Unsupervised Declarative Knowledge Induction for Constraint-Based Learning of Information Structure in Scientific Documents

Yufan Guo
DTAL
University of Cambridge, UK
yg244@cam.ac.uk

Roi Reichart
Technion - IIT
Haifa, Israel
roiri@ie.technion.ac.il

Anna Korhonen
DTAL
University of Cambridge, UK
alk23@cam.ac.uk

Abstract

Inferring the information structure of scientific documents is useful for many NLP applications. Existing approaches to this task require substantial human effort. We propose a framework for constraint learning that reduces human involvement considerably. Our model uses topic models to identify latent topics and their key linguistic features in input documents, induces constraints from this information and maps sentences to their dominant information structure categories through a constrained unsupervised model. When the induced constraints are combined with a fully unsupervised model, the resulting model challenges existing lightly supervised feature-based models as well as unsupervised models that use manually constructed declarative knowledge. Our results demonstrate that useful declarative knowledge can be learned from data with very limited human involvement.

1 Introduction

Automatic analysis of scientific text can help scientists find information from literature faster, saving valuable research time. In this paper we focus on the analysis of the *information structure (IS)* of scientific articles where the aim is to assign each unit of an article (typically a sentence) into a category that represents the information type it conveys. By information structure we refer to a particular type of discourse structure that focuses on the functional role of a unit in the discourse (Webber et al., 2011). For instance, in the scientific literature, the functional role of a sentence could be the background or motivation of the research, the methods used, the experiments carried out, the observations on the results, or the author's conclusions.

Readers of scientific literature find information in IS-annotated articles much faster than in unannotated articles (Guo et al., 2011b). Argumentative Zoning (AZ) – an information structure scheme that has been applied successfully to many scientific domains (Teufel et al., 2009) – has improved tasks such as summarization and information extraction and retrieval (Teufel and Moens, 2002; Tbahriti et al., 2006; Ruch et al., 2007; Liakata et al., 2012; Contractor et al., 2012).

Existing approaches to information structure analysis require substantial human effort. Most use feature-based machine learning, such as SVMs and CRFs (e.g. (Teufel and Moens, 2002; Lin et al., 2006; Hirohata et al., 2008; Shatkay et al., 2008; Guo et al., 2010; Liakata et al., 2012)) which rely on thousands of manually annotated training sentences. Also the performance of such methods is rather limited: Liakata et al. (2012) reported per-class F-scores ranging from .53 to .76 in the biochemistry and chemistry domains and Guo et al. (2013a) reported substantially lower numbers for the challenging *Introduction* and *Discussion* sections in biomedical domain.

Guo et al. (2013a) recently applied the Generalized Expectation (GE) criterion (Mann and McCallum, 2007) to information structure analysis using expert knowledge in the form of discourse and lexical constraints. Their model produces promising results, especially for sections and categories where

Transactions of the Association for Computational Linguistics, vol. 3, pp. 131–143, 2015. Action Editor: Masaaki Nagata.
Submission batch: 10/2014; Revision batch 1/2015; Published 3/2015.

feature-based models perform poorly. Even the unsupervised version which uses constraints under a maximum-entropy criterion without any feature-based model, outperforms fully-supervised feature-based models in detecting challenging low frequency categories across sections. However, this approach still requires substantial human effort in constraint generation. Particularly, lexical constraints were constructed by creating a detailed word list for each information structure category. For example, words such as "assay" were carefully selected and used as a strong indicator of the "Method" category: $p(\text{Method}|\text{assay})$ was constrained to be high (above 0.9). Such a constraint (developed for the biomedical domain) may not be applicable to a new domain (e.g. computer science) with a different vocabulary and writing style.

In fact, most existing works on learning with declarative knowledge rely on manually constructed constraints. Little work exists on automatic declarative knowledge induction. A notable exception is (McClosky and Manning, 2012) that proposed a constraint learning model for timeline extraction. This approach, however, requires human supervision in several forms including task specific constraint templates (see Section 2).

We present a novel framework for learning declarative knowledge which requires very limited human involvement. We apply it to information structure analysis, based on two key observations: 1) Each information structure category defines a distribution over a section-specific and an article-level set of linguistic features. 2) Each sentence in a scientific document, while having a dominant category, may consist of features mostly related to other categories. This flexible view enables us to make use of topic models which have not proved useful in previous related works (Varga et al., 2012; Reichart and Korhonen, 2012).

