हमें नहीं थी कोई आपत्ति in sharing the reign with Bhagwandas and Mohansingh .

Automatic Generation of Code-Mixed sentences

Tanish Lad and Jashn Arora

TOC

Problem Statement

Definition

The Problem

Process

Demonstration

Issues

Future Work

References

Problem Statement

Given a Parallel Corpora (Sentence Aligned Corpora), the task is to Generate Synthetic Code Mixed Data.

What are Code-Mixed Sentences?

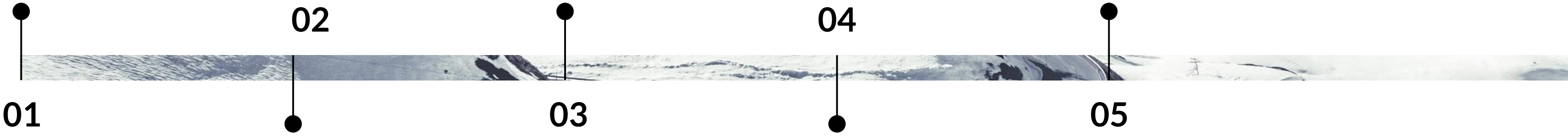
1. Code-Mixing refers to the juxtaposition of linguistic units from two or more languages in a single conversation

2. It is quite commonly observed in speech conversations of multilingual societies across the world.

3. Code-mixed sentences contain words from two or more languages.

The Procedure Used

Briefly..



01

Filter Raw Data

Sentence Align the Corpora and use RegEx to remove the inaccuracies.

02

Chunking

Chunk the Raw Data of both the Languages.

03

Giza++

We used Giza++ toolkit, to generate the word alignments from these parallel sentences

04

Head Finding

Find the Heads of each Chunk.

05

Code-Mixing

Generate Code-Mixed Sentences

Filtering Raw Data

01

Assertion of Sentence Alignment.

We removed those sentences which had no corresponding translation in the other language.

We removed the discrepancies in cases where Sentence Alignment was not there.

We used RegEx to remove inaccuracies in the sentences. For ex, sentences which had no full stop at the end.

Consider a sentence and its corresponding translation

E Every red insect is 0.75 inches long.

H प्रत्येक लाल पतंग पौन इंच लम्बा होता है।

Chunking the Raw Data

02

We used Stanford Parser (v 3.9.3) to chunk the English Sentences and LTRC Shallow Parser (v 4.0) to chunk the Hindi Sentences

English Chunks

E

```
(ROOT
 (S
      (NP (DT Every) (JJ red) (NN insect))
      (VP (VBZ is)
      (ADJP
      (NP (CD 0.75) (NNS inches))
      (JJ long)))
      (. .)))
```

Hindi Chunks

H

```
<Sentence id="8230">
1 (( NP <fs af='पतंग,n,f,sg,3,d,0,0' head='pawaMga'>
1.1 प्रत्येक QF <fs af='प्रत्येक,adj,any,any,,any,,' name='prawyeka'>
1.2 लाल NNPC <fs af='लाल,n,m,sg,3,d,0,0' name='lAla'>
1.3 पतंग NN <fs af='पतंग,n,f,sg,3,d,0,0' name='pawaMga'>
))
2 (( NP <fs af='इंच,n,m,sg,3,d,0,0' head='iMca'>
2.1 पौन NNPC <fs af='पौन,n,f,sg,3,d,0,0' name='pOna'>
2.2 इंच NN <fs af='इंच,n,m,sg,3,d,0,0' name='iMca'>
))
3 (( JJP <fs af='लम्बा,adj,m,sg,,d,,' head='lambA'>
3.1 लम्बा JJ <fs af='लम्बा,adj,m,sg,,d,,' name='lambA'>
))
4 (( VGF <fs af='हो,v,m,sg,any,,ता,wA' head='howA'>
4.1 होता VM <fs af='हो,v,m,sg,any,,ता,wA' name='howA'>
4.2 है VAUX <fs af='है,v,any,sg,2,,है,hE' name='hE'>
))
5 (( BLK <fs af='.,punc,,,,,,' head='.'>
5.1 . SYM <fs af='.,punc,,,,,,' name='.'>
))
</Sentence>
```

Giza++

03

We used Giza++ toolkit, to generate the word alignments from these parallel sentences. This step was performed simultaneously with Step 2.

