Information Extraction and Semantic Analysis

Giuseppe Bruno (Bank of Italy)

Discussant:
Monica Scannapieco | Italian National Institute of Statistics

Questions

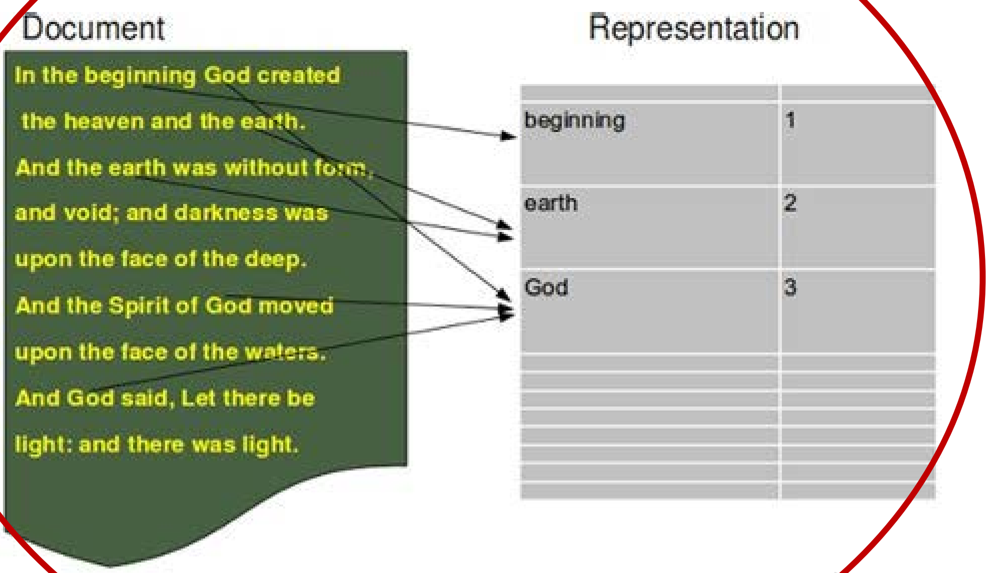
1. What are the most relevant concepts considered by the FSR (Financial Stability Report) document corpus btw 2010 and 2016?
2. What is the readability and formality level of each document?
3. How can we measure the impact of FSR on the readers by web searching?

Main Paper Contributions

- Practical experience on dealing with unstructured data
    - R code for text processing and sentiment orientation
- Data: Corpus of 58 docs (Financial Stability Report) from 2010 to 2016, each of about 40 pages
- **Most relevant concepts** in the corpus, **quality, impact**
    - Nearest Neighbours (crisis and stability)
    - Coherence (semantic similarities of sentences in the corpus)
    - Readability
    - Formality
    - Impact of FSRs by web search

Data Representation Models

Bag of words



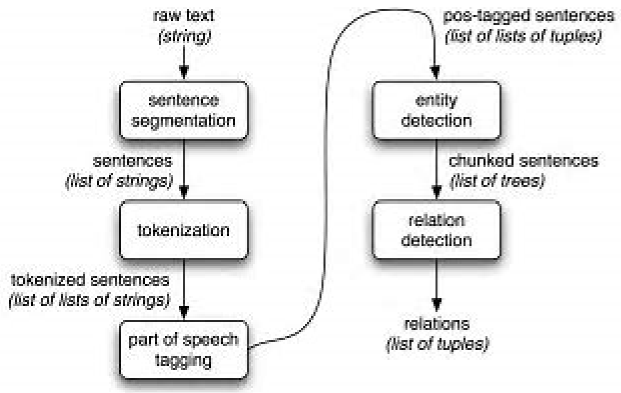
Document

In the beginning God created
the heaven and the earth.
And the earth was without form,
and void; and darkness was
upon the face of the deep.
And the Spirit of God moved
upon the face of the waters.
And God said, Let there be
light: and there was light.