We construct topic models at both the individual section and article level and apply these models to data, identifying latent topics and their key linguistic features. This information is used to constrain or bias unsupervised models for the task in a straightforward way: we automatically generate constraints for a GE model and a bias term for a graph clustering objective, such that the resulting models assign each of the input sentences to one information structure category. Both models provide high quality sentence-based classification, demonstrating the generality of our approach.

Zone	Definition
Background (BKG)	the background of the study
Problem (PROB)	the research problem
Method (METH)	the methods used
Result (RES)	the results achieved
Conclusion (CON)	the authors' conclusions
Connection (CN)	work consistent with the current work
Difference (DIFF)	work inconsistent with the current work
Future work (FUT)	the potential future direction of the research

Table 1: The AZ categorization scheme of this paper

We experiment with the AZ scheme for the analysis of the logical structure, scientific argumentation and intellectual attribution of scientific papers (Teufel and Moens, 2002), using an eight-category version of this scheme for biomedicine ((Mizuta et al., 2006; Guo et al., 2013b), Table 1). In evaluation against gold standard annotations, our model rivals the model of Guo et al. (2013a) which relies on manually constructed constraints, as well as a strong supervised feature-based model trained with up to 2000 sentences. In task-based evaluation we measure the usefulness of the induced categories for customized summarization (Contractor et al., 2012) from specific types of information in an article. The AZ categories induced by our model prove more valuable than those of (Guo et al., 2013a) and those in the gold standard. Our work demonstrates the great potential of *automatically induced* declarative knowledge in both improving the performance of information structure analysis and reducing reliance of human supervision.

2 Previous Work

Automatic Declarative Knowledge Induction Learning with declarative knowledge offers effective means of reducing human supervision and improving performance. This framework augments feature-based models with domain and expert knowledge in the form of, e.g., linear constraints, posterior probabilities and logical formulas (e.g. (Chang et al., 2007; Mann and McCallum, 2007; Mann and McCallum, 2008; Ganchev et al., 2010)). It has proven useful for many NLP tasks including unsupervised and semi-supervised POS tagging, parsing (Druck et al., 2008; Ganchev et al., 2010; Rush et al., 2012) and information extraction (Chang et al.,

2007; Mann and McCallum, 2008; Reichart and Korhonen, 2012; Reichart and Barzilay, 2012).

However, declarative knowledge is still created in a costly manual process. We propose inducing such knowledge directly from text with minimal human involvement. This idea could be applied to almost any NLP task. We apply it here to information structure analysis of scientific documents.

Little prior work exists on automatic constraint learning. Recently, (McClosky and Manning, 2012) investigated the approach for timeline extraction. They used a set of gold relations and their temporal spans and applied distant learning to find approximate instances for classifier training. A set of constraint templates specific to temporal learning were also specified. In contrast, we do not use manually specified guidance in constraint learning. Particularly, we construct constraints from latent variables (topics in topic modeling) estimated from raw text rather than applying maximum likelihood estimation over observed variables (fluents and temporal expressions) in labeled data. Our method is therefore less dependent on human supervision. Even more recently, (Anzaroot et al., 2014) presented a supervised dual-decomposition based method, in the context of citation field extraction, which automatically generates large families of constraints and learn their costs with a convex optimization objective during training. Our work is unsupervised, as opposed to their model which requires a manually annotated training corpus for constraint learning.

Information Structure Analysis Various schemes have been proposed for analysing the information structure of scientific documents, in particular the patterns of topics, functions and relations at sentence level. Existing schemes include argumentative zones (Teufel and Moens, 2002; Mizuta et al., 2006; Teufel et al., 2009), discourse structure (Burstein et al., 2003; Webber et al., 2011), qualitative dimensions (Shatkay et al., 2008), scientific claims (Blake, 2009), scientific concepts (Liakata et al., 2010), and information status (Markert et al., 2012), among others. Most previous work on automatic analysis of information structure relies on supervised learning (Teufel and Moens, 2002; Burstein et al., 2003; Mizuta et al., 2006; Shatkay et al., 2008; Guo et al., 2010; Liakata et al., 2012; Markert et al., 2012). Given the prohibitive cost of manual annotation, unsupervised and minimally supervised techniques such as clustering (Kiela et al., 2014) and topic modeling (Varga et al., 2012; Ó Séaghdha and Teufel, 2014) are highly important. However, the performance of such approaches shows a large room for improvement. Our work is specifically aimed at addressing this problem.

