A coherence model based on syntactic patterns

Annie Louis
University of Pennsylvania
Philadelphia, PA 19104, USA
lannie@seas.upenn.edu

Ani Nenkova
University of Pennsylvania
Philadelphia, PA 19104, USA
nenkova@seas.upenn.edu

Abstract

We introduce a model of coherence which captures the intentional discourse structure in text. Our work is based on the hypothesis that syntax provides a proxy for the communicative goal of a sentence and therefore the sequence of sentences in a coherent discourse should exhibit detectable structural patterns. Results show that our method has high discriminating power for separating out coherent and incoherent news articles reaching accuracies of up to 90%. We also show that our syntactic patterns are correlated with manual annotations of intentional structure for academic conference articles and can successfully predict the coherence of abstract, introduction and related work sections of these articles.

1 Introduction

Recent studies have introduced successful automatic methods to predict the structure and coherence of texts. They include entity approaches for local coherence which track the repetition and syntactic realization of entities in adjacent sentences (Barzilay and Lapata, 2008; Elsner and Charniak, 2008) and content approaches for global coherence which view texts as a sequence of topics, each characterized by a particular distribution of lexical items (Barzilay and Lee, 2004; Fung and Ngai, 2006). Other work has shown that co-occurrence of words (Lapata, 2003; Soricut and Marcu, 2006) and discourse relations (Pitler and Nenkova, 2008; Lin et al., 2011) also predict coherence.

Early theories (Grosz and Sidner, 1986) posited that there are three factors which collectively contribute to coherence: intentional structure (purpose of discourse), attentional structure (what items are discussed) and the organization of discourse segments. The highly successful entity approaches capture attentional structure and content approaches are related to topic segments but intentional structure has largely been neglected. Every discourse has a purpose: explaining a concept, narrating an event, critiquing an idea and so on. As a result each sentence in the article has a communicative goal and the sequence of goals helps the author achieve the discourse purpose. In this work, we introduce a model to capture coherence from the intentional structure dimension. Our key proposal is that syntactic patterns are a useful proxy for intentional structure.

This idea is motivated from the fact that certain sentence types such as questions and definitions have distinguishable and unique syntactic structure. For example, consider the opening sentences of two descriptive articles[1] shown in Table 1. Sentences (1a) and (2a) are typical instances of definition sentences. Definitions are written with the concept to be defined expressed as a noun phrase followed by a copular verb (is/are). The predicate contains two parts: the first is a noun phrase reporting the concept as part of a larger class (eg. an aqueduct is a water supply), the second component is a relative clause listing unique properties of the concept. These are examples of syntactic patterns related to the communicative goals of individual sentences. Similarly, sentences (1b) and (2b) which provide further details about the concept also have some distinguish-

[1] Wikipedia articles on "Aqueduct" and "Cytokine Receptors"

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1157–1168, Jeju Island, Korea, 12–14 July 2012.

1a) An aqueduct is a water supply or navigable channel constructed to convey water. b) In modern engineering, the term is used for any system of pipes, canals, tunnels, and other structures used for this purpose.
2a) Cytokine receptors are receptors that binds cytokines. b) In recent years, the cytokine receptors have come to demand more attention because their deficiency has now been directly linked to certain debilitating immunodeficiency states.

Table 1: The first two sentences of two descriptive articles

ing syntactic features such as the presence of a topicalized phrase providing the focus of the sentence. The two sets of sentences have similar sequence of communicative goals and so we can expect the syntax of adjacent sentences to also be related.

We aim to characterize this relationship on a broad scale using a coherence model based entirely on syntax. The model relies on two assumptions which summarize our intuitions about syntax and intentional structure:

1. Sentences with similar syntax are likely to have the same communicative goal.

2. Regularities in intentional structure will be manifested in syntactic regularities between adjacent sentences.

There is also evidence from recent work that supports these assumptions. Cheung and Penn (2010) find that a better syntactic parse of a sentence can be derived when the syntax of adjacent sentences is also taken into account. Lin et al. (2009) report that the syntactic productions in adjacent sentences are powerful features for predicting which discourse relation (cause, contrast, etc.) holds between them. Cocco et al. (2011) show that significant associations exist between certain part of speech tags and sentence types such as explanation, dialog and argumentation.

In our model, syntax is represented either as parse tree productions or a sequence of phrasal nodes augmented with part of speech tags. Our best performing method uses a Hidden Markov Model to learn the patterns in these syntactic items. Sections 3 and 5 discuss the representations and their specific implementations and relative advantages. Results show that syntax models can distinguish coherent and incoherent news articles from two domains with 75-90% accuracies over a 50% baseline. In addition, the syntax coherence scores turn out complementary to scores given by lexical and entity models.

We also study our models' predictions on academic articles, a genre where intentional structure is widely studied. Sections in these articles have well-defined purposes and we find recurring sentence types such as motivation, citations, description, and speculations. There is a large body of work (Swales, 1990; Teufel et al., 1999; Liakata et al., 2010) concerned with defining and annotating these sentence types (called zones) in conference articles. In Section 6, we describe how indeed some patterns captured by the syntax-based models are correlated with zone categories that were proposed in prior literature. We also present results on coherence prediction: our model can distinguish the introduction section of conference papers from its perturbed versions with over 70% accuracy. Further, our model is able to identify conference from workshop papers with good accuracies, given that we can expect these articles to vary in purpose.

