On the Quality of the Annotations with a Controlled Vocabulary

Heidi Hobel, Artem Revenko

IFIN, Florence
September 12, 2016
Introduction

Usage Example

Annotations
  Preprocessing
  Extraction
  Keywords vs Concepts Metrics
  Experiments

Conclusion
Introduction

Usage Example

Annotations
  Preprocessing
  Extraction
  Keywords vs Concepts Metrics
  Experiments

Conclusion
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.

Usage 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.
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.
Thesaurus

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

- Clearly defined *semantic* relations between annotations.
Introduction

Usage Example

Annotations
  Preprocessing
  Extraction
  Keywords vs Concepts Metrics
  Experiments

Conclusion
Annotation in PROFIT PP Server

Forex - Dollar falls on healthy global manufacturing data. Investing.com - The dollar fell against the world's major currencies in Asian trading on Thursday after a wave of healthy manufacturing figures popped up in the U.S., Europe and in China, prompting investors to short the greenback and embark in search of better returns in stocks and other currencies. The dollar dipped against the euro in Asian trading, with EUR/USD rising 0.17% and trading at 1.1812 early in the session. European and German Manufacturing Purchasing Managers' Indices came in better than expected on Wednesday, as did similar data in China. The Markit Economics manufacturing data based on a survey of purchasing managers in the euro region rose to 48.8 in January, outpacing expectations. U.K. factory data surprised on the upside as well. In the U.S., data from payroll company Automatic Data Processing (ADP) showed that the economy added 170,000 nonfarm payrolls in January, below expectations although manufacturing data offset the jobs numbers. In the U.S., the Institute for Supply Management said its January index of national factory activity rose to 54.1 from a revised 53.1 the month before, slightly below market expectations, however, the manufacturing price index exceeded expectations by rising to 55.5 from 47.5. Meanwhile, the dollar was lower against the pound, with Cable rising 0.14% to hit 1.5856. The greenback was down 0.09% against the yen, with USD/JPY trading at 76.14, and down against the Swiss franc, with USD/CHF.
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%
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%
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%
Mario Draghi was not expected to announce any changes to monetary policy.

European Central Bank leaves interest rate at record-low 1%.
Mario Draghi was not expected to announce any changes to monetary policy.

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

Usage Example

Annotations

- Preprocessing
- Extraction
- Keywords vs Concepts
- Metrics
- Experiments

Conclusion
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 Speech (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 ⇒ 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 ⇒ compare facts and expectations.

\[ C_i = \log_2(df_i+1) \times \left[ \log_2\left( \frac{N}{df_i} \right) - p(\geq 0, \frac{cf_i}{N}) \right] \]

[Maning and Schütze, 1999]
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 ⇒ compare facts and expectations.

\[ C_i = \log_2(df_i+1) \times [\log_2\left(\frac{N}{df_i}\right) - p(\geq 0, \frac{cf_i}{N})] \]  

[Manning and Schütze, 1999]

Compound Terms

For the compound ones we use Mutual Information Score.

\[ MI = \log_2\left(\frac{p(ij)}{p(i) \times p(j)}\right) \]  

[Bouma, 2009]
Concepts

Resources from the thesaurus. May have synonyms.
## Experimental Settings

### investing.com Corpora
- eurostoxx 5834
- eur/usd 19119
- crude oil 14204

### STW thesaurus

<table>
<thead>
<tr>
<th>Original</th>
<th>Extended</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concepts</td>
<td>Concepts</td>
</tr>
<tr>
<td>6521</td>
<td>6831</td>
</tr>
<tr>
<td>Broader/narrower</td>
<td>Broader/narrower</td>
</tr>
<tr>
<td>15892</td>
<td>16174</td>
</tr>
</tbody>
</table>
**Extracted Keywords and Concepts**

### Intuition

<table>
<thead>
<tr>
<th>Keywords</th>
<th>The density of information,</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concepts</td>
<td>Captured focus.</td>
</tr>
</tbody>
</table>
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 ($C > 1$)

- Scatter plot for `scatter1-stw` with correlation $r^2 = 0.683$, slope: 2.979.
- Scatter plot for `scatter1-stw-extended` with correlation $r^2 = 0.695$, slope: 2.126.
Old vs Extended

Sum of scores

![Graph showing old vs extended sum of scores](chart.jpg)
Introduction

Usage Example

Annotations
  Preprocessing
  Extraction
  Keywords vs Concepts Metrics
  Experiments

Conclusion
Conclusion

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