The Impact of Distributional Metrics in the Quality of Relational Triples

Hernani Costa, Hugo Gonçalo Oliveira\(^1\), Paulo Gomes

hpcosta@student.dei.uc.pt, \{hroliv,pgomes\}@dei.uc.pt

Cognitive & Media Systems Group
CISUC, University of Coimbra

Lisbon, August 16, 2010

\(^1\)supported by FCT scholarship grant SFRH/BD/44955/2008
1 Introduction
   - Information Extraction
   - Information Retrieval
   - Research Goals

2 Approach

3 Experimentation
   - Set-up
   - Metrics adaptation
   - Results
   - Additional experimentation

4 Concluding remarks
Knowledge bases (eg. WordNet) are useful resources for NLP.
Knowledge bases (e.g., WordNet) are useful resources for NLP.
Their creation and maintenance involves intensive human effort.

Higher coverage, easier update, but...
Precision is lower.
Evaluation requires once again intensive human labour!
Introduction

Knowledge bases (eg. WordNet) are useful resources for NLP
Their creation and maintenance involves intensive human effort
Automatic creation/enrichment from textual resources is an alternative
Knowledge bases (e.g., WordNet) are useful resources for NLP.

Their creation and maintenance involves intensive human effort.

Automatic creation/enrichment from textual resources is an alternative.

- Higher coverage, easier update, but...
Knowledge bases (e.g. WordNet) are useful resources for NLP.

Their creation and maintenance involves intensive human effort.

Automatic creation/enrichment from textual resources is an alternative.
- Higher coverage, easier update, but...
- Precision is lower.
Knowledge bases (e.g., WordNet) are useful resources for NLP.

Their creation and maintenance involves intensive human effort.

Automatic creation/enrichment from textual resources is an alternative:

- Higher coverage, easier update, but...
- Precision is lower
- Evaluation requires once again intensive human labour!
Information extraction (IE)

Automatic extraction of structured information from natural language.
Information extraction (IE)
Automatic extraction of structured information from natural language.

“Car is a vehicle with 4 wheels and an engine, used for carrying a small number of passengers.”
Information extraction (IE)

Automatic extraction of structured information from natural language.

- “Car is a vehicle with 4 wheels and an engine, used for carrying a small number of passengers.”
  - vehicle HYPERNYM_OF car
Information extraction (IE)

Automatic extraction of structured information from natural language.

“Car is a vehicle with 4 wheels and an engine, used for carrying a small number of passengers.”

- *vehicle* HYPERNYM_OF *car*
- *wheel* PART_OF *car*
- *engine* PART_OF *car*
Information extraction (IE)

Automatic extraction of structured information from natural language.

“Car is a vehicle with 4 wheels and an engine, used for carrying a small number of passengers.”

- vehicle HYPERNYM_OF car
- wheel PART_OF car
- engine PART_OF car
- carrying_people PURPOSE_OF car
Information retrieval (IR)

Locating specific information in natural language resources.
Information retrieval (IR)
Locating specific information in natural language resources.

- Approaches based on the occurrence of words in documents.
Information retrieval (IR)

Locating specific information in natural language resources.

- Approaches based on the occurrence of words in documents.
- Distributional similarity metrics
  - Cocitation (Small (1973))
  - LSA (Deerwester et al. (1990))
  - Lin's (Lin (1998))
  - PMI-IR (Turney (2001))
  - $\sigma$ (Kozima and Furugori (1993))
  - ...
Introduction

Research Goals

Goals

1. Use IR metrics to improve IE precision
   - Adapt distributional metrics to determine words similarity

Wandmacher et al. (2007) and Cederberg and Widdows (2003) used LSA to weight hypernymy triples

What about other semantic relations?

What metrics should be used?

