FSMNLP 2017

‘Word Transduction’

An Approach for Addressing the Out-of-Vocabulary Words Problem in Machine Translation for Similar Resource-Scarce Languages

Shashikant Sharma and Anil Kumar Singh
Dept. of CSE
IIT (BHU), Varanasi, India

Outline

- Background and motivation
- ‘Word transduction’
- Phonetic transcription and features
- Approaches for word transduction
- Experiments and evaluation methods
- Summary

Background

- Part of the project on building transfer-based MT systems for Bhojpuri, Maithili and Magahi
  - Languages of a part of the eastern region of India
- Hindi: A lingua-franca of (at least) northern India
- Bhojpuri, Maithili and Magahi: Sub-languages of Hindi
  - Lot of similarities
- Resource scarce languages
- Bhojpuri: Not a ‘scheduled’ language
  - Many varieties: No standard Bhojpuri
  - No NLP resources when we started

Motivation

- Exploiting the resources of a closely related similar language
  - Hindi for Bhojpuri
  - Vocabulary in particular
- Building resources, but still a lot of OOV words
  - A problem for the Bhojpuri-Hindi MT system
- Large number of cognate words
- Phonetic closeness
- Significant borrowing of words

Vocabularies of Similar Languages

- Two kinds of Hindi-Bhojpuri word pairs (with similar meaning)
- Entirely different pronunciation
  - रउआ (raua) in Bhojpuri means आप (aap: you-honorific) in Hindi
  - Not relevant for this work
- Similar pronunciation
  - भगवान → भगबान (bʰəgəvɑːnə → bʰəgəbɑːnə)
  - ज्यादा → जादा (dʒjɑːɖɑː → dʒɑːɖɑː)
  - विचार → बिचार (vicɑːrə → bicɑːrə)
  - यजमान → जजमान (jəɟəmɑːnə → ɟəɟəmɑːnə)
  - यमुना → जमुना (jəmunɑː → ɟɑːmunɑː)
  - प्रेम → परेम (preːmə → pəreːmə)

‘Word Transduction’

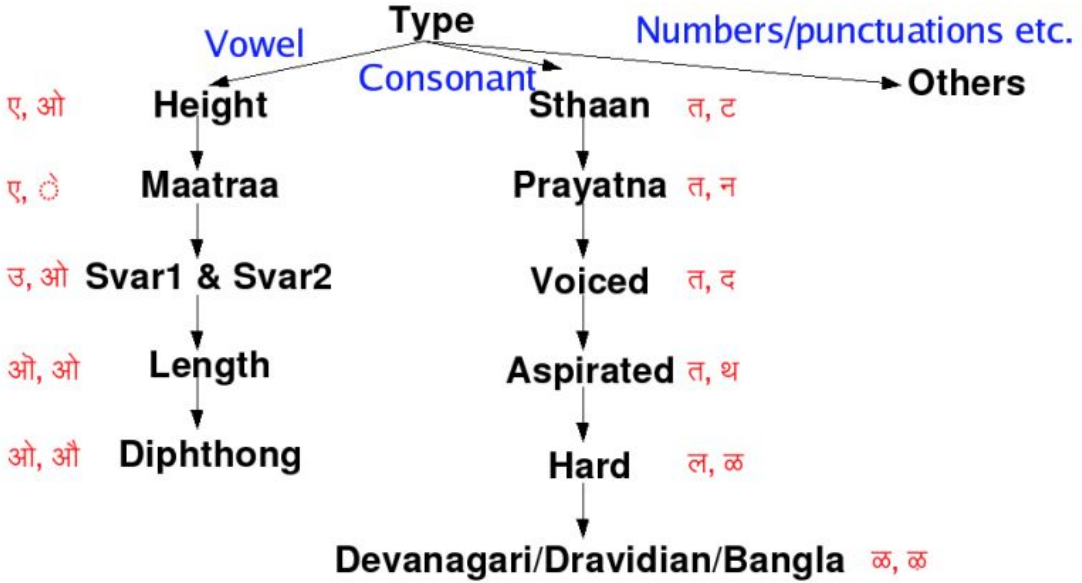
- Conversion of words from source language to (closely related) target language
  - Such that both pronunciation and meaning are the same
- Two aspects:
  - Cognate generation
  - ‘Nativization’ of borrowed words
    - Adapting the borrowed words to the phonology of the target language
    - (More or less) regular phonological changes
    - Can mostly be understood by Bhojpuri speakers
      - Even if somewhat incorrect form
  - Can also be seen as transliteration (or machine translation at character level)
    - In the same script
- ‘Transduced’ words either cognates or borrowed nativized words
- Can address the OOV problem for MT

