Reducing the Impact of Data Sparsity in Statistical Machine Translation

Karan Singla[1], **Kunal Sachdeva**[1], **Diksha Yadav**[1], **Srinivas Bangalore**[2], **Dipti Misra Sharma**[1]

[1]LTRC IIIT Hyderabad,[2]AT&T Labs-Research

Abstract

Morphologically rich languages generally require large amounts of parallel data to adequately estimate parameters in a statistical Machine Translation(SMT) system. However, it is time consuming and expensive to create large collections of parallel data. In this paper, we explore two strategies for circumventing sparsity caused by lack of large parallel corpora. First, we explore the use of distributed representations in an Recurrent Neural Network based language model with different morphological features and second, we explore the use of lexical resources such as WordNet to overcome sparsity of content words.

1 Introduction

Statistical machine translation (SMT) models estimate parameters (lexical models, and distortion model) from parallel corpora. The reliability of these parameter estimates is dependent on the size of the corpora. In morphologically rich languages, this sparsity is compounded further due to lack of large parallel corpora.

In this paper, we present two approaches that address the issue of sparsity in SMT models for morphologically rich languages. First, we use an Recurrent Neural Network (RNN) based language model (LM) to re-rank the output of a phrase-based SMT (PB-SMT) system and second we use lexical resources such as WordNet to minimize the impact of Out-of-Vocabulary(OOV) words on MT quality. We further improve the accuracy of MT using a model combination approach.

The rest of the paper is organized as follows. We first present our approach of training the baseline model and source side reordering. In Section 4, we present our experiments and results on re-ranking the MT output using RNNLM. In Section 5, we discuss our approach to increase the coverage of the model by using synset ID's from the English WordNet (EWN). Section 6 describes our experiments on combining the model with synset ID's and baseline model to further improve the translation accuracy followed by results and observations sections.We conclude the paper with future work and conclusions.

2 Related Work

In this paper, we present our efforts of re-ranking the n-best hypotheses produced by a PB-MT (Phrase-Based MT) system using RNNLM (Mikolov et al., 2010) in the context of an English-Hindi SMT system. The re-ranking task in machine translation can be defined as re-scoring the n-best list of translations, wherein a number of language models are deployed along with features of source or target language. (Dungarwal et al., 2014) described the benefits of re-ranking the translation hypothesis using simple n-gram based language model. In recent years, the use of RNNLM have shown significant improvements over the traditional n-gram models (Sundermeyer et al., 2013). (Mikolov et al., 2010) and (Liu et al., 2014) have shown significant improvements in speech recognition accuracy using RNNLM . Shi (2012) also showed the benefits of using RNNLM with contextual and linguistic features. We have also explored the use of morphological features (Hindi being a morphologically rich language) in RNNLM and deduced that these features further improve the baseline RNNLM in re-ranking the n-best hypothesis.

Words in natural languages are richly diverse so it is not possible to cover all source language words when training an MT system. Untranslated out-of-vocabulary (OOV) words tend to degrade the accuracy of the output produced by an MT model. Huang (2010) pointed to various types of OOV words which occur in a data set – seg-

Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 51–56,
October 25, 2014, Doha, Qatar.

mentation error in source language, named entities, combination forms (e.g. *widebody*) and abbreviations. Apart from these issues, Hindi being a low-resourced language in terms of parallel corpora suffers from data sparsity.

In the second part of the paper, we address the problem of data sparsity with the help of English WordNet (EWN) for English-Hindi PB-SMT. We increase the coverage of content words (excluding Named-Entities) by incorporating sysnset information in the source sentences.

Combining Machine Translation (MT) systems has become an important part of statistical MT in past few years. Works by (Razmara and Sarkar, 2013; Cohn and Lapata, 2007) have shown that there is an increase in phrase coverage when combining different systems. To get more coverage of unigrams in phrase-table, we have explored system combination approaches to combine models trained with synset information and without synset information. We have explored two methodologies for system combination based on confusion matrix(dynamic) (Ghannay et al., 2014) and mixing models (Cohn and Lapata, 2007).

