Evaluating Cub Reporter: proposals for extrinsic evaluation of journalists using language technologies to access a news archive in background research

Emma Barker and Robert Gaizauskas
Department of Computer Science, University of Sheffield

Abstract

We argue that defining an extrinsic evaluation in which systems are evaluated by how well they help a user perform a specific task is key to assessing the value of novel information access technologies. In the case of a journalist writing a background story this presupposes some way to measure the *quality* of background. We propose two approaches to assessing background quality and outline an evaluation methodology for each. The first approach relies on treating professional journalists as oracles and determining whether their professional judgements about quality concur. The other approach relies on developing a descriptive theory of the discourse level relations holding within and between background and foreground and demonstrating that predictions about quality made on the basis of nature of the discourse relations found in a given text correlates with journalists' subjective judgements of quality. A pilot study carried out to investigate the first approach, while revealing some issues with the methodology, suggests that the approach is feasible and that journalists' judgements of quality in backgrounders do indeed agree.

1 Introduction

New information access technologies – question answering, information extraction, summarisation – are being developed in the belief that they will offer better, or differently useful, access to digital text resources than conventional text search technologies. Research and development in these areas has been significantly influenced by open technology evaluation challenges set by DARPA and NIST in the US. Such challenges include the Text Retrieval Conferences (TREC), the Message Understanding Conferences (MUC) and the Document Understanding Conferences (see, e.g. [1, 2, 3]). These challenges are designed to stimulate technology R & D and as such provide excellent benchmarks for specific technologies, whose capabilities are broadly assumed to be of use in real settings. However, these evaluations do not involve real users in real settings. Furthermore they do not allow us to compare the new technologies directly with existing technologies because the outputs of the new technologies are different from those of existing technologies, and measures obtained in these evaluation exercises are not cross-comparable. How, for example, can one compare precision and recall figures for ad hoc information retrieval with mean reciprocal rank figures for question answering?

Given that outputs and measures differ for different information access technologies, real insight into their relative utility and acceptable levels of performance can only be gained by evaluating them in a context of intended use, with real users. This involves not only all the difficulties of working with human subjects, but also the methodological difficulties of determining how to evaluate the output of a *setup* [4], the combination of user + system working together to achieve a purpose.

In this paper we address this challenge in one particular scenario – that of news agency journalists in researching and writing background material to support breaking news wire stories, using a digital text news archive as an information source. The Electronic Cub Reporter is a

research project which aims to investigate the use of language technologies in this scenario. The project conjectures that recent developments in language technology in the areas of information extraction, question answering, and multi-document summarisation should be of relevance to this scenario.

To evaluate new versus conventional technologies in this scenario requires a method for comparing the setup consisting of user + system A with that of user + system B, i.e. one needs an *extrinsic* evaluation of the embedded systems. While various criteria can be proposed for this evaluation, a central one must be the *quality* of the background stories resulting from the setup (other criteria such as speed or user satisfaction are also significant, but only if one can guarantee that quality is at least preserved). Thus a key objective in a programme to carry out extrinsic evaluation of information access technologies for Cub Reporter is the establishment of method for assessing the quality of background stories. In the following, after providing more background on the scenario, we present the hypotheses we have formulated regarding assessment of background story quality and describe the methodology we propose for investigating them, including a pilot study already undertaken. While this work is more about evaluation of the output of users working with information access technology than of users interacting with such technology, we believe it lays essential foundations for understanding and comparing different sorts of information access technology, not just for the scenario we have chosen to investigate, but for any information gathering scenario whose output is a written information artifact.

2 Task, Current Practice and Proposed System

While there has been some academic work studying the information seeking behaviour of journalists in general, there has been no prior work, so far as we aware, specifically on gathering background for news as a task in its own right. Attfield and Dowell [5] propose a general model of journalistic information gathering and use in the context of the task of writing news. However, the backgrounder task is somewhat different from other news writing tasks and their model needs to be specialised for it. There is, for example, a paramount need for speed and the discovery of possible "angles". Crucially, their model does not address the specific sorts of information content typically required in writing background for a given foreground story, nor how the requirement for this sort of information influences information seeking.