2 Evidence for syntactic coherence

We first present a pilot study that confirms that adjacent sentences in discourse exhibit stable patterns of syntactic co-occurrence. This study validates our second assumption relating the syntax of adjacent sentences. Later in Section 6, we examine syntactic patterns in individual sentences (assumption 1) using a corpus of academic articles where sentences were manually annotated with communicative goals.

Prior work has reported that certain grammatical productions are repeated in adjacent sentences more often than would be expected by chance (Reitter et al., 2006; Cheung and Penn, 2010). We analyze all co-occurrence patterns rather than just repetitions.

We use the gold standard parse trees from the Penn Treebank (Marcus et al., 1994). Our unit of analysis is a pair of adjacent sentences (S_1, S_2) and we choose to use Section 0 of the corpus which has 99 documents and 1727 sentence pairs. We enumerate all productions that appear in the syntactic parse of any sentence and exclude those that appear less than 25 times, resulting in a list of 197 unique productions. Then all ordered pairs[2] (p_1, p_2) of productions are formed. For each pair, we compute

[2] (p_1, p_2) and (p_2, p_1) are considered as different pairs.

p_1, p_2	**Sentence 1**	**Sentence 2**
NP → NP NP-ADV QP → CD CD	The two concerns said they entered into a definitive merger agreement under which Ratners will begin a tender offer for all of Weisfield's common shares for [$57.50 each]$_{\text{NP}}$.	Also on the takeover front, Jaguar's ADRs rose 1/4 to 13 7/8 on turnover of [4.4 million]$_{\text{QP}}$.
VP → VB VP NP-SBJ → NNP NNP	"The refund pool may not [be held hostage through another" round of appeals]$_{\text{VP}}$," Judge Curry said.	[Commonwealth Edison]$_{\text{NP-SBJ}}$ said it is already appealing the underlying commission order and is considering appealing Judge Curry's order.
NP-LOC → NNP S-TPC-1 → NP-SBJ VP	"It has to be considered as an additional risk for the investor," said Gary P. Smaby of Smaby Group Inc., [Minneapolis]$_{\text{NP-LOC}}$.	["Cray Computer will be a concept" "stock,"]$_{\text{S-TPC-1}}$ he said.

Table 2: Example sentences for preferred production sequences. The span of the LHS of the corresponding production is indicated by [] braces.

the following: $c(p_1p_2)$ = number of sentence pairs where $p_1 \in S_1$ and $p_2 \in S_2$; $c(p_1\neg p_2)$ = number of pairs where $p_1 \in S_1$ and $p_2 \notin S_2$; $c(\neg p_1 p_2)$ and $c(\neg p_1 \neg p_2)$ are computed similarly. Then we perform a chi-square test to understand if the observed count $c(p_1p_2)$ is significantly (95% confidence level) greater or lesser than the expected value if occurrences of p_1 and p_2 were independent.

Of the 38,809 production pairs, we found that 1,168 pairs occurred in consecutive sentences significantly more often than chance and 172 appeared significantly fewer times than expected. In Table 2 we list, grouped in three simple categories, the 25 pairs of the first kind with most significant p-values.

Some of the preferred pairs are indeed repetitions as pointed out by prior work. But they form only a small fraction (5%) of the total preferred production pairs indicating that there are several other classes of syntactic regularities beyond priming. Some of these other sequences can be explained by the fact that these articles come from the finance domain: they involve productions containing numbers and quantities. An example for this type is shown in Table 2. Finally, there is also a class that is not repetitions or readily observed as domain-specific. The most frequent one reflects a pattern where the first sentence introduces a subject and predicate and the subject in the second sentence is pronominalized. Examples for two other patterns are given in Table 2. For the sequence (VP → VB VP | NP-SBJ → NNP NNP), a bare verb is present in S_1 and is often associated with modals. In the corpus, these statements often present hypothesis or speculation. The following sentence S_2 has an entity, a person or organization, giving an explanation or opinion on the statement. This pattern roughly correponds to a SPECULATE followed by ENDORSE sequence of intentions.

p_1	p_2	$c(p_1p_2)$
	— Repetition —	
VP → VBD SBAR	VP → VBD SBAR	83
QP → $ CD CD	QP → $ CD CD	18
NP → $ CD -NONE-	NP → $ CD -NONE-	16
NP → QP -NONE-	NP → QP -NONE-	15
NP-ADV → DT NN	NP-ADV → DT NN	10
NP → NP NP-ADV	NP → NP NP-ADV	7
	— Quantities/Amounts —	
NP → QP -NONE-	QP → $ CD CD	16
QP → $ CD CD	NP → QP -NONE-	15
NP → NP NP-ADV	NP → QP -NONE-	11
NP-ADV → DT NN	NP → QP -NONE-	11
NP → NP NP-ADV	NP-ADV → DT NN	9
NP → $ CD -NONE-	NP-ADV → DT NN	8
NP-ADV → DT NN	NP → $ CD -NONE-	8
NP-ADV → DT NN	NP → NP NP-ADV	8
NP → NP NP-ADV	QP → CD CD	6
	— Other —	
S → NP-SBJ VP	NP-SBJ → PRP	290
VP → VBD SBAR	PP-TMP → IN NP	79
S → NP-SBJ-1 VP	VP → VBD SBAR	43
VP → VBD NP	VP → VBD VP	31
VP → VB VP	NP-SBJ → NNP NNP	27
NP-SBJ-1 → NNP NNP	VP → VBD NP	13
VP → VBZ NP	S → PP-TMP , NP-SBJ VP .	8
NP-SBJ → JJ NNS	VP → VBP NP	8
NP-PRD → NP PP	NP-PRD → NP SBAR	7
NP-LOC → NNP	S-TPC-1 → NP-SBJ VP	6

