Identifying formal and functional zones in film reviews

Heike Bieler and **Stefanie Dipper** and **Manfred Stede**
Applied Computational Linguistics
University of Potsdam
Karl-Liebknecht-Str. 24–25
D-14476 Golm
Email: {bieler,dipper,stede}@ling.uni-potsdam.de

Abstract

We describe our system for breaking a film review (as an instance of a *semi-structured document*) into its formal and functional constituents. Based on a corpus study, we devised a set of 25 zone labels indicating the role that a unit can play within the review. We identify *formal* zones with a set of symbolic rules, while the distinction between *descriptive* and *evaluative* paragraphs is drawn with a statistical classifier. The approach achieves between 70 and 79% precision in recognizing the zones in our corpus.

1 Introduction[1]

Many text genres can be characterized as *semi-structured*: They do not display a completely conventionalized structure (as, e.g., many *weather reports* or *cooking recipes* do), but there nevertheless are some rules and tendencies that allow the reader to quickly recognize a document as an instance of the genre, and to isolate important portions. As a case in point, we are working with film reviews coming from various newspapers and web sites. While their overall structure is definitely not identical, there are similarities on what portions (henceforth *zones*) to expect, and in what order to expect them. Furthermore, in our corpus studies with English and German film reviews, we found a very clear correspondence between logical document structure (breakup in headers, lines, paragraphs) and content structure: Units playing a distinct functional role for the review are very likely to be separated in the logical structure as well. This lead us to the goal of automatically identifying the content structure of such documents. Our underlying application is automatic summarization: Identifying the zones of the film review is a prerequisite for ensuring that the summary contains information from all relevant zones (e.g., movie title, director, description of story, author's evaluation).

Following Stegert (1993), we distinguish between *formal* and *functional* elements of reviews, with the former being 'constituents' whose presence is characteristic for the genre, and the latter making contributions to the communicative goal of the author. The formal zones follow conventionalized patterns of shape and of linear order. They include the title, the name of the reviewer, list of cast, copyright notice, etc. As for the communicative goal of a film review, it is typically twofold: inform the reader about the contents of the film, and provide a subjective evaluation. The running-text paragraphs of a review belong to these two *functional* zones, and our initial corpus study had revealed that they are almost always confined to paragraphs: Authors very rarely mix description and opinion within a paragraph in their reviews. In the following, we discuss related work, then explain our approach to identifying formal zones, and finally turn to opinion classification.

Corpus The basis of our current implementation is a corpus consisting of 213 German film reviews from 7 different web sites. The reviews contain a total of 4,252 paragraphs, i.e., zones that we aim to identify.

[1]The research reported in this paper was funded by Bundesministerium für Bildung und Forschung, grant 03WKH22.

Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 75–78,
Antwerp, September 2007.

2 Related Work

The genre of film reviews has become relatively popular in computational linguistics, but the problem addressed is typically that of classifying an entire review as either positive or negative (e.g. Chaovalit and Zhou (2005)). Our work in effect takes a significant further step: We first break down the review into its various content zones, and then see opinion classification only as one subproblem, pertaining to a subset of the paragraphs.

The subtask of opinion identification has received much attention in recent years. Subjectivity in natural language encompasses a range of different phenomena, including the means to express opinions, emotions, or evaluations. Example applications are automatic classification of opinion texts (e.g. editorials) vs. factual texts (e.g. business texts or news) (Wiebe et al., 2004) or positive vs. negative ratings in reviews (Turney, 2002; Pang et al., 2002; Zhuang et al., 2006). The classification is applied to documents (e.g., Wiebe et al. (2004)) or sentences (Yu and Hatzivassiloglou, 2003).

In contrast to the above approaches, which are exclusively developed for English, we aim at learning subjectivity clues for German data. Moreover, in our classification task, paragraphs rather than documents or sentences are being classified.

