On the Quality of the Annotations with a Controlled Vocabulary

Heidi Hobel, **Artem Revenko**

SEMANTIC WEB COMPANY
linking data to knowledge

IFIN, Florence
September 12, 2016

Annotations

Annotation

An **annotation** *is a metadata (e.g. a comment, explanation, presentational markup) attached to text, image, or other data.*

Annotations

Annotation

An **annotation** *is a metadata (e.g. a comment, explanation, presentational markup) attached to text, image, or other data.*

Why?

Add another layer of data representation, thereby improve

- Search;
- Comparison;
- Comprehensibility, etc.

Controlled Annotations

Controlled Vocabularies

Fixing the list of available annotations allows for:

- Control over focus;
- Consistency;
- Translation.

Controlled Annotations

Controlled Vocabularies

Fixing the list of available annotations allows for:

- Control over focus;
- Consistency;
- Translation.

Example

Long Buy In forex trading, you are considered in a long position if you buy base currency and sell quote currency.

Short Buy The opposite of long buy. A position is considered short if you buy quote currency.

Ask The dealer has come to a decision to call on a currency quotation, he will be selling on an ask price a base currency in exchange of quote currency.

Thesauri

Thesaurus

Thesaurus is a controlled vocabulary with additional relations like broader/narrower.

- Clearly defined *semantic* relations between annotations.