Table 3: Top patterns in productions from WSJ

Similarly, in all the six adjacent sentence pairs from our corpus containing the items (NP-LOC → NNP | S-TPC-1 → NP-SBJ VP), p_1 introduces a location name, and is often associated with the title of a person or organization. The next sentence has a quote from that person, where the quotation forms the topicalized clause in p_2. Here the intentional structure is INTRODUCE X / STATEMENT BY X.

In the remainder of the paper we formalize our representation of syntax and the derived model of coherence and test its efficacy in three domains.

3 Coherence models using syntax

We first describe the two representations of sentence structure we adopted for our analysis.[3] Next, we

[3]Our representations are similar to features used for reranking in parsing. Our first representation corresponds to "rules" features (Charniak and Johnson, 2005; Collins and Koo, 2005), and our second representation is related to "spines" (Carreras et al., 2008) and edge annotation(Huang, 2008).

present two coherence models: a local model which captures the co-occurrence of structural features in adjacent sentences and a global one which learns from clusters of sentences with similar syntax.

3.1 Representing syntax

Our models rely exclusively on syntactic cues. We derive representations from constituent parses of the sentences, and terminals (words) are removed from the parse tree before any processing is done. The leaf nodes in our parse trees are part of speech tags.

Productions: In this representation we view each sentence as the set of grammatical productions, LHS → RHS, which appear in the parse of the sentence. As we already pointed out, the right-hand side (RHS) contains only non-terminal nodes. This representation is straightforward, however, some productions can be rather specific with long right hand sides. Another apparent drawback of this representation is that it contains sequence information only about nodes that belong to the same constituent.

***d*-sequence:** In this representation we aim to preserve more sequence information about adjacent constituents in the sentence. The simplest approach would be to represent the sentence as the sequence of part of speech (POS) tags but then we lose all the abstraction provided by higher level nodes in tree. Instead, we introduce a more general representation, *d-sequence* where the level of abstraction can be controlled using a parameter d. The parse tree is truncated to depth at most d, and the leaves of the resulting tree listed left to right form the d-sequence representation. For example, in Figure 1, the line depicts the cutoff at depth 2.

Next the representation is further augmented; all *phrasal* nodes in the d-sequence are annotated (concatenated) with the left-most leaf that they dominate in the full non-lexicalized parse tree. This is shown as suffixes on the S, NP and VP nodes in the figure. Such annotation conveys richer information about the structure of the subtree below nodes in the d-sequence. For example, "the chairs", "his chairs", "comfortable chairs" will be represented as NP_{DT}, $\text{NP}_{\text{PRP\$}}$ and NP_{JJ}. In the resulting representations, sentences are viewed as sequences of *syntactic words* (w_1,w_2...,w_k), $k \leq p$, where p is the length of the full POS sequence and each w_i is either POS tag or a phrasal node+POS tag combination.

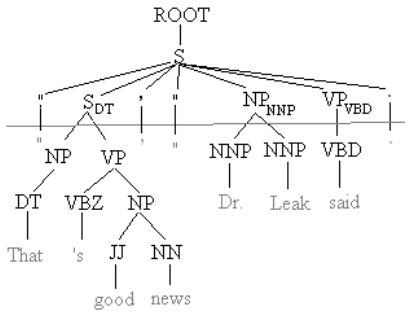

Figure 1: Example for d-sequence representation

In our example, at depth-2, the quotation sentence gets the representation (w_1=" , w_2=S_{DT} , w_3=, , w_4=" , w_5=NP_{NNP} , w_6=VP_{VBD} , w_7=.) where the actual quote is omitted. Sentences that contain attributions are likely to appear more similar to each other when compared using this representation in contrast to representations derived from word or POS sequence. The depth-3 sequence is also indicated in the figure.

The main verb of a sentence is central to its structure, so the parameter d is always set to be greater than that of the main verb and is tuned to optimize performance for coherence prediction.

3.2 Implementing the model

We adapt two models of coherence to operate over the two syntactic representations.

3.2.1 Local co-occurrence model

This model is a direct extension from our pilot study. It allows us to test the assumption that coherent discourse is characterized by syntactic regularities in adjacent sentences. We estimate the probabilities of pairs of syntactic items from adjacent sentences in the training data and use these probabilities to compute the coherence of new texts.

