Morphological Processing for English-Tamil Statistical Machine Translation

Loganathan Ramasamy Ondřej Bojar Zdeněk Žabokrtský
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague
{ramasamy, bojar, zabokrtsky}@ufal.mff.cuni.cz

ABSTRACT
Various experiments from literature suggest that in statistical machine translation (SMT), applying either pre-processing or post-processing to morphologically rich languages leads to better translation quality. In this work, we focus on the English-Tamil language pair. We implement suffix-separation rules for both of the languages and evaluate the impact of this preprocessing on translation quality of the phrase-based as well as hierarchical model in terms of BLEU score and a small manual evaluation. The results confirm that our simple suffix-based morphological processing helps to obtain better translation performance. A by-product of our efforts is a new parallel corpus of 190k sentence pairs gathered from the web.

KEYWORDS: English-Tamil Machine Translation, Parallel Corpora, Suffix Separation.

Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 113–122,
COLING 2012, Mumbai, December 2012.

1 Introduction

For any language pair, there are two main things that affect the performance of an SMT system: (i) the amount of parallel data and (ii) the language differences, mainly the morphological richness and word order differences due to syntactic divergence (Koehn et al., 2009). Indian languages (IL) in general seriously suffer both of the problems especially when they are being translated from/into English. There are very little parallel data for English and Indian languages, and English differs from IL (e.g. Tamil) in both word order (English: SVO, Tamil: SOV) as well as in morphological complexity (English: fusional, Tamil: agglutinative). While the syntactic differences contribute to the difficulties for translation models, the morphological differences contribute to data sparsity. We attempt to address both issues in this paper.

In Section 3, we propose morphological processing aimed at reducing data sparsity. In Section 4, we describe our English-Tamil parallel corpora collection and the system configurations we use. In Section 5, we report the results and analyze them in Section 6.

2 Related Work

Research into SMT involving Tamil language is not very common, the main reason perhaps being the lack of parallel corpora. Nevertheless there have been efforts for other Indian languages such as Hindi (Udupa U. and Faruquie, 2004), (Ramanathan et al., 2008) and (Bojar et al., 2008). The earliest work that appeared on English-Tamil SMT was (Germann, 2001) which described building a small English-Tamil parallel corpus as well as an SMT system. So far, the efforts for building English-Tamil parallel corpora are moderate and the readily available parallel data amount just to a few thousand sentences. One of our goals in this work is to perform experiments with a larger corpus that we collect on our own (see Section 4.1) from various web sources.

The main focus of this work is to address morphological differences between English and Tamil propose steps that improve the performance of SMT systems. Applying morphological processing to SMT is not new, the idea goes back to (Lee, 2004) for Arabic-English or (Nießen and Ney, 2004) for German-English. (Ramanathan et al., 2008) and (Ramanathan et al., 2009) are the first to experiment an Indian language, namely in English-Hindi translation. We apply similar techniques to English-Tamil pair.

3 Suffix Splitting

English and Tamil morphologies follow different inflectional patterns. While English morphology can be adequately described with a few morphological suffixes, thousands of wordforms can be built from a single root in Tamil. As expected, verbs and nouns are the main productive parts of speech in Tamil. For example, a Tamil verb, in addition to the root bearing the lexical information, can include suffixes corresponding to *person, number, gender, tense, negativity, aspect* and *mood*. Most of the additional information which a Tamil word contains can be mapped to *individual functional words* (including prepositions) in English. One type of coordination deserves a special treatment because Tamil uses suffixes instead of coordination conjunctions: '**X***um* **Y***um*' in Tamil corresponds to '**X** *and* **Y**' in English.

Our hypothesis is that separating morphological suffixes from the root and treating them as separate tokens can yield better BLEU performance. We experiment with splitting morphological suffixes on either or both English and Tamil.