3 Baseline Components

3.1 Baseline Model and Corpus Statistics

We have used the ILCI corpora (Choudhary and Jha, 2011) for our experiments, which contains English-Hindi parallel sentences from tourism and health domain. We randomly divided the data into training (48970), development (500) and testing (500) sentences and for language modelling we used news corpus of English which is distributed as a part of WMT'14 translation task. The data is about 3 million sentences which also contains MT training data.

We trained a phrase based (Koehn et al., 2003) MT system using the Moses toolkit with word-alignments extracted from GIZA++ (Och and Ney, 2000). We have used the SRILM (Stolcke and others, 2002) with Kneser-Ney smoothing (Kneser and Ney, 1995) for training a language model for the first stage of decoding. The result of this baseline system is shown in Table 1.

Number of Training Sentences	Number of Development Sentences	Number of Evaluation Sentences	BLEU
48970	500	500	20.04

Table 1: Baseline Scores for Phrase-based Moses Model

3.2 English Transformation Module

Hindi is a relatively free-word order language and generally tends to follow SOV (Subject-Object-Verb) order and English tends to follow SVO (Subject-Verb-Object) word order. Research has shown that pre-ordering source language to conform to target language word order significantly improves translation quality (Collins et al., 2005). We created a re-ordering module for transforming an English sentence to be in the Hindi order based on reordering rules provided by Anusaaraka (Chaudhury et al., 2010). The reordering rules are based on parse output produced by the Stanford Parser (Klein and Manning, 2003).

The transformation module requires the text to contain only surface form of words, however, we extended it to support surface form along with its factors such as lemma and Part of Speech (POS).

Input : the girl in blue shirt is my sister

Output : in blue shirt the girl is my sister.

Hindi : neele shirt waali ladki meri bahen hai (blue) (shirt) (Mod)(girl)(my)(sister)(Vaux)

With this transformation, the English sentence is structurally closer to the Hindi sentence which leads to better phrase alignments. The model trained with the transformed corpus produces a new baseline score of 21.84 BLEU score an improvement over the earlier baseline of 20.04 BLEU points.

4 Re-Ranking Experiments

In this section, we describe the results of re-ranking the output of the translation model using Recurrent Neural Networks (RNN) based language models using the same data which is used for language modelling in the baseline models.

Unlike traditional n-gram based discrete language models, RNN do not make the Markov assumption and potentially can take into account long-term dependencies between words. Since the words in RNNs are represented as continuous valued vectors in low dimensions allowing for the possibility of smoothing using syntactic and semantic features. In practice, however, learning long-term dependencies with gradient descent is difficult as described by (Bengio et al., 1994) due to diminishing gradients.

We have integrated the approach of re-scoring

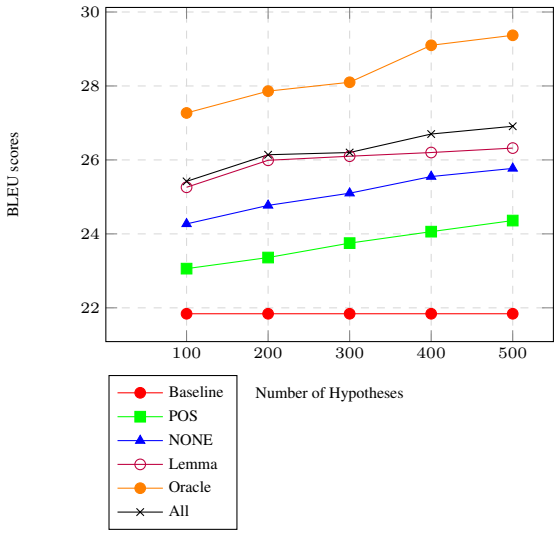

Figure 1: BLEU Scores for Re-ranking experiments with RNNLM using different feature combinations.