The coherence of a text T containing n sentences (S_1...S_n) is computed as:

$$P(T) = \prod_{i=2}^{n} \prod_{j=1}^{|S_i|} \frac{1}{|S_{i-1}|} \sum_{k=1}^{|S_{i-1}|} p(S_i^j | S_{i-1}^k)$$

where S_x^y indicates the y^{th} item of S_x. Items are either productions or syntactic word unigrams depending on the representation. The conditional probabilities are computed with smoothing:

Cluster a	Cluster b
ADJP → JJ PP | VP → VBZ ADJP	VP → VB VP | VP → MD VP
[1] This method VP-[is ADJP-[capable of sequence-specific detection of DNA with high accuracy]-ADJP]-VP .	[1] Our results for the difference in reactivity VP-[can VP-[be linked to experimental observations]-VP]-VP .
[2] The same VP-[is ADJP-[true for synthetic polyamines such as polyallylamine]-ADJP]-VP .	[2] These phenomena taken together VP-[can VP-[be considered as the signature of the gelation process]-VP]-VP .

Table 4: Example syntactic similarity clusters. The top two descriptive productions for each cluster are also listed.

$$p(w_j|w_i) = \frac{c(w_i, w_j) + \delta_C}{c(w_i) + \delta_C * |V|}$$

where w_i and w_j are syntactic items and $c(w_i, w_j)$ is the number of sentences that contain the item w_i immediately followed by a sentence that contains w_j. $|V|$ is the vocabulary size for syntactic items.

3.2.2 Global structure

Now we turn to a global coherence approach that implements the assumption that sentences with similar syntax have the same communicative goal as well as captures the patterns in communicative goals in the discourse. This approach uses a Hidden Markov Model (HMM) which has been a popular implementation for modeling coherence (Barzilay and Lee, 2004; Fung and Ngai, 2006; Elsner et al., 2007). The hidden states in our model depict communicative goals by encoding a probability distribution over syntactic items. This distribution gives higher weight to syntactic items that are more likely for that communicative goal. Transitions between states record the common patterns in intentional structure for the domain.

In this syntax-HMM, states h_k are created by clustering the sentences from the documents in the training set by *syntactic similarity*. For the productions representation of syntax, the features for clustering are the number of times a given production appeared in the parse of the sentence. For the d-sequence approach, the features are n-grams of size one to four of syntactic words from the sequence. Clustering was done by optimizing for average cosine similarity and was implemented using the CLUTO toolkit (Zhao et al., 2005). C clusters are formed and taken as the states of the model. Table 4 shows sentences from two clusters formed on the abstracts of journal articles using the productions representation. One of them, cluster (a), appears to capture descriptive sentences and cluster (b) involves mostly speculation type sentences.

The emission probabilities for each state are modeled as a (syntactic) language model derived from the sentences in it. For productions representation, this is the unigram distribution of productions from the sentences in h_k. For d-sequences, the distribution is computed for bigrams of syntactic words. These language models use Lidstone smoothing with constant δ_E. The probability for a sentence S_l to be generated from state h_k, $p_E(S_l|h_k)$ is computed using these syntactic language models.

The transition probability p_M from a state h_i to state h_j is computed as:

$$p_M(h_j|h_i) = \frac{d(h_i, h_j) + \delta_M}{d(h_i) + \delta_M * C}$$

where $d(h_i)$ is the number of documents whose sentences appear in h_i and $d(h_i, h_j)$ is the number of documents which have a sentence in h_i which is immediately followed by a sentence in h_j. In addition to the C states, we add one initial h_S and one final h_F state to capture document beginning and end. Transitions from h_S to any state h_k records how likely it is for h_k to be the starting state for documents of that domain. δ_M is a smoothing constant.

The likelihood of a text with n sentences is given by $P(T) = \sum_{h_1 \ldots h_n} \prod_{t=1}^{n} p_M(h_t|h_{t-1}) p_E(S_t|h_t)$.

All model parameters—the number of clusters C, smoothing constants δ_C, δ_E, δ_M and d for d-sequences—are tuned to optimize how well the model can distinguish coherent from incoherent articles. We describe these settings in Section 5.1.

4 Content and entity grid models

We compare the syntax model with content model and entity grid methods. These approaches are the most popular ones from prior work and also allow us to test the complementary nature of syntax with

lexical statistics and entity structure. This section explains how we implemented these approaches.

Content models introduced by Barzilay and Lee (2004) and Fung and Ngai (2006) use lexically driven HMMs to capture coherence. The hidden states represent the topics of the domain and encode a probability distribution over words. Transitions between states record the probable succession of topics. We built a content model using our HMM implementation. Clusters are created using word bigram features after replacing numbers and proper names with tags NUM and PROP. The emissions are given by a bigram language model on words from the clustered sentences. Barzilay and Lee (2004) also employ an iterative clustering procedure before finalizing the states of the HMM but our method only uses one-step clustering. Despite the difference, the content model accuracies for our implementation are quite close to that from the original.

For the entity grid model, we follow the generative approach proposed by Lapata and Barzilay (2005). A text is converted into a matrix, where rows correspond to sentences, in the order in which they appear in the article. Columns are created one for each entity appearing in the text. Each cell (i,j) is filled with the grammatical role $r_{i,j}$ of the entity j in sentence i. We computed the entity grids using the Brown Coherence Toolkit[4]. The probability of the text (T) is defined using the likely sequence of grammatical role transitions.

$$P(T) = \prod_{j=1}^{m} \prod_{i=1}^{n} p(r_{i,j}|r_{i-1,j}...r_{i-h,j})$$

for m entities and n sentences. Parameter h controls the history size for transitions and is tuned during development. When $h = 1$, for example, only the grammatical role for the entity in the previous sentence is considered and earlier roles are ignored.