Information Structure Learning with Declarative Knowledge Recently, Reichart and Korhonen (2012) and Guo et al. (2013a) developed constrained models that integrate rich linguistic knowledge (e.g. discourse patterns, syntactic features and sentence similarity information) for more reliable unsupervised or transductive learning of information categories in scientific abstracts and articles. Guo et al. (2013a) used detailed lexical constraints developed via human supervision. Whether automatically induced declarative knowledge can rival such manual constraints is a question we address in this work. While Reichart and Korhonen (2012) used more general constraints, their most effective discourse constraints were tailored to scientific abstracts and are less relevant to full papers.

3 Model

We introduce a topic-model based approach to declarative knowledge (DK) acquisition and describe how this knowledge can be applied to two unsupervised models for our task. Section 3.1 describes how topic models are used to induce topics that serve as the main building blocks of our DK. Section 3.2 explains how the resulting topics and their key features are transformed into DK – constraints in the generalized expectation (GE) model and bias functions in a graph clustering algorithm.

3.1 Inducing Information Structure Categories with Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) LDA is a generative process widely used for discovering latent topics in text documents (Blei et al., 2003). It assumes the following generative process for each document:

1. Choose $\theta_i \sim Dirichlet(\alpha), i \in \{1, ..., M\}$

2. Choose $\phi_k \sim Dirichlet(\beta), k \in \{1, ..., K\}$

3. For each word $w_{ij}, j \in \{1, ..., N_i\}$

(a) Choose a topic $z_{ij} \sim Multinomial(\theta_i)$

(b) Choose a word $w_{ij} \sim Multinomial(\phi_{z_{ij}})$,

where θ_i is the distribution of topics in document i, ϕ_k is the distribution of observed features (usually words) for topic k, z_{ij} is the topic of the j-th word in document i, and w_{ij} is the j-th word in document i. A number of inference techniques have been proposed for the parameter estimation of this process, e.g. variational Bayes (Blei et al., 2003) and Gibbs sampling (Griffiths and Steyvers, 2004) which we use in this work.

Topics and Information Structure Categories A key challenge in the application of LDA to information structure analysis is defining the observed features generated by the model. Topics are usually defined to be distributions over all the words in a document, but in our task this can lead to undesired topics. Consider, for example, the following sentences from the *Introduction* section of an article:

- First, exposure to BD-diol via inhalation causes an increase in Hprt mutation frequency in both mice and rats (25).

- Third, BD-diol is a precursor to MI, an important urinary metabolite in humans exposed to BD (19).

In a word-based topic model we can expect that most of the content words in these sentences will be generated by a single topic that can be titled as "BD-diol", or by two different topics related to "mice rat" and "human". However, information structure categories should reflect the role of the sentence in e.g. the discourse or argument structure of the paper. For example, given the AZ scheme both sentences should belong to the background (BKG) category (Table 1). The same requirement applies to the topics induced by the topic models.

Features In applying LDA to AZ, we define topics as distributions over: (a) words of particular syntactic categories; (b) syntactic (POS tag) patterns; and (c) discourse markers (citations, tables and figures). Below we list our features, among which *Pronoun*, *Conjunction*, *Adjective* and *Adverb* are novel and the rest are adapted from (Guo et al., 2013a):

Citation A single feature that aggregates together the various citation formats in scientific articles (e.g. *[20]* or *(Tudek 2007)*).

Table, Figure A single feature representing any references to tables or figures in a sentence.

Verb Verbs are central to the meaning of a sentence. Each of the base forms of the verbs in the corpus is a unique feature.

Pronoun Personal (e.g. "we") and possessive pronouns (e.g. "our") and the following adjectives (as in e.g. "our recent" or "our previous") may indicate the ownership of the work (e.g. the author's own vs. other people's work), which is important for our task. Each of the above words or word combinations is a unique feature.

Conjunction Conjunctions indicate the relationship between different sentences in text. We consider two types of conjunctions: (1) coordinating conjunctions (indicated by the POS tag "CC" in the output of the C&C POS tagger); and (2) saturated clausal modifiers (indicated by the POS tag "IN" and the corresponding grammatical relation "cmod" in the output of the C&C parser). Each word that forms a conjunction according to this definition is a unique feature.

Adjective and Adverb Adjectives provide descriptive information about objects, while adverbs may change or qualify the meaning of verbs or adjectives. Each adverb and adjective that appears in more than 5 articles in the corpus is a unique feature.[1]

Modal, Tense, Voice Previous work has demonstrated a strong correlation between tense, voice, modals and information categories (e.g. (Guo et al., 2011a; Liakata et al., 2012)). These features are indicated by the part-of-speech (POS) tag of verbs. For example, the phrase "may have been investigated" is represented as "may-MD have-VBZ be-VBN verb-VBN".