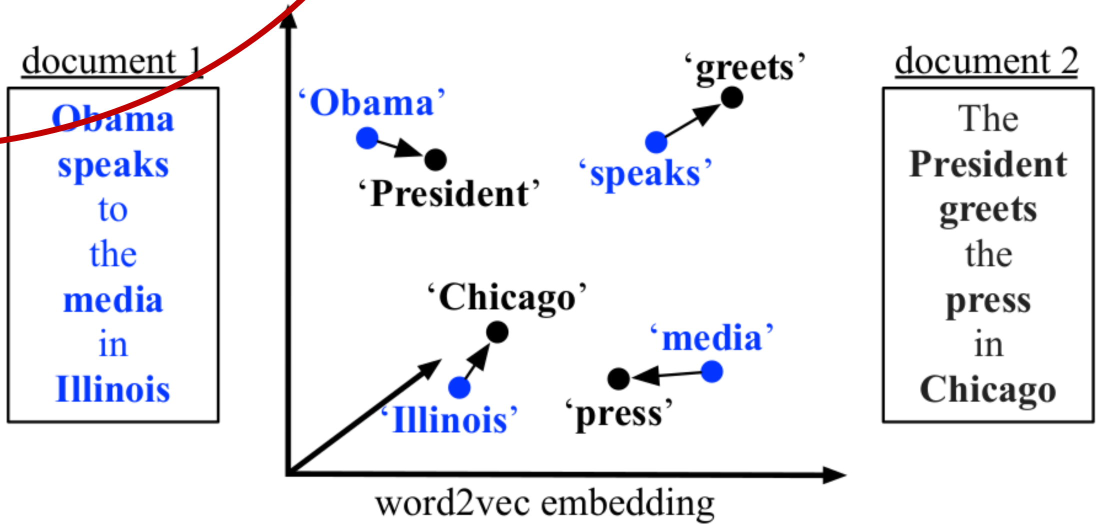
Representation

beginning	1
earth	2
God	3

Ontology Extraction



Word embeddings



document 1

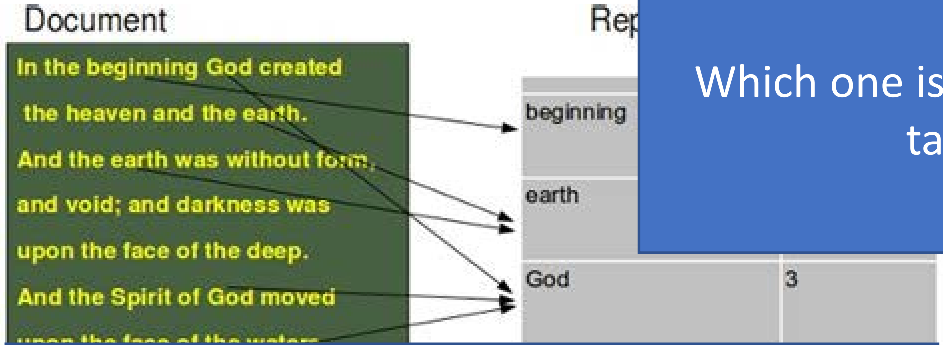
Obama
speaks
to
the
media
in
Illinois

document 2

The
President
greets
the
press
in
Chicago

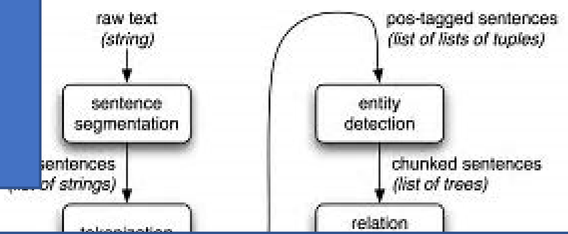
Data Representation Models

Bag of words


Document
In the beginning God created
the heaven and the earth.
And the earth was without form,
and void; and darkness was
upon the face of the deep.
And the Spirit of God moved
Rep
beginning
earth
God
3

Which one is better for the task?

Ontology Extraction

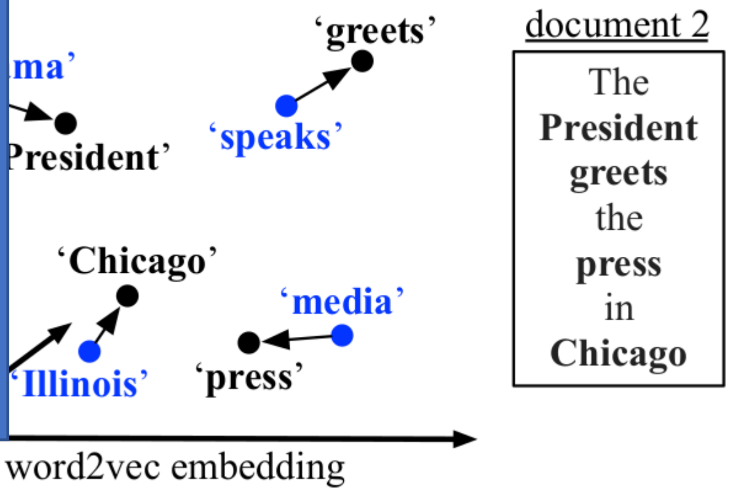


Ontology Extraction: huge projects at enterprise scale

LSA vs Word2vec: Word2Vec performance has a severe decrease, thus LSA becoming the more suitable tool*1