5 Evaluating syntactic coherence

We follow the common approach from prior work and use pairs of articles, where one has the original document order and the other is a random permutation of the sentences from the same document. Since the original article is always more coherent than a random permutation, a model can be evaluated using the accuracy with which it can identify the original article in the pair, i.e. it assigns higher probability to the original article. This setting is not ideal but has become the de facto standard for evaluation of coherence models (Barzilay and Lee, 2004; Elsner et al., 2007; Barzilay and Lapata, 2008; Karamanis et al., 2009; Lin et al., 2011; Elsner and Charniak, 2011). It is however based on a reasonable assumption as recent work (Lin et al., 2011) shows that people identify the original article as more coherent than its permutations with over 90% accuracy and assessors also have high agreement. Later, we present an experiment distinguishing conference from workshop articles as a more realistic evaluation.

We use two corpora that are widely employed for coherence prediction (Barzilay and Lee, 2004; Elsner et al., 2007; Barzilay and Lapata, 2008; Lin et al., 2011). One contains reports on airplane accidents from the National Transportation Safety Board and the other has reports about earthquakes from the Associated Press. These articles are about 10 sentences long. These corpora were chosen since within each dataset, the articles have the same intentional structure. Further, these corpora are also standard ones used in prior work on lexical, entity and discourse relation based coherence models. Later in Section 6, we show that the models perform well on the academic genre and longer articles too.

For each of the two corpora, we have 100 articles for training and 100 (accidents) and 99 (earthquakes) for testing. A maximum of 20 random permutations were generated for each test article to create the pairwise data (total of 1986 test pairs for the accident corpus and 1956 for earthquakes).[5] The baseline accuracy for random prediction is 50%. The articles were parsed using the Stanford parser (Klein and Manning, 2003).

5.1 Accuracy of the syntax model

For each model, the relevant parameters were tuned using 10-fold cross validation on the training data. In each fold, 90 documents were used for training and evaluation was done on permutations from the remaining articles. After tuning, the final model was trained on all 100 articles in the training set.

[4] http://www.cs.brown.edu/~melsner/manual.html

[5] We downloaded the permutations from `http://people.csail.mit.edu/regina/coherence/CLsubmission/`

Table 5 shows the results on the test set. The best number of clusters and depth for d-sequences are also indicated. Overall, the syntax models work quite well, with accuracies at least 15% or more absolute improvement over the baseline.

In the local co-occurrence approach, both productions and d-sequences provide 72% accuracy for the accidents corpus. For the earthquake corpus, the accuracies are lower and the d-sequence method works better. The best depth setting for d-sequence is rather small: depth of main verb (MVP) + 2 (or 1), and indicates that a fairly abstract level of nodes is preferred for the patterns. For comparison, we also provide results using just the POS tags in the model and this is worse than the d-sequence approach.

The global HMM model is better than the local model for each representation type giving 2 to 38% better accuracies. Here we see a different trend for the d-sequence representation, with better results for greater depths. At such depths (8 and 9) below the main verb, the nodes are mostly POS tags.

Overall both productions and d-sequence work competitively and give the best accuracies when implemented with the global approach.

5.2 Comparison with other approaches

For our implementations of the content and entity grid models, the best accuracies are 71% on the accidents corpus and 85% on the earthquakes one, similar to the syntactic models.

Ideally, we would like to combine models but we do not have separate training data. So we perform the following classification experiment which combines the predictions made by different models on the *test set*. Each test pair (article and permutation) forms one example and is given a class value of 0 or 1 depending on whether the first article in the pair is the original one or the second one. The example is represented as an n-dimensional vector, where n is the number of models we wish to combine. For instance, to combine content models and entity grid, two features are created: one of these records the difference in log probabilities for the two articles from the content model, the other feature indicates the difference in probabilities from the entity grid.

A logistic regression classifier is trained to predict the class using these features. The test pairs are created such that an equal number of examples have class 0 and 1, so the baseline accuracy is 50%. We run this experiment using 10-fold cross validation on the test set after first obtaining the log probabilities from individual models. In each fold, the training is done using the pairs from 90 articles and tested on permutations from the remaining 10 articles. These accuracies are reported in Table 6. When the accuracy of a combination is better than that using any of its smaller subsets, the value is bolded.

