Identifying the Discourse Function of News Article Paragraphs

W. Victor H. Yarlott[1], Cristina Cornelio[2], Tian Gao[2], Mark A. Finlayson[1]

[1]School of Computing and Information Sciences
Florida International University, Miami, FL 33199
wvyar@cs.fiu.edu, markaf@fiu.edu

[2]IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
{ccornel, tgao}@us.ibm.com

Abstract

Discourse structure is a key aspect of all forms of text, providing valuable information both to humans and machines. We applied the hierarchical theory of news discourse developed by van Dijk (1988) to examine how paragraphs operate as units of discourse structure within news articles—what we refer to here as *document-level discourse*. This document-level discourse provides a characterization of the content of each paragraph that describes its relation to the events presented in the article (such as *main events*, *backgrounds*, and *consequences*) as well as to other components of the story (such as *commentary* and *evaluation*). The purpose of a news discourse section is of great utility to story understanding as it affects both the importance and temporal order of items introduced in the text—therefore, if we know the news discourse purpose for different sections, we should be able to better rank events for their importance and better construct timelines. We test two hypotheses: first, that people can reliably annotate news articles with van Dijk's theory; second, that we can reliably predict these labels using machine learning. We show that people have a high degree of agreement with each other when annotating the theory ($F_1 > 0.8$, $\kappa > 0.6$), demonstrating that it can be both learned and reliably applied by human annotators. Additionally, we demonstrate first steps toward machine learning of the theory, achieving a performance of $F_1 = 0.54$, which is 65% of human performance. Moreover, we have generated a gold-standard, adjudicated corpus of 50 documents for document-level discourse annotation based on the ACE Phase 2 corpus (NIST, 2002).

1 Introduction

Discourse structure is a key aspect of all forms of text, providing valuable information about the contents of a given span of text. This is most obvious in academic, legal, and technical texts, which are often clearly delineated into sections containing, for example, introductory, background, or explanatory material, among others—these type of texts are designed to make it easy to find specific information within them quickly. News articles have a similarly helpful, though implicit, design: they often provide a brief, up-front summary of the important events, relevant background information, comments from both experts and the reporters, and detailed descriptions of the main events. Events are often not presented in chronological order, but rather structured by importance.

We use an established hierarchical theory of news discourse (van Dijk, 1988) to model how paragraphs operate as units of discourse structure within news articles to capture the importance of events within a story. We test two hypotheses: first, that humans can reliably annotate news articles with van Dijk's theory; second, that these discourse labels can be predicted by machine learning.

In our first hypothesis, by *reliable* we specifically mean that independent people agree with each other when applying van Dijk's theory of news discourse. We performed an annotation study to answer this question, producing a small corpus of gold-standard, adjudicated annotations in a standoff format based on the Automated Content Extraction (ACE) Phase 2 corpus (NIST, 2002). This corpus consists of 50 documents (28,236 words; 644 paragraphs) annotated at the paragraph level. Agreement was notable, with $F_1 > 0.75$ and Cohen's $\kappa > 0.60$ (see §4.3 for details). These results show that van Dijk's theory of can be both learned and reliably applied by humans to news article.

Proceedings of the Workshop on Events and Stories in the News, pages 25–33
Santa Fe, New Mexico, USA, August 20, 2018.

To address our second hypothesis, we demonstrate a machine learning approach using support vector machine (SVM) for learning to tag paragraphs with labels from van Dijk's theory. We achieve a performance of $F_1 = 0.54$, which is 65% of human performance. We also demonstrate the performance of other machine learning algorithms (decision tree and random forest) and provide the set of features that perform the best on this task.

The paper is structured as follows. We first introduce some of the existing related work (§2), and then provide a definition of van Dijk's theory as was presented to our annotators (§3). We next describe the selection of texts used in this study, provide corpus statistics, describe the training and annotation procedures for the study, and describe the results of the annotation study and provide some discussion on these results (§4). We then provide an automated method to learn and predict the discourse-structure labels of a plain document (§5), followed by discussion on the results of label prediction and some remarks of possible future directions (§6). Finally, we summarize our contributions (§7).