As a pre-processing step, each sentence in the input corpus was represented with the list of features it consists of. Consider, for example, the following sentence from a *Discussion* section in our data-set:

- In a previous preliminary study we reported that the results of a limited proof of concept human clinical trial using sulindac (1-5%) and hydrogen peroxide (25%) gels applied daily for three weeks on actinic keratoses (AK) involving the upper extremities [27]. Before running the *Discussion* section topic model (see below for the features considered by this model), this sentence is converted to the following representation:

[cite] previous preliminary we limited

The topic models we construct are assumed to gen-

[1] We collapsed adverbs ending with *-ly* into the corresponding adjectives to reduce data sparsity. Verbs were spared the frequency cut-off because rarely occurring verbs are likely to correspond to domain-specific actions that are probably indicative of the METH category.

Model	Features
Article	Verb, Table, Figure, Modal, Tense, Voice
Introduction	Citation, Pronoun, Verb, Modal, Tense, Voice
Discussion	Citation, Pronoun, Conjunction, Adjective, Adverb

Table 2: The features used in the article-level and the section-specific topic models in this paper

erate these features rather than bag-of-words.

Topic Models Construction Looking at the categories in Table 1 it is easy to see that different combinations of the features in topic model generation will be relevant for different category distinctions. For example, personal pronouns are particularly relevant for distinguishing between categories related to current vs. previous works.

Some distinctions between categories are, in turn, more relevant for some sections than for others. For example, the distinction between the background (BKG) and the definition of the research problem (PROB) is important for the *Introduction* section, but less important for the results section. Similarly the distinction between conclusions (CON) and difference from previous work (DIFF) is more relevant for the *Discussion* section than other sections.

We therefore constructed two types of topic models: *section-specific* and *article-level* models, reasoning that some distinctions apply globally at the article level while some apply more locally at the section level. Section-specific models were constructed for the *Introduction* section and for the *Discussion* section.[2] Table 2 presents the features that are used with each topic model.

A key issue in the application of topic models to our task is the definition of the unit of text for which θ_i, the distribution over topics, is drawn from the Dirichlet distribution (step 1 of the algorithm). This choice is data dependent, and the standard choice is the document level. However, for scientific articles the paragraph level is a better choice, because a paragraph contains only a small subset of information structure categories while in a full article categories are more evenly distributed. We therefore adopted the paragraph as our basic unit of text. The section-level and the article-level models are applied to the collection of paragraphs in the specific section across the test set articles or in the entire set of test articles, respectively.

[2]The *Methods* section is less suitable for a section-level topic model as 97.5% of its sentences belong to its dominant category (METH) (Table 3). Preliminary experiments with section-level topic models for the *Methods* and *Results* sections did not lead to improved performance.

3.2 Declarative Knowledge Induction

Most sentence-based information structure analysis approaches associate each sentence with a unique category. However, since the MAP assignment of topics to features associates each sentence with multiple topics, we cannot directly interpret the resulting topics as categories of input sentences.[3]

In this section we present two methods for incorporating the information conveyed by the topic models (see Section 3.1) in unsupervised models. The first method biases a graph clustering algorithm while the second generates constraints that can be used with a GE criterion.

Graph Clustering We use the graph clustering objective of Dhillon et al. (2007) which can be optimized efficiently, without eigenvalues calculations:

$$\max_{\tilde{Y}} \; trace(\tilde{Y}^T W^{-1/2} A W^{-1/2} \tilde{Y})$$

where A is a similarity matrix, W is a diagonal matrix of the weight of each cluster, and $\tilde{Y}$ is an orthonormal matrix, indicating cluster membership, which is proportional to the square root of W.

To make use of topics to bias the graph clustering towards the desired solution, we define the similarity matrix A, whose $(i,j)-th$ entry corresponds to the i-th and j-th test set sentences as follows:

$A(i,j) = f(S_i, S_j) + \gamma g(S_i, S_j, T)$, where

$S_i = \{$All the features extracted from sentence i $\}$

$T = \{T_k | T_k = \{$top N features associated with topic k$\}\}$

$f(S_i, S_j) = |S_i \cap S_j|$

$$g(S_i, S_j, T) = \begin{cases} 1 & \exists x \in S_i \exists y \in S_j \exists k \quad x \in T_k \wedge y \in T_k \\ 0 & Otherwise \end{cases}$$

where T_k consists of the N features that are assigned the maximum probability according to the k-th topic. Under this formulation, the topic model term $g(\cdot)$ is defined to be the indicator of whether two sentences share features associated with the same topic. If this is true, the algorithm is encouraged to assign these sentences to the same cluster.