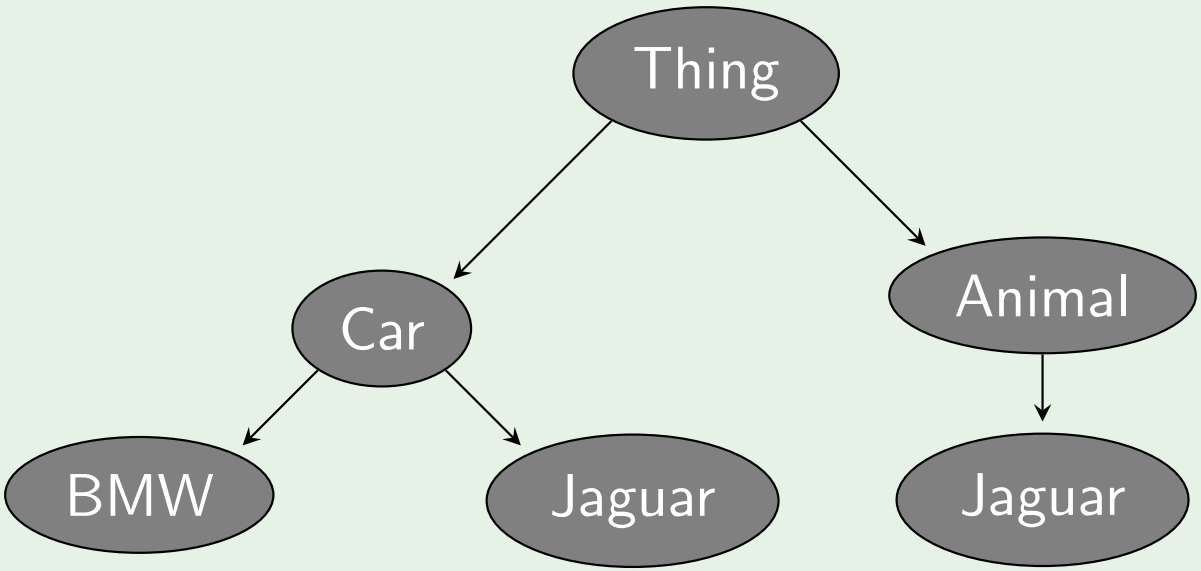
Thesauri

Thesaurus

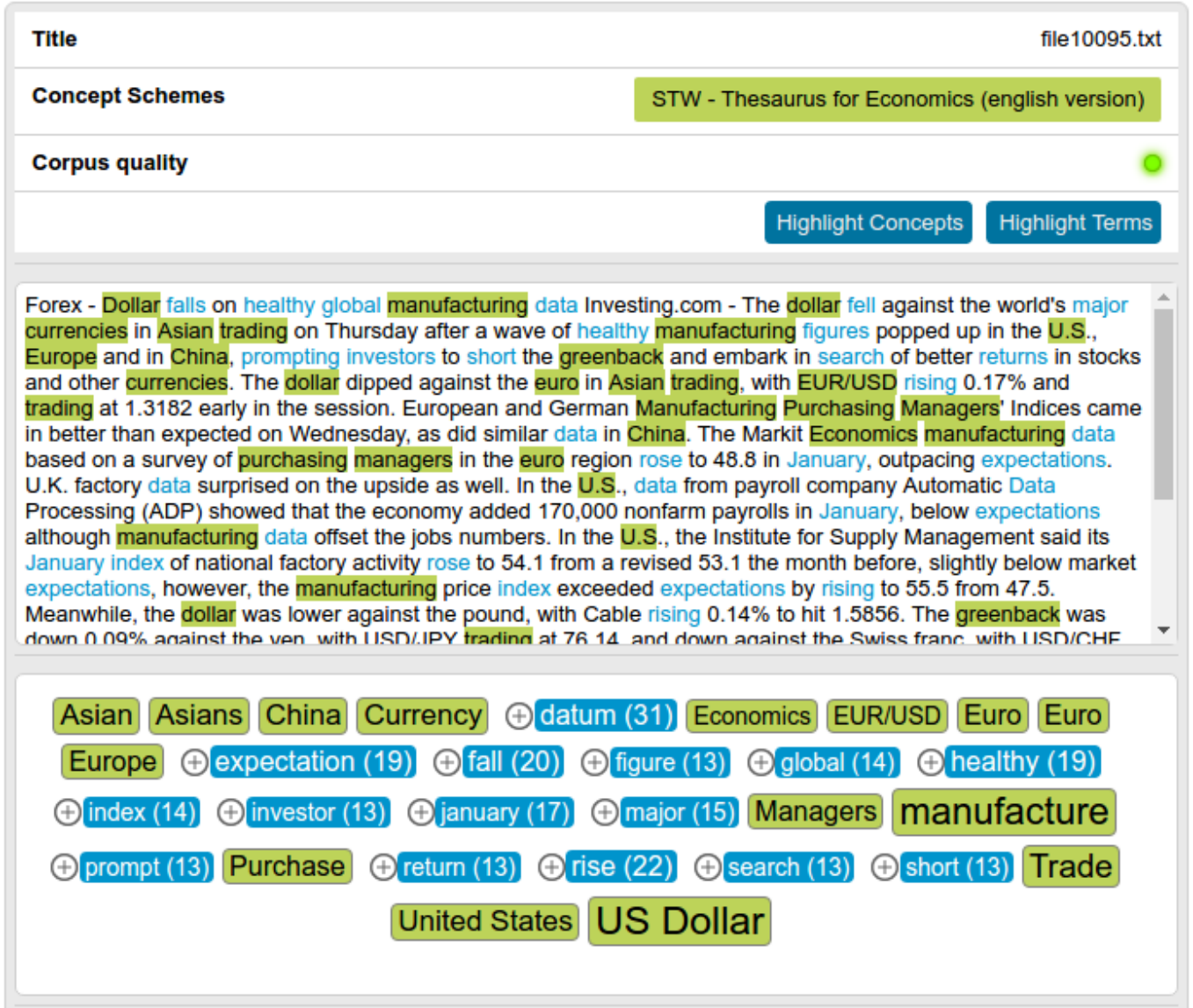
Thesaurus is a controlled vocabulary with additional relations like broader/narrower.

- Clearly defined *semantic* relations between annotations.

