Meerkat Mafia: Multilingual and Cross-Level Semantic Textual Similarity systems

Abhay Kashyap, Lushan Han, Jennifer Sleeman, Taneeya Satyapanich, Sunil Gandhi and Tim Finin
University of Maryland, Baltimore County

Roberto Yus, University of Zaragoza
Outline

• LSA word similarity
• Word similarity to text similarity
• Task 10: English and Spanish text similarity
• Conclusions
LSA Similarity

Similar words usually occur in similar contexts!

• A large clean corpus produces reliable word co-occurrence stats
• Used corpus from Stanford WebBase project
  o Feb 2007 crawl, 100M pages from > 50K websites
  o Removed non-English text
  o Segmented into paragraphs and deduplicated
• Result: 3B word corpus - [http://ebiq.org/r/351](http://ebiq.org/r/351)
LSA Similarity

• Count co-occurrences of marry_VB & wife_NN
  o POS tagging + lemmatization (Stanford POS)
  o Only content words: noun, verb, adjective, adverb

• Two window sizes: three or nine words
  o ± 1: More precise context but works only for same POS
  o ± 4: Allows for similarity computation for different POS

• Vocabulary: 22K words and phrases plus over 2K verb phrases from WordNet

• Final dimensions is a 29k x 29k matrix
LSA similarity: SVD transform

• Boost scores for for polysemous words
  o Words with many senses have lower similarity scores
• Replace frequency counts by logs
• Apply singular value decomposition to term-term matrix
• Retain 300 largest singular values
• Compute LSA similarity for two words as cosine similarity between their word vectors
• LSA word similarity

• Word similarity to text similarity

• Task 10: English and Spanish text similarity

• Conclusions
From word to text similarity

• Basic *align and penalize* approach
  ① align words to maximize LSA word similarity
  ② compute average word similarity for pairs
  ③ penalize unaligned terms

• Preprocessing: POS tag, lemmatization, REs to identify number and dates, stopword removal

• Word similarity wrapper for numbers, time expressions, pronouns and OOV words
A&P align and penalize example

Cheddar cheese was eaten by a rat
The little mouse ate some cheese

align 1 => 2
align 2 => 1

Good alignment scores

STS score = T

Allow multiple words to align with one
Outline

• LSA word similarity
• Word similarity to text similarity
• Task 10: English and Spanish text similarity
• Conclusions
Task 10: English runs

- **Pairing words (#2)** used unsupervised A&P
- **SuperSaiyan** and **Hulk** used a Support Vector Regression system with special features in addition to pairing words results
  - **Hulk (#6)** used a generic model trained on 3750 sentence pairs (1500 MSRvid, 1500 MSRpar, 750 Headlines)
  - **Super Saiyan (#5)** used domain-specific training for OnWWN (1361 pairs) and Images (1500 MSRvid pairs) and the generic model for others
Supervised Approach Details

• Words aligned by highest similarity (>0.1)

• Pairs weighted by Google word frequency

• Features for OOV words (e.g., *copasetic*)
  
  ○ Retrieve its definition from Wordnik and use the A&P score as the word similarity

• Named Entities: try to link to DBpedia entities to see if they co-refer
## English Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pairing Words</th>
<th>Hulk</th>
<th>SuperSaiyan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deft-forum (450 pairs)</td>
<td>0.4711 (7) *</td>
<td>0.4495 (15)</td>
<td>0.4918 (4)</td>
</tr>
<tr>
<td>Deft-news (300 pairs)</td>
<td>0.7628 (8)</td>
<td><strong>0.7850 (1)</strong></td>
<td>0.7712 (3)</td>
</tr>
<tr>
<td>Headlines (750 pairs)</td>
<td>0.7597 (8)</td>
<td>0.7571 (9)</td>
<td><strong>0.7666 (2)</strong></td>
</tr>
<tr>
<td>Images (750 pairs)</td>
<td>0.8013 (7)</td>
<td>0.7896 (10)</td>
<td>0.7676 (18)</td>
</tr>
<tr>
<td>OnWN (750 pairs)</td>
<td><strong>0.8745 (1)</strong></td>
<td>0.7872 (18)</td>
<td>0.8022 (12)</td>
</tr>
<tr>
<td>Tweet-news (750 pairs)</td>
<td>0.7793 (2)</td>
<td>0.7571 (7)</td>
<td>0.7651 (4)</td>
</tr>
<tr>
<td><strong>Weighted mean</strong></td>
<td><strong>0.7605 (2)</strong></td>
<td><strong>0.7349 (6)</strong></td>
<td><strong>0.7410 (5)</strong></td>
</tr>
</tbody>
</table>