2 Related Work

There has been a substantial work describing how the structure of news operates with regards to the chronology of real-world events. Much news follows an inverted chronology—called the inverted pyramid (Bell, 1998; Delin, 2000) or relevance ordering (Van Dijk, 1986)—where the most important and typically the most recent events come first. Bell claims that *"news stories... are seldom if ever told in chronological order"* (Bell, 1994, p. 105), which is demonstrated by Rafiee *et al.* for both Western (Dutch) and non-Western (Iranian) news (2017). Rafiee *et al.* also show that many stories follow a hybrid structure, which combines characteristics from both inverted and chronological structures.

In this work, we focus on van Dijk's structural approach to the structure of news discourse (van Dijk, 1988), which is organized as a tree. We choose this work as our focus due to the presentation and description of the schemata, which facilitated the quick development of an annotation guide. A more in-depth description of van Dijk's theory is presented in Section 3.

Discussing van Dijk's theory of news discourse, Bekalu states that analysis of *"the processes involved in the production of news discourses and their structures will ultimately derive their relevance from our insights into the consequences, effects, or functions for readers in different social contexts, which obviously leads us to a consideration of news comprehension"* (2006, p. 150). The theory proposed by van Dijk has also been proposed for use in annotating the global structure of elementary discourse units in Dutch news articles (van der Vliet et al., 2011).

Pan and Kosicki (1993), in a similar analysis, present a framing-based approach that provides four structural dimensions for the analysis of news discourse: syntactic structure, script structure, thematic structure, and rhetorical structure. Of these, the syntactic structure is most closely aligned with van Dijk's theory. In this paper, we chose to focus on van Dijk's theory as Pan and Kosicki do not provide a list or description of the structure that could be readily translated into an annotation scheme.

White (1998) treats the structure of news as being centered around the headline and lead. White suggests that the headline and lead, which act as a combination of both synopsis and abstract for the news story, serve as the nucleus for the rest of the text: *"the body which follows the headline/lead nucleus—acts to specify the meanings presented in the opening headline/lead nucleus through elaboration, contextualisation, explanation, and appraisal"* (1998, p. 275). We focus on van Dijk's theory for this paper as we find it to provide a higher degree of specificity: White's specification modes serve roughly the same purpose as higher-level groupings in van Dijk's theory.

For this work, we use the ACE Phase 2 corpus (NIST, 2002) as the source of our news articles. We choose this corpus because it fit three criteria: it is a widely-used news corpus, it has relevance to other tasks (entity detection and relation detection), and it was readily available to us.

3 Van Dijk's Theory of News Discourse

Van Dijk (1988) provides a hierarchical theory of news discourse, shown in Figure 1, which we apply to a subset of the news articles of the ACE Phase 2 corpus. In this section, we briefly describe the leaf categories as well as their parent categories when appropriate. We provide additional annotation details

for discourse types where van Dijk's description appeared underspecified, as we have done in the guide given to our annotators.

Summary elements express the major subject of the article, with the *headline* being a special construct that introduces a topic, and the *lead* summarizing the topic introduced by the headline. While annotators were initially instructed to annotate the headline, we do not include it in our annotations, as the ACE Phase 2 corpus has the headline separate as part of its annotation scheme.

Situation elements are the actual events that comprise the major subject of the article. *Episodes* concern *main events*, which are those events that directly relate to the major subject of the article, and the *consequences* of those events. The *background* consists of the *context*, which are any *circumstances* that contribute to understanding the subject as well as any *previous events*. Where circumstances may be non-specific, previous events refer to a specific event that has occurred recently. *History* elements are those events that have not occurred recently, typically referenced in terms of years prior, rather than months, weeks, or days. These elements of the discourse structure provide important information about the relation of each paragraph with respect to the central events of a news story.

