Improving the Identification of the Discourse Function of News Article Paragraphs

Deya Banisakher, W. Victor H. Yarlott, Mohammed Aldawsari,
Napthali D. Rishe, & Mark A. Finlayson
School of Computing and Information Sciences
Florida International University
11200 S.W. 8th Street, CASE 362, Miami, FL 33199 USA
{dbani001,wyarl001,malda021,rishe,markaf}@fiu.edu

Abstract

Identifying the discourse structure of documents is an important task in understanding written text. Building on prior work, we demonstrate an improved approach to automatically identifying the discourse function of paragraphs in news articles. We start with the hierarchical theory of news discourse developed by van Dijk (1988) which proposes how paragraphs function within news articles. This discourse information is a level intermediate between phrase- or sentence-sized discourse segments and document genre, characterizing how individual paragraphs convey information about the events in the storyline of the article. Specifically, the theory categorizes the relationships between narrated events and (1) the overall storyline (such as MAIN EVENTS, BACKGROUND, or CONSEQUENCES) as well as (2) commentary (such as VERBAL REACTIONS and EVALUATIONS). We trained and tested a linear chain conditional random field (CRF) with new features to model van Dijk's labels and compared it against several machine learning models presented in previous work. Our model significantly outperformed all baselines and prior approaches, achieving an average of 0.71 F_1 score which represents a 31.5% improvement over the previously best-performing support vector machine model.

1 Introduction

News articles usually follow strong principles of journalistic structure. By design, they often begin with a introductory summary of main events, followed by detailed exposition of the main events and consequences, interspersed in a stereotyped fashion with relevant background information, current and past evidence, and reported speech. Yarlott et al. (2018) demonstrated the feasibility of detecting this type of discourse structure for news articles using an established hierarchical theory of news discourse (van Dijk, 1988). In their study, they showed that it was feasible to identify the discourse function of news paragraphs using a support vector machine (SVM) model and a small set of simple linguistic features, with a performance of 0.54 F_1.

Similar to Yarlott et al.'s (2018) approach, we demonstrate an improved approach to automatically labeling news article paragraphs with the van Dijk discourse functions Yarlott et al. (2018) applied in their study. Our work uses a conditional random field (CRF) model, along with new features, to obtain an improved performance of 0.71 F_1. Most importantly, our model is able to precisely capture the interdependencies between the various discourse label types, which flows from our hypothesis that each paragraph in an article is dependent not only on the previous one but rather on a longer sequence of previous paragraphs.

The remainder of this paper is structured as follows. We first provide a definition of van Dijk's theory as was presented in (Yarlott et al., 2018) (§2). Second, we describe the dataset we used in training and testing our CRF model (§3). We then detail the discourse label identification methods, including the CRF model and how it captures both section ordering and section content, how the model is trained, and the features it leverages (§4). We next compare the performance of the CRF model with various baselines, demonstrating that it performs better than prior models (§5). We then discuss related work (§6), and conclude with a summary of contributions (§7).

2 Van Dijk's Theory of News Discourse

Van Dijk (1988) described a hierarchical theory of news discourse, the categories of which are shown in Figure 1, which we apply to a subset of the news articles of the ACE Phase 2 corpus. In this section,

Proceedings of the 1st Joint Workshop on Narrative Understanding, Storylines, and Events, pages 17–25

