Revenue Corpus — Documentation

Abstract

This documentation describes the “Revenue Corpus”, i.e., a German text corpus with 1,128 news articles, in which 2,075 statements on revenue were manually annotated by domain experts from the industry. The articles have been taken from general and business news websites.

1 Aim of the Corpus

The purpose of the corpus is to investigate both the structure of sentences on financial criteria and the distribution of associated information over the text. The compilation aims at being representative for target documents, a search engine returns to queries on revenue.

2 Download and Installation

The corpus is free for scientific use under the *CreativeCommons* license and can be downloaded at http://infexba.upb.de/results.html. The corpus documents are packed in *tar.gz archives*. For instructions on how to extract such files, see http://www.gzip.org.

2.1 File formats

For each article, the content comes as unicode plain text (UTF-8). The URL of the article has been appended to the text, so the original HTML source code can be accessed.

Annotations are specified in a standard XMI file. These files are preformatted for *Apache UIMA*, which is an implementation of the *Unstructed Information Management Architecture* (Ferrucci and Lally, 2004) for the development of natural language processing applications. However, the annotations can be easily imported into arbitrary applications as well.

Source websites	Total	Training	Val./Test
produktion.de	139	127	12
heise.de	139	127	12
golem.de	129	117	12
wiwo.de	123	112	11
boerse-online.de	111	100	11
spiegel.de	104	93	11
capital.de	87	76	11
tagesschau.de	73	0	73
finanzen.net	37	0	37
www.vdma.org	37	0	37
de.news.yahoo.com	37	0	37
faz.net	19	0	19
vdi.de	16	0	16
zdnet.de	13	0	13
handelsblatt.com	13	0	13
zvei.org	11	0	11
sueddeutsche.de	7	0	7
boerse.ard.de	7	0	7
it-business.de	5	0	5
manager-magazin.de	5	0	5
sachen-machen.org	5	0	5
swissinfo.ch	4	0	4
hr-online.de	1	0	1
nachrichten.finanztreff.de	1	0	1
tognum.com	1	0	1
pcgameshardware.de	1	0	1
channelpartner.de	1	0	1
cafe-future.net	1	0	1
pokerzentrale.de	1	0	1
	1128	752	188/188

Table 1: Distribution of source websites, the corpus documents have been taken from.

3 Description

The corpus consists of 1,128 German news articles from the years 2003 to 2009, which were taken from 29 general and business news websites, as shown in Table 1.

In each article, statements on the revenue of companies or markets were manually annotated, i.e., sentences and entities that refer to a statement are tagged and linked to each other. The annotation scheme is given below. Altogether, 2,075 statements are annotated in this way.

Statements	Total	Forecasts	Declarations
Complete corpus	2075	523 (25.2%)	1552 (74.8%)
Training set	1366	306 (22.4%)	1060 (77.6%)
Validation set	362	113 (31.2%)	249 (68.8%)
Test set	347	104 (30.0%)	243 (70.0%)

Table 2: Statements on revenue in the corpus.

3.1 Split for Evaluation

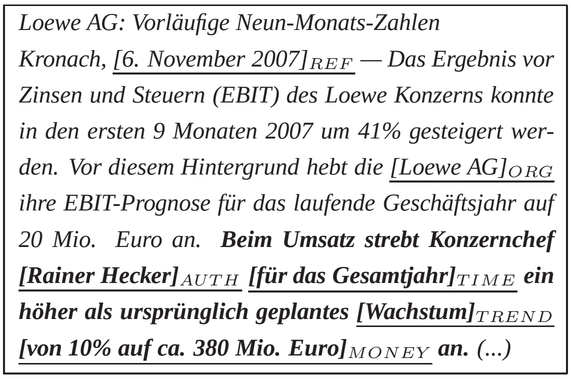
We created a split, in which two third of the documents constitute the training set and each one sixth refers to the validation and test set. To simulate real conditions, the training documents were randomly chosen from only the seven most represented websites, while the validation and test data both cover all 29 sources. Table 2 shows some corpus statistics, which give a hint that the validation and test set differ significantly from the training set.