Conclusions are those *comments* made by the journalistic entity (the newspaper, reporter, etc.) regarding the subject. These can be *expectations* about the resolution or consequences of an event, or *evaluations* of the current situation. In contrast, *verbal reactions* are *comments* solicited from an external source, such as a person involved in the events of the article, an expert, etc. These elements of the discourse provide further supporting context for the central events of an article.

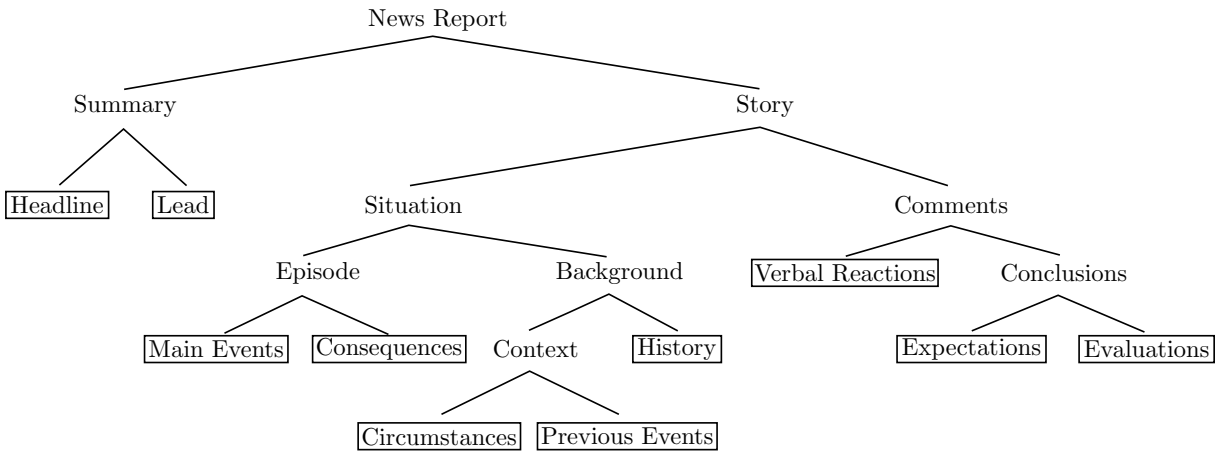

Figure 1: The hierarchical discourse structure of news proposed by van Dijk (van Dijk, 1988).

4 Data & Annotation

One of the major contributions of this paper is the generation of a gold-standard corpus of document-level discourse structure based on the existing ACE Phase 2 corpus. This new dataset comprises 50 documents containing 28,236 words divided in 644 paragraphs. This annotation is, to the best of knowledge, the first of this kind, and provides additional information about the corpus that, until now, is not considered in any knowledge extraction method. We provide, in the following sections, the details of our corpus.

4.1 Selection of Texts

We selected the ACE Phase 2 corpus because it is a major standard corpora of news articles that satisfied three criteria: it is widely-used, has relevance to other tasks, and was readily available to us. We annotated 50 randomly selected news articles from the development set, divided into ten sets of five documents each. Within these sets, documents were swapped or replaced in order to obtain uniform sets in terms of total document lengths. Table 1 shows the corpus-wide statistics for the number of words and paragraphs, where each paragraph is given a single type in accordance to van Dijk's theory. The majority of texts were already divided into paragraphs in an obvious manner, either with empty lines or with indentation.

The remaining texts were divided by the adjudicator based on either contextual or structural clues, such as abrupt change in topic or unnatural line breaks.

	Words	Paragraphs
Total	28,236	644
Average	564.7	12.9
Standard Deviation	322.1	4.9

Table 1: Corpus-wide statistics on the relevant lexical features for annotating the news articles.

4.2 Annotation

Annotation was done in a double-blind manner by three annotators, one of whom also acted as the adjudicator. All three annotators are Ph.D. students in computer science with a focus on natural language processing, with experience in both annotating and running annotation studies.