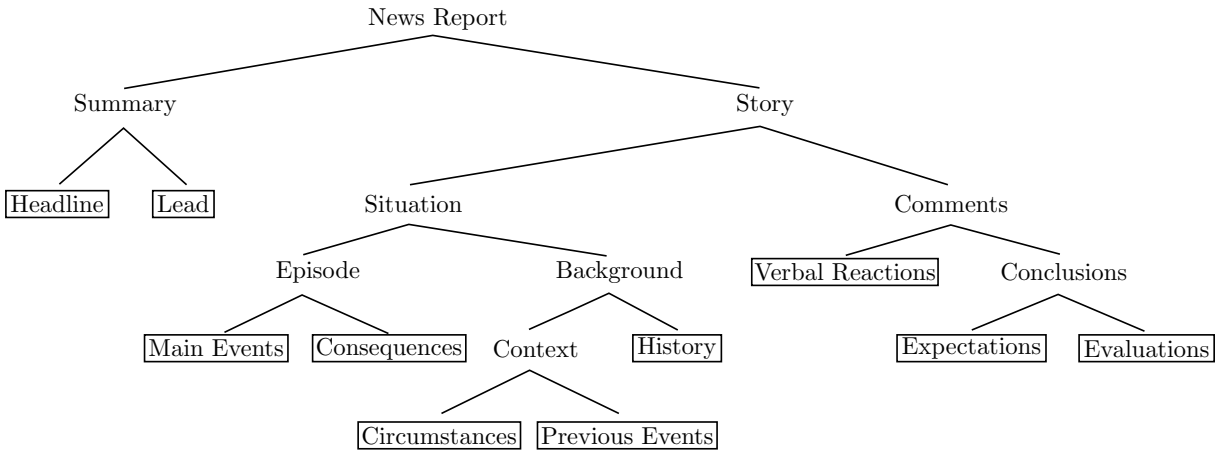

Figure 1: The hierarchical discourse structure of news proposed by van Dijk (van Dijk, 1988). Boxes indicate labels that were directly annotated on the documents; other labels can be inferred. From Yarlott et al. (2018), Figure 1.

we repeat our descriptions of the leaf categories from our prior paper, as well as their parent categories when appropriate, for ease of reference.

SUMMARY elements express the major subject of the article, with the HEADLINE being the actual headline of the article, and the LEAD being the first sentence, which is often a summary of the main events of the article.

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 provides important information about the relation of each paragraph with respect to the central events of a news story. Background includes the CONTEXT, of which CIRCUMSTANCES are temporally or spatially non-specific states that contribute to understanding the subject, while PREVIOUS EVENTS are specific recent events that enhance understanding of the main events. HISTORY paragraphs are another type of Background describing events that have not occurred recently, typically referenced in terms of years prior, rather than months, weeks, or days.

COMMENTS provide further supporting context for the central events of an article. Comments may include VERBAL REACTIONS solicited from an external source, such as a person involved in the events, or an expert. CONCLUSIONS, by contrast, are comments made by a journalistic entity (the newspaper, reporter, etc.) regarding the subject. Conclusions can be separated into EXPECTATIONS about the resolution or consequences of an event, or EVALUATIONS of the current situation.

3 Dataset

We used a gold-standard corpus previously developed by Yarlott et al. (2018) of van Dijk's labels applied to a subset of the Automated Content Extraction (ACE) Phase 2 corpus (NIST, 2002). The ACE Phase 2 corpus is a major standard corpora of news articles that boasts three advantages: it is widely-used, has relevance to other tasks, and was readily available to researchers. This dataset comprises 50 documents containing 28,236 words divided in 644 paragraphs.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.

	Words	Paragraphs
Total	28,236	644
Average	564.7	12.9
Std. Dev.	322.1	4.9

Table 1: Corpus-wide statistics for the annotated data. Adapted from Yarlott et al. (2018), Table 1.

Yarlott et al. (2018) doubly annotated 50 randomly selected news articles, 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. 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. The authors report an all-around high agreement with the gold standard ($F_1 = 0.85$, $\kappa = 0.75$) which demonstrates that the gold-standard was not dominated by a single annotator.

Although the dataset discussed was annotated for all labels discussed here, the HEADLINE label could be computed automatically from the structure of ACE Phase 2 corpus, as the files has the headline separate as part of its markup scheme.

Table 2 provides the resulting distribution of van Dijk's labels. Verbal reactions and circumstances dominate the labels. 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.

Label	Count	Label	Count
HEADLINE	50	LEAD	42
MAIN EVENTS	60	CONSEQUENCES	19
CIRCUMSTANCES	103	PREVIOUS EVENTS	64
HISTORY	27	VERBAL REACTIONS	252
EXPECTATIONS	21	EVALUATIONS	56

Table 2: Distribution of the labels within the annotated corpus, with 644 labels total. The majority of paragraphs fall under the categories of verbal reactions or circumstances. From (Yarlott et al., 2018)