New combined metrics?
Goals

1. Use IR metrics to improve IE precision
   - Adapt distributional metrics to determine words similarity
   - Wandmacher et al. (2007) and Cederberg and Widdows (2003) used LSA to weight hypernymy triples
Use IR metrics to improve IE precision

- Adapt distributional metrics to determine words similarity
- Wandmacher et al. (2007) and Cederberg and Widdows (2003) used LSA to weight hypernymy triples
- What about other semantic relations?
Goals

1. Use IR metrics to improve IE precision
   - Adapt distributional metrics to determine words similarity
   - Wandmacher et al. (2007) and Cederberg and Widdows (2003) used LSA to weight hypernymy triples
   - What about other semantic relations?
   - What metrics should be used?
#### Goals

1. Use IR metrics to improve IE precision
   - Adapt distributional metrics to determine words similarity
   - Wandmacher et al. (2007) and Cederberg and Widdows (2003) used LSA to weight hypernymy triples
   - What about other semantic relations?
   - What metrics should be used?
   - New combined metrics?
Goals

1. Use IR metrics to improve IE precision
   - Adapt distributional metrics to determine words similarity
   - Wandmacher et al. (2007) and Cederberg and Widdows (2003) used LSA to weight hypernymy triples
   - What about other semantic relations?
   - What metrics should be used?
   - New combined metrics?

2. Help manual evaluation
IE system

- Corpus
- Extraction of relational triples
- Removal of triples with stopwords
- Lemmatisation
- Additional extraction of triples
- Metrics application
Experimentation set-up

- CETEMPúblico\(^2\) corpus (annotated version)
  - 28,000 documents
  - 30,100 unique context words (nouns, verbs and adjectives)
  - *term-document* matrix

\(^2\)http://www.linguateca.pt/cetemppublico/
Experimentation set-up

- CETEMPúblico\(^2\) corpus (annotated version)
  - 28,000 documents
  - 30,100 unique context words (nouns, verbs and adjectives)
  - *term-document* matrix

- Triples obtained
  - Extracted: 20,308
  - Discarded: 5,844
  - Inferred: 2,492
  - Final triple set: \textbf{16,956}

\(^2\)http://www.linguateca.pt/cetempublico/
Similarity between two documents

For instance, Cocitation:

- First presented as a similarity metric between scientific papers (Small (1973))

\[
Cocitation(d_i, d_j) = \frac{P(d_i \cap d_j)}{P(d_i \cup d_j)}
\]  

\(d_i, d_j\) represent two documents

\(P(d_i \cap d_j)\), is the number of documents in the collection referring both documents

\(P(d_i \cup d_j)\), is the number of documents referring at least to one of the documents
Similarity between two documents

For instance, Cocitation:

- First presented as a similarity metric between scientific papers (Small (1973))

\[
\text{Cocitation}(d_i, d_j) = \frac{P(d_i \cap d_j)}{P(d_i \cup d_j)}
\]  

- \(d_i, d_j\) represent two documents
Similarity between two documents

For instance, Cocitation:

- First presented as a similarity metric between scientific papers (Small (1973))

\[
Cocitation(d_i, d_j) = \frac{P(d_i \cap d_j)}{P(d_i \cup d_j)}
\]  

- \(d_i, d_j\) represent two documents
- \(P(d_i \cap d_j)\), is the number of documents in the collection referring both documents
Similarity between two documents

For instance, Cocitation:

- First presented as a similarity metric between scientific papers (Small (1973))

\[
\text{Cocitation}(d_i, d_j) = \frac{P(d_i \cap d_j)}{P(d_i \cup d_j)}
\]  

- \(d_i, d_j\) represent two documents
- \(P(d_i \cap d_j)\), is the number of documents in the collection referring both documents
- \(P(d_i \cup d_j)\), is the number of documents referring at least to one of the documents
Adaptation to measure word similarity

\[
\text{Cocitation}(e_i, e_j) = \frac{P(e_i \cap e_j)}{P(e_i \cup e_j)}
\]
Adaptation to measure word similarity

\[ \text{Cocitation}(e_i, e_j) = \frac{P(e_i \cap e_j)}{P(e_i \cup e_j)} \]  

