Project: Generating monolingual dictionary

Faculty Mentor: Mrs. Soma Paul
Student Mentor: Harsh Vardhan GSV
December 18, 2015

Approach: Corpus based paradigm matching

Faculty Mentor: Mrs. Soma Paul
Student Mentor: Harsh Vardhan GSV
December 18, 2015

Objective

The aim of this project was to expand the Bangla dictionary. Bangla dictionaries already exist but they are pretty small right now and need expansion.

This project uses an already existing dictionary as a reference dictionary to add new words to the dictionary.

Why?

Dictionary acts as a useful tool in many domains of linguistics. If we already know the forms of a word, it becomes easier to annotate the data based on contextual information.

Why paradigm-matching?

case/num	singular	plural	vocative
direct	ladak**A**	ladake	ladake
oblique	ladake	ladak**oM**	ladak**o**

Noun Paradigm

person	singular	plural
first	jA**UMgA**	jAyeMge
second(tU/tuma)	jA**egA**	jA**oge**
second(tuma/Apa)	jA**oge**	jAeMge
third	jA**yegA**	jAyeMge

Verb Paradigm

Why paradigm-matching?

The dictionary that we are trying to make lists down all the forms of a word and also lists down the various properties of the various forms of that word thereby making annotation easier.

Very helpful for parsers.

This is Why.

Sample entries in the pardef section

```
<pardef n="jA/__v">
<e><p>
<l>UMgA</l>
<r><s n="verb" /><s n="singular" /><s n="first" />
<s n="masc" /><s n="future" /></r>
</p></e>
<e><p>
<l>eMge</l>
<r><s n="verb" /><s n="plural" /><s n="first" />
<s n="masc" /><s n="future" /></r>
</p></e>
```

```
<sdefs>
<sdef n="verb" />
<sdef n="singular" />
<sdef n="plural" />
<sdef n="first" />
<sdef n="second" />
<sdef n="second_h" />
<sdef n="third" />
<sdef n="masc" />
<sdef n="future" />
</sdefs>
```

Snapshots of our paradigm based dictionary.

1. pardef section
2. sdef section
3. new-word addition

Sample entries in the dict section

```
<section id="main" type="standard">
<e lm="jA"><i>jA</i><par n="jA/__v" />
<e lm="KA"><i>KA</i><par n="jA/__v" />
<e lm="gA"><i>gA</i><par n="jA/__v" />
</section>
```

Procedure

- First, we found a large Bangla corpus.

Procedure

- Our next task was to annotate the corpus. For that, we used the POS tagger by LTRC.

Procedure

- To suit the input format for the tagger, we had to tokenize the data by a tokenizer, downloaded form ltrc.iiit.ac.in.

Procedure

- Before using the tokenizer, we removed special characters(#*=-) from the corpus using a bash script.

Procedure

- Next task was to extract the Nouns from the tagged data. We wrote a python script for that.

Procedure

- Given, an input, we first try to look for it in our existing dictionary. If we find an entry for the word in our dictionary, we generate its word forms based on the referenced paradigm table and print them.

Procedure

- If we encounter a word that we can't find in our existing dictionary, we follow a simple RULE to predict the paradigm table for that word.

A slight detour: The Rule

We categorize the words based on their endings, i.e. we say that if a word ends with a sound, say 'a', we can assume that its paradigm table will match with one of the already existing tables for words that end with 'a'.

This simple rule shortens the list of tables that we need to try out for our word.

A slight detour: The Rule

The ASSUMPTION that we make in this rule is that we consider the set of paradigm tables in our existing dictionary to be exhaustive, i.e we assume that all the words in the corpus will find a paradigm match in the given dictionary.

This rule is very useful in Bangla because of the nature of the language.

Procedure

- This simple rule shortens the list of tables that we need to try out for our word.

Procedure

- In order to narrow down on one paradigm table for the word, we generate the word forms corresponding to all the paradigm tables and then for each one of them, we go back to the original corpus and start looking for entries of the generated forms in our corpus.

Procedure

- The paradigm table that has the maximum matchings in the corpus is assumed to be the one for the word and the user is notified that it is most probable that the given word follows that paradigm table.

Procedure

- In case, where multiple tables get the the same number of matchings in the corpus, we display all the tables, because that is the maximum narrowing down we could do with our given data.

Procedure

- In case we cannot match any of the generated forms to the ones in our original corpus, we can say that we cannot predict the paradigm table for the given word.

A major limitation

Problem:

A major limitation of our approach is our assumption that the set of paradigm tables in our existing dictionary is exhaustive. This restricts us to the given set of paradigm tables and hence for words which cannot be mapped to one of these tables, we need another set of paradigm tables.

Solution:

When we are provided with a substantially large corpus, we may get all the forms for the word in our corpus only and hence we can create a paradigm table with the help of a native speaker for that language.

Thank you