4 Identifying Discourse Labels

In contrast to the approach reported by (Yarlott et al., 2018), we the treated label identification for paragraphs as a sequence modeling task. Formally, the task is as follows: given a news report with n discourse labels and m paragraphs, where the paragraphs are unlabeled, identify the optimal sequence (order) of discourse labels $H^* = (L_1^*, \ldots, L_n^*)$ from among all possible label sequences, and assign every paragraph a discourse label $H^* = (H_1, \ldots, H_m)$ consistent with L^*. Sequence labeling problems in NLP, medical informatics, and discourse parsing have been studied by both generative and discriminative approaches, including Hidden Markov Models (HMMs; generative) and Conditional Random Fields (CRFs; discriminative). Li et al. (2010) used HMM and n-gram models to detect the orders or labels of sections within clinical reports, while modeling the observation probabilities at the section level. Sherman and Liu (2008) used HMMs as well as n-gram models to detect topic shifts in meeting minutes, and, in contrast to Li et al., modeled the observation probabilities on the sentence level.

Our approach was inspired by the method described in Banisakher et al. (2018) which identifies section labels in clinical psychiatric reports. Their approach combined a *Hierarchical* Hidden Markov Model (HHMM)—which used section statistics as the model's transition probabilities—with n-grams for the observation probabilities of words. In this paper we substitute a CRF for the HHMM. Generative models such as HMMs have more explanatory power when compared with their discriminative counterparts such as CRFs. However, HMMs, rely on the assumption that observations are statistically independent from one another. For our problem, this means that an HMM assumes that the presence of certain paragraphs corresponding to a certain discourse label or function A is independent from other paragraphs within another section B. In practice, however, this is not the case: for example a paragraph following the MAIN EVENTS are often either CONSEQUENCES or CIRCUMSTANCES.

4.1 Linear Chain Conditional Random Fields

Conditional Random Fields (CRFs) are undirected graphical models (Lafferty et al., 2001; Konkol and Konopík, 2013) that can be used for discriminative sequence labeling. CRFs have proved useful for many sequence labeling problems in NLP and computer vision (Lin and Wu, 2009), including Named Entity Recognition (NER) and image classification. There are several CRF variations such as the tree CRF and the hierarchical CRF which are mostly used for computer vision related tasks.

We built and trained a linear chain CRF modeled on Banisakher et al.'s HHMM approach. In contrast to an HHMM, the CRF encodes labels as nodes in the CRF graphical representation (instead of HMM states), and uses weighted feature functions for transitions between nodes (instead of the HMM transition and emission probabilities). Additionally, the CRF model captures the "true" desired probability distribution, that is the *conditional distribution* of labels given the observations $P(Y|X)$, instead of modeling the joint distribution of observations and labels $P(X, Y)$. This a known advantage of CRFs in general over HMMs

and is mainly due to, again, removing the independence assumption. Thus, CRFs can have an arbitrary number of dependencies as opposed to the limited dependency structure of HMMs. Our model benefits from this as it does not only record the dependence of a discourse label only on its predecessor and observations, but on additional dependencies given the entire sequence of labels (i.e., paragraph discourse functions) and observations (i.e., paragraphs).

The CRF probability distribution is defined by Equation 1. Let $\bar{l}$ be the sequence of discourse labels, $\overline{p}$ be the sequence of paragraphs (i.e., the observations) in a given report, and L be the set of all possible label sequences. Our model follows a typical linear chain CRF where the conditional distribution is:

$$P(\bar{l}|\overline{p},\lambda) = \frac{\exp(\sum_i \sum_j \lambda_j F_j(l_{i-1}, l_i, \overline{p}, i))}{\sum_{l' \in L} \exp(\sum_i \sum_j \lambda_j F_j(l'_{i-1}, l'_i, \overline{p}, i))} \quad (1)$$

where λ is a set of model parameters, and each λ_j is a weight associated with each feature function F_j. Each feature function represents a dependency within the model. We used the L-BFGS method to estimate each λ_j (Nocedal, 1980). The model's probability distribution is thus generated by summing over the entire observation sequence, where each observation is indexed by the variable i and the entire feature function space index by the variable j. The denominator sums over all possible label sequences L.

The most critical component in the design of CRF models is the feature function space. In our model, each feature function is:

$$\begin{aligned} &F_j(l_{i-1}, l_i, \overline{p}, i) = \\ &H_j(l_{i-1}, l_i, \overline{p}, i) \cdot SF_j(l_{i-1}, l_i, \overline{p}, i) \end{aligned} \quad (2)$$

where H_j models the discourse labels' order, and SF_j models the labels' content. These are similar to an HMM's transition and emission probability distributions, respectively. In contrast to HMMs, however, the feature functions are evaluated over the entire observation sequence $\overline{p}$ taking into account the neighboring labels l_i and l_{i-1}. This conditions the probability of a given discourse label type on the content and order of the entire sequence. We outline the intuition behind and implementation of our feature functions in the following sections.

