Semantic Textual Similarity using Universal Networking Language

Avishek Dan
Pushpak Bhattacharyya
{avishekdan, pb}@cse.iitb.ac.in

Center for Indian Language Technology
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
India
Goal: Semantic Textual Similarity score generation

Approach: Semantic graph extraction and matching
Introduction

• The system generates an unambiguous, syntax agnostic graphical representation from the sentences

• Matching of graph through lexical and semantic relations done to generate semantic score
Motivation

• Structurally different sentences can have same meaning
  – *The cat chased the rat*
  – *The rat was chased by the cat*

• Lexical and semantic relations between words control the degree of similarity
  – Synonyms, antonyms, hypernyms, meronyms
Universal Networking Language

“The boy chased the dog”
Semantics Extraction Architecture
Graph Matching

• The edges of the graph are compared
  – Score between 0 and 1 based on matching of edge label and degree of match of universal words

• Universal word matching based on WordNet relations
  – Synonymy, antonymy, hypernymy, entailment and meronymy relations used
  – Score inversely proportional to distance between concepts

• Final score is sum of scores divided by the average of the nodes in the two graphs
Graph Matching Example
Additional Similarity Measures

• UNL system has poor accuracy for very long sentences whose dependency parse cannot be generated accurately
  – Other similarity measures are used to stabilize the score
• Jacard Similarity
  – Matching of lemmas
  – Matching using lexical and semantic similarity measures similar to graph matching
• Dependency graph match (direct and WordNet relation based)
Training

- All scores combined through linear regression to generate composite similarity score
- Weights trained on 2012 STS shared task gold standard data
- Weka library API used

\[ score(S1, S2) = c + \sum_{i=1}^{5} \lambda_i score_i(S1, S2) \]
Evaluation (Source of errors)

- Imperfect structure of sentences
  - and showings such as "king of the beasts", "18 genres of martial arts of China" intoxicated audience members; whereas "3 champions combat lu-bu" and
  - fill, attach, fix with, or as if with gum

- Word sense disambiguation
  - "Cricket" tagged as insect instead of game
  - Leads to incorrect graph matching
Evaluation (Source of errors)

- Relationship among quantities
  - No damage, tsunami after 6.6 quake hits off Indonesia's Sumatra
  - 5.9 quake hits Indonesia's Sumatra: USGS
  - The quantities 6.6 ans 5.9 are quite close

- Named entity representation
  - Afghan and S. Afghanistan
  - UK soldiers and British soldiers
  - TSE and Tokyo Stock Exchange
Evaluation (Source of errors)

• Relation generation errors
  - aoj-agt confusion occurs many times since they have the same dependency

• Elaborate and summarized forms of same information
  - while in motion, an impactor makes sudden, forcible contact with the impactee, or two impactors both move, mutually making forcible contact.
  - hit against; come into sudden contact with;
## Results

<table>
<thead>
<tr>
<th>Corpus</th>
<th>CFILT</th>
<th>Best Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Headlines</td>
<td>0.5336</td>
<td>0.7642</td>
</tr>
<tr>
<td>OnWN</td>
<td>0.2381</td>
<td>0.7529</td>
</tr>
<tr>
<td>FNWN</td>
<td>0.2261</td>
<td>0.5818</td>
</tr>
<tr>
<td>SMT</td>
<td>0.2906</td>
<td>0.3804</td>
</tr>
<tr>
<td>Mean</td>
<td>0.3531</td>
<td>0.6181</td>
</tr>
</tbody>
</table>
Future Work

• Addition of resources
  – Use resources such as VerbOcean which gives other similarity measures between words
  – Use of resources such as Freebase for named entity matching

• Parameter learning
  – Currently penalties for partial mismatch are intuitively fixed, the values can be learnt from the data
References


English UNL Enconverter Demo
http://www.cfilt.iitb.ac.in/UNL_Enco