Running the tool both ways, first considering English as Base Language and then Hindi, we then select only those outputs which were common in both the cases (to generate more accurate Code-Mixed Sentences) .

Head Finding and Extraction

04

The output of LTRC Parser contained the head of each chunk. We extracted them. But this was not the case in Stanford Parser's output. We used a list of possible tags that can act as a head, and used this list to find head of each chunk. In case of multiple matches, we use the match that occurs last in that chunk.

English Heads

E

```
Sentence_id=8230
H   NP   insect
T   DT   Every
T   JJ   red
T   NN   insect
H   NULL   NULL
T   VBZ   is
H   NP   inches
T   CD   0.75
T   NNS   inches
H   NULL   NULL
T   JJ   long
H   NULL   NULL
T   SYM   .
#
```

Hindi Heads

```
Sentence_id=8230
H   NP   पतंग
T   QF   प्रत्येक
T   NNPC   लाल
T   NN   पतंग
H   NP   इंच
T   NNPC   पौन
T   NN   इंच
H   JJP   लम्बा
T   JJ   लम्बा
H   VGF   हो
T   VM   होता
T   VAUX   है
H   BLK   .
T   SYM   .
#
```

Generate Code-Mixed Sentences

05

We first use English as the Base Language. We replace chunks whose both the heads are present as a pair in the output of Giza++. We allow replacement of every chunk that is possible to replace. We don’t restrict the maximum number of chunks that are to be replaced. We do the same process again considering Hindi as the Base Language.

We generate Code-Mixed in Chunked form. Then we flatten the chunks to get Code-Mixed Sentences.

Code-Mixed English Chunks

E

H	NP	पतंग
T	QF	प्रत्येक
T	NNPC	लाल
T	NN	पतंग
H	NULL	NULL
T	VBZ	is
H	NP	इंच
T	NNPC	पौन
T	NN	इंच
H	NULL	NULL
T	JJ	long
H	NULL	NULL
T	SYM	.

Code-Mixed Hindi Chunks

E Sentence_id=8230

H	NP	insect
T	DT	Every
T	JJ	red
T	NN	insect
H	NP	inches
T	CD	0.75
T	NNS	inches
H	JJP	लम्बा
T	JJ	लम्बा
H	VGF	हो
T	VM	होता
T	VAUX	है
H	BLK	.
T	SYM	.

Code-Mixed Sentences

E Every red insect 0.75 inches लम्बा होता है .

H प्रत्येक लाल पतंग is पौन इंच long .

Issues

1 Redundancy of Postposition/Preposition (Solved) in English

The suspicion of ‘बीज सत्याग्रह यात्रा’ is wide ranged .

```
Sentence_id=835
H   NP   suspicion
T   DT   The
T   NN   suspicion
H   NULL   NULL
T   IN   of
T   ''   '
H   NP   यात्रा
T   SYM   '
T   NNPC   बीज
T   NNPC   सत्याग्रह
T   NN   यात्रा
T   SYM   '
T   PSP   का
H   NULL   NULL
T   VBZ   is
H   NULL   NULL
T   JJ   wide
H   VP   ranged
T   VBD   ranged
H   NULL   NULL
T   SYM   .
```

Issues

2 Absence of any case markers in Hindi Based Code-Mixed Sentences

It शुरूआत Japan हुई .

3 Redundancy of words due to difference in number of chunks

It **enforced** **लागू** हो चुका है .

Tatbuni the fruits छिलके से **extracted** **निकाला** गया .

Future Work

We plan on Running a Language Model to extract the most Natural Sentences.

References

1 Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, Kalika Bali. ***Language Modeling for Code-Mixing: The Role of Linguistic Theory based Synthetic Data** (2018)*

2 Rafiya Begum, Kalika Bali, Monojit Choudhury, Koustav Rudra, Niloy Ganguly. ***Functions of Code-Switching in Tweets: An Annotation Scheme and Some Initial Experiments** (2016)*

3 Kalika Bali, Jatin Sharma, Monojit Choudhury, Yogarshi Vyas. ***”I am borrowing ya mixing ?” An Analysis of English-Hindi Code Mixing in Facebook** (2014)*

Thank you.

Checkout our GitHub Repository on the work:
https://github.com/destinyson7/CL-1-Project