4.2 Modeling the Discourse Labels' Order

The feature function F_j incorporates section ordering through the section ordering function $H(l_{i-1}, l_i, \overline{p}, i)$. As discussed above, there is a feature function for each of the dependencies defined in the model. We encode the interdependent order of labels (i.e., which labels depend upon each other) using a binary matrix. To achieve this, we first used the van Dijk discourse labels shown in Tables 2 and discussed in §2. Then we created a binary matrix V_{l_{i-1},l_i} whose entries represent whether a label follows another or not. For example if label HISTORY (indexed as label 6) was observed in the data directly before VERBAL REACTIONS (indexed as label 7), then the entry $V_{6,7}$ would contain a value of 1. The matrix contained N^2 entries, where N is the total number of labels. Thus our CRF models contained 9 nodes in total. We formulated the section order feature function as follows:

$$H_j(l_{i-1}, l_i, \overline{s}, i) = V_{l_{i-1},l_i} \quad (3)$$

Note that for each label s_i, the model sums the total entries for the entire sequence of labels and observations as shown in Equation 1, thus conditioning each label on the entire sequence.

4.3 Modeling the Discourse Labels' Content

Similarly, the feature function F_j incorporates the discourse label type content via the feature function $SF(l_{i-1}, l_i, \overline{p}, i)$. These functions model the dependency between a label type and its content. Importantly, the feature function should not be confused with the linguistic features that are extracted from the text and input into the section feature function. To capture label content (i.e., to model discourse label type-specific language) we extracted the following set of features:

Features from Yarlott et al. (2018): *Unigrams* (i.e., bag of words), the *tf-idf* count vector of the top 3 words (across the corpus) per label type, bag-of-words, and *paragraph vectors* using the Doc2Vec approach (Le and Mikolov, 2014). As pointed out by Yarlott et al., the *tf-idf* and *paragraph vectors* approximate topics within a given paragraph. Yarlott et al. also used the previous paragraph's label as an explicit feature; this is included by default in the CRF model.

Lexical: *Bigrams* to capture the type of language per discourse label type.

Positional: *Size of paragraphs* represented by number sentences present. As well as the *paragraph position* relative to the document head.

Syntactic: A *POS count vector* which encodes the number of times each part of speech (POS)

(specifically, nouns, verbs, adjectives, and adverbs) appears in the paragraph.

Semantic: Here we incorporated four additional features: a *reported speech* feature, a *majority event tense* feature, a *subevent relation* count vector, and *NER vectors* representing a select set of named entities. For the *reported speech* feature, we extracted quotations and sentences with tagged as reported speech by the `textacy` library (DeWilde, 2020) and labeled the containing paragraph as VERBAL REACTIONS. For the *majority event tense* feature, we extracted the events in each paragraph using the CAEVO event extraction system (Chambers et al., 2014), noted their POS tags using a dependency tree, and recorded the majority verb tense in that paragraph. For the *subevent relation* feature, we used Aldawsari and Finlayson's subevent extraction system (2019) to capture relationships between paragraphs. For this, we used a vector for each paragraph corresponding to the number of paragraphs of the article with the maximum number of paragraphs in the corpus. Aldawsari and Finlayson (2019) presented a supervised model for automatically identifying when one event is a subevent of another using narrative and discourse features. For each event relation found by this system between two distinct paragraphs, we recorded a +1 in that corresponding vector cell, while we discarded relationships found within a single paragraph. For the *NER vectors*, we applied Named Entity Recognition (NER) and extracted the first 13 named entity types found by the Spacy library (AI, 2020) including PERSON, LOCATION, DATE, and TIME. These 13 types were represented in a numerical vector for each discourse label type such that, for each type, we recorded the number of entity occurrences.

4.4 Inference

We applied the usual inference process for linear chain CRFs operating at the paragraph level (Forney, 1973). Inference in linear chain CRFs follows a similar algorithm to Viterbi, which is used in decoding HMM models. While not stated explicitly in the Equation 1 above, the normalization factor $Z(S)$ is calculated as is often done using the Gaussian prior as it was introduced in (Chen and Rosenfeld, 1999).

