M. Tech. Thesis End-Term Project

SHASHWAT CHANDRA
DEBATE STANCE IDENTIFICATION USING UNSUPERVISED TECHNIQUES
Given a debate topic and a paragraph, is the paragraph for the debate topic, or against the debate topic?
What this problem entails

- Semantic Analysis
- Discourse Analysis
- Information Extraction
- Contradiction/Relationship Detection
Previous Work

- Stance Detection in Congressional Floor Debates
  - Sentiment Analysis (Document-level sentiment-polarity classification)
  - Supervised, Uses Stemmer, Parser
- Recognizing stance in Ideological On-Line debates
  - Opinion - Target pairs: Supervised, Uses Stemmer
  - Positive arguing, negative arguing: Trigger expressions
  - n-gram matching
- Both of these approaches use language resources that are unavailable for most languages apart from English.
Approaches

- **Emotion Detection**
  - Link stance to emotion
  - Conduct a survey?
    - Positive and negative reactions to a given situation.
    - Emotionally-driven stance analysis.

- **Text Analysis**
  - TF-IDF
  - pLSA: Detect keywords identifying stance
### Example and Results

**This House believes that children should be allowed to own and use mobile phones**

<table>
<thead>
<tr>
<th>'Yes', ['parent', 'ag', 'child', 'call', 'technolog']</th>
<th>'No', ['parent', 'health', 'expens', 'cell', 'mani']</th>
</tr>
</thead>
</table>
| Mobile phones keep children safer, as it is easier for parents to stay in touch with their children and for children to contact someone in an emergency. | There are possible potential long-term health risks from using mobile phones.  
Mobile phones are too expensive for children. |
- Dimensions of Argument in Social Media
- Language as context for the perception of emotion
- Recognizing stance in Ideological On-Line debates
- Emotion Detection from Text
- Detecting Emotion in Text
- Emotion modulates language production during covert picture naming
- Categorizing Emotion in Spoken Language: An analysis of Semantic and Prosodic Contributions to Emotional Communication
- Emotion characterization across Languages
Is there a relationship between the sound a word makes and the emotions attached with it?

- **Bouba/Kiki effect**
- **SentiWordNet**
  - [http://sentiwordnet.isti.cnr.it](http://sentiwordnet.isti.cnr.it)
  - Word, Sense, Synset, PosScore, NegScore, Glossary
  - 213,702 different word senses.
  - Wiktionary was used to extract pronunciation rules.
Results

- Looking for a connection between phonemes and emotion.
  - International Phonetic Alphabet for English
- Some syllables do appear to be in positive words more frequently than negative words. Most are neutral (+ve/-ve percentage within 10% of 50%)

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Meaning</th>
<th>Emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td>Aspirated h</td>
<td>84.6% +ve (52)</td>
</tr>
<tr>
<td>a</td>
<td>nut</td>
<td>81.8% +ve (22)</td>
</tr>
<tr>
<td>Λ</td>
<td>gut</td>
<td>61.7% -ve (1118)</td>
</tr>
<tr>
<td>Ω</td>
<td>not</td>
<td>62.1% -ve (1060)</td>
</tr>
<tr>
<td>f</td>
<td>fill</td>
<td>60.3% -ve (1280)</td>
</tr>
</tbody>
</table>
Extending pLSA approach to Hindi

Major Requirement: Hindi Stemmer

Major Requirement: Hindi Debate Database

- Lok Sabha/Rajya Sabha Database:
  - In PDF format. Hard to process
  - Often off-topic.
- News Pages
- Academic Organizations
Hindi Stemmer Creation

- No good stemmer exists
- Partially a rule-based approach
- Inflections:
  - Prefixes
  - Suffixes
  - Infixes
- Hindi WordNet
  - Has entries for derivatives, not inflections
  - Acts as oracle.
Previous Work

- **An Unsupervised Hindi Stemmer with Heuristic Improvements**
  - Unsupervised
  - Splits each word into Stem + Suffix
  - Uses Naïve Bayes approach to find optimal split.
  - Creates suffix list. Removes multiple suffixes, if present

- **A Lightweight Stemmer for Hindi**
  - Entirely Rule-based
  - Lists possible inflections and removes them
  - Removes longest possible suffix.
Problems Faced

- **Assimilation**
  - The sound at the end of the word changes with a suffix addition (Example, 35 = पॅंतीस, not पांचतीस).

- **Surface-Form Similarity: Morphophonology**
  - Words are often spelt like they’re pronounced.
  - No fixed spelling.
    - Fuzzy.

- **Multiple, unordered prefixes and suffixes.**
  - Should all map to the same root word.
Procedure Followed

- Create a list of prefixes and suffixes using a corpus of Hindi Literature.
- Using grounded knowledge, select useful prefixes and suffixes and divide them into two categories on the basis of their prefix/suffix frequency.
  - Calculate the number of times the prefix occurs as a prefix versus the number of times it occurs in a normal word
- The categories are: Safe Prefix/Suffix List and Unsafe Prefix/Suffix List
Procedure Followed (continued)

- **Unsafe Prefix/Suffix List:**
  - These prefixes/suffixes are those that occur as prefixes/suffixes with a low degree of confidence.
  - Words formed on removal of these prefixes/suffixes are checked using Hindi WordNet before being returned.

- **Safe Prefix/Suffix List:**
  - These prefixes/suffixes are those that occur as prefixes with a high percentage.
  - Words formed on removal of these prefixes/suffixes, if they exist in Hindi WordNet, are returned instantly.
  - If the removal of a Safe/Unsafe prefix/suffix does not lead to a word present in Hindi WordNet, the Safe Prefixes/Suffixes are removed from that word, and the result is returned.
## Results of the Stemmer

<table>
<thead>
<tr>
<th>जो पुस्तके हमें सोचने के लिए विवश करती हैं</th>
<th>वे</th>
</tr>
</thead>
<tbody>
<tr>
<td>जो पुस्तक हम सोच के लिए विवश कर हैं</td>
<td>वे</td>
</tr>
<tr>
<td>हमारी सबसे अधिक सहायक हैं</td>
<td></td>
</tr>
<tr>
<td>हमारी सबसे अधिक सहायक हैं</td>
<td></td>
</tr>
</tbody>
</table>

**Approximate Accuracy: 86%**

*Green words are correctly stemmed*

*Red words are incorrectly stemmed*
Next Steps

- Create a dictionary of small words, since they are likely to give false positives in the process of removing unsafe prefixes and leaving a single letter.
- Deal with Assimilation.
- Deal with Infixes.
- Remove spurious punctuation.
- Use sentence-structure information

- Test the stemmer for use in real scenarios (*Shouvik Sachdeva*)
Future Work

- Improve on the accuracy of the Stemmer
- Include POS Tag information in the Stemmer.
- Perform the unsupervised stance classification algorithm on a dataset in Hindi, or bilingual parallel corpora.
- Perform survey and get relevant results for use in database creation for emotion/cognition-based approach.
Thank You