Generalized Expectation A generalized expectation (GE) criterion is a term in an objective function

[3]Our preliminary experiments demonstrated that assigning the learned topics to the test sentences performs poorly.

that assigns a score to model expectations (Mann and McCallum, 2008; Druck et al., 2008; Bellare et al., 2009). Given a score function $g(\cdot)$, a discriminative model $p_\lambda(y|x)$, a vector of feature functions $\mathbf{f}^*(\cdot)$, and an empirical distribution $\tilde{p}(x)$, the value of a GE criterion is:

$$g(E_{\tilde{p}(x)}[E_{p_\lambda(y|x)}[\mathbf{f}^*(x,y)]])$$

A popular choice of $g(\cdot)$ is a measure of distance (e.g. L^2 norm) between model and reference expectations. The feature functions $\mathbf{f}^*(\cdot)$ and the reference expectations of $\mathbf{f}^*(\cdot)$ are traditionally specified by experts, which provides a way to integrate declarative knowledge into machine learning.

Consider a Maximum Entropy (MaxEnt) model $p_\lambda(y|x) = \frac{1}{Z_\lambda} exp(\lambda \cdot \mathbf{f}(x,y))$, where $\mathbf{f}(\cdot)$ is a vector of feature functions, λ the feature weights, and Z_λ the partition function. The following objective function can be used for training MaxEnt with GE criteria on unlabeled data:

$$\max_\lambda -g(E_{\tilde{p}(x)}[E_{p_\lambda(y|x)}[\mathbf{f}^*(x,y)]]) - \sum_j \frac{\lambda_j^2}{2\sigma^2}$$

where the second term is a zero-mean σ^2-variance Gaussian prior on parameters.

Let the k-th feature function $f_k^*(\cdot)$ be an indicator function:

$$f_k^*(x,y) = \mathbb{1}_{\{x_{i_k}=1 \wedge y=y_k\}}(x,y)$$

where x_{i_k} is the i_k-th element/feature in the feature vector $\mathbf{x}$. The model expectation of $f_k^*(\cdot)$ becomes:

$$E_{\tilde{p}(x)}[E_{p_\lambda(y|x)}[f_k^*(x,y)]] = \tilde{p}(x_{i_k}=1)p_\lambda(y_k|x_{i_k}=1)$$

To calculate $g(\cdot)$, a reference expectation of $f_k^*(\cdot)$ can be obtained after specifying (the upper and lower limits of) $p(y_k|x_{i_k}=1)$:

$$l_k \leq p(y_k|x_{i_k}=1) \leq u_k$$

This type of constraints, for example, $0.9 \leq p(\text{CON}|\textit{suggest}) \leq 1$, have been successfully applied to GE-based information structure analysis by Guo et al. (2013a). Here we build on their framework and our contribution is the *automatic induction* of such constraints by topic modeling.

The association between features and topics can be transformed into constraints as follows. Let W_z be a set of top N key features associated with topic z – the N features that are assigned the maximum probability according to the topic. We compute the following topic-specific feature sets:

$A_z = \{w|w \in W_z \wedge \forall t \neq z \; w \notin W_t\}$ – the set of features associated with topic z but not with any of the other topics;

$B_z = \bigcup_{t \neq z} W_t$ – the set of features associated with at least one topic other than z.

For every topic-feature pair (z_k, w_k) we therefore write the following constraint:

$$l_k \leq p(z_k|w_k=1) \leq u_k$$

We set the probability range for the k-th pair as follows:

$$\begin{aligned}
&\text{If } w_k \in A_{z_k} \text{ then } l_k = 0.9, u_k = 1,\\
&\text{If } w_k \in B_{z_k} \text{ then } l_k = 0, u_k = 0.1,\\
&\text{In any other case} \quad l_k = 0, u_k = 1.
\end{aligned}$$

The values of l_k and u_k were selected such that they reflect the strong association between the key features and their topics. Our basic reasoning is that if a sentence is represented by one of the key *unique* features of a given topic, it is highly likely to be associated with that topic. Likewise, a sentence is unlikely to be associated with the topic of interest if it has a key feature for any other topics.

3.3 Summary of Contribution

Learning with declarative knowledge is an active recent research avenue in the NLP community. In this framework feature-based models are augmented with domain and expert knowledge encoded most often by constraints of various types. The human effort involved with this framework is the *manual specification of the declarative knowledge*. This requires deep understanding of the domain and task in question. The resulting constraints typically specify detailed associations between lexical, grammatical and discourse elements and the information to be learned (see, e.g., tables 2 and 3 of (Guo et al., 2013a) and table 1 of (Chang et al., 2007)).