5 Results and Discussion

In order to test our model, we randomly split each corpus into training and testing sets in a cross-validation setup, using five folds, resulting in 40 news reports for training and 10 for testing in each fold. Our model was trained to learn a total of 9 distinct discourse label types as represented in 2 (all leaf labels minus HEADLINE). In this section we describe our baseline comparisons and overall experiments and results.

5.1 Baseline Methods

We followed Yarlott et al. (2018) in their baseline comparisons. We compared our model's performance against five other methods: two baselines including the most frequent class (MFC) and a support vector machine using bag-of-words (SVM+BoW); third, a decision tree classifier; fourth, a random forest classifier; and fifth, Yarlott et al. (2018)'s best performing model, a support vector machine. As described above, the latter three models incorporate a the following set of four features: bag-of-words, *tf-idf*, paragraph vectors, and previous paragraph labels. We used the same experimental setup for all of these models. Yarlott et al. (2018) obtained the best experimental results using grid search to maximize the micro-averaged performance of each classifier, as measured across five folds. Following Yarlott et al. (2018), 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.

5.2 Results

Our CRF model outperformed all other classifiers and baselines achieving a 0.71 F_1 score. Table 3 shows the micro-averaged precision (P), recall (R), and F_1 scores for the five models from (Yarlott et al., 2018) as well as our current CRF approach. Our experimental results show that our CRF approach is a substantial improvement over the previously best performing model.

For CRF, we performed 8 feature combination experiments (shown in Table 3) to evaluate the effect of feature classes as well as the individual semantic features. As discussed before, the SVM as well as the decision tree and random forest classifiers only leveraged Yarlott et al.'s original four

Model	Features	P	R	F_1
MFC	-	0.39	0.39	0.39
HHMM	Bigrams	0.42	0.45	0.43
SVM	BoW	0.46	0.46	0.46
DT	Yarlott et al.	0.41	0.41	0.41
RDF	Yarlott et al.	0.43	0.43	0.43
SVM	Yarlott et al.	0.54	0.54	0.54
CRF	Yarlott et al.	0.58	0.60	0.59
CRF	+Lexical	0.61	0.63	0.62
CRF	+Positional	0.62	0.66	0.64
CRF	+Syntactic	0.65	0.69	0.67
CRF	+*subevent relation*	0.65	0.0.70	0.67
CRF	+*majority event tense*	0.67	0.71	0.68
CRF	+*reported speech*	0.68	0.72	0.70
CRF	All (+Remaining Sem.)	0.69	0.73	**0.71**

Table 3: Experimental results for discourse label identification. All results are micro-averaged across instances, including precision (P), recall (R), and balanced F-measure (F_1). The Decision Tree, Random Forest, and SVM classifiers used the features outlined in (Yarlott et al., 2018). For the middle three lines of the CRF section, these indicate features groups added to the previous line's model. We present the results for the smenatic features individually. The CRF model in the last line (CRF with ALL features) includes all the features from the previous lines as well as all remaining semantic features.

features: bag-of-words, *tf-idf*, paragraph vectors, and previous paragraph labels. While our CRF approach uses a more sophisticated set of features leveraging additional syntactic and semantic features as outlined in 4.3. Most importantly, our model treats the problem as a sequence labeling task and therefore captures the sequential dependencies between the paragraphs as well as the labels within each report. This is evidenced by the CRF model that uses only Yarlott et al.'s features, which achieves a higher performance than the original SVM classifier.

Our CRF model achieved the largest increase in performance after adding the semantic features. This was expected: we anticipated a boost in performance on the VERBAL REACTIONS class given detection of reported speech, and a similar increase in performance on the MAIN EVENTS and PREVIOUS EVENTS classes given the addition of event and subevent features. Of the semantic features, the *reported speech* feature had the biggest impact on the model's performance as the verbal reactions section was predominant in the dataset. Here `textacy` performed quite well in automatically identifying reported speech as the model achieved a 0.91 F_1 score for the VERBAL REACTIONS class. The *subevent relation* and *majority event tense* features improved the performance by about one point F_1 each, with the second contributing slightly more to the overall performance. The *majority event tense* feature contributed heavily to the PREVIOUS EVENTS and HISTORY, we suspect due to the relatively more frequent use of past tense verbs in paragraphs belonging to those classes. As discussed before, we used automated systems to detect events and subevent relations. Naturally, these systems do not boast a perfect performance and therefore error propagation is expected. Thus, we expect that our model can further achieve higher performance using more refined event detection solutions, as well as a larger corpus.

