for English↔Hindi in Section 2.. We also briefly describe the existing MT systems for translation to Hindi in Section 3.. We describe the proposed pipeline for linguistic preprocessing of texts in Section 4.. Finally, we conclude the paper by presenting a few benchmarked models for SMT using resources created through the pipeline in Section 5..

2. Corpora

In this section, we introduce the different parallel corpora datasets available for English↔Hindi. Bojar et al. (2010) mention three previous datasets for the language pair. One of the first known corpus comes from the EMILLE/CIIL corpus created by a collaboration between Lancaster University and Central Institute of Indian Languages, India through the EMILLE project. The parallel corpora consists of texts in English along with their translations in Hindi, Bengali and three other Indian languages. The corpus contains texts

[1] http://anusaaraka.iiit.ac.in/

from different domains such as *education*, *health*, *legal texts*. A subset of this parallel corpus was validated and released as part of the ACL (2005) shared task on word-alignment (Mihalcea and Pedersen, 2003).[2] Another corpus that came into use for English↔Hindi is the DARPA-TIDES corpus. The corpus was released as part of language contest on SMT in 2002. After manual refinement and cleaning, a subset of this corpus was released for the NLP Tools Contest(Venkatapathy, 2008) on SMT for English→Hindi.

Apart from these two datasets, efforts to create large-scale mutlilingual parallel corpora for English, Hindi and several other Indian languages have been part of two projects: English to Indian languages MT (EILMT)[3] and Indian Languages Corpora Initiative (ILCI)[4]. Both the projects (*till date*) have focussed on collecting resources for two particular domains: *tourism* and *health*. In case of the EILMT project, bilingual lexica have been additionally created for both these domains containing domain-specific term translations and multi-word expressions.

On the other hand, the ILCI project provides parallel corpora with part-of-speech tags created by linguistic annotators. Both the EILMT and ILCI projects are initiatives by the Department of Information and Technology (DIT) of India, handled by a consortia of different participating institutions. The distributional effort to create these resources facilitates quick creation of large-scale parallel corpora.

In the rest of this paper, we refer to the resources created from the EILMT project as Tourism-EILMT, Health-EILMT and ILCI project as Tourism-ILCI and Health-ILCI.

Though the above mentioned projects have led to the creation of parallel corpora in multiple Indian languages, the current work focusses on English↔Hindi portion of these resources to present them along with other existing resources. Table 1 shows the statistics of the datasets in their current form.

Corpus	# sents	# En tok	# Hn tok
EMILLE-ACL05	3,556	57,118	70,932
TIDES-ICON08	52,000	12,43,815	13,38,994
Tourism-EILMT	15,198	3,83,992	3,65,163
Health-EILMT	7,484	1,37,396	1,69,039
Tourism-ILCI	25,000	4,25,646	4,23,711
Health-ILCI	25,000	4,22,436	4,40,764
NCERT	9,340	1,73,129	1,98,264
Total	137,578	-	-

Table 1: Statistics about English-Hindi parallel corpora: # sents- sent counts in the corpora; # En tokens- token count in English sentences; # Hn tokens- token count in Hindi sentences

An additional resource that was created at IIIT-Hyderabad is a corpora made up of a small portion of physics text-books taught at the highschool level. Individual chapters were extracted from the text books and aligned at the sentence level using an automatic aligner. The automatically aligned corpus was validated and evaluated by a small group of native speakers to prune out erroneous alignments. This corpus is atypical from the other datasets in our work, it is a sample of technical writing. For example, a significant portion of this corpus contains mathematical equations and formulae.

Another notable effort towards creating parallel corpora for Indian languages has been carried out through the use of crowd-sourcing (Post et al., 2012). The resource was created by employing large crowd of cheap translators to translate texts in Indian languages to English. To allow for translation variation, they provide multiple alternate translations for each sentence in Indian language.