Example



Annotation in PROFIT PP Server



SEMANTIC WEB COMPANY

Document Similarities

Document 1

Mario Draghi was not expected to announce any changes to monetary policy

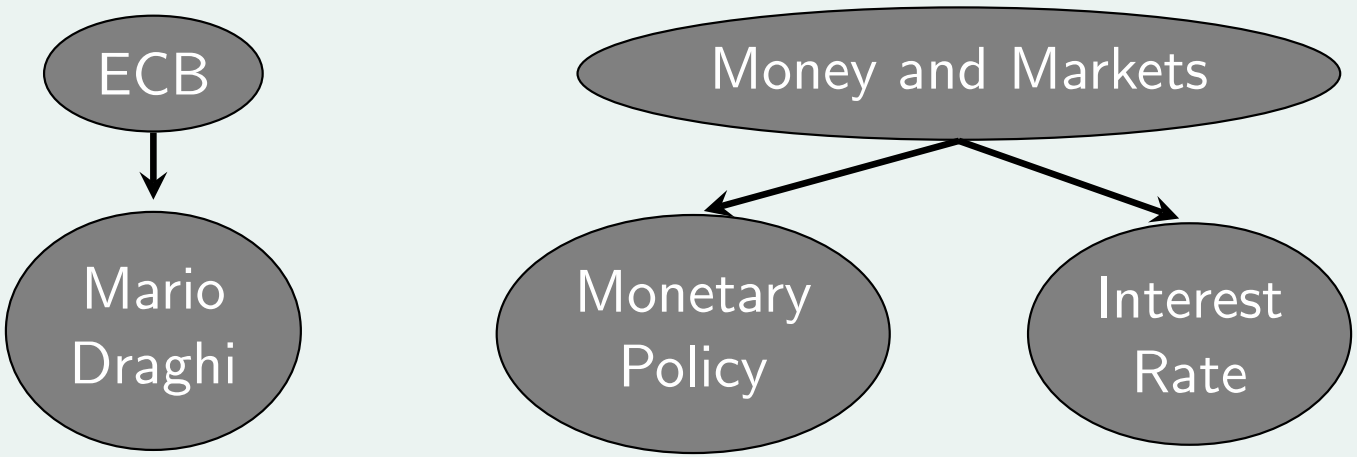
Document 2

European Central Bank leaves interest rate at record-low 1%

SEMANTIC WEB COMPANY

Document Similarities

Thesaurus



Document 1

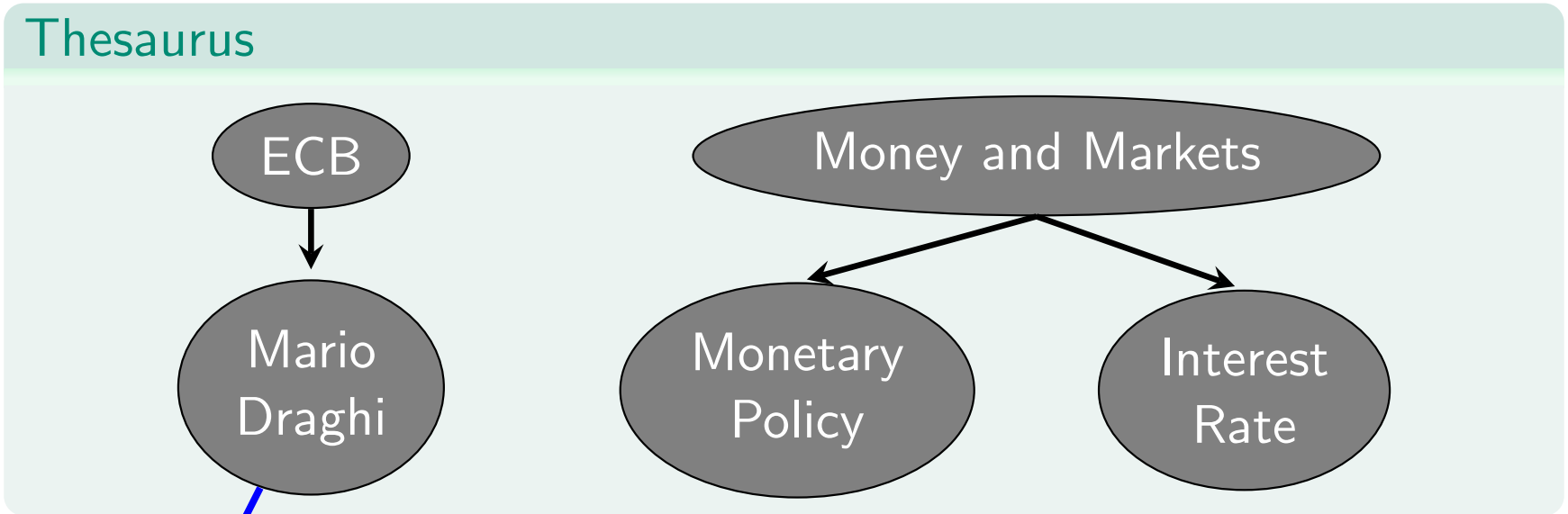
Mario Draghi was not expected to announce any changes to monetary policy