Our work to date has involved the study of journalists who either work for or with materials produced by the Press Association, the major UK domestic newswire service which provides copy to all major national daily newspapers. For breaking news stories, in contrast to say feature stories, the news cycle unfolds as follows: when breaking news is received a journalist writes one or two sentences summarising it and then passes this text, called a snap to a sub-editor for checking. When satisfied, the sub-editor "moves" the copy on the wire, marking the first instalment of a new story. The story is then published as a series of instalments, where each instalment contains a new and updated account of the news. While background figures in a number of ways, including simple descriptive phrases interjected into the current story (e.g. *former Chancellor of the Exchequer*) and fact sheets listing similar or relevant occurrences (e.g. a listing of previous train crashes), we shall focus on the most significant form of background material only, the so-called "backgrounder". Backgrounders may be regarded as stories in their own right, i.e. they typically have a simple narrative structure comprising elements such as an abstract or introduction, a central argument or "angle" (the point of the story) and propositions which support or elaborate upon the argument. Thus they may be easily distinguished from the list format of a fact-sheet. A background is typically written when a news editor deems a particular story worthy of dedicated background material. They are usually not released till sometime after the first instalments as time is needed both to determine whether a story merits a backgrounder, but also for the research to be carried out to assemble the material. Their function is not to continue to report details of new events, but rather to provide text that supports and contextualises these events.

Our analysis of background information in news stories has revealed that background is typically made up from the following four types of materials: (1) accounts of similar events in the

past (e.g. other train crashes, scandals of similar nature, etc.); (2) accounts of events which have lead up to the current event (e.g. a chronology of company takeovers, store openings, price cuts and profit warnings in the months leading up to a supermarket's announcement of low annual profits); (3) profiles of persons or organisations or locations (usually role players in the new event) comprising some highly structured factual information about the role player, for example date and place of birth, career appointments; spouse etc; accounts of the role player in events leading up to the event and accounts of the role player in similar events to the current event; and (4) comment (quotes) on any of the preceding by notable individuals.

Currently the PA and journalists with access to the PA archive access the archive via a conventional free text search, i.e information retrieval (IR), system. To complete the background writing task the journalist searches the archive and possibly other sources, using his world knowledge in addition to the information given about the new event to guide his search. The system the PA currently use allows boolean queries to be formed from single words or phrases. Searches may be restricted to story text, headlines or byline and they may be date range restricted as well. Results may be requested sorted by weighting (a relevance ranking of some sort) or date.

While IR is the workhorse of the news world, other language technologies whose aim is to facilitate information access would also appear to be of relevance. Question answering (QA), multi-document summarisation (MDS) and information extraction (IE) all have potential for the backgrounder scenario. Of particular relevance are the *Who is ...?* questions figuring in the definition component of the TREC QA track [6] and the entity- and event-focussed MDS tasks within DUC [3].

To explore the space of possible systems ranging from primitive document retrieval to fully automated backgrounder production we are developing a platform which will permit us to configure various systems from different technologies. Work on the project to date has involved the construction of a prototype which incorporates a standard information retrieval engine, an existing question answering system and an existing multi-document system. We are also building on existing information extraction technology to perform shallow syntactic and semantic analysis of texts to extract representations of key events plus participants, so as to support a "search for similar events" capability. These information access technologies are being embedded in a browser-based graphical user interface which will allow users to combine them flexibly, as well as to save documents found while accessing the archive and to assemble a new document from retrieved materials.

Given this research platform, how can we establish which configuration is best for the task?

3 Research Hypotheses

In the introduction we argued that existing intrinsic evaluations of language technologies are insufficient to enable us to answer the deeper question of whether these technologies, at their current level of development, are of use to a journalist in the task of writing a backgrounder for a breaking news story. To establish this, we need an extrinsic evaluation which allows us to evaluate indirectly one system A against another system B by showing that setup 1 consisting of user + system A is superior another setup 2 consisting of user + system B, assuming both setups to have the same purpose and to be given the same input. Thus, if a journalist using existing tools produces more timely and accurate coverage of background material for a breaking news story (and hence more adequately fulfills the purpose of the setup) than a journalist using novel language technologies, then we have an extrinsic evaluation which demonstrates the inferiority of the language technologies

For such an extrinsic evaluation to be possible, however, we need to operationalise the criteria implicit in the comparison, namely speed and quality of output, in terms of observable measures. This is easy enough for speed – simply measure elapsed time in production of backgrounders. However, measuring the quality of backgrounders is not so straightforward. For an extrinsic evaluation relying on quality assessment to be possible at all it must be the case that quality of backgrounders is something that can be objectively assessed. This is an empirical question the

answer to which might be established in various ways. We might suppose that there is a set of quality criteria for backgrounders which could (1) be agreed by journalists and (2) be consistently applied. However, we do not want to assume that such criteria can be written down or agreed in the abstract. Rather we need only determine whether experts can form quality judgements about which they concur; i.e. we treat them as oracles and determine if they agree.