4.2.1 Annotator Training

The annotators that took part in this project were given minimal training outside of their individual experience with annotation studies. Annotators were provided with a guide describing van Dijk's theory. A single adjudication meeting was held after annotation for the first two sets of documents was completed. The primary purpose of this adjudication meeting was to resolve any questions the annotators had, discover any uncertainty in the annotation guide, and revise the annotation guide to address these questions. The annotation guide contains descriptions of each discourse label in addition to an example of a fully-annotated news article, shown in Figure 2.

4.2.2 Annotation Procedure

Annotation was performed over the course of a month, as time allowed. The adjudicator performed annotation of all ten sets of documents, while the other two annotators performed annotation of six sets each. Figure 3 illustrates this division of work. Annotation of each set took approximately 45 minutes to an hour, resulting in roughly ten hours of annotation work for the adjudicator and six hours for the other two annotators. The annotations were performed using Microsoft Word's built-in comment feature, to eliminate the need for any tool-based annotator training. When confronted with multiple labels that seemed to fit, annotators were instructed to choose the label that seemed the most applicable.

The adjudication procedure took a further hour for each set of documents, resulting in another ten hours of work for the adjudicator and another two hours for the other two annotators, who were only required to participate in adjudication of the first two sets of documents. The purpose of this group adjudication meeting was to resolve any outstanding questions or confusions regarding the annotation procedure. The annotation resulted in triple annotation for the first ten documents, and double annotation for the remaining forty documents. The multiple annotations were merged into a gold standard for every document. Additionally, although annotators were instructed to annotate the headline for each document, these labels are not included as part of the gold standard because within the ACE Phase 2 dataset, the headlines themselves are clearly annotated.

4.3 Annotation Results

This annotation study had two goals: first, to produce a benchmark dataset of document-level discourse annotations to evaluate the impact of document-level discourse on information extraction. Second, to evaluate whether or not humans can reliably apply van Dijk's theory to actual documents. By *reliable* we mean that annotators have a high degree of agreement with respect to each other. To measure agreement, we use the standard F_1 score (van Rijsbergen, 1979), treating one of the annotators as the correct labels, as well as Cohen's kappa coefficient for inter-rater agreement (Cohen, 1968).

The results of the annotation study are shown in Table 2. Inter-annotator agreement between annotators A1 and A2 was measured over ten documents; inter-annotator agreement between the annotators and the