- Pairing continues to do very well
- Hulk trained on data from all genres and SuperSaiyan on genre-specific data
- Supervised systems will do much better with more training data
Task 10 Spanish in a nutshell

“Un cojín es una funda de tela […]”
“A cushion is a fabric cover […]”

“Una almohada es un cojin […]”
“A pillow is a rectangular pad […]”

Pairing words & Hulk

3.7!
Improving the simple idea

- Direct Spanish to English translations good, yielding moderate results with our systems:
  - Paring, rank #13
  - Hulk, rank #5
- Handle translation anomalies to improve scores
  - *Las costas o costa de un mar* ...
    - Costs or the cost of a sea ...
    - Coasts or the coast of a sea ...
  - *Una almohada es un cojín en forma rectangular* ...
    - A pillow is a rectangular pad ...
    - A pillow is a rectangular cushion ...
Improving direct translation

• Generate candidates by considering alternate translations for each word and combining them
  o Costs or the cost of a sea, lake or wide river is the land along the edge of these.
  o Coasts or the cost of a sea, lake or wide river is the land along the edge of these.
  o The coasts or the cost of a sea, lake or wide river is the land along the edge of these.
  o The shores or the cost of a sea, lake or wide river is the land along the edge of these.

• Control combinatorics by using at most 20 alternatives with Google score > 65
  o News and Wikipedia tests went from 480 & 324 sentence pairs to 5756 & 1776
Scoring candidate pairs

- \( I_1 \) and \( I_2 \) (a pair of Spanish sentences)
- Possible translations generated for each sentence:
  - \( TI_1 = \{ T_{11}, T_{12}, T_{13}, \ldots, T_{1n} \} \)
  - \( TI_2 = \{ T_{21}, T_{22}, \ldots, T_{2m} \} \),
- Compute similarity by using:
  \[
  Sim_{SPA}(I_1, I_2) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} Sim_{ENG}(T_{1i}, T_{2j})}{n \times m}
  \]
Results on Spanish task

Three Spanish similarity runs submitted

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pairing Words</th>
<th>PairingAvg</th>
<th>Hulk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia (324 pairs)</td>
<td>0.6682 (12)</td>
<td>0.7431 (6)</td>
<td>0.7382 (8)</td>
</tr>
<tr>
<td>News (480 pairs)</td>
<td>0.7852 (12)</td>
<td><strong>0.8454 (1)</strong></td>
<td>0.8225 (6)</td>
</tr>
<tr>
<td><strong>Weighted mean</strong></td>
<td><strong>0.7380 (13)</strong></td>
<td><strong>0.8042 (2)</strong></td>
<td><strong>0.7885 (5)</strong></td>
</tr>
</tbody>
</table>

- Best run had weighted correlation of 0.8042, behind 1st place by only 0.003
- Wikipedia scores worse than news: more Spanish names with non-English characters that caused problems to our English-trained STS system
Outline

- LSA word similarity
- Word similarity to text similarity
- Task 10: English and Spanish text similarity
- Conclusions
Conclusions

• LSA word similarity model is strength of our systems
• Algorithms for computing text similarity are relatively simple and work well
• There are good sources of data to support specialized text similarity tasks, e.g. Wordnik
• Web search queries helped for OOV words
• Using Google translate worked surprisingly well for Spanish STS