Case	Tamil/Transliteration	English Tr.
Accusative	மரத்தை/mara**TTai**	tree
	மரத்தினை/mara**TTinai**	tree
	மரங்களை/mara**ngkaLai**	tree**s**
Instrumental	மரத்தால்/mara**TTAl**	**by a/the** tree
	மரத்தினால்/mara**TTinAl**	**by a/the** tree
	மரங்களால்/mara**ngkaLAl**	**by** tree**s**
Locative	மரத்தில்/mara**TTil**	**in the** tree
	மரத்தினில்/mara**TTinil**	**in the** tree
	மரங்களில்/mara**ngkaLil**	**in the** tree**s**
Dative	மரத்துக்கு/mara**TTukku**	**to/for** tree
	மரத்திற்கு/mara**TTiRku**	**to/for** tree
	மரங்களுக்கு/mara**ngkaLukku**	**to/for** tree**s**
Genitive	மரத்தின்/mara**TTin**	tree**'s**
	மரங்களின்/mara**ngkaLin**	tree**s'**
Ablative	மரத்திலிருந்து/mara**TTiliruwTu**	**from a/the** tree
	மரங்களிலிருந்து/mara**ngkaLiliruwTu**	**from** tree**s**

Tamil/Transliteration	English Translation
படிக்கிறேன்/pati**kkiREn**	**I** study
படிக்கிறோம்/pati**kkiROm**	**We** study
படிக்கிறாய்/pati**kkiRAy**	**You** study
படிக்கிறீர்/pati**kkiRIr**	**You** study (formal)
படிக்கிறீர்கள்/pati**kkiRirkaL**	**You** study (plural)
படிக்கிறான்/pati**kkiRAn**	**He** studies
படிக்கிறாள்/pati**kkiRAL**	**She** studies
படிக்கிறார்/pati**kkiRAr**	**He/She** studies (formal)
படிக்கிறது/pati**kkiRaTu**	**It** studies
படிக்கிறார்கள்/pati**kkiRArkaL**	**They** study (human)
படிக்கின்றன/pati**kkinRana**	**They** study (non human)
படிக்கிறவன்/pati**kkiRavan**	**He who** studies
படிக்காதவன்/pati**kkATavan**	**He who does not** study
படிக்கிறவள்/pati**kkiRavaL**	**She who** studies
படிக்காதவள்/pati**kkATavaL**	**She who does not** study
படிக்கிறது/pati**kkiRaTu**	**That which** studies
படிக்காதது/pati**kkATaTu**	**That which does not** study
படிக்கிறவர்/pati**kkiRavar**	**He/she who** studies
படிக்காதவர்/pati**kkATavar**	**He/sho who does not** study
படிக்கிறவர்கள்/pati**kkiRavarkaL**	**Those who** study
படிக்காதவர்கள்/pati**kkATavarkaL**	**Those who do not** study

Figure 1: Various forms of Tamil noun root: *'maram'* ('tree') and the verb root: *'pati'* ('study').

3.1 Rules

For suffix separation, we identify a number of linguistic rules for both Tamil and English. Each linguistic rule has a form of a regular expression in our system and operates on the wordform. The rules for Tamil operate based on solely the word endings whereas the rules for English also make use of the parts of speech (POS). For Tamil, we have identified **716** inflectional rules for nouns and **519** rules for verbs. Since the number of the rules for Tamil is large, we use three strategies to avoid repeated or often spurious splitting on a wordform: (i) a rule for separating a large suffix (in the number of characters) takes precedence over a rule for a smaller suffix (ii) at most one rule is applied to any wordform and (iii) no rule is applied for wordforms of less than 5 characters after transliteration. At present, our Tamil suffix splitter only works with transliterated data. So, the Tamil side of the parallel corpus must be transliterated from UTF-8 encoding to Latin. Once the suffix splitting is done, the corpus is transliterated back to UTF-8.

3.2 Suffix Splitting: Tamil

Only verbs and nouns are the major parts of speech (Lehmann, 1989) in Tamil that undergo various morphological processes. Although Tamil is an agglutinative language (i.e. suffixes bringing separate morphological features are concatenated one after another), instead of splitting each morpheme into a separate token, we split only suffixes that are often a functional word or a separate token in English. This approach avoids too much spurious splitting.

Tamil nouns mainly inflect for various *case markers* which mostly correspond to individual functional words in English. For example, *'palkalaikkazaka***TTil**' (**in the** *university*) where the *locative* case marker *'TTil'* corresponds to the preposition *'in'* in English. In the same way, the verb suffixes are separated from the inflected verb.