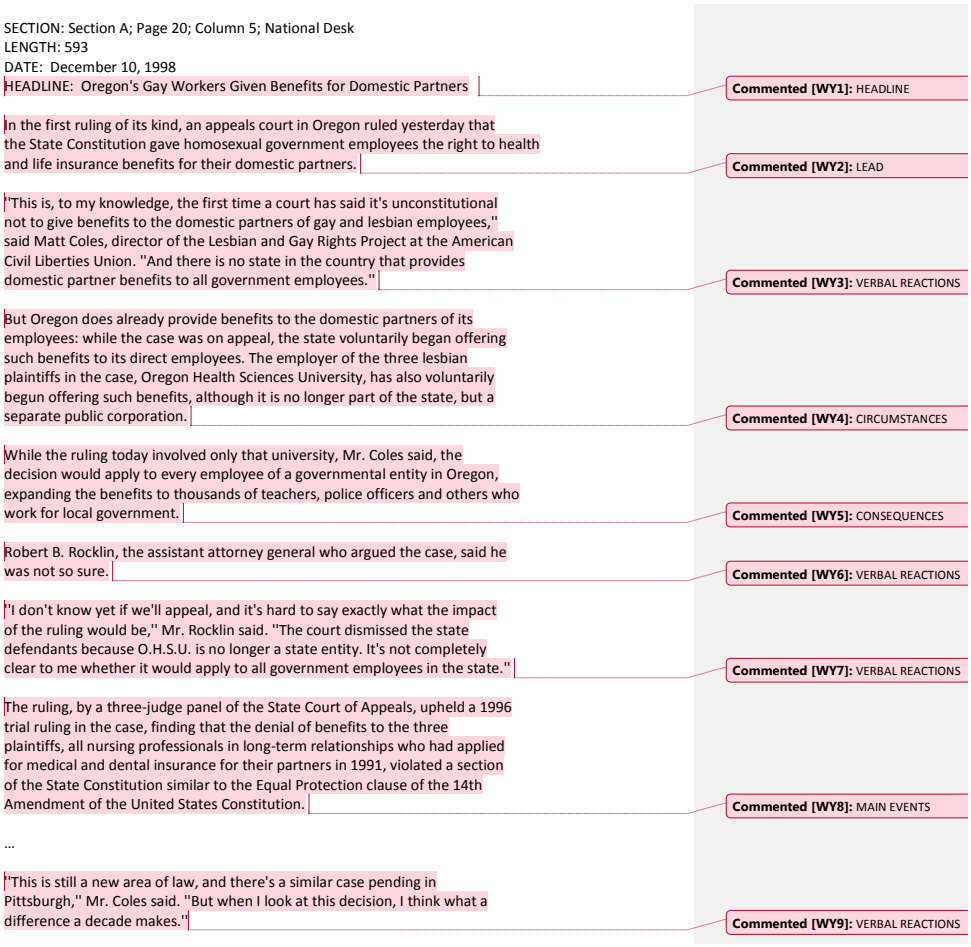

Figure 2: Example annotation included in the annotation guide. Some parts of the annotation have been omitted for brevity.

Figure 3: Division of work for the annotation study.

adjudicator, as well as the annotators and the gold standard, was measured over 30 documents. The comparison between the adjudicator and the gold standard was measured over the entire collection of 50 documents.

Table 3 provides the distribution of van Dijk's labels (sans headlines, of which there are 50: one for each document, annotated within the ACE Phase 2 corpus). Verbal reactions and circumstances dominate the labels.

4.3.1 Discussion

We observed that the inter-annotator agreement between the adjudicator and the individual annotators is high ($F_1 = 0.8$, $\kappa = 0.6$). Moreover, the results in Table 2 indicate that annotators, even with minimal training, can reliably apply van Dijk's theory.

Inter-annotator agreement between the two annotators is also high, although lower than agreement with the adjudicator ($F_1 = 0.75$, $\kappa = 0.5$). One possible reason is that the adjudicator was also responsible for the annotation guide: since the adjudicator is the source of the initial examples and instructions for

Comparison	# Docs	P	R	F_1	p_0	p_e	κ
A1 vs. A2	10	0.76	0.79	0.77	0.63	0.18	0.55
Adj. vs. A1	30	0.81	0.85	0.83	0.71	0.19	0.64
Adj. vs. A2	30	0.80	0.83	0.82	0.69	0.18	0.62
A1 vs. Gold	30	0.93	0.92	0.92	0.86	0.19	0.83
A2 vs. Gold	30	0.92	0.90	0.91	0.83	0.19	0.80
Adj. vs. Gold	50	0.93	0.87	0.90	0.81	0.18	0.77

Table 2: Microaveraged agreement measures between the annotators (A1, A2), adjudicator (Adj.), and the merged gold standard (Gold)—including precision (P), recall (R), balanced F-measure (F_1), relative observed agreement among raters (p_0), probability of chance agreement (p_e), and Cohen's kappa (κ, derived from p_0 and p_e).

Label	Count
Lead	42
Main	60
Consequences	19
Circumstances	103
Previous Events	64
History	27
Verbal Reactions	252
Expectations	21
Evaluations	56
Total	644

Table 3: Distribution of the labels within the annotated corpus. The majority of paragraphs fall under the categories of verbal reactions or circumstances.

annotation, is reasonable that the annotators would agree more strongly with the adjudicator than with each other.

Comparisons with the gold-standard are included for completeness: the all-around high agreement with the gold standard ($F_1 = 0.85$, $\kappa = 0.75$) demonstrate that the gold-standard is not dominated by a single annotator.

Although the distribution of labels is highly skewed, we find that this is roughly in-line with the style of reporting featured in the ACE Phase 2 corpus, which seeks comments and analysis from experts within the field as well as explaining the immediate context that has an effect on the main event.

