Phrase Based SMT Using Pivot Language

Ankith MS 143059007
Eeshan Malhotra 14305R001

Motivation for using Pivot Language

Reason 1

Lack of parallel corpora between source and target language.

In this scenario, we would like to use a resource intensive language such as English as pivot

Reason 2

Improvement of translation.

In this scenario we would like to use a language that is closely related to either the source or the target language. There is some evidence that the pivot being close to the target language helps improve translation.

This hypothesis will also be verified in our experiments

Methodology for Using Pivot

1. **Cascading**
   Use two translations as a pipeline. This is a direct and simple method

2. **Triangulation**
   Construct a phrase-table for source-target using phrase table for source-pivot and phrase table for pivot-target (Not a MOSES functionality, requires coding)

Mathematical Modeling

We want to model

$$p(e|f)$$

This is calculated as the (weighted) product of four different models

Phrase Translation Model	Language Model	Distortion Model	Word Penalty Model
$\phi(f\|e)$	LM	$D(e,f)$	$W(e)$

Since it is a multiplicative model, we use weights in the exponent

$$p(e|f) = \phi(f|e)^{\text{weight}_\phi} \times \mathrm{LM}^{\text{weight}_{\mathrm{LM}}} \times D(e,f)^{\text{weight}_d} \times W(e)^{\text{weight}_W}$$

Steps in Training

- The first step is preparation of corpora - This involves steps like cleaning, tokenization, proper casing (in appropriate language).
  This is done using inbuilt MOSES scripts

- **Language Model**
  For this, we utilize IRSTLM. We need the language model for the target language only

- **Word Alignment**
  This is done using Giza++. The methodology is the expansion-technique discussed in class

- **Phrase Extraction and Scoring**
- **Lexicalized reordering Model**
- **Combined Generation Model**
  These steps are performed using Moses

Steps in Training (contd.)

- Weights determination
  Now, we need to determine the weighting of each parameter

  (Recall $p(e|f) = \phi(f|e)^{\text{weight}_\phi} \times \text{LM}^{\text{weight}_{\text{LM}}} \times D(e,f)^{\text{weight}_d} \times W(e)^{\text{weight}_w}$)

  For this we use a *separate* tuning corpus we have already set aside

  This is also done using a functionality in MOSES

Project Progress

1. Become familiar with Moses, GIZA++, IRST-LM Toolkit, including preprocessing steps (tokenization, cleaning, etc.). Train language model and phrase table, experiment with tuning - Done ✓

   We decided to use the three languages as:

   Source: English

   Target: Hindi

   Pivot: Marathi

   (This is in line with the use of pivot as described on first slide)

2. Establish Baseline for English-Hindi translation, for varying corpora sizes - Done ✓ (Results included)

Project Progress

3. Create phrase table for source-target, using phrase tables for source-pivot, and pivot-target through Python code - Done.

    3.a Combine Phrase translation probabilities- Done.

    3.b Calculated lexical probabilities- Done.

    3.c Merge alignments - (merge via highest likelihood pivot) - Done.

Phrase tables used from www.cfilt.iitb.ac.in/~moses/shata_anuvaadak

4. Run MOSES on combined phrase table - facing some issues

    srilm issue - Resolved

    MOSES code-base compatibility issues - Resolved

    Special characters issue - Resolved

5. Compare triangulation results with baseline results calculated- Done

Reduction Techniques

1. **Sampling**

   Sample from the set of *f*s, and retain all *e*|*f* pairs for the chosen *f*

2. **Probability Pruning**

   Discard *lines* of *e*|*f* pairs if *P(e|f)* < *T*, where the threshold T is decided based on the desired size of table

3. **Relative Threshold Pruning**

   Discard phrases that are far worse than the best target phrase for a given source phrase, i.e. if

$$P(e|f) < T * \max_{e} \{ P(e|f) \}$$

In cases 2 and 3, the probabilities need to be normalized such that P(E|F) sums to 1

Results - Baselines

Source: English; Pivot: Hindi; Target: Marathi

Method	BLEU	Multi-BLEU
Direct (No Pivot)	10.33	36.6/14.7/6.7/3.2
Direct - Probability Pruning	10.45	36.9/14.8/6.8/3.2
Direct - Relative Pruning	10.51	37.3/14.9/6.8/3.2

All results are reported for tuned parameters.

For the next set of results on pivot-based translation, a small Source-Target corpus was assumed available and used for tuning

Results - Pivot Based

Method	BLEU	Multi-BLEU
Direct (No Pivot)	10.33	36.6/14.7/6.7/3.2
Direct - Probability Pruning	10.45	36.9/14.8/6.8/3.2
Direct - Relative Pruning	10.51	37.0/14.9/6.8/3.2
Cascading	10.73	36.9/15.5/7.0/3.3

Results - Pivot Based

Method	BLEU	Multi-BLEU
Direct (No Pivot)	10.33	36.6/14.7/6.7/3.2
Direct - Probability Pruning	10.45	36.9/14.8/6.8/3.2
Direct - Relative Pruning	10.51	37.0/14.9/6.8/3.2
Cascading	10.73	36.9/15.5/7.0/3.3
Triangulation + Augmentation: Sampling (400k/1.5mm)	1.2	10.8/1.8/0.6/0.2
Triangulation + Augmentation: Probability Pruning (at source)	10.18	35.5/14.3/6.6/3.2
Triangulation + Augmentation: Relative Pruning (at source)	10.66	37.1/15.3/7.0/3.2

Results - To Do

Method	BLEU	Multi-BLEU
Triangulation + Augmentation: Probability Pruning (on final)	?	?
Triangulation + Augmentation: Relative Pruning (on final)	?	?

Challenge:

Extremely slow to create lexical probabilities on un-pruned Phrase Table.

Triangulation of un-pruned S-P (1.5mm) and P-T (1.5mm) tables gives a S-T table of size > 20mm!

Next Steps

1. Attempt to overcome the lexical probability problem with use of better techniques
2. Complete the same analysis using English as Pivot language
3. If time permits, try significance based pruning

References

1. Hua Wu and Haifeng Wang. 2007. **Pivot Language Approach for Phrase-Based Statistical Machine Translation.** In Proceedings of 45th Annual Meeting of the Association for Computational Linguistics, pages 856-863.
2. Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. **Phrase-Based Statistical Machine Translation with Pivot Languages**. In Proceedings of the International Workshop on Spoken Language Translation, pages 143- 149.
3. M. Paul, H. Yamamoto, E. Sumita and S. Nakamura, **On the Importance of Pivot Language Selection for Statistical Machine Translation**, 2009, Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics
4. Zens, Richard, Daisy Stanton, and Peng Xu. A **systematic comparison of the phrase table pruning techniques**, Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012.