QuoteMine: A Repository of Newsworthy Quotes

Alan Akbik and Martin Schenck

Technische Univeristät Berlin,
Database Systems and Information Management Group,
Einsteinufer 17, 10587 Berlin, Germany
{firstname.lastname}@tu-berlin.de

Abstract. We present QuoteMine, a repository of newsworthy German language quotes that is automatically gathered from newswire text on the Web. The project is available online through a Web interface that allows users to query for all quotes of certain persons, all quotes that contain specific keywords, or both. Each quote is displayed with a link to the news article in which it was found, as well as an automatically computed sentiment value. Our system runs continuously, keeping up-to-date with the current stream of online news. In this demonstration, we illustrate the design of our system and show how relatively simple Information Extraction methods can be used on large amounts of text to create a large repository of structured data. We discuss present challenges and future ideas.

Keywords: Information Extraction, Newswire Articles, Quote Extraction, Speaker Attribution

1 Introduction

Comments and statements from politicians, celebrities and other people commonly feature in new media text to give a direct account of their point of views on different topics. In the fast-paced breaking-news environment of online newspapers, entire articles can be devoted to statements made by people of prominence. In the domain of politics, the use of quoted speech has even become such an important and often abused rhetorical device that the Urban Dictionary defines the verb *quote mining* as "the repeated use of quotes out of context in order to skew or contort the meaning of a passage or speech by an author on a controversial subject" [1].

Nevertheless, a direct search engine-like access to quotes and their speakers can be valuable for many domains. For instance, a journalist might want to know statements from world leaders with regards to global policy elements. In an election year, voters might want to browse statements of politicians on divisive issues. People might also be looking for quotes that contain certain words for the purpose of inspiration.

[1] http://www.urbandictionary.com/define.php?term=quote%20mining

In this system demonstration, we present QuoteMine, a continuously running system that crawls online German newswire articles and extracts quotes and speakers to address this information need. QuoteMine features an online search engine-like interface that allows for queries of quotes and speakers. It is publicly available online at `www.textmining.tu-berlin.de/quotemine/`[2] and at time of writing has gathered a repository of over 140.000 quotes.

QuoteMine is one of a number of quote repositories available online, but differs in philosophy, scale and language. Many projects, such as WikiQuote[3] and BrainyQuote[4], rely on manually curated content, either by a small group of contributors or a more open community-based effort. Such projects usually focus on quotes of high importance or impact, as manual curation does not scale to the amount of quotes contained in the large amount of news articles and blogs published every day. In contrast, our goal is to collect all quotes deemed *newsworthy* enough to feature in at least one news article.

Some related projects also use Information Extraction technologies on crawled Web data similar to QuoteMine. The largest such project known to us is NewsExplorer [2], which performs automatic detection of quotations in eleven languages and is, to our knowledge, the only system besides ours that extracts quotes in the German language. However, the patterns they apply are very restrictive, limiting extractions only for quotes found for one of a list of pre-defined speakers and only if found with certain verbs (says, said, etc.). QuoteMine on the other hand has no such restrictions, finding quotes for any entity identified as a person. Another difference between the two projects is that NewsExplorer focuses on news aggregation while QuoteMine is a dedicated tool for searching quotes and speakers.

In this demonstration, we illustrate the system outline, our information extraction methodology and discuss the structured data we collect in terms of potential for future projects.

2 System Outline

The system consists of three components that work independently of each other. The first is a *crawling component* that crawls newswire text from the ten biggest German online news sites. Using the RRS feeds of these sites, the crawler is informed when a new article is released. Each new article is crawled and the full HTML code stored locally for analysis. AJAX calls and other dynamic content loaded by a browser are not taken into account.

The downloaded Web pages are passed to an *information extraction component*, which executes a pipeline of pre-processing and extraction steps. First, a boilerplating method [1] is applied to remove all HTML and template code from the downloaded sites, leaving only plain text in natural language. This text is passed to two taggers; the first is a Named Entity detection method for

[2] http://www.textmining.tu-berlin.de/quotemine/ accessed on April 11th 2013
[3] http://en.wikiquote.org/wiki/Main_Page accessed on April 11th
[4] http://www.brainyquote.com/ accessed on April 11th