At this point, we are also familiar with another recent effort to create large corpora for MT between English↔Hindi (Bojar et al., 2014). The effort resulted in a resource containing about 287,000 translations, 25% of which have been included from the TIDES, Tourism-EILMT and EMILLE-ACL05 corpora. The preliminary version of the corpora released for the shared task on SMT for English→Hindi was reported [5] to have issues due to quality of sources datasets from which the resource was created.

The work in this paper is an independent attempt to improve the quality of existing parallel corpora for English↔Hindi and provide uniform linguistic annotations to these resources using start-of-art NLP technologies. We hope that the resources from this attempt will allow for easy and more accurate comparison of the on-going attempts in MT for English↔Hindi.

3. Machine Translation for English→Hindi

While approaches for SMT have improved greatly in the recent years, work focussed on using SMT techniques for Indian languages has only begun recently. There has been a surge in recent years on developing general-purpose SMT systems for translating from English to Indian languages (Venkatapathy and Bangalore, 2009; Ramanathan et al., 2008; Ramanathan et al., 2009; Venkatapathy et al., 2010). English to Hindi machine translation, in addition to the lack of large-scale training corpora, also grapples with a number of issues owing to the typological divergence between the two languages.

Ramanathan et al. (2008) and Ramanathan et al. (2009) discussed methods to handle the morphological complexity of Indian languages, while translation both to and from Indian languages. Venkatapathy and Bangalore (2009) present a context based approach for translating from English to Hindi in the framework on

[2]We refer to this released dataset as EMILLE-ACL05 corpus in the rest of this paper.

[3]http://www.cdacmumbai.in/e-ilmt

[4]http://sanskrit.jnu.ac.in/projects/ilci.jsp?proj=ilci

[5]By the authors on http://ufallab.ms.mff.cuni.cz/~bojar/hindencorp/

Global Lexical Selection models. Also, Venkatapathy et al. (2010) proposed a dependency-based SMT system for translation from English to Indian languages. The dependency based framework is best suited for translation between languages with free-word order, another characteristic of a few Indian languages like Hindi, Telugu and Marathi. The framework also allows use of large set of features functions, with a flexible feature design from using discriminative models. There are ongoing efforts into building larger collections of parallel corpus for Indian languages as outlined in Section 2. to help in creating general-purpose SMT systems.

On the other hand, long and steady efforts on both research and engineering fronts to develop general-purpose MT systems have been going on for a long time. One of the key features of these systems is that they come from different paradigms in MT as opposed to purely statistical approaches to MT. Transfer-based MT systems separate the task of translation into three steps– analysis of the source sentence, a transfer step followed by a generation module to compose translations in the target language. In the source analysis phase, a sentence in the source language is analyzed using a syntactic parser combined with other modules such as word-sense disambiguation. The role of the transfer component is to translate the words in the source language using a bilingual dictionary and to carry out syntactic transformations to reflect the word order of the target language. Finally, the generation module generates accurate word forms in the target language along with handling agreement phenomena. Typically, both bilingual dictionaries for a specific language pair and transfer grammars used for syntactic transformations are hand-crafted by bilingual experts in both languages.

One of the earliest attempts to develop MT systems for English to Indian languages, *Anusaaraka* was created using a transfer-based approach. The system makes use of a combination of multiple state-of-the-art parsers for analyzing the source sentences, and has other components to detect multi-word expressions, generate right inflections in the translation. At the same time, the EILMT project has led to development of both *Shakti*, a transfer-based system and another example-based translation system.

Both statistical and transfer-based make use of different pipelines setup for analyzing texts in English and Hindi. The variations in these pipelines cause difficulty in replicating experiments and accurately comparing the results from different systems. As such, we propose a standard pipeline for linguistic analysis of English and Hindi texts that can be used across different translation systems in the next section.