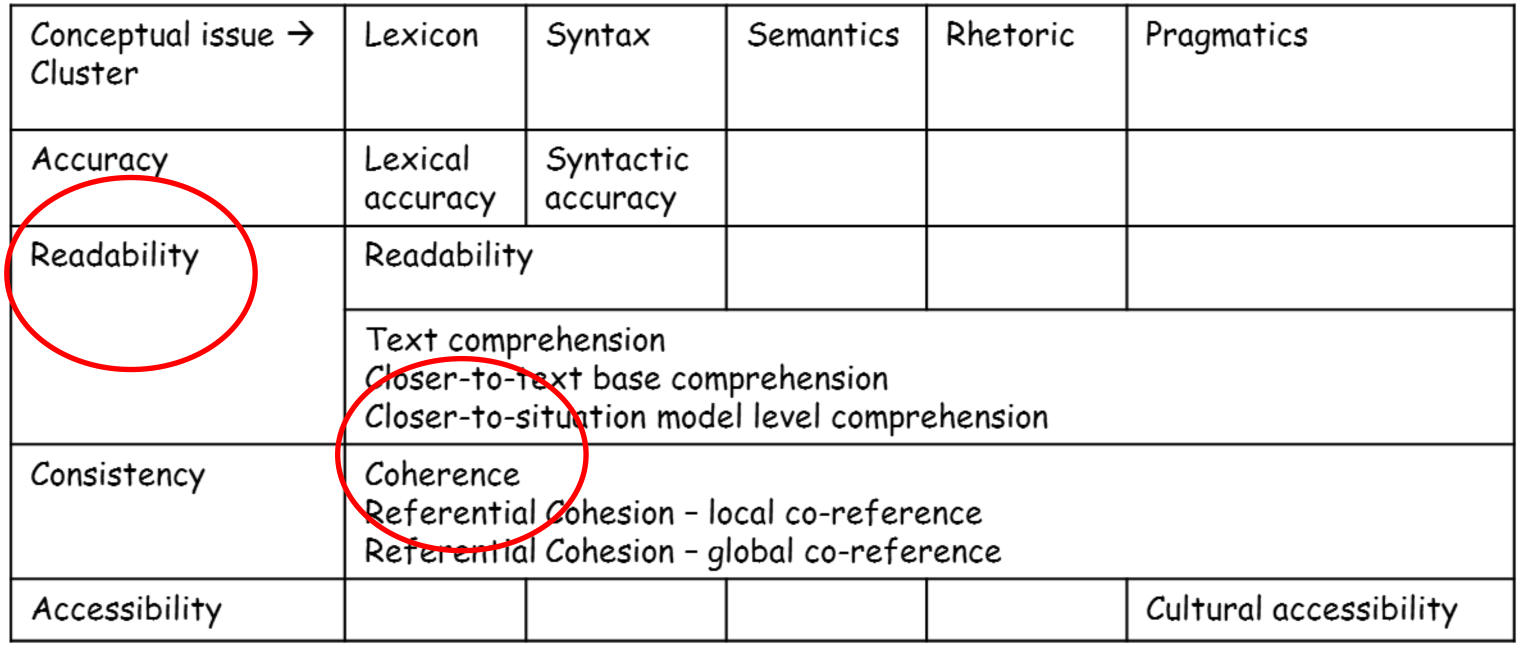
*1 Altszyler et al.: Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database, 2017

ord embeddings



Coherence, Readability

- Two of the quality dimensions of texts + formality



Conceptual issue → Cluster	Lexicon	Syntax	Semantics	Rhetoric	Pragmatics
Accuracy	Lexical accuracy	Syntactic accuracy			
Readability	Readability				
	Text comprehension Closer-to-text base comprehension Closer-to-situation model level comprehension				
Consistency	Coherence Referential Cohesion – local co-reference Referential Cohesion – global co-reference				
Accessibility					Cultural accessibility

Batini, Scannapieco: «Data and Information Quality», Springer 2016

Questions: Full quality evaluation framework? Semi-structureness to be considered?