Table 4 presents the per-label results from our experiments. The relatively strong performance on CIRCUMSTANCES and VERBAL REACTIONS is not surprising, given their relative prevalence in our corpus. Similarly it is not surprising that we have low performance on labels that occur, on average, about once (or less) a document (HISTORY, EXPECTATIONS). However, these label types saw a significant performance boost in our model compared to the previous approaches as our features have captured more of their distinct language. For CONSEQUENCES HISTORY, EXPECTATIONS, and EVALUATIONS, the syntactic and positional features were most helpful. Similar to (Yarlott et al., 2018), we observe an unexpected–but not surprising–level of performance on LEAD paragraphs, given their relative scarcity in the dataset: we find that leads, with a single exception, occur once at the start of the document.

Again, similar to (Yarlott et al., 2018), we expected the tree-oriented methods—decision trees and random forests—to at least outperform the SVM classifier. However, this was not the case in practice and they were outperformed by one of the baselines. We believe that this partially attributed to the fact that these models did not leverage the full set of hierarchical labels in van Dijk's discourse theory: they were only presented with the leaf labels.

6 Related Work

There has been 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

Label Type	F_1	Label Type	F_1
HEADLINE	-	LEAD	0.95
MAIN EVENTS	0.69	CONSEQUENCES	0.29
CIRCUMSTANCES	0.72	PREVIOUS EVENTS	0.51
HISTORY	0.24	VERBAL REACTIONS	0.91
EXPECTATIONS	0.26	EVALUATIONS	0.51
		Macro Average	0.56

Table 4: Per-label F_1 results. The last row shows the macro average over all label types. Best performance occurs for the LEAD, MAIN EVENTS, CIRCUMSTANCES, and VERBAL REACTIONS.

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 (2018). Rafiee et al. also show that many stories follow a hybrid structure, which combines characteristics from both inverted and chronological structures.

Our approach was inspired by Banisakher et al. (2018)'s HHMM approach to section identification in clinical notes. In turn, their work extend an earlier study on section identification of psychiatric evaluation reports that combined the work of Li et al. (2010) on identifying section types within clinical reports and that of Sherman and Liu (2008) on text segmentation of meeting minutes. Li et al. modeled HMM emissions at the section level using bigrams, while Sherman and Liu modeled the emissions at the sentence level and used unigrams and trigrams. Other approaches followed similar strategies in segmenting story text and in creating generative models for detecting story boundaries (Mulbregt et al., 1998; Yamron et al., 1998). More recently, Yu et al. (2016) used a hybrid deep neural network combined with a Hidden Markov Model (DNN-HMM) to segment speech transcripts from broadcast news to a sequence of stories. Similar to our approach, (Sprugnoli et al., 2017) used CRFs and SVMs for the classification of automatic classification of Content Types, a novel task that was introduced to provide cues to access the structure of a document's types of functional content.

Discussing van Dijk's theory of news discourse, Bekalu stated 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, presented 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.

While 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.

7 Contributions

We extend earlier work on news paragraph discourse function labeling. We built a linear chain CRF model incorporating various lexical, positional, syntactic, and semantic features that improves detection of the order of discourse labels in a news article at the paragraph level as well as models the paragraph content of each label type. We evaluated our model's performance against two baselines and three existing models with various subsets of features. We showed that the CRF model represents a significant improvement in this task. Most importantly, our work demonstrated the importance of modeling paragraph and discourse label type inter-dependencies.

Acknowledgments

Mr. Banisakher was supported by a Dissertation Year Fellowship from FIU. Mr. Yarlott was sup-

ported by DARPA Contract FA8650-19-C-6017. Mr. Aldawsari was supported by a doctoral fellowship from Prince Sattam Bin Abdulaziz University, and thanks Dr. Sultan Aldossary for his advice and support. Dr. Finlayson was partially supported by NSF Grant IIS-1749917.

References

Explosion AI. 2020. Annotation Specifications-SpaCy API Documentation.