Our key contribution is the *automatic induction* of declarative knowledge that can be easily integrated into unsupervised models in the form of constraints and bias functions. Our model requires minimal domain and task knowledge. We do not specify lists of words or discourse markers (as in (Guo et al., 2013a)) but, instead, our model automatically associates latent variables both with linguistic features, taken from a very broad and general feature set (e.g.

	BKG	PROB	METH	RES	CON	CN	DIFF	FUT
Article (8171)	16.9	2.8	34.8	17.9	22.3	4.3	0.8	0.2
Introduction (1160)	74.8	13.2	5.4	0.6	5.9	0.1	-	-
Methods (2557)	0.5	0.2	97.5	1.4	0.2	0.2	0.1	-
Results (2054)	4.0	2.1	11.7	68.9	12.1	1.1	0.1	-
Discussion (2400)	16.9	1.1	0.7	1.5	63.5	13.3	2.4	0.7

Table 3: Distribution of sentences (shown in percentages) in articles and individual sections in the AZ-annotated corpus. The total number of sentences in each section appears in parentheses below the section name.

all the words that belong to a given set of POS tags), and with sentences in the input text. In the next section we present our experiments which demonstrate the usefulness of this declarative knowledge.

4 Experiments

Data and Models We used the full paper corpus earlier employed in (Guo et al., 2013a) which includes 8171 annotated sentences (with reported inter-annotator agreement: $\kappa = .83$) from 50 biomedical journal articles from the cancer risk assessment domain. One third of this corpus was saved for a development set on which our model was designed and its hyperparameters were tuned (see below). The corpus is annotated according to the Argumentative Zoning (AZ) scheme (Teufel and Moens, 2002; Mizuta et al., 2006) described in Table 1. Table 3 shows the distribution of AZ categories and the total number of sentences in each individual section. Since section names vary across articles, we grouped similar sections before calculating the statistics (e.g. *Materials* and *Methods* sections were grouped under *Method*). The table demonstrates that although there is a dominant category in each section (e.g. BKG in *Introduction*), up to 36.5% of the sentences in each section fall into other categories.

Feature Extraction We used the C&C POS tagger and parser trained on biomedical literature (Curran et al., 2007; Rimell and Clark, 2009) in the feature extraction process. Lemmatization was done with Morpha (Minnen et al., 2001).

Baselines We compared our models (TopicGC and TopicGE) against the following baselines: (a) an unconstrained unsupervised model – the unbiased version of the graph clustering we use for TopicGC (i.e. where $g(\cdot)$ is omitted, GC); (b) the unsupervised constrained GE method of (Guo et al., 2013a) where the constraints were created by experts (ExpertGE); (c) supervised unconstrained Maximum Entropy models, each trained to predict categories in a particular section using 150 sentences from that section, as in the lightly supervised case in (Guo et al., 2013a) (MaxEnt); and (d) a baseline that assigns all the sentences in a given section to the most frequent gold-standard category of that section (Table 3). This baseline emulates the use of section names for information structure classification.

Our constraints, which we use in the TopicGE and TopicGC models, are based on topics that are learned on the test corpus. While having access to the raw test text at training time is a standard assumption in many unsupervised NLP works (e.g. (Klein and Manning, 2004; Goldwater and Griffiths, 2007; Lang and Lapata, 2014)), it is important to quantify the extent to which our method depends on its access to the test set. We therefore constructed the TopicGE* model which is identical to TopicGE except that the topics are learned from another collection of 47 biomedical articles containing 9352 sentences. Like our test set, these articles are from the cancer risk assessment domain - all of them were published in the Toxicol. Sci. journal in the years 2009-2012 and were retrieved using the PubMed search engine with the key words "cancer risk assessment". There is no overlap between this new dataset and our test set (Guo et al., 2013a).

Models and Parameters For graph clustering, we used the Graclus software (Dhillon et al., 2007). For GE and MaxEnt, we used the Mallet software (McCallum, 2002). The γ parameter in the graph clustering was set to 10 using the development data. Several values of this parameter in the range of $[10, 1000]$ yielded very similar performance. The number of key features considered for each topic, N, was set to 40, 20 and 15 for the article, *Introduction* section, and *Discussion* section topic models, respectively. This difference reflects the number of feature types (Table 2) and the text volume (Table 3) of the respective models.