n-best output using RNNLM which has also been shown to be helpful by (Liu et al., 2014). Shi (2012) also showed the benefits of using RNNLM with contextual and linguistic features. Following their work, we used three type of features for building an RNNLM for Hindi : lemma (root), POS, NC (number-case). The data used was a Wikipedia dump, MT training data, news articles which had approximately 500,000 Hindi sentences. Features were extracted using paradigm-based Hindi Morphological Analyzer [1]

Figure 1 illustrates the results of re-ranking performed using RNNLM trained with various features. The *Oracle* score is the highest achievable score in a re-ranking experiment. This score is computed based on the best translation out of n-best translations. The best translation is found using the cosine similarity between the hypothesis and the reference translation. It can be seen from Figure 1, that the LM with only word and POS information is inferior to all other models. However, morphological features like lemma, number and case information help in re-ranking the hypothesis significantly. The RNNLM which uses all the features performed the best for the re-ranking experiments achieving a BLEU score of 26.91, after re-scoring 500-best obtained from the pre-order SMT model.

[1]We have used the HCU morph-analyzer.

	System	BLEU
	Baseline	21.84
Rescoring 500-best with RNNLM		
Features	NONE	25.77
	POS	24.36
	Lemma(root)	26.32
	ALL(POS+Lemma+NC)	26.91

Table 2: Rescoring results of 500-best hypotheses using RNNLM with different features

5 Using WordNet to Reduce Data Sparsity

We extend the coverage of our source data by using synonyms from the English WordNet (EWN). Our main motivation is to reduce the impact of OOV words on output quality by replacing words in a source sentence with their corresponding synset IDs. However, choosing the appropriate synset ID based upon its context and morphological information is important. For sense selection, we followed the approach used by (Tammewar et al., 2013), which is also described further in this section in the context of our task. We ignored words that are regarded as Named-Entities as indicated by Stanford NER tagger, as they should not have synonyms in any case.

5.1 Sense Selection

Words are ambiguous, independent of their sentence context. To choose an appropriate sense according to the context for a lexical item is a challenging task typically termed as word-sense disambiguation. However, the syntactic category of a lexical item provides an initial cue for disambiguating a lexical item. Among the varied senses, we filter out the senses that are not the same POS tag as the lexical item. But words are not just ambiguous across different syntactic categories but are also ambiguous within a syntactic category. In the following, we discuss our approaches to select the sense of a lexical item best suited in a given context within a given category. Also categories were filtered so that only content words get replaced with synset IDs.

5.1.1 Intra-Category Sense Selection

First Sense: Among the different senses,we select the first sense listed in EWN corresponding to the POS-tag of a given lexical item. The choice is motivated by our observation that the senses of a

lexical item are ordered in the descending order of their frequencies of usage in the lexical resource.

Merged Sense: In this approach, we merge all the senses listed in EWN corresponding to the POS-tag of the given lexical item. The motivation behind this strategy is that the senses in the EWN for a particular word-POS pair are too finely classified resulting in classification of words that may represent the same concept, are classified into different synsets. For example : *travel* and *go* can mean the same concept in a similar context but the first sense given by EWN is different for these two words. Therefore, we merge all the senses for a word into a super sense (synset ID of first word occurred in data), which is given to all its synonyms even if it occurs in different synset IDs.

5.2 Factored Model

Techniques such as factored modelling (Koehn and Hoang, 2007) are quite beneficial for Translation from English to Hindi language as shown by (Ramanathan et al., 2008). When we replace words in a source sentence with the synset IDâs, we tend to lose morphological information associated with that word. We add inflections as features in a factored SMT model to minimize the impact of this replacement.