The left part of Figure 1 shows various case inflections for the Tamil noun *'maram'* ('tree'). After the suffix splitting, all the case markers will be separated from the root. Note that the Tamil noun *'maram'* ('tree') is not preserved in full in the declension. Instead, only the stem *'mara'* is recovered. The right part of Figure 1 shows some of the conjugations of verb *'pati'* ('to study').

For example: the gloss *'he who studies'* tries to mimic a relative construct which is represented as one word in Tamil. The pronominal information in *'patikkiRavan'* ('he who studies') indicates that the word refers to a masculine antecedent while the verbal part *'patikkiRavan'* adds the new information: that the person is studying. Syntactically, the word *'patikkiRavan'* behaves as a noun. After our suffix splitting, the verbal part will be separated from the nominal part.

Apart from major inflectional paradigms, we also implemented rules handling *nouns + postpositions* and *sandhis*. It is very common in Tamil to concatenate *postpositions* to the preceding nouns. But in the English translation, they correspond to separate prepositions or other functional words. For example in *'uLwOkkam***illAmal***'* ('**without** *an ulterior motive*'), the suffix **'illAmal'** will be separated from the original wordform to better match with the English translation. So far **70** rules have been identified to split such combinations of *nouns + postpositions*.

One more phenomenon is the *external sandhi*, i.e. the situation when a stop consonant (*k,c, T, p* in Tamil) is added to the end of a word if the following word starts with a stop consonant. For example: in *'aTai***k** *kotukka'* ('to give that'), *'aTai***c** *ceyya'* ('to do that'), *'aTai***T** *Tota'* ('to touch that') and *'aTai***p** *patikka'* ('to read that'), the English word *'that'* is mapped to four different forms in Tamil each differing by the last character. To avoid this data sparsity issue, we add a simple rule that separates this *external sandhi* from Tamil wordforms.

3.3 Suffix Splitting: English

Although the English morphology is not as complex as Tamil, we perform a similar rule-based suffix splitting for English. In English, we proceed in two steps: (i) tag the corpus using Stanford tagger (ii) and apply suffix splitting rules on the tagged data. This process allows us to perform suffix splitting only for certain *word ending - tag* combinations, thus avoiding spurious splitting. Our suffix splitting for English uses 34 *tag-suffix* rules. The rule has the format *'tag (T) - suffix (S)'*, which means that we will separate the suffix (S) from any wordform that has the tag (T).

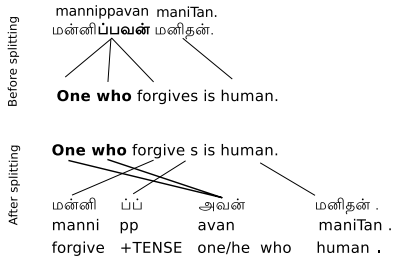

Figure 2: Suffix separation in participial nouns

Figures 2 and 3 illustrate suffix separation for both Tamil and English. The alignment links shown in the figures are not automatically aligned but actual translation links to demonstrate the possibility of better alignment (thanks to the reduced sparsity) after the suffix separation. Figure 2 shows how a participial noun *'mannippavan'* ('one who forgives') can be splitted so that many to one alignment is reduced. Figure 3 illustrates how the separated negative suffix *'Ata'* and the postposition *'il'* correspond directly to individual words in English.

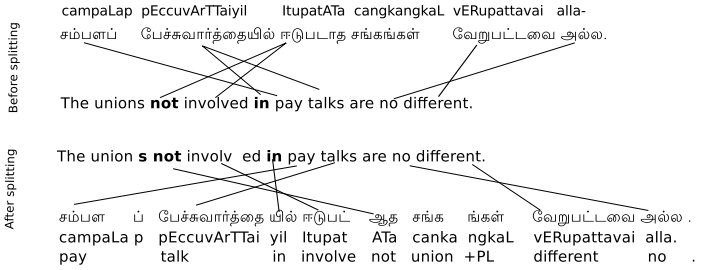

Figure 3: Separation of negative suffixes and postpositions

4 Experimental Setup

4.1 Data

(Germann, 2001) built a small English-Tamil parallel corpus (around 5000 sentences in total) by hiring translators. A focused effort to build a parallel corpus for English and Indian languages including Tamil was initiated by the EILMT[1] consortium project. This attempt too was a manual effort and the parallel corpora were constructed for the health and tourism domain. Recently, (Post et al., 2012) released manually constructed parallel corpora for six Indian languages by crowd-sourced translation.

