Aggregating opinions: Explorations into Graphs and Media Content Analysis

Gabriele Tatzl
SORA
Institute for Social Research & Analysis
Vienna, Austria
gt@sora.at

Christoph Waldhauser
SORA
Institute for Social Research & Analysis
Vienna, Austria
chw@sora.at

Abstract

Understanding, as opposed to reading is vital for the extraction of opinions out of a text. This is especially true, as an author's opinion is not always clearly marked. Finding the overall opinion in a text can be challenging to both human readers and computers alike. Media Content Analysis is a popular method of extracting information out of a text, by means of human coders. We describe the difficulties humans have and the process they use to extract opinions and offer a formalization that could help to automate opinion extraction within the Media Content Analysis framework.

1 Introduction

When humans read, they try to not only decode the written language, but also link it with external information. This gives them access to meaning and opinion of a text, that remain hidden from a mere decoder. This process of reading can be organized scientifically within the framework of Media Content Analysis (MCA). Reading, however, is expensive in terms of time and money. Yet the volume of textual data that is available for research grows seemingly without bounds. Automating reading, indeed doing MCA – at least to some degree – is a very desirable advance for any practitioner in the field.

The purpose of this short positional paper is to introduce MCA as we use it in our day-to-day lives and discuss challenges and possible solutions for them, with regards to automation.

The remainder of this paper is organized as follows. First we give a brief introduction to Media Content Analysis and it's applications in the social sciences in general. We will then focus on opinion mining as an important task within the general MCA framework. Special emphasis will be put on the challenges humans (and computers alike) face, when extracting opinions from a document. As a contribution to the effort of overcoming these obstacles, we offer a formalized interpretation of the MCA opinion extraction process in section 4. Finally, some concluding remarks and suggestions for an algorithmic implementation are made.

2 Media Content Analysis

Media Content Analysis from a social science perspective is driven by research questions (e.g. *How does the perception of migrant groups vary in different media?*) and practical questions of private and public clients (e.g. *In which context do negative opinions about a corporation occur?*) in order to investigate and evaluate the content of communication.

Media Content analysis can be generally described as "systematic reading of a body of texts, images, and symbolic matter" (Krippendorf, 2004). It "is applied to a wide variety of printed matter, such as textbooks, comic strips, speeches, and print advertising" (Krippendorf, 2004) or more generally to any cultural artifact[1]. Additionally, Content Analysis is defined as an empirical method for (I) systematic and inter-subjective understandable description of textual and formal characteristics and (II) for inquiring into social reality that consists of inferring features of a non-manifest context from features of a manifest written text and other meaningful matters (Merten, 1995; Krippendorf, 2004; Früh, 2007).

There is a wide range of methods of research,

> "(...) from simple and extensive classifications of types of content for organizational or descriptive purposes to

[1] MCA is e.g. also used for comparing representations of groups, issues and events to their real-world occurrences.

Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, pages 93–97,
Uppsala, Sweden, 16 July 2010. ©2010 Association for Computational Linguistics

> deeply interpretative enquiries into specific examples of content, designed to uncover subtle and hidden potential meanings” (McQuail, 2005).

The methodology we use is based upon a broad foundation of recent and widely approved literature (Riffe et al., 1998; Franzosi, 2008; Kaplan, 2004; Merten, 1995; Roberts, 2001; Krippendorf, 2004; Neuendorf, 2007; Rössler, 2005; Früh, 2007; Weerakkody, 2009): The analysis typically starts from the formulation of some specific research questions, in terms of topics, actors and patterns of interpretation that need to be investigated. Based on theoretical foundations and operationalisation, categories (theoretically or empirically grounded) and indicators are defined. All categories together make up the codebook, which is the instrument for the manual coding of text. The codebook consists of different characteristics for every variable and of instructions for the manual coding. One can compare the codebook to the perhaps more familiar questionnaire used in empirical quantitative social science. In this understanding, the codebook is little more than questions on the text and some hints on how to answer them. For instance, a question might concern a statement’s speaker or subject actor and the way she is arguing her opinion: Is the argumentation of SACT in the statement rational?; possible answer codes are 1—the argumentation is consistent and rational, 2—the argumentation is not consistent and not well explained, and 3—no valuation possible.