Mohammed Aldawsari and Mark Finlayson. 2019. Detecting Subevents using Discourse and Narrative Features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4780–4790, Florence, Italy. Association for Computational Linguistics.

Deya Banisakher, Naphtali Rishe, and Mark A. Finlayson. 2018. Automatically Detecting the Position and Type of Psychiatric Evaluation Report Sections. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 101–110, Brussels, Belgium. Association for Computational Linguistics.

Mesfin Awoke Bekalu. 2006. Presupposition In News Discourse. *Discourse & Society*, 17(2):147–172.

Allan Bell. 1994. Telling stories. In David Graddol and Oliver Boyd-Barrett, editors, *Media texts: Authors and readers*, pages 100–118. Multilingual Matters, Clevedon, U.K.

Allan Bell. 1998. The Discourse Structure of News Stories. In *Approaches to Media Discourse*, pages 64–104. Blackwell Oxford.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense Event Ordering with a Multi-Pass Architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Stanley F Chen and Ronald Rosenfeld. 1999. A Gaussian Prior for Smoothing Maximum Entropy Models. Technical report, Carnegie-Mellon University, School of Computer Science, Pittsburgh, Pennsylvania, USA.

Judy Delin. 2000. *The Language of Everyday Life: An Introduction*. Sage, London, UK.

Burton DeWilde. 2020. textacy.

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

G.D. Forney. 1973. The Viterbi Algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Michal Konkol and Miloslav Konopík. 2013. CRF-Based Czech named Entity Recognizer and Consolidation of Czech NER Research. In *Text, Speech, and Dialogue*, pages 153–160, Berlin, Heidelberg. Springer Berlin Heidelberg.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196, Beijing, China.

Ying Li, Sharon Lipsky Gorman, and Noémie Elhadad. 2010. Section Classification in Clinical Notes Using Supervised Hidden Markov Model. In *Proceedings of the 1st ACM International Health Informatics Symposium IHI*, pages 744–750, Arlington, Virgina, USA.

Dekang Lin and Xiaoyun Wu. 2009. Phrase Clustering for Discriminative Learning. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1030–1038, Suntec, Singapore. Association for Computational Linguistics.

Paul van Mulbregt, Ira Carp, Lawrence Gillick, Steve Lowe, and Jon Yamron. 1998. Text Segmentation and Topic Tracking on Broadcast News Via a Hidden Markov Model Approach. In *Fifth International Conference on Spoken Language Processing*, ICSLP '98, Sydney, Australia.

NIST. 2002. Ace phase 2.

Jorge Nocedal. 1980. Updating Quasi-Newton Matrices with Limited Storage. *Mathematics of computation*, 35(151):773–782.

Zhongdang Pan and Gerald M Kosicki. 1993. Framing Analysis: An Approach to News Discourse. *Political Communication*, 10(1):55–75.

Afrooz Rafiee, Wilbert Spooren, and José Sanders. 2018. Culture and Discourse Structure: A Comparative Study of Dutch and Iranian News Texts. *Discourse & Communication*, 12(1):58–79.

M. Sherman and Yang Liu. 2008. Using Hidden Markov Models for Topic Segmentation of Meeting Transcripts. In *Proceedings of the 2008 IEEE Spoken Language Technology Workshop*, pages 185–188, Goa, India.

Rachele Sprugnoli, Tommaso Caselli, Sara Tonelli, and Giovanni Moretti. 2017. The content types dataset: a new resource to explore semantic and functional characteristics of texts. In *Proceedings of the 15th*

Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 260–266, Valencia, Spain. Association for Computational Linguistics.

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. Sage, Beverly Hills, California, USA.

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.

Peter R White. 1998. *Telling Media Tales: The News Story as Rhetoric*. Department of Linguistics, Faculty of Arts, University of Sydney, Sydney, Australia.

J. P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. 1998. A Hidden Mrkov Model Approach to Text Segmentation and Event Tracking. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '98, pages 333–336 vol.1, Seattle, Washington, USA.

W. Victor Yarlott, Cristina Cornelio, Tian Gao, and Mark Finlayson. 2018. Identifying the discourse function of news article paragraphs. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 25–33, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jia Yu, Xiong Xiao, Lei Xie, Chng Eng Siong, and Haizhou Li. 2016. A DNN-HMM Approach to Story Segmentation. In *INTERSPEECH 2016*, San Francisco, California, USA.