Tag	Description
<audience-restriction>	Age restrictions for viewing (in the U.S.: MPAA rating)
<author>	Author of review
<author_place>	Author of review and source of publication
<author_rating>	Author of review and overall rating
<cast>	List of actors, possibly with their roles
<credits>	Credits (Producer, Camera, etc.)
<country_year>	Country and year of production
<date>	Date of review
<director>	Director of film
<format>	Technical format of film (16:9, 4:3, PAL, black/white, etc.)
<genre>	Genre of film (Comedy, Thriller, Documentary, etc.)
<language>	Language of film
<language-subtitles>	Language of subtitles
<legal-notice>	Copyright statement for review
<note>	Various meta-notes (e.g., review has been published earlier at different source)
<quote>	Quotation taken from film or other source
<rating>	Overall rating (5 stars, etc.)
<runtime>	Length of film
<show-loc_date>	Screening locations and dates
<structure>	Explicitly-structuring element, usually a single-word headline
<tagline>	Very short “grabbing” headline
<title>	Title of film
<DATA>	Mixed information, enumerated (credits, cast, etc.)
<dvd-DATA>	DVD release information

Table 1: Tag set for formal zones

3 Formal zones

The inventory of formal zones we determined in the corpus study is shown in Table 1. Recall that we are tagging zones paragraph-wise, which is warranted by the aforementioned relatively clean layout–function correspondence in the genre; at the same time, this decision leads to the occasional need for zones that combine different information. We thus found that `author` is often given together with the `place` of publication, and often with his or her overall `rating` for the film. The other frequent case of “mixing” information are enumerations of cast and contributors (credits); for these, we use the tag `DATA`, which also has a variant for DVD-related information (see bottom of the table).

Our corpus for evaluation (see below) contains a total of 1,156 zones. Zones that occur most often are `DATA` (which make up 18% of all zones), `title` (16%) and `structure` (15%). The zones that are highly relevant for text summarization certainly include the `title` zone, but also zones that are considerably less frequent, like `director` (3%), `rating` (0.4%) or `author_rating` (1%).

3.1 Identifying formal zones

After hand-annotating portions of our corpus, we inspected the various instances of the formal zones and found that they display striking formal characteristics that can quite well be captured in regular expressions. A very simple case is `legal-notice`, which invariably contains the copyright symbol or the word itself. Less simple yet tractable is a zone like `author`, since person names can be recognized by the number of words, capital letters, optional middle initials. Also, information about the position of the text span plays an important role here: the author is always given toward the beginning or

the end of the text. The same holds for `title`, which in addition regularly occurs in neighbourhood to `author` (but the order can vary). What we are *not* exploiting for the time being is layout information such as HTML tags of the original documents. Instead, we convert all input to plain text, and thus our approach operates in the same way for both internet and newspaper material.

Given the observations on regularities in the formal zones, we decided to follow a symbolic approach for them, i.e., we wrote recognition rules encoding features like the ones just mentioned. As a convenient tool for this purpose, we used LAPIS (Miller, 2002), a toolbox for "lightweight text processing". The data set for developing these rules (i.e., for first taking inspiration and then fine-tuning the rules), consisted of 101 film reviews. The evaluation was then performed on a set of 112 unseen reviews.

3.2 Evaluation

The symbolic rules perform excellently on the zones `rating`, `author_rating`, `audience-restriction` and `format` (all with 100% precision and 100% recall). Results for other zones relevant for summarization are: `title` (P: 61%, R: 65%), `director` (P: 42%, R: 78%). Average performance of the rules is 70% precision and 63% recall.

An error analysis of the automatic `title` zone classifications reveals that zones that erroneously get classified as `title` are `DATA` (33% of the misclassifications), `tagline` (25%), and `structure` (17%). On the other hand, `title` is often misclassified as `tagline` (53%) or `director` (15% — this happens with 2-words film titles like *Brokeback Mountain*). Very often, indeed, none of the rules matched a `title` zone, and the rules did not come up with a classification at all (28%). To overcome such problems, we are currently adding a post-processing step that reconsiders all the tag assignments in the light of the overall situation — in this step we can use non-local information like the corpus observations that `author` or `title` (as a single text span) appears at least once in the document but no more than twice (see Section 5).