This leads to the first research hypothesis in our programme:

> **Hypothesis 1:** *Independent expert assessors can (1) consistently rank backgrounders according to quality; (2) consistently categorise backgrounders according to quality.*

The first part of this hypothesis addresses the question of whether experts have a shared notion of *relative* quality; the second allows us to see whether they have a shared notion of *categorical quality*, e.g. very good, fair, poor, etc. – this would allow not merely the ranking of systems, but some notion of distance between them. Note that in neither case are we proposing or even supposing the existence of a shared definition of quality.

Supposing this hypothesis in either or both forms were proved to be true. This would be sufficient to allow us to carry out an extrinsic evaluation of two or more systems in the background setup: that system which consistently lead to higher quality backgrounders being written (controlling for topic and writer) would be superior. While useful such an evaluation would not be terribly informative. This is because it gives a quality assessment to a backgrounder overall and does not give any insight into the value assigned to components of the backgrounder. Such finer-grained assignment of credit is necessary to gain insight into whether information access system is delivering appropriate material.

How could a finer-grained evaluation be designed? There are numerous possibilities. One question to be addressed is the grain-size of the elements whose contribution to the backgrounder to to be assessed. One could, for example, ask journalists to assess the value of each sentence in a backgrounder; or one could choose content units based on some theory of functional or rhetorical structure of background text. A second question is what judgement they will be asked to make about each unit. For example, an attempt could be made, in the spirit of assessing information "nuggets" in the TREC QA track definition task, to get journalists to classify each sentence in a backgrounder as essential, optional or irrelevant (coarser or finer classifications could be imagined). This proposal invites the question of whether it makes sense to treat the assessment of content units independently. For example, the inclusion of further examples of prior similar events may not, beyond some point, add anything to a backgrounder, especially if they are included at the cost of omitting some other information, such as profile information about a key role player. In other words it may not make sense to assess background content units independently of each other, but only in the context of the overall background piece or with respect to some prescriptive theory about the content of backgrounders. Further, if judgements about content units are made independently then there is the additional question of how these judgements are to be combined to arrive at an overall quality score for a backgrounder or a comparative judgement between two backgrounders.

An alternative proposal would be to articulate a theory of background based on the semantic relation of content units in the background in relation to the foreground – for example we indicated above that background material can be classified as reporting events similar to the current event, events leading up to the current event, profiles of participants in the current event, or comment by significant persons on the foregoing. Such a theory could have both a descriptive and a prescriptive component. The descriptive component would provide a set of categories for characterising the relation between content units in the background and the foreground and perhaps also between units within the background; the prescriptive component would offer suggestions as to the appropriate mix of background content types and perhaps be dependent on the type of foreground event and on some notion of what constitutes "common knowledge" at a given time.

Our view, informed by interviews with journalists, observations of journalists carrying out background writing exercises and an analysis of archive backgrounders, is that the hypothesis that content units in backgrounders could be independently assessed and an overall quality assessment composed from these unit assessments in such a fashion that it correlated well with global quality

assessments of the sort proposed in hypothesis 1 is unlikely. Hence we intend to explore the alternative just advanced and therefore propose the following hypothesis:

> **Hypothesis 2:** *A theory of discourse-level semantic relations between propositions in background and foreground news can be developed such that (1) independent human annotators can consistently mark these relations between propositions in background and foreground news; (2) quality criteria for backgrounders can be stated in terms of the presence, absence and balance of propositions in the background standing in these specific semantic relations to each other and to the foreground* [1]*; (3) judgements made in terms of these quality criteria will correlate positively with global quality judgements whose existence is asserted by hypothesis 1.*

Were hypothesis 2 be proved true a number of interesting consequences would follow. First we would have a an explanatory theory of why one backgrounder was better than another. Secondly we would have a means to determine quality of a backgrounder that did not depend on the expert judgement of a journalist. Thirdly we would have a means to assess the output of an information access system in relation to the background task.

4 Methodology

To investigate the research hypotheses developed in the preceding section an experimental programme needs to be defined and carried out. In this section we discuss this programme and preliminary work carried out to date.