4. Linguistic Preprocessing

We mentioned in Section 1. the need for corpora with standard linguistic annotations. In this section, we explain in detail the pipeline that was setup for processing texts in English and Hindi to annotate them with different levels of linguistic analysis. The pipeline is made up of state-of-art tools for syntactic analysis in English and Hindi, set-up in an incremental fashion. Before we describe the pipeline setup, we mention our efforts to clean the datasets discussed in Section 2..

4.1. Corpora Cleaning

Apart from the EMILLE and TIDES datasets which were released publicly earlier, we noticed several errors in the case of remaining datasets while setting up our pipeline. The cases we observed frequently repeated across the datasets are reported below.

1. **Tokenization errors**: In the ILCI corpus, with manually tagged part-of-speech tags, we noticed several tokens left untagged mostly due to errors in tokenization. Presence of non-uniform delimiter between a word and its part-of-speech tag also caused issue while extracting raw tokens.

   Eg: the\DT person\NN who\WP **has/VBZ** got\VBN

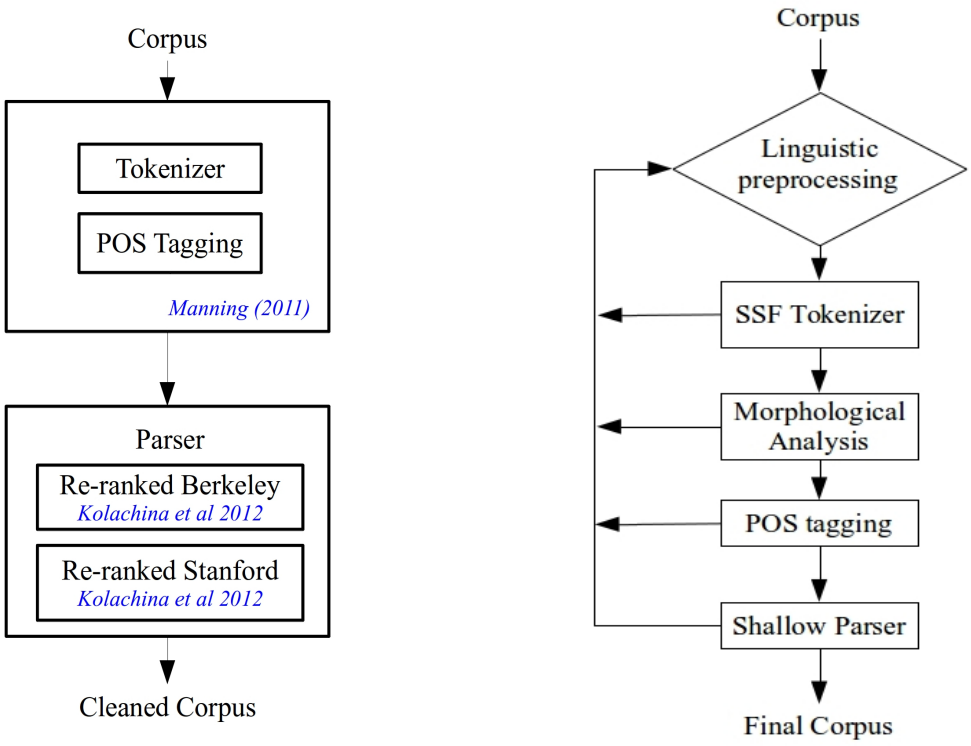
   The presence of different delimiter for **has** is one example from the corpus.

   We also noticed a variation in tokenization across different datasets, for e.g the case of hypenated compound modifiers in English. The difference between "small appliance industry" and "small-appliance industry" is nullified during corpus preparation. All such irregularities were corrected to reflect correct and uniform tokenization across different datasets.

2. **Misalignment**: Some instances in the ILCI datasets were cases of translations being *misaligned* (or *mistranslated*). The topic across translations in these sentences were different. For e.g. the English sentence talks about the human-organ *heart*, its respective translation is focussed on *liver*, an easily detectable error by human verification. However, these errors are difficult to detect automatically. We manually verified the entire corpus and pruned such erroneous sentences out of the datasets.