In particular, variables are extracted on different levels of the documents: some address the whole document (article) and its source, some focus on claims to be able to answer all the different research questions. A core point in conducting empirical research is the demand for validity (external and internal) and reliability[2] (pre-tests). These quality checks have to be done carefully (Krippendorf, 2004).

The work proceeds with the identification (the manual annotation) of specific variables and indicators by turning text into numbers and fill out the codebook’s answer sheet (data entry mask). The turning of text into numbers (coding process) is at the moment a very cumbersome task, as it is done manually. Humans, so called coders (usually trained junior researchers), have to read each article and de facto answer questions (the codebook) on the text afterwards. Last but not least, the final data file (cleaned manual codings) is used in statistical analysis in order to answer the research questions. The significance of this methodology lies precisely in its capacity to describe the mediated public discourse and various forms and aspects of diversity (i.e. diversity of opinions).

It should be considered that we conduct neither discourse analysis (e.g. Hajer and Versteeg, 2005) nor linguistic analysis (e.g. Livesey, 2001). Our approach is an analysis of mediated public discourse (see inter alia Gerhards et al., 2007), which implies certain methodological differences. This methodology is especially useful for the analysis of web media content and can be combined with other approaches. In the LivingKnowledge project[3], the analysis of the mediated public discourse is combined with Multimodal Genre Analysis (Baldry and Thibault, 2005).

3 Opinion Mining in MCA

Determining the degree to which a whole article (entire content) or a statement in a text (part of content) is positive, negative or neutral is not the only but a very essential reason for conducting Media Content Analysis. Applying the kind of Media Content Analysis mentioned above, we are able to describe the polarity of an opinion and the degree of correlation between the polarity of an opinion and the context of the opinion holder. An opinion holder could be considered as the speaker (person or organization) of a statement in the text. The human coders are instructed by the codebook (rules for coding) how opinions should be detected and ranked (five point-scale[4]). We are firmly convinced that it is not possible to detect opinions across different use cases only by means of polar words or opinion bearing words, because meaning of these words is always dependent on the con-

[2] Reliability in Content Analysis is the amount of agreement or correspondence among two or more coders (Krippendorf, 2004; Neuendorf, 2007).

[3] The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement n°231126 Living Knowledge: Living Knowledge – Facts, Opinions and Bias in Time.

[4] Rate the opinion according to your interpretation of the article: The overall opinion is very positive, if the topic is mentioned with positive attributes and/or if a really positive outcome of an event is reported and not criticized and/or if the author of the article or more than half of the speakers talking about a certain topic evaluates it as very positive (1 = very positive).

tent's context. If you only have a short view on parts of the text, it can result in narrow incomplete interpretations. Besides that, additional information (which is not in the text) is often required to interpret an opinion and to understand the elements of social structure. It must be pointed out that when human coders read an article, there is a process of automatic inference. The proverbial concept of reading vs. understanding captures this notion with surprising accuracy. Correspondingly, sentiment analysis is a rather challenging process for humans as well as for computers.

4 Structuring opinions

In the following we will try to formalize what usually happens inside a human coder, coding an article. A typical research question in this sense might be: *is the opinion of article X, Θ_x positive, neutral, or negative towards a topic Y^5?* The tricky part lies in the fact, that very few articles state their opinions expressis verbis. Rather, articles contain a number of statements on diverse facets of the article's topic. These statements in turn are again composed of reported actions or speech of subject actors[6] (SACTs). All these elements can be thought of as nodes in a tree: article being the root node containing M statement nodes and N SACT nodes. Note, that the N SACT nodes need not be uniformly distributed between the M statement nodes. Figure 1 displays the tree structure inherent to Media Content Analysis.

Each node has a number of attributes, variables in the codebook terminology, such as the name of the author or SACT. Next to these obvious attributes there are also latent ones, which are only accessible by analyzing all child nodes and aggregating the results (possibly with using external information). Opinions of articles are one example of latent attributes in Media Content Analysis. The process of aggregating all of a statement's SACTs' opinions (θ_{mn}) into a single statement opinion (θ_m), and further aggregating all of an article's statement opinions into a single article opinion, lies at the hearth of opinion mining within the Media Content Analysis framework. Figure 2 depicts the aggregating of SACTs' opinions into a statement opinion as a subtree.