4.1 Data: Assembling Backgrounders for Evaluation

We begin with the presumption that to test our hypotheses about the assessment of background quality we require a set of backgrounders written in support of the same news story and of comparable word length.

Newswire archives are not good sources for multiple backgrounds to the same story since while a newswire service might produce more than one background for a particular story, there are typically important differences in content. One might focus on the background to the event, another on a role player in the event. Thus, they are not suitable for comparison. Another source of backgrounders is background stories written by journalism students. Student backgrounds offer the advantages of being written on the same topic and under similar constraints of resources, time, word length etc. It is also easy to obtain relatively large numbers of them. On the other hand, a limitation of student backgrounds is that they are, of course, not written by 'real users', experts in writing news stories, using systems in a real workplace and in response to professional demands, e.g. deadlines. A further option is to ask a group of professional news agency journalists to each write a background to the same news story under controlled production constraints. Although this method eliminates the problems associated with non-experts it is costly, since backgrounds take about an hour to produce using current technology, and news agency journalists are unlikely to donate time sufficient to create a corpus of adequate size for evaluation purposes.

In the light of these considerations we decided to use a collection of student assignments, since they were immediately available, for a pilot study, described below. Based on the experience gained in this study our current plan is to construct a more tightly controlled corpus, including a range of background types (e.g. natural disasters, political resignations, etc.), using professional journalists.

The student corpus used in the pilot study consisted of the two sets of stories written by different year cohorts to contextualise a news story about urban regeneration in the City of Sheffield. Cohort A comprised 15 backgrounds and Cohort B comprised 17 backgrounds.

[1] Note that we presume the factual accuracy of propositions in the background here. Clearly factual inaccuracy would negatively affect quality.

Bground id	j1	j2	j3	|j1 - j2|	|j2 - j3|	|j1- j3|	total pd	mean pd
101	fair	poor	fair	1	1	0	2	0.67
102	good	very good	good	1	1	0	2	0.67
103	fair	fair	fair	0	0	0	0	0.00
104	poor	poor	poor	0	0	0	0	0.00
105	poor	good	poor	1	1	0	2	0.67
106	good	very good	fair	1	2	1	4	1.33
107	very poor	poor	poor	1	0	1	2	0.67
108	very good	good	good	1	0	1	2	0.67
109	very good	good	fair	1	1	2	4	1.33
110	poor	good	very poor	2	3	1	6	2.00
111	fair	good	good	1	0	1	2	0.67
112	fair	good	good	1	0	1	2	0.67
113	poor	fair	very poor	1	2	1	4	1.33
114	fair	fair	poor	0	1	1	2	0.67
115	fair	fair	very poor	0	2	2	4	1.33
							Overall mean pd	0.84

Table 1: Categorical assessments of backgrounders with pairwise differences

4.2 Investigating Hypothesis 1

To investigate the hypothesis that journalists can consistently rank and categorise a set of backgrounders with respect to quality we need to get a number of journalists to rank a set of backgrounders. For reasons of availability and in order to uncover problems in our experimental approach we decided to first carry out a pilot study using the student data described above and academic journalists as judges. This study and some initial results are described below. A further study building on the pilot study using professional journalists as authors and judges is planned.

4.2.1 Pilot Study: Methodology

We asked three independent evaluators, members of the Sheffield University Department of Journalism with experience as practising journalists, to judge two sets of backgrounds in terms of their respective quality.

First we gave the participants the foreground story for which the students had been asked to write a background. Then we asked them to arrange each set of backgrounds into a ranked list according to their quality as background for the given foreground story. Here the judges were not asked to indicate any notion of absolute quality for an individual background. They were allowed up to a period of one and a half hours to complete the task. Next, in order to acquire some idea of the scale of the perceived differences in quality between any pair of backgrounds in the ranked list, we asked the judges to place each background into one of five categories: *very poor*, *poor*, *fair*, *good* and *very good*. They were given an hour to complete this task, but having previously ranked the documents the participants were able to complete this task in less than 30 minutes. We then investigated levels of agreement between judges by comparing the results. We did not provide a definition of the five categories beyond the labels given above, nor did we list any criteria that had to be present for a particular grade. Thus the judgements were subjective and we expected the results to reveal possible differences between journalists in their interpretation of categories of quality. The exercise finished with the judges being asked to comment on the assessment experience.

4.2.2 Pilot Study: Initial Results

We have yet to apply appropriate statistical tests, such as Spearman's rho, to the ranked lists in order to ascertain to the degree of agreement between the judges' rankings.