4 Functional zones

Functional zones are paragraphs with free text. We distinguish two main types of functional zones: descriptive zones (`describe`) and comment zones (`comment`). Descriptions are paragraphs that describe the story, different aspects or peculiarities of the film, without commenting about it. They therefore can be considered as 'objective' information. In contrast, comment zones are paragraphs that contain expressions of opinions by the author, i.e., 'subjective' information. In our application (text summarization), it is very important to be able to reliably distinguish between the two types.In our data, there are slightly more `comment` paragraphs (54%) than `describe` paragraphs (46%).

4.1 Identifying functional zones

Feature set For classifying the functional zones, we used as training features a bag-of-"words" approach. In a detailed evaluation of tf*idf measures used as relevance weights, we found that 5-grams perform best for German data, so our bag of "words" consists of weighted character 5-grams. All 5-grams occurring in the paragraph that is to be classified are weighted according to the $tf * idf$ measure, where tf is the frequency of the 5-gram in the paragraph, and idf is the inverse document (i.e., paragraph) frequency according to a reference corpus: a large collection of internet film reviews.

Training procedure Pang et al. (2002) compare different machine learning methods and achieved accuracies between 72.8% and 82.9%, depending on the training features and the method. In their evaluation, Support Vector Machines (SVM) perform best for many of the feature combinations.

In our approach, we also use SVM. Our feature sets, however, do not consist of words or POS tags but 5-grams. We used the tool SVMLight (Joachims, 1999) and performed a threefold-crossvalidation on the 213 reviews, which contain 1,159 functional zones..

4.2 Evaluation

The table below presents the results from the functional zone classification. Overall accuracy is quite satisfactory, at 79.34%. `Comment` zones are classified more successfully than `describe` zones.

Zone type	Precision	Recall	Accuracy
`comment`	81.60%	79.69%	79.34%
`describe`	76.83%	78.94%	

5 Conclusion and outlook

For many applications, including summarization, but also question–answering and others, the range of portions and their relative relevance for the application heavily depends on the *genre*. For the example discussed here, film reviews, it is evident that information about the *content structure* of a document can be of immense help for creating a balanced summary, for choosing zones in which the answer to a question is sought, etc.

Based on a corpus study, we have developed an inventory of zone labels for the genre *film review* and implemented a system for automatically identifying these zones, i.e., for breaking up a document into its content structure. The precision currently ranges from 70% for formal zones to 79% for the two functional zones. Our approach is hybrid: it utilizes both symbolic rules and a statistical classifier. The overall algorithm first decides heuristically whether to invoke the symbolic rules or the classifier (the functional zones are longer-text paragraphs that occur in the middle of the document and are not interrupted by formal zones), and then each paragraph of the document receives its label by either module. Recognition is based on merely local information so far.

Our current work aims at improving the results by taking two different routes. For one thing, we are integrating layout information, in particular HTML tags, into the identification of formal zones. To that end, the input to the system will no longer be plain text but a canonical, XML-based representation of the logical document structure, which is produced from HTML. The other line is to make more extensive use of knowledge about zone neighbourhood. To this end, we are revising the rules for formal zones so that they output probabilistic judgements, and these will be combined with a trigram model capturing the zone sequences in our corpus. Thus, all information about zone locations will be removed from the rules and incorporated into a single, separate knowledge source. Finally, we are currently adapting our implemented text summarizer (Stede et al., 2006) to utilize the zone information so that the quality of summaries for the particular genre of film reviews will be improved considerably.

References

P. Chaovalit and L. Zhou. 2005. Movie review mining: A comparison between supervised and unsupervised classification approaches. In *Proc. of the 38th Hawaii Int'l Conference on System Sciences.*

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*. MIT Press.

Robert C. Miller. 2002. *Lightweight Structure in Text*. Ph.D. thesis, Computer Science Department, School of Computer Science, Carnegie Mellon University.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. of EMNLP-02*, pages 79–86.

M. Stede, H. Bieler, S. Dipper, and A. Suryiawongkul. 2006. Summar: Combining linguistics and statistics for text summarization. In *Proc. of ECAI-06*, Riva del Garda.

G. Stegert. 1993. *Filme rezensieren in Presse, Radio und Fernsehen*. München: TR-Verlagsunion.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. of the ACL-02*, pages 417–424.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proc. of EMNLP-03*.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proc. of the 15th ACM international conference on Information and knowledge management*.