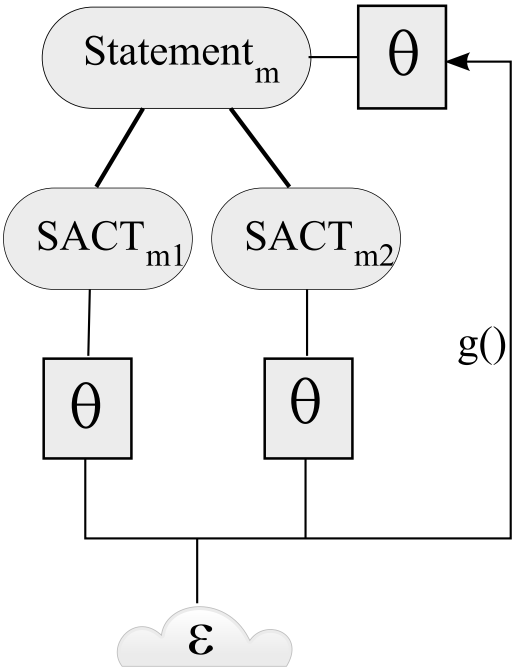

Figure 2: Aggregating SACTs' opinions into a statement opinion within the MCA framework

To return to the more formalized notation introduced above, $\Theta_x = f(g_1, g_2, \ldots, g_m)$, with $g_k(\theta_{m1}, \theta_{m2}, \ldots, \theta_{mn}, \epsilon)$. A description of these two classes of functions is not trivial. A function (f) that aggregates statement opinions (g_k, themselves aggregates of their SACTs' opinions) into an overall article opinion (θ) requires to take into account not only the opinion attributes of its statement arguments, but also their relationships, an assessment of their equal presentation and take hints at the author's intentions. This function will typically be a weighted mean of the values for the opinion variable for the contained statements:

$$\widehat{\Theta_x} = \frac{\sum_{k=1}^{M} w_k g_k}{\sum_{k=1}^{M} w_k}$$

Estimating the weights w_k needs to include the aforementioned interstatement relationships and presentation. For instance, in the aggregation of two mildly negative statements and a very positive one, do these opinions really cancel out? Difficult as this may be, aggregating SACTs' opinions into a single statement opinion is even more difficult. Here, external information (ϵ) plays a crucial role, e.g. can the three SACTs Bill Gates, Linus Torvalds and an unnamed undergraduate computer science student be equal contributors to any

[5] Selecting only statements that deal with a certain topic Y is beyond the scope of this paper. However, automating topic selection is rather feasible by including background knowledge on the topic itself. Background knowledge that is readily available at a very early stage of MCA research question formulation.

[6] A subject actor is the person that effects a claim, e.g. if the claim is a statement, it is the speaker