- \( e_i, e_j \) represent two entities (uni or multiword)
Adaptation to measure word similarity

\[ \text{Cocitation}(e_i, e_j) = \frac{P(e_i \cap e_j)}{P(e_i \cup e_j)} \]  \hspace{1cm} (2)

- \( e_i, e_j \) represent two entities (uni or multiword)
- \( P(e_i \cap e_j) \), is the number of documents containing both entities
Adaptation to measure word similarity

\[ \text{Cocitation}(e_i, e_j) = \frac{P(e_i \cap e_j)}{P(e_i \cup e_j)} \]  

- \(e_i, e_j\) represent two entities (uni or multiword)
- \(P(e_i \cap e_j)\), is the number of documents containing both entities
- \(P(e_i \cup e_j)\), is the number of documents containing at least one of the entities
## Triples and metrics

<table>
<thead>
<tr>
<th>Triple</th>
<th>Manual</th>
<th>Coc</th>
<th>LSA (oc)</th>
<th>LSA (tf-idf)</th>
<th>PMI</th>
<th>Lin</th>
<th>σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>nação SINONIMO_DE povo nation SYNONYM_OF people</td>
<td>2</td>
<td>4.21</td>
<td>7.92</td>
<td>8.21</td>
<td>66.65</td>
<td>55.12</td>
<td>35.79</td>
</tr>
<tr>
<td>violência CAUSADOR_DE estrago violence CAUSE_OF damage</td>
<td>2</td>
<td>1.60</td>
<td>4.38</td>
<td>4.47</td>
<td>63.90</td>
<td>29.51</td>
<td>43.82</td>
</tr>
<tr>
<td>palavra HIPERONIMO_DE beato</td>
<td>1</td>
<td>0.16</td>
<td>1.75</td>
<td>1.78</td>
<td>61.83</td>
<td>0</td>
<td>48.25</td>
</tr>
<tr>
<td>jogo FINALIDADE_DE preparar game PURPOSE_OF prepare</td>
<td>1</td>
<td>1.61</td>
<td>3.53</td>
<td>3.62</td>
<td>50.89</td>
<td>48.22</td>
<td>25.52</td>
</tr>
<tr>
<td>sofrer SINONIMO_DE praticar suffer SYNONYM_OF practice</td>
<td>0</td>
<td>0.73</td>
<td>1.34</td>
<td>1.37</td>
<td>52.04</td>
<td>27.77</td>
<td>34.25</td>
</tr>
<tr>
<td>atender FINALIDADE_DE moderno</td>
<td>0</td>
<td>0.69</td>
<td>1.81</td>
<td>1.82</td>
<td>55.22</td>
<td>13.84</td>
<td>41.24</td>
</tr>
</tbody>
</table>
Manual validation of the results

![Bar graph showing the distribution of results for different categories: purpose_of, hypernymy, causation, part_of, and synonymy. The categories are further divided into three levels: 0, 1, and 2. The graph includes the following counts:

- purpose_of: 108, 41, 30
- hypernymy: 261, 96, 146
- causation: 83, 34, 16
- part_of: 43, 15, 13
- synonymy: 203, 37, 30]
## Manual evaluation vs. Distributional metrics

<table>
<thead>
<tr>
<th></th>
<th>Cocitation</th>
<th>LSA (oc)</th>
<th>LSA (tfidf)</th>
<th>PMI</th>
<th>Lin</th>
<th>σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>synonym</td>
<td>-0.06</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.11</td>
<td>0.00</td>
<td>0.19</td>
</tr>
<tr>
<td>part_of</td>
<td>0.15</td>
<td>0.19</td>
<td>0.19</td>
<td>0.21</td>
<td>0.39</td>
<td>-0.22</td>
</tr>
<tr>
<td>Causation</td>
<td>0.09</td>
<td>0.01</td>
<td>0.01</td>
<td>-0.01</td>
<td>-0.07</td>
<td>0.18</td>
</tr>
<tr>
<td>hypernymy</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
<td>0.46</td>
<td>0.41</td>
<td>0.08</td>
</tr>
<tr>
<td>purpose_of</td>
<td>0.02</td>
<td>0.09</td>
<td>0.07</td>
<td>0.18</td>
<td>0.00</td>
<td>0.10</td>
</tr>
</tbody>
</table>