4 Annotations

Every text span that covers sentence, includes a temporal entity T and a monetary entity M and that represents a *forecast* or *declaration* on the revenue of an organization or market is marked as such. If a sentence comprises more than one statement, it is annotated multiple times.

T and M are annotated themselves and linked to the sentence. Relative money information is preferred against absolute amounts in case they are separated. Accordingly, the *subject* is tagged (and linked) within the sentence boundaries if available, otherwise its last mention in the preceding text. The same holds for optional entities, namely a possibly *referenced time*, a *trend indicator* and the *author* of a statement. As can be seen in the example of Figure 1, only information that refers to a statement on revenue (typed in bold face) is annotated. These annotations may be spread across the text and are represented in the XMI files as follows:

- **Forecast/Declaration:** A sentence that represents a forecast or declaration on revenue.
- **Organization/Market:** The subject, i.e., either an organization or a market.
- **TimeExpression:** The temporal entity.
- **ReferencePoint:** The reference time, the temporal entity refers to.
- **MoneyExpression:** The monetary entity.
- **Trend:** A word that indicates the trend of the monetary entity.
- **Author:** The author of the statement.



Loewe AG: Vorläufige Neun-Monats-Zahlen

Kronach, [6. November 2007]$_{REF}$ — Das Ergebnis vor Zinsen und Steuern (EBIT) des Loewe Konzerns konnte in den ersten 9 Monaten 2007 um 41% gesteigert werden. Vor diesem Hintergrund hebt die [Loewe AG]$_{ORG}$ ihre EBIT-Prognose für das laufende Geschäftsjahr auf 20 Mio. Euro an. ***Beim Umsatz strebt Konzernchef [Rainer Hecker]$_{AUTH}$ [für das Gesamtjahr]$_{TIME}$ ein höher als ursprünglich geplantes [Wachstum]$_{TREND}$ [von 10% auf ca. 380 Mio. Euro]$_{MONEY}$ an.*** *(...)*

Figure 1: An annotated document of the corpus. The text is taken from *www.boerse-online.de*, but has been modified for clarification.

Subject	Total	Forecasts	Declarations
Organization	1425 (68.7%)	312 (40.3%)	1113 (71.7%)
Market	650 (31.3%)	211 (59.7%)	439 (28.3%)

Table 3: Corpus distribution of subjects.

Table 3 shows that the majority of annotated statements discusses organizations. However, markets occur comparatively often in forecasts.

5 Compilation and Annotation Process

The source documents were manually selected and prepared by four employees of one of our industrial partners. Afterwards, two of their employees annotated the plain document text with the *CAS Editor* provided by Apache UIMA. Annotation guidelines were written before.

Each document was annotated once. With respect to the statement annotations, a preceding pilot study yielded substantial inter-annotator agreement, as indicated by the value $\kappa = 0.79$ of the conservative measure *Cohen's Kappa* (Carletta, 1996). Additionally, we performed a manual correction process for each annotated document to improve consistency.

6 Participants

The corpus was planned by Henning Wachsmuth from the University of Paderborn, Germany and Peter Prettenhofer from the Bauhaus University Weimar, Germany. The process of collecting and manually annotating the documents was done by Resolto Informatik GmbH, a company from the semantic technology field based in Herford, Germany.

The main publication connected to the project is Wachsmuth et al. (2010).

References

Carletta, Jean. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22: 249–254.

Ferrucci, David and Adam Lally. 2004. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3–4): pages 327–348.

Wachsmuth, Henning, Peter Prettenhofer, and Benno Stein. 2010. Efficient Statement Identification for Automatic Market Forecasting. In *Proceedings of the 23rd International Conference on Computational Linguistics*, to appear.