We feel that there is a decent amount of parallel English-Tamil data available in the web that are largely unnoticed and we thus collect our own corpus quickly and at no cost for translation.

We mainly collect parallel corpora from three sources: (i) www.wsws.org (*News* - news website) (ii) www.cinesouth.com (*Cinema* - Tamil cinema articles) and (iii) biblephone.intercer.net (*Bible*). The above three sources are either multilingual or contain exclusive English and Tamil contents. To collect the *News* corpus, we downloaded only URLs that have matching *file names* on both English and Tamil sides. The collection of *Cinema* corpus was simple: all the English articles had a link to the corresponding Tamil translation on the same page. The collection of *Bible* corpus followed a similar pattern.

After downloading the English URLs and the corresponding Tamil URLs, we stripped all the HTML tags. At this stage, the *Bible* corpus was already sentence aligned. The *News* and *Cinema* articles had similar paragraph structures but they were not sentence aligned. We used *hunalign* (Varga et al.) to sentence align them.

Table 1 summarizes the data sizes. Apart from the development set (1000 sentences), we partition the remaining data into the training portion (90%) and testing portion (10%). The table also includes the statistics of word tokens of both English and Tamil corpora. The corpus `All` combines everything into one big corpus. The Tamil side is encoded in `UTF-8`.

The domain difference between the corpora is reflected in the average sentence length (English, before suffix separation) of 15 for the `Cinema` compared to 26 and 25 for `News` and `Bible`, respectively. The corpora also vary in terms of language style: the style in `Bible` is very different from that of `News` and `Cinema`.

[1]English to Indian Languages Machine Translation (EILMT) is a Government of India funded project.

Corpus	Sentences			Training data		Test data		Dev data	
	Training	Test	Dev	English	Tamil	English	Tamil	English	Tamil
News	108,332	12,037	1000	2.9M	2.1M	328K	247K	27K	20K
				3.4M	3.8M	386K	447K	32K	37K
Cinema	34,690	3854	1000	529K	353K	60K	40K	15K	10K
				605K	610K	68K	69K	18K	18K
Bible	26,884	2987	1000	668K	352K	94K	50K	22K	12K
				733K	731K	103K	103K	24K	24K
All	171,706	19,078	1000	4.1M	2.9M	459K	318K	23K	16K
				4.8M	5.3M	534K	586K	27K	30K

Table 1: Corpus statistics. For each corpus, the *upper* and *lower row* correspond to the number of tokens before and after the suffix splitting.

4.2 Systems Used

We use phrase-based and hierarchical (Chiang, 2005) MT systems as implemented by Koehn et al. (2007) for our experiments. We use the default system settings for all experiments and report results for individual datasets as well as for the entire training data, `All`.

4.3 Examined Configurations

Our experiments consist of the following settings for both phrase based and hierarchical systems:

- **baseline:** The default, no suffix splitting.
- **target**$_{mor}$: No change in English side of the data. Our suffix splitter is run on Tamil.
- **source+target**$_{mor}$: Both the English and Tamil suffix splitters are run on the respective sides of the data.

For each settings, we report BLEU (Papineni et al., 2002) scores in three variations: BLEU$_{suff_sep}$, BLEU$_{suff_rej}$ and BLEU$_{stem_only}$.

In the case of BLEU$_{suff_sep}$ evaluation, both the reference and hypothesis translations are suffix-separated before the evaluation, allowing a better match with the reference but also risking more false positives. The BLEU$_{suff_rej}$ evaluation corresponds to what Tamil readers would like to see: the suffixes are rejoined (if they were separated) prior to evaluation. BLEU$_{stem_only}$ ignores suffixes altogether, both hypothesis and reference translations contain only stem words.