We show the results of the processing steps on an example sentence below.
Original Sentence : Ram is going to market to buy apples
New Sentence : Ram is Synset(go.v.1) to Synset(market.n.0) to Synset(buy.v.1) Synset(apple.n.1)
Sentence with synset ID: Ram_E is_E Synset(go.v.1)_ing to_E Synset(market.n.0)_E to_E Synset(buy.v.1)_E Synset(apple.n.1)_s
Then English sentences were reordered to Hindi word-order using the module discussed in Section 3.
Reordered Sentence: Ram_E Synset(apple.n.1)_s Synset(buy.v.1)_E to_E Synset(market.n.0)_E to_E Synset(go.v.1)_ing is_E

In Table 3, the second row shows the BLEU scores for the models in which there are synset IDs for the source side. It can be seen that the factored model also shows significant improvement in the results.

6 Combining MT Models

Combining Machine translation (MT) systems has become an important part of Statistical MT in the past few years. There are two dominant approaches. (1) a system combination approach based on confusion networks (CN) (Rosti et al., 2007), which can work dynamically in combining the systems. (2) Combine the models by linearly interpolating and then using MERT to tune the combined system.

6.1 Combination based on confusion networks

We used the tool MANY (Barrault, 2010) for system combination. However, since the tool is configured to work with TERp evaluation metric, we modified it to use METEOR (Gupta et al., 2010) metric since it has been shown by (Kalyani et al., 2014), that METEOR evaluation metric is better correlated to human evaluation for morphologically rich Indian Languages.

6.2 Linearly Interpolated Combination

In this approach, we combined phrase-tables of the two models (Eng (sysnset) - Hindi and Baseline) using linear interpolation. We combined the two models with uniform weights – 0.5 for each model, in our case. We again tuned this model with the new interpolated phrase-table using standard algorithm MERT.

7 Experiments and Results

As can be seen in Table 3, the model with synset information led to reduction in OOV words. Even though BLEU score decreased, but METEOR score improved for all the experiments based on using synset IDs in the source sentence, but it has been shown by (Gupta et al., 2010) that METEOR is a better evaluation metrics for morphologically rich languages. Also, when synset IDâs are used instead of words in the source language, the system makes incorrect morphological choices. Example : *going* and *goes* will be replaced by same synset ID âSynset(go.v.1)â, so this has lead to loss of information in the phrase-table but METEOR catches these complexities as it considers features like stems, synonyms for its evaluation metrics and hence showed better improvements compared to BLEU metric. Last two rows of Table 3 show results for combination experiments and Mixture Model (linearly interpolated model) showed best

System		#OOV words	BLEU	Meteor
Baseline		253	21.8	.492
Eng(Synset ID)-Hindi	Baseline	237	19.2	.494
	*factor(inflections)	225	20.3	.506
Ensembled Decoding		213	21.0	.511
Mixture Model		210	21.2	.519

Table 3: Results for the model in which there were Synset ID's instead of word in English data

results with significant reduction in OOV words and also some gains in METEOR score.

8 Observations

In this section, we study the coverage of different models by categorizing the OOV words into 5 categories.

- **NE(Named Entities)** : As the data was from Health & Tourism domain, these words were mainly the names of the places and medicines.
- **VB** : types of verb forms
- **NN** : types of nouns and pronouns
- **ADJ** : all adjectives
- **AD** : adverbs
- **OTH** : there were some words which did not mean anything in English
- **SM** : There were some occasional spelling mistakes seen in the test data.

Note : There were no function words seen in the OOV(un-translated) words

Cat.	Baseline	Eng(synset)-Hin	MixtureModel
NE	120	121	115
VB	47	37	27
NN	76	60	47
ADJ	22	15	12
AD	5	5	4
OTH	2	2	2
SM	8	8	8

Table 4: OOV words in Different Models

As this analysis was done on a small dataset and for a fixed domain, the OOV words were few in number as it can be seen in Table 4. But the OOV words across the different models reduced as expected. The NE words remained almost the same for all the three models but OOV words from category VB,NN,ADJ decreased for Eng(synset)-Hin model and Mixture model significantly.