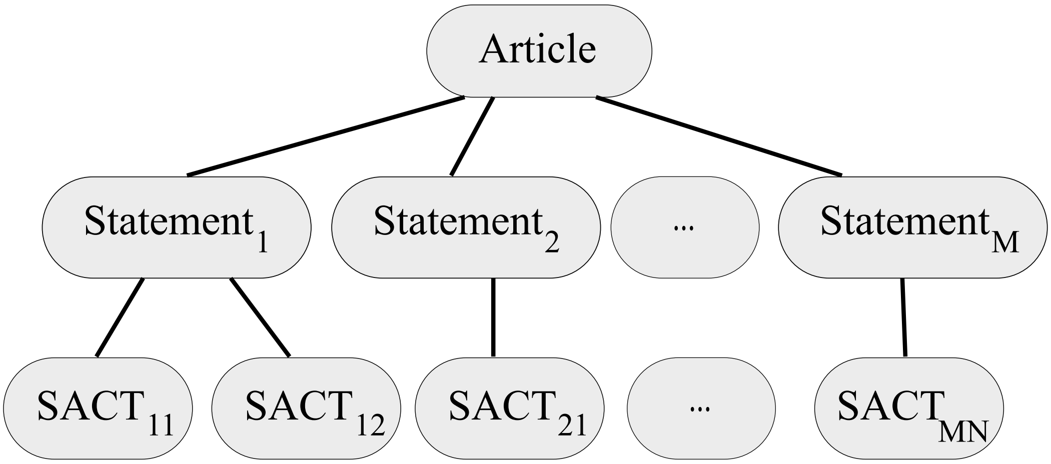

Figure 1: Relationship among levels of a document

given statement. In structure, this class of functions is also based on the weighted mean concept. However, in estimating the weights, notions of speaker interaction, speaker significance and effectiveness come into play. Many of these concepts cannot be sufficiently included by means of analyzing the text. Further, external information is required. This information can be thought of as an ontology or metadata, giving meaning to the actions and speech of a SACT. In a manual coding process, this information has been learned by the human coders through their past experience in reading texts. This is one of the reasons junior researchers, and not e.g. unskilled laborers, are used for this task. External knowledge, quite often to a substantial part, is critical in understanding a text.

5 Conclusion

Reading and understanding text is daunting task for humans. It requires years if not decades of training and experience to uncover hidden meanings and latent opinions. However, the process of reading is rather simple. We formalized this process by focusing on the example of extracting and aggregating opinions of an article. By rethinking reading and understanding opinions as a tree, we were able to structure the way humans use automatic inference to weight arguments and form opinions. The aggregating functions are simple themselves, however, estimating the right arguments is tricky. It requires the inclusion of massive amounts of external knowledge. In our opinion, this knowledge is currently not available in machine accessible form. With the ever increasing diffusion of semantic web data and ongoing efforts to create substantial ontologies of external knowledge, the future certainly will show interesting developments in this field.

In the meantime, thinking opinion extracting as traversing a tree might help to create software that helps human coders in their work. Also, large training sets of manually coded articles could be used to estimate the weights required to aggregate opinions on higher levels of analysis. However, achieving acceptable performance across diverse topics and usecases seems unlikely at this time.

References

Anthony Baldry and Paul J Thibault. 2005. *Multimodal Transcription and Text Analysis*. Equinox, London and Oakville.

Roberto Franzosi. 2008. *Content analysis*, volume 1 of *Sage benchmarks in social research methods*. Sage, Los Angeles.

Werner Früh. 2007. *Inhaltanalyse. Theorie und Praxis*. UVK, Konstanz, 6. rev. ed. edition.

Jürgen Gerhards, Anke Offerhaus, and Jochen Roose. 2007. The public attrobution of responsibility. developing an instrument for content analysis. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 59:105–125.

Marteen Hajer and Wytske Versteeg. 2005. A decade of discourse analysis of environmental politics: Achievements, challenges, perspectives. *Journal of Environmental Policy and Planning*, 7(3):175–184.

David Kaplan, editor. 2004. *The SAGE handbook of quantitative methodology for the social sciences*. Sage, Thousand Oaks.

Klaus Krippendorf. 2004. *Content analysis. An introduction to its methodology*. Sage, London, 2. ed edition.

Sharon M Livesey. 2001. Eco-identity as discursive struggle: Royal dutch/shell, brent spar and nigeria. *Journal of Business Communication*, 38(1):58–91.

Denis McQuail. 2005. *McQuail's Mass Communication Theory*. Sage, London, 5. ed edition.

Klaus Merten. 1995. *Inhaltsanalyse. Einführung in Theorie, Methode und Praxis*. Westdt. Verlag, Opladen.

Kimberly A Neuendorf. 2007. *The content analysis guidebook*. Sage, Thousand Oaks.

Daniel Riffe, Stephen Lacy, and Frederick Fico. 1998. *Analyzing media messages: using quantitative content analysis in research*. Erlbaum, Mahwah.

C W Roberts. 2001. Content analysis. In Smelser and Baltes (Smelser and Baltes, 2001).

Patrick Rössler. 2005. *Inhaltsanalyse*. UVK, Konstanz.

Neil J Smelser and Paul B Baltes, editors. 2001. *International Encyclopedia of the Social & Behavioral Science*. Elsevier, Amsterdam.

Niranjala Weerakkody. 2009. *Research Methods for Media and Communication*. Oxford University Press, Oxford.