Document 2

European Central Bank leaves interest rate at record-low 1%

SEMANTIC WEB COMPANY

Document Similarities



Document 1

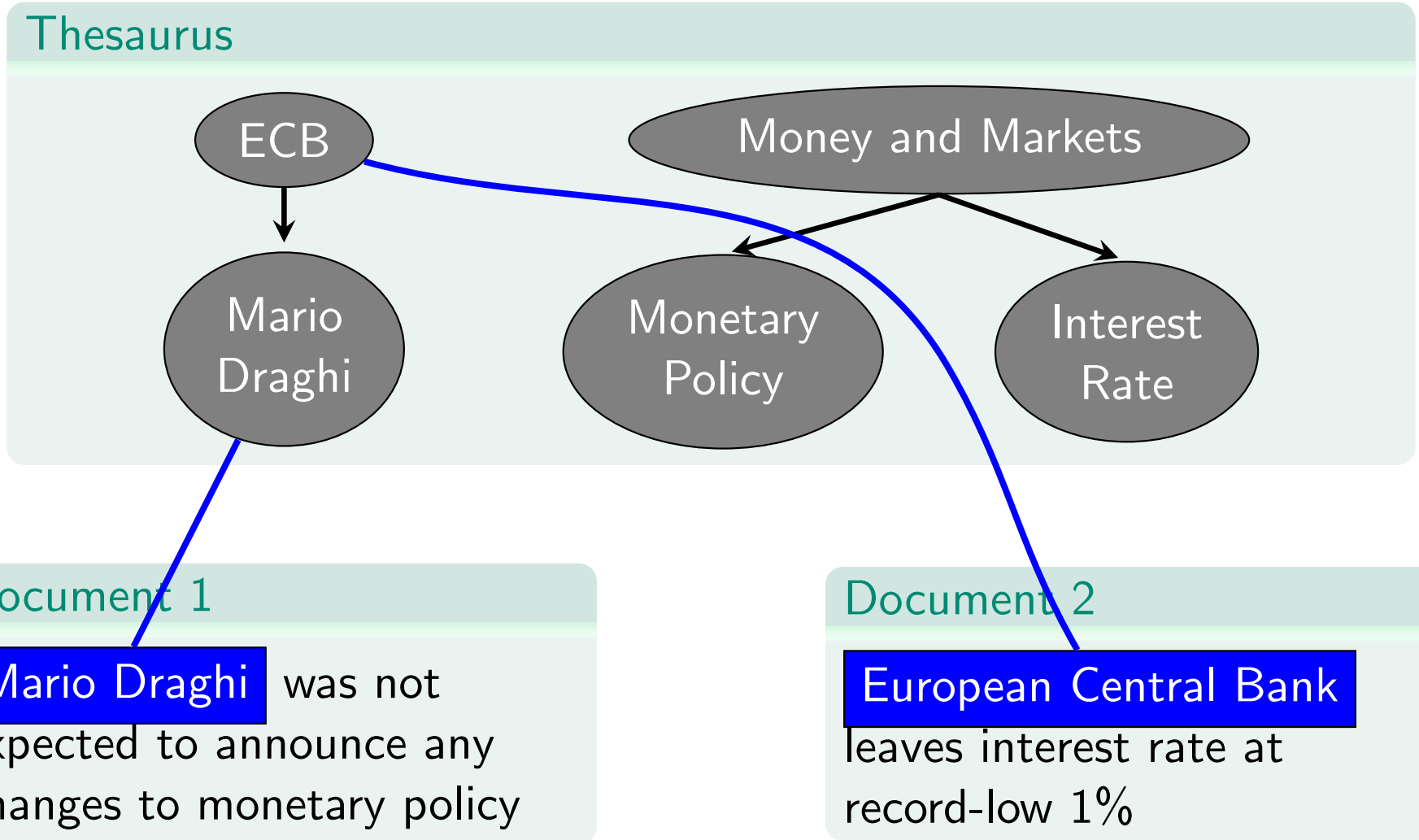
Mario Draghi was not expected to announce any changes to monetary policy