9 Future Work

In the future, we will work on using the two approaches discussed: Re-Ranking & using lexical resources to reduce sparsity together in a system. We will work on exploring syntax based features for RNNLM and we are planning to use a better method for sense selection and extending this concept for more language pairs. Word-sense disambiguation can be used for choosing more appropriate sense when the translation model is trained on a bigger data data set. Also we are looking for unsupervised techniques to learn the replacements for words to reduce sparsity and ways to adapt our system to different domains.

10 Conclusions

In this paper, we have discussed two approaches to address sparsity issues encountered in training SMT models for morphologically rich languages with limited amounts of parallel corpora. In the first approach we used an RNNLM enriched with morphological features of the target words and show the BLEU score to improve by 5 points. In the second approach we use lexical resource such as WordNet to alleviate sparsity.

References

Loïc Barrault. 2010. Many: Open source machine translation system combination. *The Prague Bulletin of Mathematical Linguistics*, 93:147–155.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *Neural Networks, IEEE Transactions on*, 5(2):157–166.

Sriram Chaudhury, Ankitha Rao, and Dipti M Sharma. 2010. Anusaaraka: An expert system based machine translation system. In *Natural Language Processing*

and Knowledge Engineering (NLP-KE), 2010 International Conference on, pages 1–6. IEEE.

Narayan Choudhary and Girish Nath Jha. 2011. Creating multilingual parallel corpora in indian languages. In *Proceedings of Language and Technology Conference*.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 728. Citeseer.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 531–540. Association for Computational Linguistics.

Piyush Dungarwal, Rajen Chatterjee, Abhijit Mishra, Anoop Kunchukuttan, Ritesh Shah, and Pushpak Bhattacharyya. 2014. The iit bombay hindi-english translation system at wmt 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 90–96, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Sahar Ghannay, France Le Mans, and Loıc Barrault. 2014. Using hypothesis selection based features for confusion network mt system combination. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)@ EACL*, pages 1–5.

Ankush Gupta, Sriram Venkatapathy, and Rajeev Sangal. 2010. Meteor-hindi: Automatic mt evaluation metric for hindi as a target language. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*.

Chung-chi Huang, Ho-ching Yen, and Jason S Chang. 2010. Using sublexical translations to handle the oov problem in mt. In *Proceedings of The Ninth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Aditi Kalyani, Hemant Kamud, Sashi Pal Singh, and Ajai Kumar. 2014. Assessing the quality of mt systems for hindi to english translation. In *International Journal of Computer Applications*, volume 89.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *EMNLP-CoNLL*, pages 868–876. Citeseer.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

X Liu, Y Wang, X Chen, MJF Gales, and PC Woodland. 2014. Efficient lattice rescoring using recurrent neural network language models.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics.

Ananthakrishnan Ramanathan, Jayprasad Hegde, Ritesh M Shah, Pushpak Bhattacharyya, and M Sasikumar. 2008. Simple syntactic and morphological processing can help english-hindi statistical machine translation. In *IJCNLP*, pages 513–520.

Majid Razmara and Anoop Sarkar. 2013. Ensemble triangulation for statistical machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 252–260.

Antti-Veikko I Rosti, Spyridon Matsoukas, and Richard Schwartz. 2007. Improved word-level system combination for machine translation. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 312. Citeseer.

Yangyang Shi, Pascal Wiggers, and Catholijn M Jonker. 2012. Towards recurrent neural networks language models with linguistic and contextual features. In *INTERSPEECH*.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*.

Martin Sundermeyer, Ilya Oparin, J-L Gauvain, Ben Freiberg, R Schluter, and Hermann Ney. 2013. Comparison of feedforward and recurrent neural network language models. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8430–8434. IEEE.

Aniruddha Tammewar, Karan Singla, Srinivas Bangalore, and Michael Carl. 2013. Enhancing asr by mt using semantic information from hindiwordnet. In *Proceedings of ICON-2013: 10th International Conference on Natural Language Processing*.