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

Introduction

- Knowledge bases (eg. WordNet) are useful resources for NLP

Introduction

- Knowledge bases (eg. WordNet) are useful resources for NLP
- Their creation and maintenance involves intensive human effort

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

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
  - Higher coverage, easier update, but...

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
  - Higher coverage, easier update, but...
  - Precision is lower

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
  - 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 resouces.

Information retrieval (IR)

Locating specific information in natural language resouces.

- Approaches based on the occurrence of words in documents.

Information retrieval (IR)

Locating specific information in natural language resouces.

- 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))
  - σ (Kozima and Furugori (1993))
  - ...

Goals

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

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

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?

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?

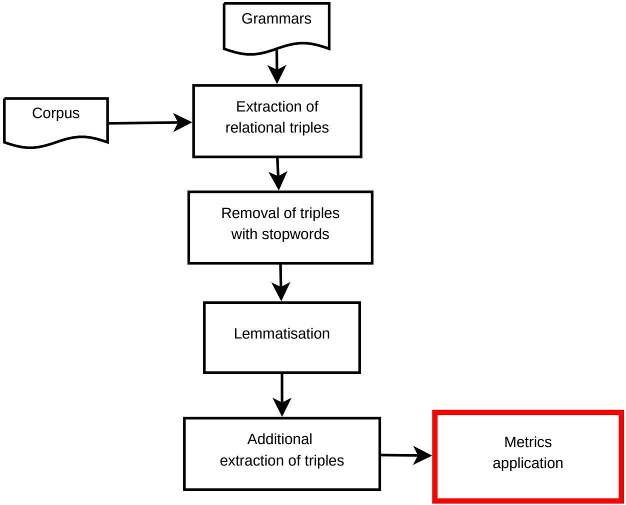
Goals

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


Grammars
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/cetempublico/

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: **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))

$$Cocitation(d_i, d_j) = \frac{P(d_i \cap d_j)}{P(d_i \cup d_j)} \quad (1)$$

Similarity between two documents

For instance, Cocitation:

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

$$Cocitation(d_i, d_j) = \frac{P(d_i \cap d_j)}{P(d_i \cup d_j)} \tag{1}$$

  - 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))

$$Cocitation(d_i, d_j) = \frac{P(d_i \cap d_j)}{P(d_i \cup d_j)} \tag{1}$$

  - 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))

$$Cocitation(d_i, d_j) = \frac{P(d_i \cap d_j)}{P(d_i \cup d_j)} \quad (1)$$

  - 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

$$Cocitation(e_i, e_j) = \frac{P(e_i \cap e_j)}{P(e_i \cup e_j)} \tag{2}$$

Adaptation to measure word similarity

$$Cocitation(e_i, e_j) = \frac{P(e_i \cap e_j)}{P(e_i \cup e_j)} \quad (2)$$

- e_i, e_j represent two entities (uni or multiword)

Adaptation to measure word similarity

$$Cocitation(e_i, e_j) = \frac{P(e_i \cap e_j)}{P(e_i \cup e_j)} \quad (2)$$

- e_i, e_j represent two entities (uni or multiword)
- $\mathrm{P}(e_i \cap e_j)$, is the number of documents containing both entities

Adaptation to measure word similarity

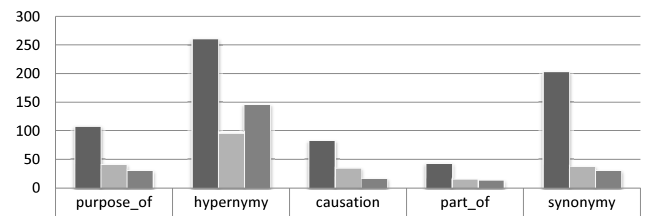
$$Cocitation(e_i, e_j) = \frac{P(e_i \cap e_j)}{P(e_i \cup e_j)} \quad (2)$$

- e_i, e_j represent two entities (uni or multiword)
- $\mathrm{P}(e_i \cap e_j)$, is the number of documents containing both entities
- $\mathrm{P}(e_i \cup e_j)$, is the number of documents containing at least one of the entities

Triples and metrics

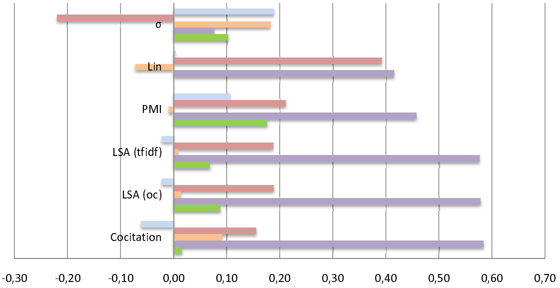
Triple	Manual	Coc	LSA (oc)	LSA (tf-idf)	PMI	Lin	σ
nação SINONIMO_DE *povo* *nation* SYNONYM_OF *people*	2	4.21	7.92	8.21	66.65	55.12	35.79
violência CAUSADOR_DE *estrago* *violence* CAUSE_OF *damage*	2	1.60	4.38	4.47	63.90	29.51	43.82
palavra HIPERONIMO_DE *beato* *word* HYPERNYM_OF *pietist*	1	0.16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

