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]International Institute of Information Technology - Hyderabad, [2]ATT Labs-Research

at&t

IIIT, HYDERABAD

Introduction

- 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.
- We explore two strategies for circumventing sparsity caused by lack of large parallel corpora.

Approach

- First, we use a Recurrent Neural Network (RNN) based language model (LM) to rerank the output of a phrase based SMT (PB-SMT) system.
- 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.

Baseline Components

- We have used ILCI corpora for our experiments.
- We trained a phrase based MT system using the Moses toolkit with word alignments extracted from GIZA++.
- We then used the SRILM with Kneser-Ney smoothing for training a language model for the first stage of decoding.

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

English Transformation Module

- We created a re-ordering module for transforming an English sentence to be in the Hindi order based on reordering rules provided by Anusaaraka (open-source MT system) .
- The reordering rules are based on parse output produced by the Stanford Parser.
- With this transformation,the English sentence is structurally closer to Hindi sentence which leads to better phrase alignments.
- The model trained with transformed corpus produces a new baseline score of 21.84 BLEU

Examples

Original -
Traditionally, the Governor can decide on the fortnightly report to the President .

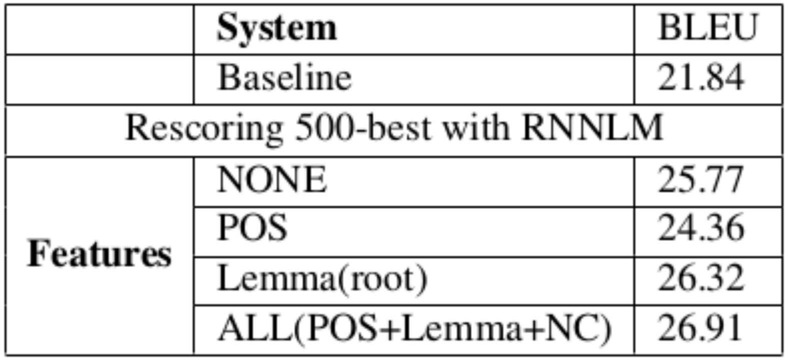
Upon reordering -
Traditionally, the Governor the President to the fortnightly report on decide can .

Why Recurrent Neural Network

- RNN do not make the Markov assumption and potentially can take into account long-term dependencies between words.
- Furthermore, with the words in RNNs are represented as continuous valued vectors in low dimensions allowing for the possibility of smoothing using syntactic and semantic features.