3. **Incomplete translations**: In some of the corpora, only partial translations of Hindi sentences were noticed on English side. In others, Hindi sentences contained partial translations retaining English text. Also sentences with translator comments and doubts were noticed. Cases like this were identified using heurisitics on sentence length and cross-language length ratio. We chose to prune instances based on the heurisitc given below

   X = avg * 0.3 - diff
   if X < 0 prun sentence
   where 'avg' is average of English and Hindi sentence lengths
   and 'diff' is positive difference in lengths.



Pipeline for English **Pipeline for Hindi**

Figure 1: Pipeline for linguistic analysis of English and Hindi texts

While in some of the above cases the errors were corrected, several sentence pairs were pruned completely. A log of such errored and pruned sentences is maintained, using which we plan to manually correct these cases and add them back to their respective corpora.

4. **Word formatting**: As the datasets were created in different environments on different platforms, presence of marking characters like dos-based endmarkers, font conversion residue characters are noticed in several datasets. The presence of these characters causes deviations in MT by marking the token as unseen token. Such cases were automatically corrected and manually verified.

   Eg: the person who has got^**M**

   The ^M present at the end of sentence makes "got" occuring in this sentence as an unseen token "got^M".
   Also multiple forms of same characters create deviation of probabilities. Normalization of variants of quotes, hyphens, mathematical symbols are performed across all the corpora.

   Eg: Smart quotes “ , ” are normalized to ""

4.2. Pipeline for Linguistic annotations

Kolachina and Kolachina (2012) conducted an extensive study on benchmarking state-of-art statistical parsers for analyzing English text. The main motivation for their study comes from the need to identify a high quality parser for English that can be used in an English-to-Indian language MT system. The study identifies a reranked variant of the Berkeley parser to perform better over datasets from varied domains. Kolachina (2012) later extended the study to identify a high quality dependency parser by combining multiple parsers to return a consensus analysis for sentences. Following their results, we set up the pipeline for English using the reranked Berkeley parser for producing the syntactic annotation of the sentences.

Figure 1 shows the pipeline that we used to annotate the parallel corpora with different levels of linguistic annotations. For texts in English, the tokenizer and part-of-speech tagger were part of Stanford NLP pipeline (Manning, 2011). The part-of-speech tagged text is parsed using reranked variants of Berkeley and Stanford parsers (Kolachina and Kolachina, 2012). For sentences where the reranked Berkeley parser fails, the reranked Stanford parser is used. The syntactic structures are then converted into dependency representations using the same method outlined in Kolachina and Kolachina (2012).

In the case of Hindi texts, the pipeline is made up of independent modules developed for use in the Indian Languages Machine Translation (ILMT) project. Modules used in the Hindi pipeline are morphological analyser, part-of-speech tagger and a shallow parser. The morphological analyzer gives multiple possible

analysis for each word in the sentence, which are disambiguated using a pruning module before tagging the sentence with part-of-speech tags. The shallow parser breaks the sentence into *chunks* and assigns to each chunk a *head* word. This essentially reduces the problem of parsing a sentence to parsing these *chunks*, as relations inside a chunk are assigned deterministically based on part-of-speech tags.

As mentioned previously in Section 4. due to the presence of errors, several modules of Hindi and English pipeline experienced hindrances and crashes. In order for the pipeline to proceed forward these errors either had to corrected, normalized or the sentences had to be pruned. We thus augmented the pipeline with a feedback loop (shown in Figure 1) allowing us to examine sentences that needed cleaning at the tokenization and formatting levels.

The final output of the pipeline for English contain syntactically annotated sentences with full parses. The Hindi pipeline provides a corpus with morphological analysis, part-of-speech tags and shallow parse information. The head of each chunk is also marked in the sentence.