We find that syntax supplements both content and entity grid methods. While on the airplane corpus syntax only combines well with the entity grid, on the earthquake corpus, both entity and content approaches give better accuracies when combined with syntax. However, adding all three approaches does not outperform combinations of any two of them. This result can be due to the simple approach that we tested for combination. In prior work, content and entity grid methods have been combined generatively (Elsner et al., 2007) and using discriminative training with different objectives (Soricut and

Model	**Accidents**		**Earthquake**	
	Parameter	**Acc**	**Parameter**	**Acc**
A. Local co-occurrence				
Prodns		72.8		55.0
d-seq	dep. MVP+2	71.8	dep. MVP+1	65.1
POS		61.3		42.6
B. HMM-syntax				
Prodns	clus. 37	74.6	clus. 5	93.8
d-seq	dep. MVP+8	82.2	dep. MVP+9	86.5
	clus. 8		clus. 45	
C. Other approaches				
Egrid	history 1	67.6	history 1	82.2
Content	clus. 48	71.4	clus. 23	84.5

Table 5: Accuracies on accident and earthquake corpora

Model	**Accid.**	**Earthq.**
Content + Egrid	**76.8**	**90.7**
Content + HMM-prodn	74.2	**95.3**
Content + HMM-d-seq	82.1	**90.3**
Egrid + HMM-prodn	**79.6**	93.9
Egrid + HMM-d-seq	**84.2**	**91.1**
Egrid + Content + HMM-prodn	79.5	95.0
Egrid + Content + HMM-d-seq	84.1	92.3
Egrid + Content + HMM-prodn + HMM-d-seq	83.6	95.7

Table 6: Accuracies for combined approaches

Marcu, 2006). Such approaches might bring out the complementary strengths of the different aspects better and we leave such analysis for future work.

6 Predictions on academic articles

The distinctive intentional structure of academic articles has motivated several proposals to define and annotate the communicative purpose (argumentative zone) of each sentence (Swales, 1990; Teufel et al., 1999; Liakata et al., 2010). Supervised classifiers were also built to identify these zones (Teufel and Moens, 2000; Guo et al., 2011). So we expect that these articles form a good testbed for our models. In the remainder of the paper, we examine how unsupervised patterns discovered by our approach relate to zones and how well our models predict coherence for articles from this genre.

We employ two corpora of scientific articles.

ART Corpus: contains a set of 225 Chemistry journal articles that were manually annotated for intentional structure (Liakata and Soldatova, 2008). Each sentence was assigned one of 11 zone labels: *Result*, *Conclusion*, *Objective*, *Method*, *Goal*, *Background*, *Observation*, *Experiment*, *Motivation*, *Model*, *Hypothesis*. For our study, we use the annotation of the introduction and the abstract sections. We divide the data into training, development and test sets. For abstracts, we have 75, 50 and 100 for these sets respectively. For introductions, this split is 75, 31, 82.[6]

ACL Anthology Network (AAN) Corpus: Radev et al. (2009) provides the full text of publications from ACL venues. These articles do not have any zone annotations. The AAN corpus is produced from OCR analysis and no section marking is available. To recreate these, we use the Parscit tagger[7] (Councill et al., 2008). We use articles from years 1999 to 2011. For training, we randomly choose 70 articles from ACL and NAACL main conferences. Similarly, we obtain a development corpus of 36 ACL-NAACL articles. We create two test sets: one has 500 ACL-NAACL conference articles and another has 500 articles from ACL-sponsored workshops. We only choose articles in which all three sections—abstract, introduction and related work—could be successfully identified using Parscit.[8]

This data was sentence-segmented using MxTerminator (Reynar and Ratnaparkhi, 1997) and parsed with the Stanford Parser (Klein and Manning, 2003).

For each corpus and each section, we train all our syntactic models: the two local coherence models using the production and d-sequence representations and the HMM models with the two representations. These models are tuned on the respective development data, on the task of differentiating the original from a permuted section. For this purpose, we created a maximum of 30 permutations per article.

6.1 Comparison with ART Corpus zones

We perform this analysis using the ART corpus. The zone annotations present in this corpus allow us to directly test our first assumption in this work, that sentences with similar syntax have the same communicative goal.

For this analysis, we use the the HMM-prod model for abstracts and the HMM-d-seq model for introductions. These models were chosen because they gave the best performance on the ART corpus development sets.[9] We examine the clusters created by these models on the training data and check whether there are clusters which strongly involve sentences from some particular annotated zone.

For each possible pair of cluster and zone (C_i, Z_j), we compute $c(C_i, Z_j)$: the number of sentences in C_i that are annotated as zone Z_j. Then we use a chi-square test to identify pairs for which $c(C_i, Z_j)$ is significantly greater than expected (there is a "positive" association between C_i and Z_j) and pairs where $c(C_i, Z_j)$ is significantly less than chance (C_i is not associated with Z_j). A 95% confidence level was used to determine significance.

The HMM-prod model for abstracts has 9 clusters (named Clus0 to 8) and the HMM-d-seq model for introductions has 6 clusters (Clus0 to 5). The pairings of these clusters with zones which turned out to be significant are reported in Table 7. We also report for each positively associated cluster-zone pair, the following numbers: matches $c(C_i, Z_j)$, precision $c(C_i, Z_j)/|C_i|$ and recall $c(C_i, Z_j)/|Z_j|$.

[6] Some articles did not have labelled 'introduction' sections resulting in fewer examples for this setup.

[7] http://aye.comp.nus.edu.sg/parsCit/

[8] We also exclude introduction and related work sections longer than 50 sentences and those shorter than 4 sentences since they often have inaccurate section boundaries.

[9] Their test accuracies are reported in the next section.