5 Discourse Label Prediction

We build on top of our annotation study to demonstrate the automated learning of document-level discourse on a per-paragraph basis. We use machine learning algorithms included in Scikit-learn (Pedregosa et al., 2011) for our classifiers: in particular, we use the `svm.SVC` implementation of support vector machines (SVM), the `tree.DecisionTreeClassifier` implementation of decision trees, and the `ensemble.RandomForestClassifier` implementation of random forests. We include decision tree and random forest results despite their lower performance because they are particularly interesting for this experiment, as the theory itself is hierarchical. In addition to features from Scikit-learn, we also use the paragraph vectors (Mikolov et al., 2013) implementation in Gensim (Řehůřek and Sojka, 2010).

5.1 Feature Selection

In this section, we briefly describe the features we use and explain our rationale behind them.

Bag of Words We use Scikit-learn's `text.CountVectorizer` class with the standard English stop-words to provide a count of the tokens in each paragraph. This feature was selected based on the idea that paragraphs from different types of discourse would use different language.

TF-IDF We use Scikit-learn's `text.TfidfTransformer` class with standard parameters. TF-IDF was selected as one method to approximate topics within a given paragraph.

Paragraph Vectors We use Gensim's `models.doc2vec.Doc2Vec` class using the distributed bag-of-words model, with a minimum α of 0.01, a minimum word occurrence of five, and 50 steps (`dm=0, min_alpha=0.01, min_count=5, steps=50`). We use this as a second method of approximating the topic of a given paragraph.

Previous Paragraph's Label We also include the label from the previous paragraph. This feature is based on the idea that there is, to some degree, some sequential ordering or restriction in discourse type. One simple example is that a lead paragraph is never followed by another lead paragraph.

The bag of words, TF-IDF, and paragraph vector models are built across the entire training corpus and roughly measure what topics and words correspond to specific label types.

5.2 Results

Our best experimental results were obtained using grid search to maximize the micro-averaged performance of the classifier, as measured across five folds. The SVM classifier uses a linear kernel with $C = 10$ and the class weights balanced based on the training data; the decision tree classifier uses the default parameters with the class weights balanced; the random forest uses 50 estimators with balanced class weights.

Feature Groups	P	R	F_1
Baseline #1 (Most Freq. Class)	0.39	0.39	0.39
Baseline #2 (SVM + Bag of Words)	0.46	0.46	0.46
Decision Tree	0.41	0.41	0.41
Random Forest	0.43	0.43	0.43
SVM	0.54	0.54	**0.54**

Table 4: Results from label prediction using SVM. All results are micro-averaged across instances, including precision (P), recall (R), and balanced F-measure (F_1). For the final three classifiers, all four features are described in §5.1.

Table 4 presents the results from our experiments, showing that our classifier is a substantial improvement over the most-frequent-class and bag-of-words baselines. We note that because these features are fairly generic, and do not include potentially more informative and semantically and syntactically rich features (such as, e.g., event-, coreference-, dialog-, or discourse-specific features), these results give us hope of much better performance with further experimentation.

Table 5 presents the per-label results from our experiments. The relatively strong performance on *circumstances* and *verbal reactions* is not surprising, given their predominance. Similarly it is not surprising that we have low performance on labels that occur, on average, about once a document. We observe an unexpected level of performance on *lead* paragraphs, given their relative scarceness in the dataset. Within the data, we find that leads, with a single exception, occur at the start of the document: this accounts for the high performance, given that the first paragraph's previous paragraph is represented as -1, allowing the classifier to take advantage of their strong positional tendency.

6 Discussion

We find that using our SVM classifier, we achieve reasonable performance (65% of human performance). We suspect that an increase in performance can be gained by additional feature engineering. Moreover,

Label Type	F_1
Lead	0.87
Main	0.23
Consequences	0.13
Circumstances	0.46
Previous Events	0.18
History	0.05
Verbal Reactions	0.76
Expectations	0.08
Evaluations	0.19

Table 5: Per-label F_1 results. Best performance occurs for the lead, circumstances, and verbal reactions.

we expect that including high-precision rule-based prediction will further improve the performance of the system: this is based on comments during adjudication from annotators, who stated that they relied heavily on lexical clues such as quotation marks and specific words ("said," "commented," etc.) to select certain categories (in this case, *verbal reactions*).