Evaluation We evaluated the overall accuracy as well as the category-level precision, recall and F-score for each section. TopicGC, TopicGE, TopicGE* and the baseline GC methods are unsuper-

	Introduction						*Method*						*Result*						*Discussion*					
	GC	TGC	TGE	TGE*	EGE	MFC	GC	TGC	TGE	TGE*	EGE	MFC	GC	TGC	TGE	TGE*	EGE	MFC	GC	TGC	TGE	TGE*	EGE	MFC
F1																								
BKG	.78	.83	**.89**	.86	.87	.86	-	-	-	-	**.07**	-	-	-	-	-	**.46**	-	.47	.47	.45	**.49**	.46	-
PROB	**.34**	.16	.31	.19	.24	-	-	-	-	-	**.33**	-	-	-	-	-	**.04**	-	-	-	-	-	**.32**	-
METH	-	.16	.12	.16	**.35**	-	.98	.98	.98	.98	.93	**.99**	.29	-	.25	**.32**	.29	-	-	-	-	-	**.14**	-
RES	-	-	-	-	**.07**	-	-	-	-	-	**.27**	-	.67	**.82**	.81	.77	.80	**.82**	-	-	-	-	**.14**	-
CON	-	.10	.26	.03	**.28**	-	-	-	-	-	-	-	.39	.28	.27	.29	**.42**	-	.82	**.83**	.82	.82	.71	.78
CN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**.25**	-	-	.21	**.23**	.11	.20	-
DIFF	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.12	-
FUT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.36	-
Acc.	.61	.68	**.77**	.74	.72	.75	**.97**	**.97**	**.97**	**.97**	.87	**.97**	.51	.68	.67	.62	.64	**.69**	.66	**.67**	**.67**	**.67**	.56	.63

Table 4: Performance (class based F1-score and overall accuracy (Acc.)) of unbiased Graph Clustering (GC), Graph Clustering with declarative knowledge learned from topic modeling (TopicGC model, TGC column), Generalized Expectation using constraints learned from topic modeling (TopicGE, TGE) and the same model where constraints are learned using an external set of articles (TopicGE*, TGE*), GE with constraints created by experts (ExpertGE, EGE - a replication of (Guo et al., 2013a)) and the most frequent gold standard category of the section (MFC)

vised and therefore induce unlabeled categories. To evaluate their output against the gold standard AZ annotation we first apply a standard greedy many-to-one mapping (naming) scheme in which each induced category is mapped to the gold category that shares the highest number of elements (sentence) with it (Reichart and Rappoport, 2009). The total number of induced topics was 9 with each topic model inducing three topics.[4] For light supervision, a ten-fold cross-validation scheme was applied.

In addition, we compare the quality of the automatically induced and manually constructed declarative knowledge in the context of customized summarization (Contractor et al., 2012) where summaries of specific types of information in an article are to be generated (we focused on the article's conclusions). While an intuitive solution would be to summarize the *Discussion* section of a paper, only 63.5% of its sentences belong to the gold standard *Conclusion* category (Table 3).

For our experiment, we first generated five sets of sentences. The first four sets consist of the article sentences annotated with the CON category according to: TopicGE or TopicGC or ExpertGE or the gold standard annotation. The fifth set is the *Discussion* section. We then used Microsoft AutoSummarize (Microsoft, 2007) to select sentences from each of the five sets such that the number of words in each summary amounts for 10% of the words in the input.

[4]The number of gold standard AZ categories is 8. However, we wanted each of our topic models to induce the same number of topics in order to reduce the number of parameters to the required minimum.

For evaluation, we asked an expert to summarize the conclusions of each article in the corpus. We then evaluated the five summaries against the gold-standard summaries written by the expert in terms of various ROUGE scores (Lin, 2004).

5 Results

We report here the results for our constrained unsupervised models compared to the baselines. We start with quantitative evaluation and continue with qualitative demonstration of the topics learned by the topic models and their key features which provide the substance for the constraints and bias functions used in our information structure models.

Unsupervised Learning Results Table 4 presents the performance of the four main unsupervised learning models discussed in this paper: GC, TopicGC, TopicGE, and ExpertGE of (Guo et al., 2013a). Our models (TopicGC and TopicGE) outperform the ExpertGE when considering category based F-score for the dominant categories of each section. ExpertGE is most useful in identifying the less frequent categories of each section (Table 3), which is in line with (Guo et al., 2013a). The overall sentence-based accuracy of TopicGE is significantly higher than that of ExpertGE for all four sections (bottom line of the table). Furthermore, for all four sections it is one of our models (TopicGC or TopicGE) that provides the best result under this measure, among the unsupervised models.