Phonetic Transcription

- Indic (Brahmic) scripts:
  - Highly phonetic: Written form corresponds almost perfectly with the pronounced form
    - For most practical purposes
- Bhojpuri, Maithili and Magahi use Devanagari
  - Like Hindi
- Trivial to convert words in Devanagari to IPA
  - Again, for practical purposes
  - Not strictly correct
- Minor issues:
  - क (/kə/) is a single Devanagari letter, but is equivalent to क् (/k/)+ अ (/ə/), which is easily reflected in its IPA representation (/kə/)

Hierarchical Phonetic/Orthographic Features

- To differentiate between two phonemes [Singh, 2006]
- Used to calculate phonetic and orthographic distance between two strings
- Also for representing Devanagari strings in terms of features



Approaches

- Statistical machine translation
- Factored statistical machine translation
- Phoneme transcription based approach
  - Effectively using WFST

Statistical Machine Translation

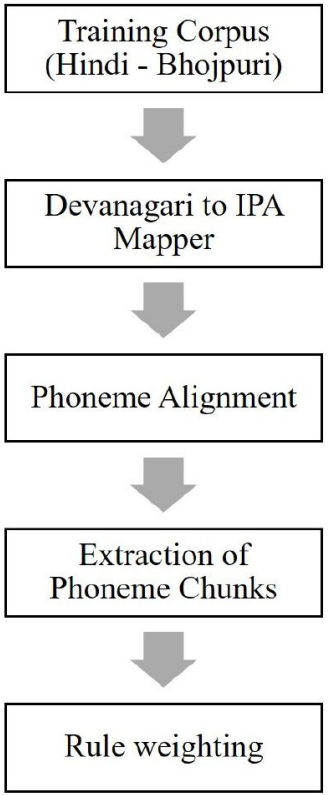
- Using parallel list of words (Hindi-Bhojpuri)
  - As the parallel corpus at character-level
- Translation model, language model and the decoder
- Character-level model developed using Moses

Factored Statistical Machine Translation

- Standard SMT does not use any linguistics information
- Factored SMT allows this
  - One ‘factor’ for each kind of information
- Hierarchical phonetic and orthographic features as the factors
- Tried two different sets of factors
  - FSMT1: Type, Height and Prayatna (manner)
  - FSMT2: Type, Height and Sthaan (place)

Phoneme Transcription Based Approach

- Rewrite rules of all possible chunks (n-grams), weighted according to the frequency of occurrence in the dataset
  - As opposed to single phonemes [Koo, 2011]
  - Context-based rewrite rules
- **Training the model:**
- Conversion from Devanagari to IPA
- Phonetic alignment from source to target language
  - Insert an extra symbol ‘ɛ’ until unit-wise length of both source and target becomes equal, and have maximum phonetic similarity
  - E.g. (sɑːmәne, sәmәnәʊeː) after alignment becomes (<s>sɛɑːɛmәneː</s>, <s>sәmәnәʊeː</s>)
  - Same unit-wise length ensures that only one operation needed: Substitution



Phoneme Transcription Based Approach (contd.)