While we expected the tree-oriented methods—decision trees and random forests—to outperform the SVM classifier, this was not the case in practice and they were outperformed by one of the baselines. We believe that this is because the features currently used fail to capture the higher-level semantic ideas that van Dijk used to group together the discourse types. While people understand that verbal reactions from experts, expectations, and evaluations are all types of comments, our current features do not capture these relations.

We anticipate several avenues for future work: first, there is much to be explored towards improving the performance of discourse label prediction; second, high-performance discourse label prediction enables the creation of larger corpora using automated methods; finally, following our annotation model, we anticipate further work in discourse-based corpora.

We also believe that we can improve performance on specific under-performing label types by implementing high-precision classification rules to be applied prior to statistical classification, and annotating additional data with these under-performing types to obtain further representation of them within the dataset. Furthermore, we currently do not exploit the hierarchical nature of van Dijk's theory: doing so may provide additional performance gain by allowing specific classifiers for higher-level types.

7 Contributions

In this paper, we made several key contributions. First, we have demonstrated that humans can reliably learn and annotate news articles with van Dijk's theory of news discourse with a high degree of agreement. Second, we have developed a system that can predict the document-level discourse labels for paragraphs within a news article with reasonable performance (65% of human performance). Third, we have generated a gold-standard corpus of these labels, along with an annotation guide, to support future work.

Given that our corpus is based on the ACE Phase 2 data, this work will provide a foundation for interesting discoure-based approaches to information in news, provide a benchmark for testing extraction of document-level discourse extraction, and promote research related to discourse and the news. While entity detection and relation detection are directly supported by the corpus, we also see connections to event detection and coreference given the event-central nature of van Dijk's theory.

8 Acknowledgments

This research was made possible by an FIU's Presidential Fellowship and FIU's SCIS's Director's Fellowship, both awarded to Victor Yarlott. This work was also partially supported by Office of Naval Research award # N00014-17-1-2983. We would like to thank our two annotators, Deya Banisakher and

Joshua Eisenberg, for their hard work and attention to detail.

References

Mesfin Awoke Bekalu. 2006. Presupposition in news discourse. *Discourse & Society*, 17(2):147–172.

Allan Bell. 1994. Telling stories. *Media texts: Authors and readers*, pages 100–118.

Allan Bell. 1998. The discourse structure of news stories. *Approaches to media discourse*, pages 64–104.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Judy Delin. 2000. *The language of everyday life: An introduction*. Sage.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

NIST. 2002. Ace phase 2.

Zhongdang Pan and Gerald M Kosicki. 1993. Framing analysis: An approach to news discourse. *Political communication*, 10(1):55–75.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Afrooz Rafiee, Wilbert Spooren, and José Sanders. 2017. Culture and discourse structure: A comparative study of dutch and iranian news texts. *Discourse & Communication*, page 1750481317735626.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Nynke van der Vliet, Ildikó Berzlánovich, Gosse Bouma, Markus Egg, and Gisela Redeker. 2011. Building a discourse-annotated dutch text corpus. *Bochumer Linguistische Arbeitsberichte*, 3:157–171.

Teun A Van Dijk, 1986. *Studying Writing: Linguistic Approaches. Written Communication Annual: An International Survey of Research and Theory Series, Volume 1.*, chapter News Schemata, pages 155–185. ERIC.

Teun A van Dijk, 1988. *News as Discourse*, chapter Structure of News, pages 52–57. Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey.

Cornelis J. van Rijsbergen. 1979. *Information retrieval*. Butterworths, London Boston.

Peter R White. 1998. *Telling media tales: The news story as rhetoric*. Department of Linguistics, Faculty of Arts, University of Sydney.