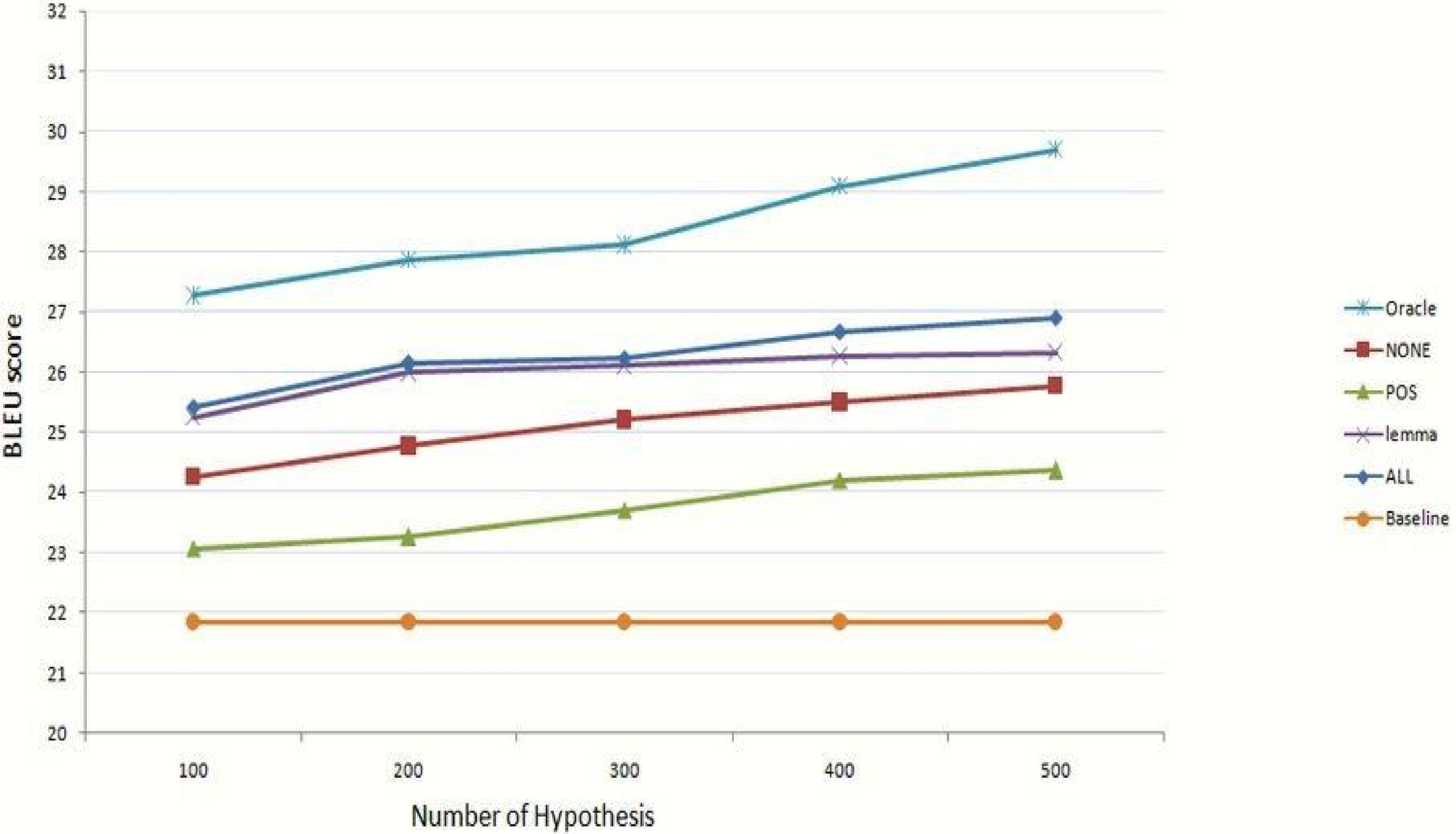
Re-Ranking

- We Reranked the n-best list output using RNN language model using RNNLM tool kit [Mikolov et al, 2012] .
- We used three type of features for building an RNNLM for Hindi : Lemma(root), POS, NC (number-case).



	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

Re-Ranking Experiment



WordNet to Reduce Data Sparsity

- We extend the coverage of our source data by replacing words in the source sentences with their Synset IDs.
- 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.
- Only content words get replaced with synset IDs.
- We follow two approaches to select the sense with the same syntactic category.

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.

Intra-Category Sense Selection

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.
- 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 sysnet IDs.

Factored Model

- When we replace words in a source sentence with the Synset IDs, 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.
  - **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
  - **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

Combining MT models

Combination based on confusion networks

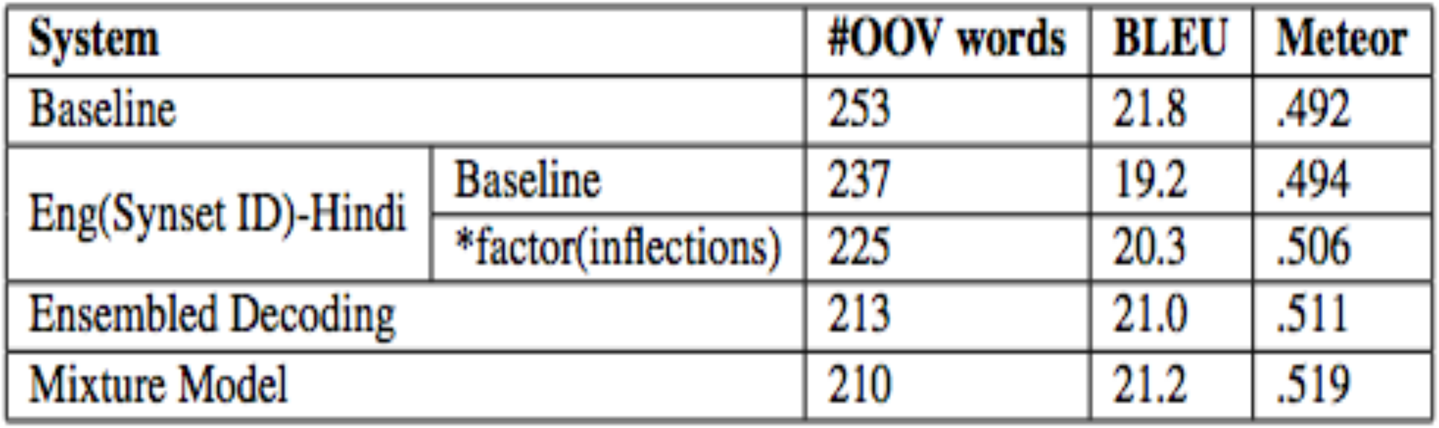
- We used the tool MANY to combine the systems.
- Since the tool is configured to work with TERp evaluation metric,we modified it to use METEOR since METEOR evaluation metric is better correlated to human evaluation for morphologically rich Indian Languages.

Combining MT models

Linearly Interpolated Combination

- In this approach, we combined phrase-tables of the two models (Eng (synset) - 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.

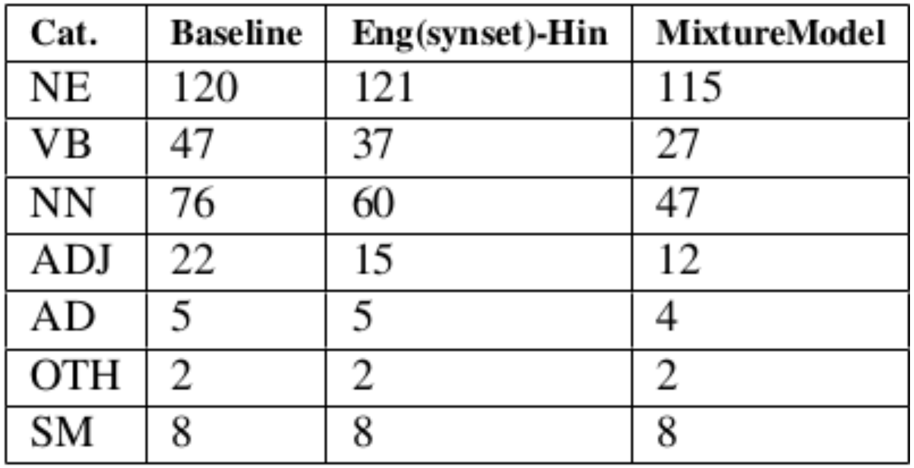
Experiments & Results



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

Observations

Coverage of different models on Out of Vocabulary (OOV) 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

The OOV words across the different models reduced as expected except NE since named entities need not be substituted.