Manual sentence level ranking: We use the WMT-style manual ranking technique (Callison-Burch et al., 2010; Bojar et al., 2011) for a sample of 100 sentences from the test set of `'All'` translated by each of the examined configurations. Without knowing which is which, we rank hypotheses from best to worst for each sentence, allowing ties. The overall score for each system is calculated by considering all pairwise comparisons implied by the rankings. We report three flavours: (i) how often the system was ranked better or equal than other systems ('$\geq others$' $= \frac{wins+ties}{wins+losses+ties}$), (ii) not favoring ties ('$> others$' $= \frac{wins}{wins+ties+losses}$) and (iii) ignoring ties altogether ('$no\ ties$' $= \frac{wins}{wins+losses}$).

5 Results

The results for phrase-based and hierarchical MT systems are given in Tables 2 and 3, respectively. Comparing **baseline** scores for both the phrase-based and the hierarchical systems, the

hierarchical system performs better or equally well across all the corpora.

System	BLEU$_{suff_sep}$				BLEU$_{suff_rej}$			
	News	Cinema	Bible	All	News	Cinema	Bible	All
baseline	10.97	8.54	12.87	12.45	**6.15**	**7.13**	**6.10**	**7.44**
target$_{mor}$	**13.79**	10.40	18.27	**14.30**	4.91	7.01	5.82	6.05
source+target$_{mor}$	13.69	**10.56**	**18.30**	14.15	4.74	7.23	5.79	5.98

Table 2: Results for phrase based SMT

System	BLEU$_{suff_sep}$				BLEU$_{suff_rej}$			
	News	Cinema	Bible	All	News	Cinema	Bible	All
baseline	10.98	8.62	14.59	13.09	**6.20**	**7.41**	6.92	**7.78**
target$_{mor}$	14.01	**10.92**	**19.56**	14.82	4.94	7.33	**7.25**	6.40
source+target$_{mor}$	**14.17**	8.84	19.29	**15.12**	4.85	5.87	6.79	6.43

Table 3: Results for hierarchical SMT

Evaluating suffixes separately (BLEU$_{suff_sep}$), we see big jumps in the scores when the target or both sides of the training data were splitted. Note that for BLEU$_{suff_sep}$, the baseline system output is subjected to suffix splitting, but only *after* the translation. In the **baseline** of both Table 2 and 3, the BLEU score sharply increases for `Bible` than the `News` and `Cinema` when we compare BLEU$_{suff_sep}$ and BLEU$_{suff_rej}$. One reason could be, the increase in the number of tokens (in the reference data) after the suffix separation of `Bible` (85.4%) is larger than the `News` (71.6%) and `Cinema` (57.1%), in other words, the `Bible` has morphologically more complex forms than the other two corpora. We also observe from Table 2 and 3 that the BLEU differences between **target**$_{mor}$ and **source+target**$_{mor}$ is narrow compared to their **baseline** counterparts in BLEU$_{suff_sep}$ evaluation. Rejoining the suffixes appears detrimental for BLEU$_{suff_rej}$ (in both Table 2 and 3) but we feel that the observed loss is caused rather by the properties of BLEU. This is because, even a small change in wordforms are treated as separate tokens in the BLEU evaluation.

System	Phrase based				Hierarchical			
	News	Cinema	Bible	All	News	Cinema	Bible	All
baseline	7.60	8.08	7.87	8.96	7.68	8.40	9.17	9.40
target$_{mor}$	**8.50**	8.62	**9.33**	**9.22**	**8.60**	**9.06**	**10.69**	**9.77**
source+target$_{mor}$	8.36	**8.98**	9.17	9.13	**8.60**	7.70	10.43	9.73

Table 4: BLEU$_{stem_only}$ evaluation results

In the stem only evaluation (Table 4), splitting suffixes on the target side of the training data helps in all cases except the `Cinema` domain translated with the phrase based system. The target-only vs. both-sides splitting are incomparably close.