Document 2

European Central Bank leaves interest rate at record-low 1%

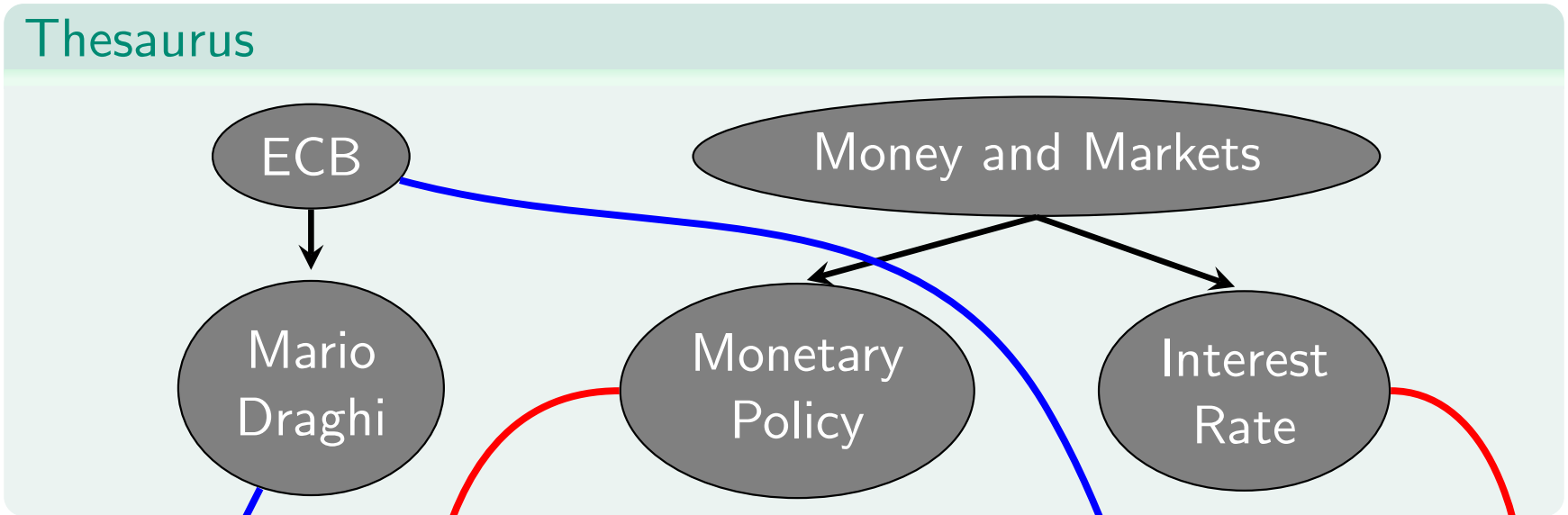
SEMANTIC WEB COMPANY

Document Similarities



SEMANTIC WEB COMPANY

Document Similarities



Document 1

Mario Draghi was not expected to announce any changes to monetary policy

⇔

Document 2

European Central Bank leaves interest rate at record-low 1%

SEMANTIC WEB COMPANY

Lemmatization

The same word may appear in different inflected forms:

- Plural or singular,
- Gender,
- Tense,
- Part of Speach (example: “race”).