Manual validation of the results



	purpose_of	hypernymy	causation	part_of	synonymy
0	108	261	83	43	203
1	41	96	34	15	37
2	30	146	16	13	30

Manual evaluation vs. Distributional metrics



	Cocitation	LSA (oc)	LSA (tfidf)	PMI	Lin	σ
synonymy	-0,06	-0,02	-0,02	0,11	0,00	0,19
part_of	0,15	0,19	0,19	0,21	0,39	-0,22
causation	0,09	0,01	0,01	-0,01	-0,07	0,18
hypernymy	0,58	0,58	0,58	0,46	0,41	0,08
purpose_of	0,02	0,09	0,07	0,18	0,00	0,10

- Some observations:
  - Hypernymy is highly correlated with all metrics except σ

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

- Some observations:
  - Hypernymy is highly correlated with all metrics except σ
  - 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 σ
  - 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 σ
  - 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 σ
  - 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 σ
  - 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 σ
  - 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

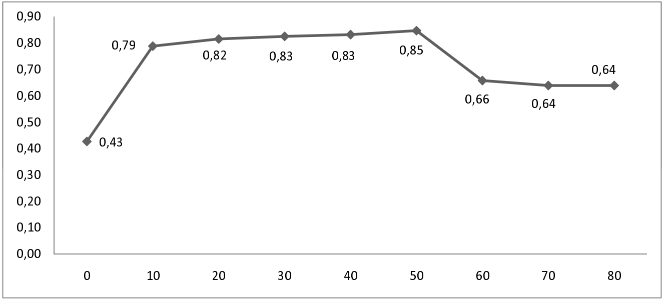
- Threshold based on the Cocitation value

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.

Relation	Simple Linear	Corel	Isotonic	Corel
cause_of	$(0.01*\sigma+0.05)$	0.12	-	-
purpose_of	(0.02*Pmi-0.6)	0.22	Pmi	0.24
hypernymy	(0.02*Cocitation+0.49)	0.56	Cocitation	0.66
part_of	(0.01*Lin+0.26)	0.28	Cocitation	0.38
synonymy	-	-	σ	0.22

New combined metrics?

- Metrics learned with Weka

Table: Metrics with higher correlation coefficient.

Relation	Simple Linear	Corel	Isotonic	Corel
cause_of	$(0.01*\sigma+0.05)$	0.12	-	-
purpose_of	(0.02*Pmi-0.6)	0.22	Pmi	0.24
hypernymy	(0.02*Cocitation+0.49)	0.56	Cocitation	0.66
part_of	(0.01*Lin+0.26)	0.28	Cocitation	0.38
synonymy	-	-	σ	0.22

- Best correlation selects the measure which minimises the squared error

Discrete classification

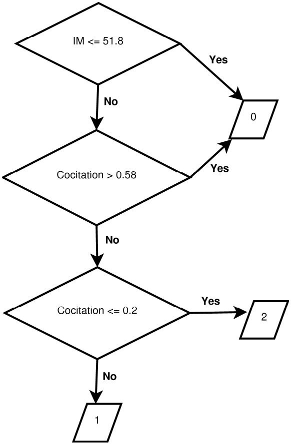
- Models obtained using a 10-fold cross-validation test

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

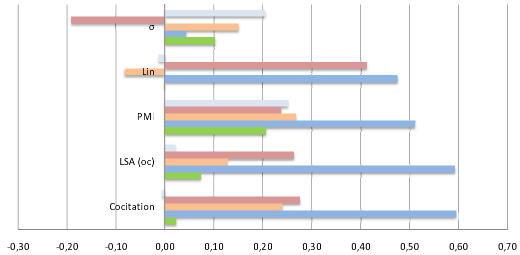


Instead of a *term-document* matrix...

- 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%



	Cocitation	LSA (oc)	PMI	Lin	σ
synonymy	-0,01	0,02	0,25	-0,01	0,20
part_of	0,27	0,26	0,24	0,41	-0,19
causation	0,24	0,13	0,27	-0,08	0,15
hypernymy	0,59	0,59	0,51	0,47	0,04
purpose_of	0,02	0,07	0,21	0,00	0,10

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

Cederberg, S. and Widdows, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proc. 7th (CoNLL)*, pages 111–118. Association for Computational Linguistics.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Kozima, H. and Furugori, T. (1993). Similarity between words computed by spreading activation on an english dictionary. In *Proc. 6th EACL*, pages 232–239. ACL.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. 15th ICML*, pages 296–304. Morgan Kaufmann.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269.

Turney, P. D. (2001). Mining the web for synonyms: PMI–IR versus LSA on TOEFL. In *Proc. 12th ECML*, volume 2167, pages 491–502. Springer.

Wandmacher, T., Ovchinnikova, E., Krumnack, U., and Dittmann, H. (2007). Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. In Meyer, T. and Nayak, A. C., editors, *Proc. 3rd Australasian Ontology Workshop (AOW 2007)*, volume 85 of *CRPIT*, pages 61–69. ACS.

Thank you!

Questions?