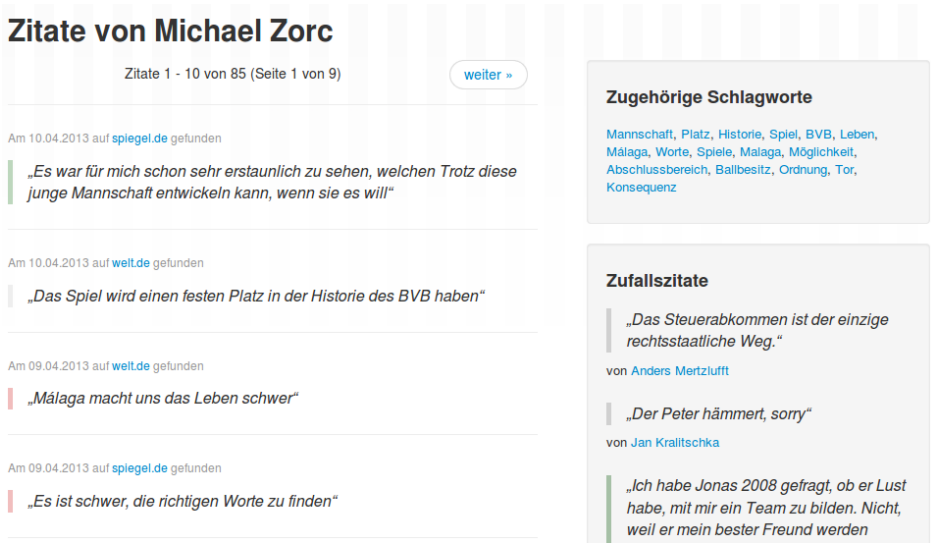

Fig. 1. A screenshot of the Web interface showing the results of a query for all quotes of “Michael Zorc”. The quotes are ordered by the date they were found by the system. The colored line left to every quote is an automatically computed sentiment value, where red indicates a negative and green a positive quote. In the lower right panel, a list of random quotes are displayed to encourage user to browse the repository in an exploratory fashion.

finding speakers in the text and the resolution of proper noun coreferences. The second is a tagger that identifies text within quotation marks as candidates for quoted speech. Using a set of manually crafted heuristics, quotes are attributed to speakers, yielding a list of quote-speaker tuples for each news article. In a post-processing step, quotes are analyzed using a sentiment analysis module that uses SentiWS [3], a list that assigns sentiment values to German words. Words such as “*schwer*” are assigned a negative value, while words such as “*erstaunlich*” a positive one. All structured information generated in the information extraction module (quotes, the document of quote discovery, speakers and sentiment values) is stored in a relational database.

The third component is the *front end* (Figure 1). It features full access to all extracted quotes and allows users to search for all quotes of a speaker, all quotes that contain a keyword, or both. Quotes are displayed along with their speaker, the name of the online newspaper where the quote was found and a sentiment value. The latter is presented in the form of a beam thats either red, green or colorless, indicating negative, positive or neutral sentiment respectively. Clicking on a quote leads to a detailed page where users can retrieve the Web page in which the quote was found. It is possible to browse through the repository by clicking on person names or keywords; For example, clicking on a person name

leads to a page that lists all known quotes of this person. In addition, random quotes are presented in a widget to encourage browsing the database.

3 Current and Future Work

We analyzed the present system with regards to quality of quote extraction and investigated additional functionality that we wish to add.

For the former, we conducted a preliminary evaluation of the system by manually annotating 200 randomly chosen news articles with quotes and speakers. We used our information extraction component on this corpus and found very high f-measure in detecting quotes (93%) and generally high quality detection of speakers (84%). Our heuristics of quote and speaker attribution performed well at an f-measure of 83%, but with potential for further improvement. Difficulties include quotes of unspecified speakers, such as people speaking on the condition of anonymity or unnamed bystanders. When more than one person name is named in the immediate vicinity of quoted text, the algorithm sometimes selects the wrong speaker. Currently, we are expanding the rule set for quote and speaker extraction to account for identified error classes and increase overall extraction quality.

In terms of functionality, we believe in line with user feedback that two kinds of structured information are missing from the prototype system. The first is the *topic* of a quote, which often is not explicitly stated within the quote text. Currently, we are investigating methods for detecting and assigning topic words to quotes. This information will be used to enable two new search types: A search for all quotes of a certain topic, and a search for all people that are often cited in the context of a certain topic.

We are also building a new extractor that finds *appositions* for people, which often indicate attributional information such as their profession. This will enable searches for quotes of all persons belonging to certain groups.

References

1. Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 441–450, New York, NY, USA, 2010. ACM.
2. Bruno Pouliquen, Ralf Steinberger, and Clive Best. Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing*, pages 487–492, 2007.
3. R. Remus, U. Quasthoff, and G. Heyer. Sentiws – a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, 2010.