Impact through Web searches

- Semantic orientation: historical word (Turney 2002)

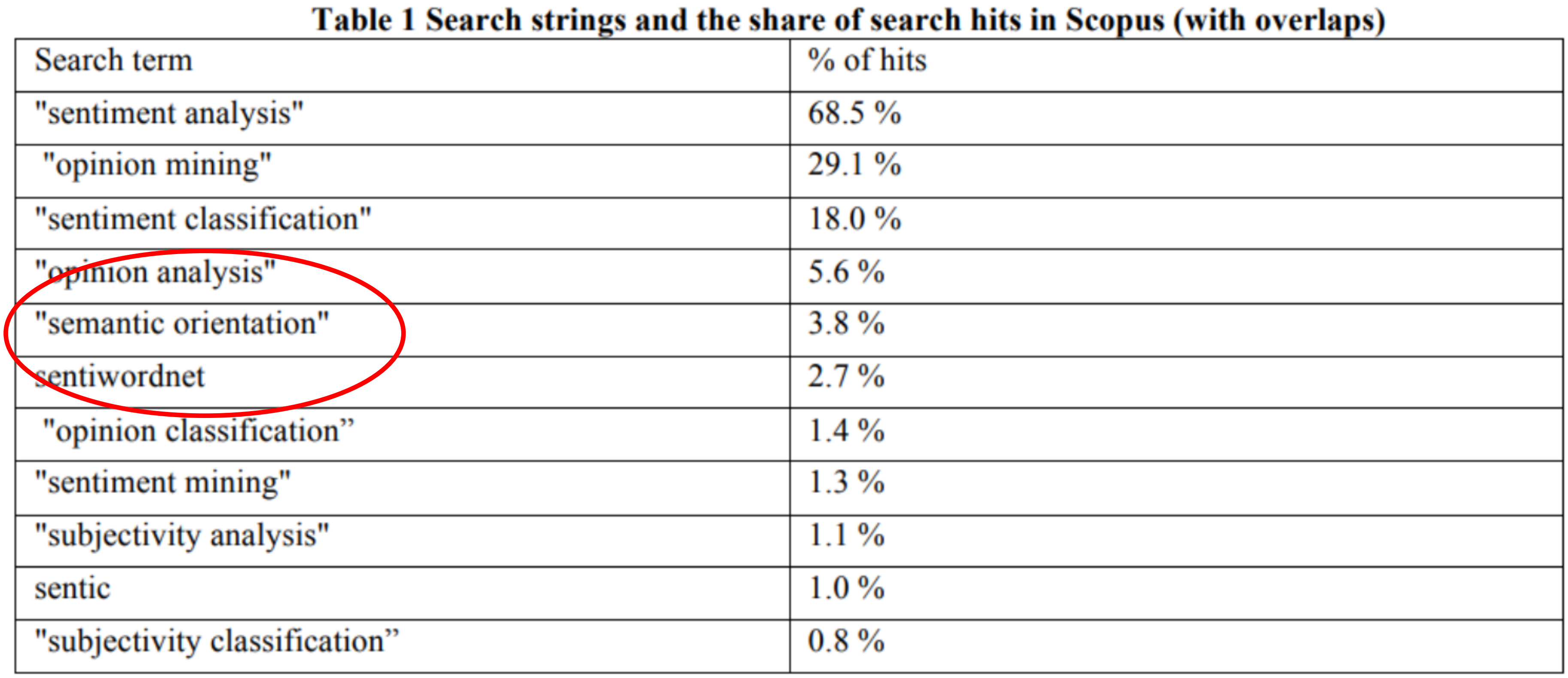


Table 1 Search strings and the share of search hits in Scopus (with overlaps)

Search term	% of hits
"sentiment analysis"	68.5 %
"opinion mining"	29.1 %
"sentiment classification"	18.0 %
"opinion analysis"	5.6 %
"semantic orientation"	3.8 %
sentiwordnet	2.7 %
"opinion classification"	1.4 %
"sentiment mining"	1.3 %
"subjectivity analysis"	1.1 %
sentic	1.0 %
"subjectivity classification"	0.8 %

Mäntylä et al: The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers
https://arxiv.org/ftp/arxiv/papers/1612/1612.01556.pdf, 2016

Impact through Web searches

- Whole sets of statements of RSF
- PMI (Pointwise Mutual Information) approximated by Web search hits in the web search of the statement associated with two antonyms
- Computation for three antonyms:
    - Stabilità/instabilità
    - Crisi/espansione
    - Vulnerabilità/solidità
- On the results:
    - Initial issue of 2010, 2014, 2016
    - Conclusions only sketched

Answers in the paper

1. What are the most relevant concepts considered by the FSR (Financial Stability Report) document corpus btw 2010 and 2016?

2. What is the readability and formality level of each document? Possible expansions, e.g. further quality dimensions

3. How can we measure the impact of FSR on the readers by web searching?
   - Some more work to evaluate if FSRs have an impact: e.g. expansion of the corpus, comparison with other search engines, possibly filter of «news» web hits in Google