5. Statistical Machine Translation

In this section, we describe our setup of Statistical Machine Translation (SMT) systems and the relevant experimental details. We use Moses (Koehn et al., 2007), a toolkit for experimenting with different classes of SMT models. In our experiments, we included phrase-based SMT (PBSMT) and hierarchical SMT (Hiero) for translation from English→Hindi. These classes of models are implemented in the Moses toolkit and thus provide a singular framework for carrying out experiments with different types of SMT models.

In our experiments, we divide the datasets into three partitions for all corpora: training (to extract bilingual information i.e. phrase-tables or synchronous grammars), tuning (to tune parameters of the statistical model) and an evaluation dataset. In the case of the Tourism-EILMT and TIDES dataset, we replicate the partitions provided during the NLP tools contest 2008 (Venkatapathy, 2008) to allow for comparsions with previous results from the shared task. We carried out experiments on EMILLE, Tourism-ILCI, Health-ILCI and NCERT datasets. The statistics of the paritions used in the SMT models are shown in Table 2.

Dataset	Training	Tune	Evaluation
EMILLE-ACL05	3,441	25	90
Tourism-ILCI	23,448	750	750
Health-ILCI	23,018	750	750
Tourism-EILMT	14,192	500	500
Health-Merged	30,498	750	750
NCERT	8,286	500	500

Table 2: Datasets parition statistics

The settings used to train Moses models are the same as suggested for baseline models in the WMT shared tasks [6]. We used both the MERT (Och, 2003) and MIRA (Hasler et al., 2011) algorithms to tune parameters of the statistical models. While creating a large language model by combining all the target language texts seemed like a more efficient option, we chose to create the target language model using only the target side of the training corpus. The evaluation of the SMT models were done using BLEU (A. Papineni et al., 2002) which is the widely used MT metric today. Table 3 shows the BLEU scores obtained from both PBSMT and Hiero models for all the datasets. We also report the Translation Edit rate scores from the sames in Table 4.

Dataset	Phrase-Based		Hiero	
	MERT	MIRA	MERT	MIRA
EMILLE-ACL05	43.57	46.73	44.14	46.30
Tourism-ILCI	18.37	18.41	19.73	19.70
Health-ILCI	17.43	17.81	19.39	19.09
Tourism-EILMT	9.04	9.32	6.83	8.25
Health-Merged	17.53	17.66	18.87	19.19
NCERT	17.16	17.28	12.54	12.88

Table 3: BLEU scores obtained from different SMT models

Dataset	Phrase-Based		Hiero	
	MERT	MIRA	MERT	MIRA
EMILLE-ACL05	0.4739	0.4709	0.4936	0.4641
Tourism-ILCI	0.6328	0.6354	0.6229	0.6143
Health-ILCI	0.6382	0.6353	0.6215	0.6288
Tourism EILMT	0.8270	0.8177	0.9504	0.9081
Health-Merged	0.6414	0.6349	0.6334	0.6303
NCERT	0.7390	0.7281	0.9123	0.8971

Table 4: Translation error rates (TER) obtained from different SMT models

The BLEU evaluation scores show that the Hiero models perform significantly between the phrase-based models for a 4 of the 6 datasets. This is expected since hierarchical SMT models are more flexible in reordering translations, a phenomena common in the case of English←Hindi. However, there is a significant drop in the performance of the Hiero models for the Tourism-EILMT and NCERT corpora. The same pattern is noticed from the Translation edit rate scores for all the datasets. The exceptions in the case of Tourism-EILMT is puzzling given the variation in the behavior when compared with the Tourism-ILCI dataset. Additionally, the low scores might seem puzzling to most as high BLEU scores have been reported in previous at-

[6] `http://www.statmt.org/wmt11/baseline.html`

tempts to MT for English←Hindi. However, Arafat et al. (2010) previously provided an explanation for these significantly low scores compared to previous results. The variation in the case of NCERT corpora is less puzzling given the small size of the corpus. However, it is interesting to analyze the performance of SMT models on this dataset given the atypical nature of the corpora. We are currently looking into the NCERT corpus and conducting a manual analysis of the translations to better understand these results.