System	Phrase based			Hierarchical		
	$\geq$ *others*	$>$ *others*	*no ties*	$\geq$ *others*	$>$ *others*	*no ties*
baseline	49.0	31.0	37.8	43.5	36.0	38.9
target$_{mor}$	62.5	44.5	54.3	**60.5**	**51.5**	**56.6**
source+target$_{mor}$	**64.5**	**48.5**	**57.7**	58.5	50.0	54.6

Table 5: Manual evaluation on 100 sentences (sentence-level ranking).

In the case of manual evaluation (Table 5), the **source+target**$_{mor}$ is ranked as the best in phrase based system whereas in hierarchical system the **target**$_{mor}$ performs better. Comparing different evaluation systems, both **target**$_{mor}$ and **source+target**$_{mor}$ helps achieving better performance than the **baseline**. But, as the results suggest, the **target**$_{mor}$ performs only marginally better than **source+target**$_{mor}$ in general.

6 Observations & Error Analysis

From the previous section, we observed that morphological suffix splitting improves the performance. Following are some of the observations and possible suggestions in general to improve the performance further.

- Interpretation of automatic scores like BLEU deserves a great care as the results are heavily affected by tokenization.
- Suffix splitting reduces only the sparsity problem. This does not solve the agreement problem such as adding subject's gender suffixes on the target side verb.
- Suffix splitting reduces the sparsity by allowing more one-to-one word alignments but that could lead to complex reordering scenarios, see Figure 3.
- Coordination is one of the difficult phenomenon in Tamil. In most cases, both phrase-based and hierarchical system translations produced the English style coordination instead of adding coordination suffixes to all the conjuncts. This behaviour could be tweaked by preprocessing English and adding fake tokens to serve as "placeholders" for Tamil coordination suffixes.
- Both *News* and *Cinema* corpora are sentence aligned automatically using a sentence aligner. Although a strict threshold has been set to eliminate improbable alignments, there could be a minor percentage of misaligned sentences.

Conclusion

In this work, we described our experiments with separation of morphological suffixes in English and Tamil to improve translation quality of phrase-based and hierarchical machine translation systems. We demonstrated that suffix separation helps in reducing the data sparsity and improves translation quality.

We also documented our efforts to collect parallel corpora for English and Tamil from web sources, obtaining about 190,000 sentence pairs in total. To our knowledge, this is currently the largest amount of data available for the English-Tamil language pair.

Acknowledgments

The research leading to these results has received funding from the European Commission's 7th Framework Program (FP7) under grant agreement n° 238405 (CLARA) and from the Czech Grant Agency project number P406/10/P259.

References

Bojar, O., Ercegovčević, M., Popel, M., and Zaidan, O. (2011). A grain of salt for the wmt manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland. Association for Computational Linguistics.

Bojar, O., Straňák, P., and Zeman, D. (2008). English-Hindi Translation in 21 Days. In *Proceedings of the 6th International Conference On Natural Language Processing (ICON-2008) NLP Tools Contest*, Pune, India. NLP Association of India.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and Zaidan, O. F. (2010). Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 17–53, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 263–270, Stroudsburg, PA, USA. Association for Computational Linguistics.

Germann, U. (2001). Building a statistical machine translation system from scratch: How much bang for the buck can we expect? In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, pages 63–70.

Koehn, P., Birch, A., and Steinberger, R. (2009). 462 Machine Translation Systems for Europe. In *MT Summit XII*.

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 *ACL*.

Lee, Y.-S. (2004). Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lehmann, T. (1989). *A Grammar of Modern Tamil*. Pondicherry Institute of Linguistics and Culture.

Nießen, S. and Ney, H. (2004). Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204.

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

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. Association for Computational Linguistics.

Ramanathan, A., Bhattacharyya, P., Hegde, J., Shah, R. M., and M, S. (2008). Simple syntactic and morphological processing can help english-hindi statistical machine translation. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP)*.

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 Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 800–808, Stroudsburg, PA, USA. Association for Computational Linguistics.

Udupa U., R. and Faruquie, T. A. (2004). An english-hindi statistical machine translation system. In *Proceedings of the First international joint conference on Natural Language Processing*, IJCNLP'04, pages 254–262, Berlin, Heidelberg. Springer-Verlag.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590–596.