Abstracts (HMM-prod 9 clusters)			
Positive associations	**matches**	**prec.**	**recall**
Clus5 - Model	7	17.1	43.8
Clus7 - Objective	27	27.6	32.9
Clus7 - Goal	16	16.3	55.2
Clus0 - Conclusion	15	50.0	12.1
Clus6 - Conclusion	27	51.9	21.8
Not associated: Clus7 - Conclusion, Clus8 - Conclusion			
Introductions (HMM-d-seq 6 clusters)			
Positive associations	**matches**	**prec.**	**recall**
Clus2-Background	161	64.9	14.2
Clus3-Objective	37	7.9	38.5
Clus4-Goal	29	9.8	32.6
Clus4-Hypothesis	12	4.1	52.2
Clus5-Motivation	61	12.9	37.4
Not associated: Clus1 - Motivation, Clus2 - Goal, Clus4 - Background, Clus 5 - Model			

Table 7: Cluster-Zone mappings on the ART Corpus

The presence of significant associations validate our intuitions that syntax provides clues about communicative goals. Some clusters overwhelmingly contain the same zone, indicated by high precision, for example 64% of sentences in Clus2 from introduction sections are background sentences. Other clusters have high recall of a zone, 55% of all goal sentences from the abstracts training data is captured by Clus7. It is particularly interesting to see that Clus7 of abstracts captures both objective and goal zone sentences and for introductions, Clus4 is a mix of hypothesis and goal sentences which intuitively are closely related categories.

6.2 Original versus permuted sections

We also explore the accuracy of the syntax models for predicting coherence of articles from the test set of ART corpus and the 500 test articles from ACL-NAACL conferences. We use the same experimental setup as before and create pairs of original and permuted versions of the test articles. We created a maximum of 20 permutations for each article. The baseline accuracy is 50% as before.

For the ART corpus, we also built an oracle model of annotated zones. We train a first order Markov Chain to record the sequence of zones in the training articles. For testing, we assume that the oracle zone is provided for each sentence and use the model to predict the likelihood of the zone sequence. Results from this model represent an upper bound because an accurate hypothesis of the communicative goal is available for each sentence.

The accuracies are presented in Table 8. Overall, the HMM-d-seq model provides the best accuracies. The highest results are obtained for ACL introduction sections (74%). These results are lower than that obtained on the earthquake/accident corpus but the task here is much harder: the articles are longer and the ACL corpus also has OCR errors which affect sentence segmentation and parsing accuracies. When the oracle zones are known, the accuracies are much higher on the ART corpus indicating that the intentional structure of academic articles is very predictive of their coherence.

6.3 Conference versus workshop papers

Finally, we test whether the syntax-based model can distinguish the structure of conference from workshop articles. Conferences publish more complete and tested work and workshops often present preliminary studies. Workshops are also venues to discuss a focused and specialized topic. So the way information is conveyed in the abstracts and introductions would vary in these articles.

We perform this analysis on the ACL corpus and no permutations are used, only the original text of the 500 articles each in the conference and workshop test sets. While permutation examples provide cheap training/test data, they have a few unrealistic properties. For example, both original and permuted articles have the same length. Further some permutations could result in an outstandingly incoherent sample which is easily distinguished from the original articles. So we use the conference versus workshop task as another evaluation of our model.

We designed a classification experiment for this task which combines features from the different syntax models that were trained on the ACL conference training set. We include four features indicating the perplexity of an article under each model (Local-prod, Local-d-seq, HMM-prod, HMM-d-seq). We use perplexity rather than probability because the length of the articles vary widely in contrast to the previous permutation-based tests, where both permutation and original article have the same length. We compute perplexity as $P(T)^{-1/n}$, where n is the number of words in the article. We also obtain the most likely state sequence for the article under

Data	Section	Test pairs	Local-prod	Local-d-seq	HMM-prod	HMM-d-seq	Oracle zones
ART Corpus	Abstract	1633	57.0	52.9	**64.1**	55.0	80.8
	Intro	1640	44.5	54.6	58.1	**64.6**	94.0
ACL Conference	Abstract	8815	44.0	47.2	58.2	**63.7**	
	Intro	9966	54.5	53.0	64.4	**74.0**	
	Rel. wk.	10,000	54.6	54.4	57.3	**67.3**	

Table 8: Accuracy in differentiating permutation from original sections on ACL and ART test sets.

HMM-prod and HMM-d-seq models using Viterbi decoding. Then the proportion of sentences from each state of the two models are added as features.

We also add some fine-grained features from the local model. We represent sentences in the training set as either productions or d-sequence items and compute pairs of associated items (x_i, x_j) from adjacent sentences using the same chi-square test as in our pilot study. The most significant (lowest p-values) 30 pairs (each for production and d-seq) are taken as features.[10] For a test article, we compute features that represent how often each pair is present in the article such that x_i is in S_m and x_j is in S_{m+1}.

We perform this experiment for each section and there are about 90 to 140 features for the different sections. We cast the problem as a binary classification task: conference articles belong to one class and workshop to the other. Each class has 500 articles and so the baseline random accuracy is 50%. We perform 10-fold cross validation using logistic regression. Our results were 59.3% accuracy for distinguishing abstracts of conference verus workshop, 50.3% for introductions and 55.4% for related work. For abstracts and related work, these accuracies are significantly better than baseline (95% confidence level from a two-sided paired t-test comparing the accuracies from the 10 folds). It is possible that introductions in either case, talk in general about the field and importance of the problem addressed and hence have similar structure.