Shall those be counted separately?

Lemmatization

The same word may appear in different inflected forms:

- Plural or singular,
- Gender,
- Tense,
- Part of Speach (example: “race”).

Shall those be counted separately?

Lemmatization

The process of bringing the word to its “base” form.
Special rules for compound terms.

Keywords

Keyword

A word which occurs in a text more often than we would expect to occur by chance alone.

How to find them?

From the explanation above $\Rightarrow$ compare facts and expectations.

Keywords

Keyword

A word which occurs in a text more often than we would expect to occur by chance alone.

How to find them?

From the explanation above $\Rightarrow$ compare facts and expectations.

$$C_i = \log_2(df_i+1)*[\log_2(\frac{N}{df_i})-p(\geq 0, \frac{cf_i}{N})]$$ [Manning and Schütze, 199

Keywords

Keyword

A word which occurs in a text more often than we would expect to occur by chance alone.

How to find them?

From the explanation above $\Rightarrow$ compare facts and expectations.

$$C_i = \log_2(df_i+1)*[\log_2(\frac{N}{df_i})-p(\geq 0, \frac{cf_i}{N})]$$ [Manning and Schütze, 199

Compound Terms

For the compound ones we use Mutual Information Score.

$$MI = \log_2(\frac{p(ij)}{p(i)*p(j)})$$ [Bouma, 2009]

Concepts

Concepts

Resources from the thesaurus. May have synonyms.

Experimental Settings

investing.com Corpora

eurostoxx	5834
eur/usd	19119
crude oil	14204

STW thesaurus

Original		Extended	
Concepts	6521	Concepts	6831
Broader/narrower	15892	Broader/narrower	16174

Extracted Keywords and Concepts

Intuition

Keywords The density of information,

Concepts Captured focus.