6. Conclusion

In this paper, we presented different parallel corpora available for English↔Hindi and discussed the nature of these datasets. We also proposed a standard pipeline for processing texts in English and Hindi to annotate them with different levels of linguistic analysis. The main motivation of this work was to create uniformly annotated corpora resources for English↔Hindi MT. Additionally, we also present baseline statistical machine translation systems that can be used as reference to future work in MT for Hindi. These systems are trained using the resources created from the pipeline. The purpose of creating both the annotated datasets and baseline systems is to serve as benchmarks for future translation systems.

Acknowledgements

The research conducted in this project has been supported by the funds provided as part of EILMT consortium project by Department of Information Technology (DIT), India. We thank Dr. Vineet Chaithanya for his guidance and support throughout this work. We thank Dr. Radhika Mamidi for her encouragement and help with the resources. We would also like to thank TDIL for distributing the resources from the projects. A special mention of gratitute towards NCERT for granting permission to work on CBSE text books. We thank the students of Banasthali Vidyapeth for their work on validating the EILMT corpora and helping in extraction of the NCERT corpus. We thank all the consortia project members for distributing the state-of-art resources.

7. References

A. Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Arafat, A., Kolachina, P., Kolachina, S., Sharma, D. M., and Sangal, R. (2010). Coupling Statistical Machine Translation with Rule-based Transfer and Generation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas (AMTA)*.

Bojar, O., Straňák, P., and Zeman, D. (2010). Data Issues in English-to-Hindi Machine Translation. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Bojar, O., Diatka, V., Rychlý, P., Straňák, P., Tamchyna, A., and Zeman, D. (2014). Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the Ninth International Language Resources and Evaluation Conference (LREC'14)*, Reykjavik, Iceland, May. ELRA, European Language Resources Association. in prep.

Hasler, E., Haddow, B., and Koehn, P. (2011). Margin Infused Relaxed Algorithm for Moses. In *The Prague Bulletin of Mathematical Linguistics*, volume 96, pages 69–78, September.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Kolachina, S. and Kolachina, P. (2012). Parsing Any Domain English text to CoNLL Dependencies. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Kolachina, S. (2012). Non-local Features in Syntactic Parsing. Hyderabad, India.

Manning, C. D. (2011). Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In *Computational Linguistics and Intelligent Text Processing - 12th International Conference (CICLing)*, volume 6608 of *Lecture Notes in Computer Science*, pages 171–189. Springer.

Mihalcea, R. and Pedersen, T. (2003). An Evaluation Exercise for Word Alignment. In Mihalcea, R. and Pedersen, T., editors, *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10.

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada, June. Association for Computational Linguistics.

Ramanathan, A., Hegde, J., Shah, R. M., Bhat-

tacharyya, P., and Sasikumar, M. (2008). Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, pages 513–520, Hyderabad, India, January. Asian Federation of Natural Language Processing (AFNLP).

Ramanathan, A., Choudhary, H., Ghosh, A., and Bhattacharyya, P. (2009). Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint inproceedings on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 800–808, Suntec, Singapore, August. Association for Computational Linguistics.

Venkatapathy, S. and Bangalore, S. (2009). Discriminative Machine Translation Using Global Lexical Selection. *ACM Transactions on Asian Language Information Processing*, 8(2).

Venkatapathy, S., Sangal, R., Joshi, A., and Gali, K. (2010). A Discriminative Approach for Dependency Based Statistical Machine Translation. In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*, pages 66–74, Beijing, China, August. Coling 2010 Organizing Committee.

Venkatapathy, S. (2008). NLP Tools Contest-2008: Machine Translation for English to Hindi . In *Proceedings of the International COnference on Natural Language Processing*.