![Graph showing comparison between manual evaluation and distributional metrics](image-url)
Some observations:

- Hypernymy is highly correlated with all metrics except $\sigma$

- Part-of is less, but also correlated with the former metrics

- For purpose triples, PMI has a 0.18 correlation coefficient

- Hyponyms and hypernyms tend to co-occur more frequently than causes/effects or means/purposes

- No conclusions taken for causation

- Synonymy has low or negative correlation coefficients with the metrics

- Few correct triples

- In corpora, synonymous words do not co-occur frequently...
Some observations:

- Hypernymy is highly correlated with all metrics except $\sigma$
- Part-of is less, but also correlated with the former metrics
Some observations:

- Hypernymy is highly correlated with all metrics except $\sigma$
- Part-of is less, but also correlated with the former metrics
- For purpose triples, PMI has a 0.18 correlation coefficient
Some observations:

- Hypernymy is highly correlated with all metrics except $\sigma$
- Part-of is less, but also correlated with the former metrics
- For purpose triples, PMI has a 0.18 correlation coefficient
  - Hyponyms and hypernyms tend to co-occur more frequently than causes/effects or means/purposes
Some observations:

- Hypernymy is highly correlated with all metrics except $\sigma$
- Part-of is less, but also correlated with the former metrics
- For purpose triples, PMI has a 0.18 correlation coefficient
  - Hyponyms and hypernyms tend to co-occur more frequently than causes/effects or means/purposes
- No conclusions taken for causation
Some observations:

- Hypernymy is highly correlated with all metrics except $\sigma$
- Part-of is less, but also correlated with the former metrics
- For purpose triples, PMI has a 0.18 correlation coefficient
  - Hyponyms and hypernyms tend to co-occur more frequently than causes/effects or means/purposes
- No conclusions taken for causation
  - Few correct triples
Some observations:

- Hypernymy is highly correlated with all metrics except $\sigma$
- Part-of is less, but also correlated with the former metrics
- For purpose triples, PMI has a 0.18 correlation coefficient
  - Hyponyms and hypernyms tend to co-occur more frequently than causes/effects or means/purposes
- No conclusions taken for causation
  - Few correct triples
- Synonymy has low or negative correlation coefficients with the metrics
Some observations:

- Hypernymy is highly correlated with all metrics except $\sigma$
- Part-of is less, but also correlated with the former metrics
- For purpose triples, PMI has a 0.18 correlation coefficient
  - Hyponyms and hypernyms tend to co-occur more frequently than causes/effects or means/purposes
- No conclusions taken for causation
  - Few correct triples
- Synonymy has low or negative correlation coefficients with the metrics
  - Few correct triples
  - In corpora, synonymous words do not co-occur frequently...
Metrics-based threshold

- Threshold based on the Cocitation value
Metrics-based threshold

- Threshold based on the Cocitation value
- Increased gradually for **hypernymy** triples
Metrics-based threshold

- Threshold based on the Cocitation value
- Increased gradually for hypernymy triples
- 50 seems to be a good cut-point
## New combined metrics?