Our accuracies are not as high as on permutations examples because the task is clearly harder. It may also be the case that the prediction is more difficult for certain papers than for others. So we also analyze our results by the confidence provided by the classifier for the predicted class. We consider only the examples predicted above a certain confidence level and compute the accuracy on these predictions.

[10] A cutoff is applied such that the pair was seen at least 25 times in the training data.

Conf.	Abstract	Intro	Rel wk
$>= 0.5$	59.3 (100.0)	50.3 (100.0)	55.4 (100.0)
$>= 0.6$	63.8 (67.2)	50.8 (71.1)	58.6 (75.9)
$>= 0.7$	67.2 (32.0)	54.4 (38.6)	63.3 (52.8)
$>= 0.8$	74.0 (10.0)	51.6 (22.0)	63.0 (25.7)
$>= 0.9$	91.7 (2.0)	30.6 (5.0)	68.1 (7.2)

Table 9: Accuracy (% examples) above each confidence level for the conference versus workshop task.

These results are shown in Table 9. The proportion of examples under each setting is also indicated.

When only examples above 0.6 confidence are examined, the classifier has a higher accuracy of 63.8% for abstracts and covers close to 70% of the examples. Similarly, when a cutoff of 0.7 is applied to the confidence for predicting related work sections, we achieve 63.3% accuracy for 53% of examples. So we can consider that 30 to 47% of the examples in the two sections respectively are harder to tell apart. Interestingly however even high confidence predictions on introductions remain incorrect.

These results show that our model can successfully distinguish the structure of articles beyond just clearly incoherent permutation examples.

7 Conclusion

Our work is the first to develop an unsupervised model for intentional structure and to show that it has good accuracy for coherence prediction and also complements entity and lexical structure of discourse. This result raises interesting questions about how patterns captured by these different coherence metrics vary and how they can be combined usefully for predicting coherence. We plan to explore these ideas in future work. We also want to analyze genre differences to understand if the strength of these coherence dimensions varies with genre.

Acknowledgements

This work is partially supported by a Google research grant and NSF CAREER 0953445 award.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of NAACL-HLT*, pages 113–120.

Xavier Carreras, Michael Collins, and Terry Koo. 2008. Tag, dynamic programming, and the perceptron for efficient, feature-rich parsing. In *Proceedings of CoNLL*, pages 9–16.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of ACL*, pages 173–180.

Jackie C.K. Cheung and Gerald Penn. 2010. Utilizing extra-sentential context for parsing. In *Proceedings of EMNLP*, pages 23–33.

Christelle Cocco, Raphaël Pittier, François Bavaud, and Aris Xanthos. 2011. Segmentation and clustering of textual sequences: a typological approach. In *Proceedings of RANLP*, pages 427–433.

Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31:25–70.

Isaac G. Councill, C. Lee Giles, and Min-Yen Kan. 2008. Parscit: An open-source crf reference string parsing package. In *Proceedings of LREC*, pages 661–667.

Micha Elsner and Eugene Charniak. 2008. Coreference-inspired coherence modeling. In *Proceedings of ACL-HLT, Short Papers*, pages 41–44.

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of ACL-HLT*, pages 125–129.

Micha Elsner, Joseph Austerweil, and Eugene Charniak. 2007. A unified local and global model for discourse coherence. In *Proceedings of NAACL-HLT*, pages 436–443.

Pascale Fung and Grace Ngai. 2006. One story, one flow: Hidden markov story models for multilingual multidocument summarization. *ACM Transactions on Speech and Language Processing*, 3(2):1–16.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 3(12):175–204.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of EMNLP*, pages 273–283.

Liang Huang. 2008. Forest reranking: Discriminative parsing with non-local features. In *Proceedings of ACL-HLT*, pages 586–594, June.

Nikiforos Karamanis, Chris Mellish, Massimo Poesio, and Jon Oberlander. 2009. Evaluating centering for information ordering using corpora. *Computational Linguistics*, 35(1):29–46.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*, pages 423–430.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *Proceedings of IJCAI*.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL*, pages 545–552.

Maria Liakata and Larisa Soldatova. 2008. Guidelines for the annotation of general scientific concepts. *JISC Project Report*.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of LREC*.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of EMNLP*, pages 343–351.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of ACL-HLT*, pages 997–1006.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of EMNLP*, pages 186–195.

Dragomir R. Radev, Mark Thomas Joseph, Bryan Gibson, and Pradeep Muthukrishnan. 2009. A Bibliometric and Network Analysis of the field of Computational Linguistics. *Journal of the American Society for Information Science and Technology*.

David Reitter, Johanna D. Moore, and Frank Keller. 2006. Priming of Syntactic Rules in Task-Oriented Dialogue and Spontaneous Conversation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 685–690.

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the fifth conference on Applied natural language processing*, pages 16–19.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of COLING-ACL*, pages 803–810.

John Swales. 1990. *Genre analysis: English in academic and research settings*, volume 11. Cambridge University Press.

Simone Teufel and Marc Moens. 2000. What's yours and what's mine: determining intellectual attribution in scientific text. In *Proceedings of EMNLP*, pages 9–17.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of EACL*, pages 110–117.

Ying Zhao, George Karypis, and Usama Fayyad. 2005. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10:141–168.