Our results from the categorisation evaluation for cohort A[2], shown in Table 1, consist of three categorical judgements $j1, j2, j3$ for each background. A simple measure of agreement between these judgements can be calculated as follows. First we calculate the pairwise differences between any two judgements as the distance the categories are apart on the 5-point scale. For example, if $j1$ is *very good*, $j2$ is *fair* and $j3$ is *good* we have the pairwise differences $|j1 - j2| = 2$, $|j1 - j3| = 1$ and $|j2 - j3| = 1$. These differences are then used to calculate the total and mean pairwise difference for each background.

For three independent judgements on a 5-point scale the maximum possible sum of pairwise differences ("total pairwise difference") is 8, thus the maximum mean pairwise difference possible is 2.67, which indicates greatest divergence in judgements. If there are no differences in judgements, i.e. we have total agreement, the mean pairwise difference score is 0. We can interpret scores of <1 as indicating small differences in judgement, as this shows at least 2 judges assigned the same category. The results in table 1 suggest that there is reasonably high agreement between judges, the mean pairwise difference for cohort A being 0.84. The figures also show there was only strong disagreement for one of the 15 backgrounds, with a mean pairwise difference of 2.

4.2.3 Discussion

While initial results from the pilot study show strong agreement between judges they also revealed weaknesses in the data used in the experiment which need to be addressed before carrying out any further study. In particular significant differences in the rhetorical form and content of some of the backgrounders made it difficult for the judges to feel confident in all of their judgements. Two judges complained that they were not comparing like with like. For example, a number of the backgrounders adopted a "microcosm" approach, where one past event is explored in some detail and used to throw light on the current event. Other backgrounds, however, presented wider coverage of past events relevant to the current event. The judges indicated that the microcosm approach is not typical for news agency backgrounders, where one would instead expect to see the wider coverage approach. This suggests that in a future evaluation we should be careful to control the instructions to participants writing the background: it should be in the style of a newswire background, perhaps with examples to illustrate typical backgrounds.

The judges also noted that the authors were students who did not have professional experience of background writing for newswires and that consequently the background cohorts were susceptible to significant digressions from standard journalistic style. Some of the texts were judged poor because they did not ask the right questions or revealed that the students simply did not understand what was required in a news backgrounder. In comparing work of professionals we would expect to see narrower differences in the quality of backgrounders. As a consequence it may be more difficult for judges to assess the differences in quality between professional backgrounders.

4.3 Investigating Hypothesis 2

In order to explore hypotheses 2 we require a theory of background news based on the semantic relations of content units in a background to a foreground story, relations that allow a text to fulfill the function of a backgrounder. Bell [7]and van Dijk [8] have addressed the semantic structure of news texts and propose that they are hierarchically structured, with background figuring as part of the structure. However, neither of these authors addresses background in news in detail, though van Dijk notes that background is a complex category, implying it merits further treatment.

For background discourse a set of relations is needed to characterise the functional relation between propositions in the background and events in the foreground, relations which indicate what role a set of background propositions is playing in the backgrounder. For example, in response to a train crash a backgrounder will probably mention previous train crashes. The background discourse segment describing the previous train crash stands in the relation of "similar event" to the foreground story. While we speculate that the distribution of these relations between the background and foreground text contributes in some way to the objective quality of a backgrounder,

[2]We are awaiting judgements from the final judge for cohort B.

they alone are unlikely to be sufficient to allow one to predict the quality judgements that would be made by independent judges. Such an approach does not account for the arrangement of the propositions *within* the background text, i.e. the structuring of propositions and the relations that hold between them. These relations clearly play a role in assessments of quality. For example, a factsheet containing a simple list of similar train crashes is likely to be judged of lower quality as a backgrounder than a text in which several different generalisations are made about the class of previous train crashes, with details of individual train crashes used to provide evidence for these generalisations. Thus, we propose that in order to assess the richness of the information content in a backgrounder a descriptive framework is required that captures both discourse level relations between background and foreground texts and between discourse units within the background text.

Much work has been devoted to the analysis of coherence relations between discourse units in text (see, e.g., [9], [10], [11]). While there is disagreement about whether the structures be represented as graphs or trees, and on what comprises a discourse unit in text and differences between semantic accounts of informational relations and intensional relations, there is compatibility at least between the different taxonomies for semantic relations (differences are usually in terms of granularity). Wolf and Gibson [11] have shown that it is possible to specify a simple set of coherence relations for discourse segments that are easy to code. These include generalisation (where one discourse segment states a generalisation of what is stated by another discourse statement), cause and effect, temporal sequence, condition, elaboration, and so on.