- Metrics learned with Weka

### Table: Metrics with higher correlation coefficient.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Simple Linear</th>
<th>Corel</th>
<th>Isotonic</th>
<th>Corel</th>
</tr>
</thead>
<tbody>
<tr>
<td>cause_of</td>
<td>(0.01*σ + 0.05)</td>
<td>0.12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>purpose_of</td>
<td>(0.02*Pmi - 0.6)</td>
<td>0.22</td>
<td>Pmi</td>
<td>0.24</td>
</tr>
<tr>
<td>hypernymy</td>
<td>(0.02*Cocitation + 0.49)</td>
<td>0.56</td>
<td>Cocitation</td>
<td>0.66</td>
</tr>
<tr>
<td>part_of</td>
<td>(0.01*Lin + 0.26)</td>
<td>0.28</td>
<td>Cocitation</td>
<td>0.38</td>
</tr>
<tr>
<td>synonymy</td>
<td>-</td>
<td>-</td>
<td>σ</td>
<td>0.22</td>
</tr>
</tbody>
</table>
New combined metrics?

- Metrics learned with Weka

Table: Metrics with higher correlation coefficient.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Simple Linear</th>
<th>Corel</th>
<th>Isotonic</th>
<th>Corel</th>
</tr>
</thead>
<tbody>
<tr>
<td>cause_of</td>
<td>((0.01\sigma + 0.05))</td>
<td>0.12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>purpose_of</td>
<td>((0.02\text{Pmi} - 0.6))</td>
<td>0.22</td>
<td>\text{Pmi}</td>
<td>0.24</td>
</tr>
<tr>
<td>hypernymy</td>
<td>((0.02\text{Cocitation} + 0.49))</td>
<td>0.56</td>
<td>\text{Cocitation}</td>
<td>0.66</td>
</tr>
<tr>
<td>part_of</td>
<td>((0.01\text{Lin} + 0.26))</td>
<td>0.28</td>
<td>\text{Cocitation}</td>
<td>0.38</td>
</tr>
<tr>
<td>synonymy</td>
<td>-</td>
<td>-</td>
<td>(\sigma)</td>
<td>0.22</td>
</tr>
</tbody>
</table>

- Best correlation selects the measure which minimises the squared error
Discrete classification

- Models obtained using a 10-fold cross-validation test
Discrete classification

- Models obtained using a 10-fold cross-validation test
  - J48 decision tree learned for purpose_of
Discrete classification

- Models obtained using a 10-fold cross-validation test
  - J48 decision tree learned for purpose_of
  - Classifies 59.1% of the purpose_of triples correctly
Instead of a *term-document* matrix...

- If a *term-term* matrix was used
- Context = sentence
Instead of a *term-document* matrix...

- If a *term-term* matrix was used
- Context = sentence
- Statistical dominance (considering hypernymy and part_of):
  - *term-document* vs. *term-term* = 89%
  - *term-term* vs. *term-document* = 72%
Conclusions

- IE may benefit from the application of IR metrics
Conclusions

- IE may benefit from the application of IR metrics
  - At least concerning hypernymy and part-of relations
Conclusions

- IE may benefit from the application of IR metrics
  - At least concerning hypernymy and part-of relations
- Using either a *term-document* or a *term-term* matrix seems to suit our purpose.
Conclusions

- IE may benefit from the application of IR metrics
  - At least concerning hypernymy and part-of relations
- Using either a *term-document* or a *term-term* matrix seems to suit our purpose.
- What if the triples and the matrix were extracted from different sources?
Conclusions

- IE may benefit from the application of IR metrics
  - At least concerning hypernymy and part-of relations
- Using either a term-document or a term-term matrix seems to suit our purpose.
- What if the triples and the matrix were extracted from different sources?
- Future:
  - Use more documents of the corpus
  - Use another corpus
  - Web distributional metrics
Conclusions

- IE may benefit from the application of IR metrics
  - At least concerning hypernymy and part-of relations
- Using either a *term-document* or a *term-term* matrix seems to suit our purpose.
- What if the triples and the matrix were extracted from different sources?
- Future:
  - Use more documents of the corpus
  - Use another corpus
  - Web distributional metrics
  - Weight triples in available Portuguese lexical resources (eg. PAPEL)
References


Thank you!

Questions?