Extracted Keywords and Concepts

Intuition

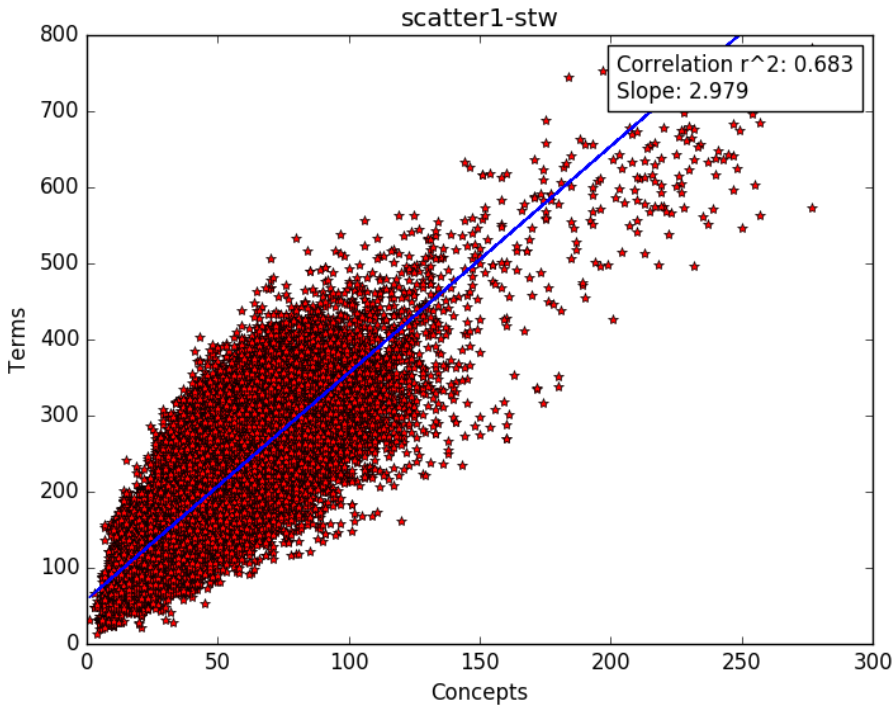
Keywords The density of information,

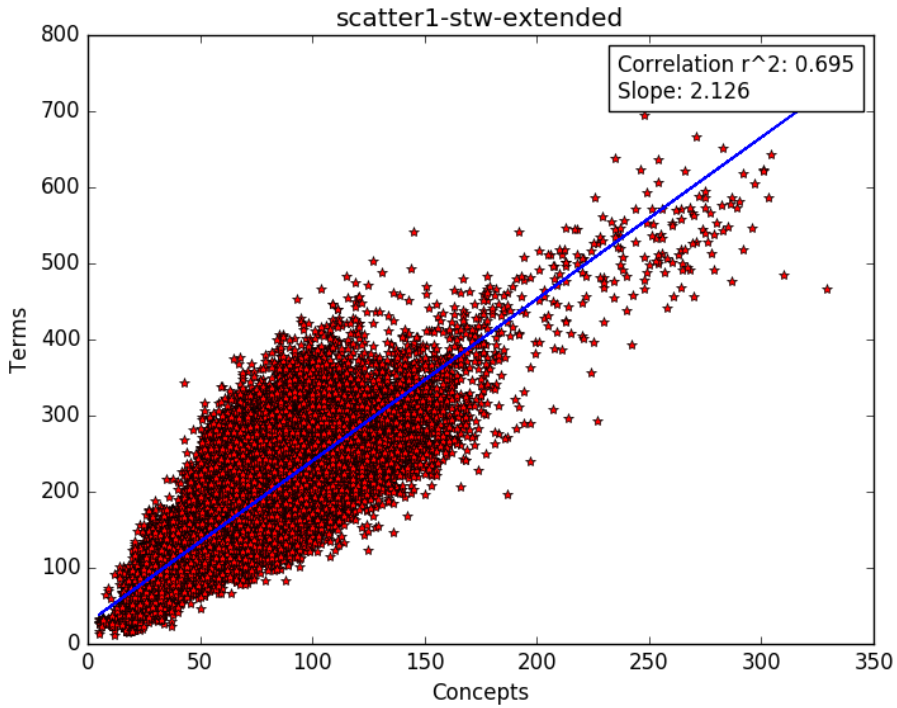
Concepts Captured focus.

It is better if they correlate.

Old vs Extended

Number of terms, threshold 1 ($C > 1$)