- Extract phoneme chunks
- For example, after alignment with its Bhojpuri translation, प्रेम (/pre:mə/), परेम (/pəre:mə/) constituent phoneme chunks can be generated as:

<s> p | ɛre:mə <s>	<s> p | əre:mə <s>
<s> pɛ | re:mə <s>	<s> pə | re:mə <s>
<s> pɛr | e:mə <s>	<s> pɛr | e:mə <s>
<s> pɛre: | mə <s>	<s> pɛre: | mə <s>
<s> pɛre:m | ə <s>	<s> pəre:m | ə <s>
<s> p | ɛ | re:mə <s>	<s> p | ɛ | re:mə <s>
<s> ..... <s>	<s> ..... <s>

Phoneme Transcription Based Approach (contd.)

- **Rule weighing:**
- Rewrite rules: $\alpha \rightarrow \beta$
- Weights: $W(\alpha \rightarrow \beta) = (p(\alpha \rightarrow \beta))^2 * plen(\alpha)$
- $p(\alpha \rightarrow \beta) = C(\alpha \rightarrow \beta) / C(\alpha)$
- Probability p considered only if p >= 0:50
  - Based on several experiments

Phoneme Transcription Based Approach (contd.)

- **Estimating Bhojpuri pronunciation:**
- Composed of two steps
- Using weighted phoneme chunks, assign a rank to each possible candidate
  - From aggregated weights of phoneme chunks
- Then treat phoneme representation of the highest ranked word from these outputs as an input to the general rewrite rule system

Some Rewrite Rules

- $kʂ \rightarrow c^h$
- $ɳ \rightarrow n$
- $ɕ \rightarrow s$
- $ʋ \rightarrow b$
- $ɻj \rightarrow ʝ$
- $<s>j \rightarrow <s>ʝ$ (<s> is a boundary marker for the start of a word)
- $ʂ \rightarrow s$

Experiments and Evaluation Methods

- Dataset
- Word Accuracy
- Normalized Phonetic Distance (NPD)
- BLEU Score

$$WA = \frac{Number\ of\ correct\ translation}{Total\ number\ of transduced\ words}$$

$$NPD(T,B) = \frac{PD(T,B) - PD_{min}}{PD_{max} - PD_{min}}$$

Dataset

- A list of Hindi-Bhojpuri word pairs was prepared as part of the MT project
- Language model was prepared using the 19532 words, compiled from a Bhojpuri newspaper and (a very limited) Hindi-Bhojpuri parallel corpus
- The proposed model trained and tested using a dataset consisting of 4220 Hindi-Bhojpuri word pairs
- The dataset was randomly split into a training set and a test set in a three-to-one ratio

Sample Word Pairs

Hindi	Bhojpuri
किस्मत (kismata)	किसमत (kisamata)
ढिंढोरा (dhindhoraa)	ढिनढोरा (dhinadhoraa)
सताता (sataataa)	सतावल (sataavala)
विकसित (vikasita)	बिकसित (bikasita)
स्कूल (skUla)	इसकूल (iskUla)
कौआ (kau-aa)	कउआ (ka-u-aa)

Word Accuracy

- Word accuracy defined as the percentage of the number of correctly transduced words divided by total number of generated transductions

Method	Accuracy
SMT	53.02
FSMT1	54.75
FSMT2	54.99
Proposed method	64.41

BLEU Score

- Papineni et al., 2002
- Results relatively similar to accuracy

Method	BLEU Score
SMT	75.05
FSMT1	75.05
FSMT2	76.18
Proposed method	79.82

Normalized Phonetic Distance

- This test has physical significance in terms of pronunciation difference between generated output and the correct result
- Same pattern observed as for other measures
  - Proposed method > FSMT > SMT

Summary

- Proposed an approach (‘word transduction’) for addressing the OOV word problem for similar languages
  - Of which one is more resource-scarce
- Aimed at guessing the pronunciation or the orthographic form of the target word, given the source word
  - Assuming similar meaning
- Learn to do this from a parallel list of cognate words
- Can also be useful for adapting borrowed words to the phonology of the target language
- Current implementation inefficient, but required only offline

Future Directions

- Larger dataset
  - Currently going on
- More optimized implementation
  - Say, using OpenFST
  - Rather than regular expressions for rewrite rules
- Trying out neural translation at character level

Thank You