We propose to adopt a framework such at that of Wolf and Gibson, perhaps further simplified, to analyse discourse level relations within and between background and foreground texts in order to investigate hypothesis 2. Given such a descriptive framework the next step is to invite human judges (non-journalists) to annotate a corpus of background and foreground texts, according to this framework. Since we will already have quality judgements for the corpus created for investigating hypothesis 1, in relation to which we will want to compare the annotations, we will use this corpus as the corpus for annotation. An iterative procedure of refining the framework and annotation guidelines and testing for inter-annotator agreement between human judges will be followed until an acceptably high level of agreement can be obtained.

Once the background corpus has been acceptably annotated with the discourse relations we can then move to investigating whether a correlation between the discourse annotations and the quality judgements made in the investigation of hypothesis 1 can be established. If it can hypothesis 2 will be established; if not then either the hypothesis is false, our descriptive framework is inadequate, or our data is flawed.

5 Conclusion

In the foregoing we have argued that defining an extrinsic evaluation in which systems are evaluated by how well they help a user perform a task is key to assessing the value of novel information access technologies. Assessing how well information access technologies help in the specific task of a journalist writing a background story, however, presupposes some way to measure quality of background. In this paper we have proposed two approaches to assessing background quality and outlined a methodology for carrying them out. One approach relies on treating professional journalists as oracles and determining whether their professional, unarticulated judgements about quality, both relative and categorical, concur. The other, more speculative, but if successful more informative, approach relies upon developing a descriptive theory of the discourse level relations holding within and between background and foreground and demonstrating that predictions about quality made on the basis of nature of the discourse relations found in a given text correlates with journalists' subjective judgements of quality. A pilot study carried out to investigate the first approach, while revealing some issues with the methodology, suggests that the approach is feasible and that journalists' judgements of quality in backgrounders do indeed agree. A more controlled study in support of the first approach as well as the work required to realise the second approach now needs to be carried out.

Acknowledgements

The authors would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council, research grant: R91465.

References

[1] Voorhees, E.: Overview of TREC 2003. In: Proceedings of the Twelfth Text Retrieval Conference (TREC 2003). NIST Special Publication 500-255 (2004) Available at: http://trec.nist.gov/pubs/trec12/papers/OVERVIEW.12.pdf.

[2] Defense Advanced Research Projects Agency: Proceedings of the Seventh Message Understanding Conference (MUC-7), Defense Advanced Research Projects Agency (1998) Available at: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/.

[3] Over, P., Yen, J.: Introduction to DUC-2004: An intrinsic evaluation of generic news text summarization systems. In: Proceedings of the HLT/NAACL 2004 Document Understanding Workshop (DUC-2004). (2004) Available at: http://www-nlpir.nist.gov/projects/duc/pubs/2004slides/duc2004.intro.pdf.

[4] Sparck Jones, K., Galliers, J.R.: Evaluating Natural Language Processing Systems. Springer, Berlin (1996)

[5] Attfield, S., Dowell, J.: Information seeking and use by newspaper journalists. Journal of Documentation **59** (2003) 187–204

[6] Voorhees, E.: Overview of the TREC 2003 question answering track. In: Proceedings of the Twelfth Text Retrieval Conference (TREC 2003). NIST Special Publication 500-255 (2004) Available at: http://trec.nist.gov/pubs/trec12/papers/QA.OVERVIEW.pdf.

[7] Bell, A.: The Language of News Media. Blackwell, Oxford (1991)

[8] van Dijk, T.: Structures of news in the press. In van Dijk, T., ed.: Discourse and Communication. De Gruyter, Berlin (1985) 69–93

[9] Mann, W., Thompson, S.: Rhetorical structure theory: Toward a functional theory of text organization. Text **8** (1988) 243–281

[10] Marcu, D.: The Theory and Practice of Discourse Parsing and Summarization. MIT Press, Cambridge, Mass (2000)

[11] Wolf, F., Gibson, E.: A response to Marcu (2003). Discourse structure: trees or graphs? (2004) Available at: http://web.mit.edu/fwolf/www/discourse-annotation/Wolf_Gibson-coherence-representation.pdf.