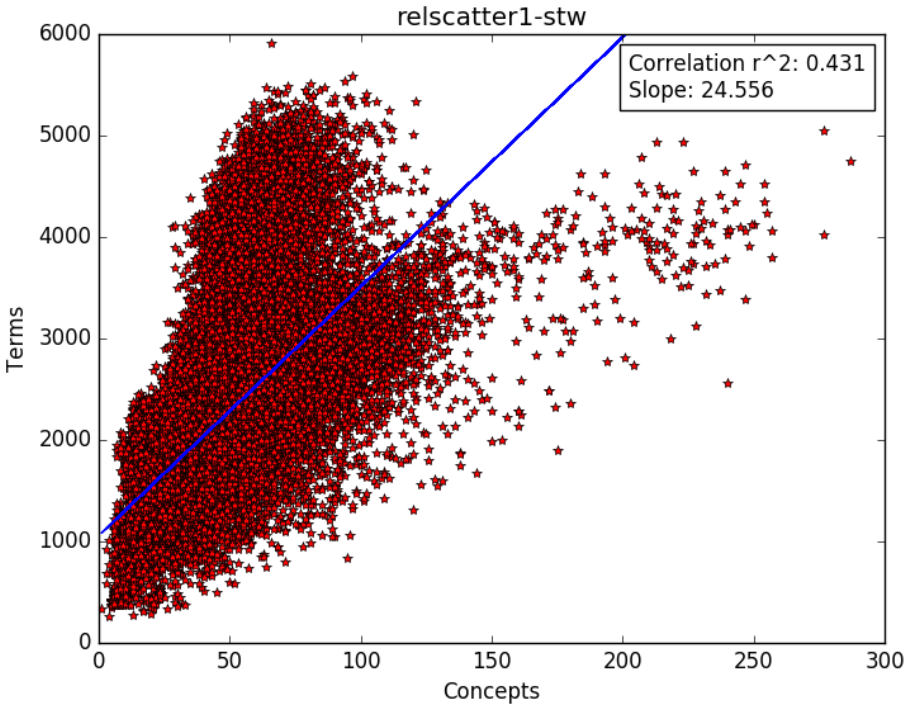


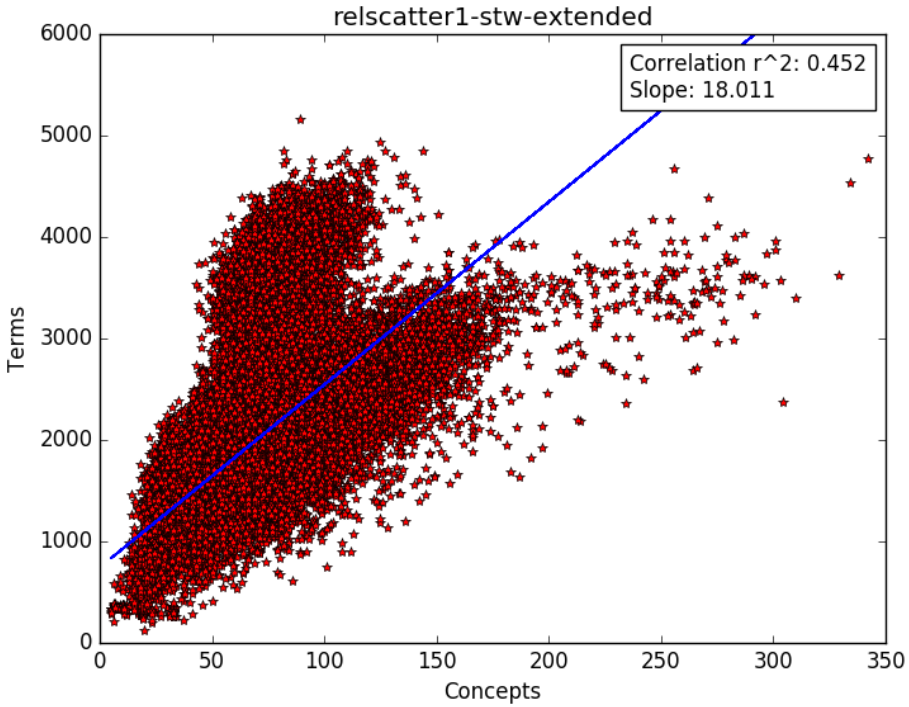


SEMANTIC WEB COMPANY

Old vs Extended

Sum of scores

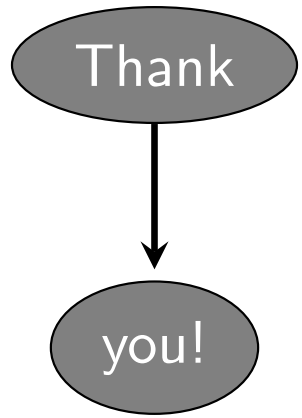




SEMANTIC WEB COMPANY

Conclusion

- The usefulness of annotations depends on the underlying tokens;
- Keywords provide a useful measure of the density of indormation, but they are not always good for annotations;
- The improvement of the thesaurus and annotation quality may be measured wrt the keywords.


Thank
you!


Bouma, G. (2009).
Normalized (pointwise) mutual information in collocation extraction.
Proceedings of GSCL, pages 31–40.


Manning, C. D. and Schütze, H. (1999).
Foundations of statistical natural language processing, volume 999.
MIT Press.