The table further provides a comparison of the unsupervised models to the MFC baseline which assigns all the sentences of a section to its most fre-

	Introduction						*Method*						*Result*						*Discussion*					
	TopicGE			Light			TopicGE			Light			TopicGE			Light			TopicGE			Light		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
BKG	.84	.95	**.89**	.78	.99	.87	-	-	-	-	-	-	-	-	-	-	-	-	.41	.51	**.45**	.38	.19	.25
PROB	.33	.30	**.31**	.57	.11	.18	-	-	-	-	-	-	-	-	-	.25	.02	.04	-	-	-	-	-	-
METH	.40	.07	.12	.50	.21	**.30**	.97	1	.98	.97	1	.98	.34	.20	**.25**	.62	.14	.23	-	-	-	-	-	-
RES	-	-	-	-	-	-	-	-	-	-	-	-	.74	.90	.81	.71	.98	**.82**	-	-	-	-	-	-
CON	.44	.18	**.26**	.80	.06	.11	-	-	-	-	-	-	.30	.25	**.27**	.57	.16	.25	.78	.87	**.82**	.69	.96	.80
CN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.32	.18	**.23**	.35	.06	.10
DIFF	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FUT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Acc.	0.77			0.77			0.97			0.97			0.67			0.70			0.67			0.66		

Table 5: Performance (class based Precision, Recall and F-score as well as overall accuracy (Acc.)) of the TopicGE model and of an unconstrained MaxEnt model trained with Light supervision (total of 600 sentences - 150 training sentences for each section-level model). The same pattern of results holds when the MaxEnt is trained with up to 2000 sentences (500 sentences for each section-level model).

	TopicGE			**TopicGC**			**ExpertGE**			**Section**			**Gold**		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
ROUGE-1	45.2	54.0	**46.8**	43.5	55.1	46.1	43.7	49.1	43.8	**46.7**	43.8	42.6	43.3	**55.4**	46.2
ROUGE-2	**30.0**	**35.8**	**30.8**	28.4	35.7	29.8	25.5	28.2	25.2	28.6	26.3	25.8	27.8	35.1	29.3
ROUGE-L	43.3	51.6	**44.8**	41.6	**52.6**	44.1	41.3	46.2	41.3	**44.2**	41.3	40.3	41,1	52.3	43.7

Table 6: ROUGE scores of zone (TopicGE, TopicGC, ExpertGE or gold standard) and *Discussion* section based summaries. TopicGE provides the best summaries. TopicGC outperforms ExpertGE and the *Discussion* section systems and in two measures the gold categorization based system as well. Result patterns with ROUGE(3,4,W-1.2, S* and SU*) are very similar to those of the table. The differences between TopicGE and ExpertGE are statistically significant using t-test with $p < 0.05$. The differences between TopicGE and gold, as well as between ExpertGE and gold are not statistically significant.

quent category according to the gold standard. This baseline sheds light on the usefulness of section names for our task. As is evident from the table, while this baseline is competitive with the unsupervised models in terms of accuracy, its class-based F-score performance is quite poor. Not only does it lag behind the unsupervised models in terms of the F-score of the most frequent classes of the *Introduction* and *Discussion* sections, but it does not identify any of the classes except from the most frequent ones in any of the sections - a task the unsupervised models often perform with reasonable quality.

Finally, the table also presents the performance of the TopicGE* model for which constraints are leaned from an external data set - different from the test set. The results show that there is no substantial difference between the performance of the TopicGE and TopicGE* models. While TopicGE achieves better F-scores in five of the cases in the table, TopicGE* is better in four cases and the performance is identical in two cases. Section level accuracies are better for TopicGE in two of the four sections, but the difference is only 3-5%.

Comparison with Supervised Learning Table 5 compares the quality of unsupervised constrained-based learning with that of lightly supervised feature-based learning. Since our models, TopicGC and TopicGE, perform quite similarly, we included only TopicGE in this evaluation. The lightly supervised models (MaxEnt classifiers) were trained with a total of 600 sentences - 150 for each section-specific classifier. The table demonstrates that TopicGE outperforms MaxEnt with light supervision in terms of class based F-scores in the *Introduction* and *Discussion* sections. In the *Methods* section, where 97.5% of the sentences belong to the most frequent category, and in the *Results* section, the models perform quite similarly. Overall accuracy numbers are quite similar for both models with MaxEnt doing better for the *Results* section and TopicGE for the *Discussion* section. These results further demonstrate that unsupervised constrained learning provides a practical solution to information structure analysis of scientific articles.

Extractive Summarization Evaluation Table 6 presents the average ROUGE scores for zone-based (TopicGE, TopicGC, ExpertGE and gold) and section-based summaries across our test set articles.

Topic	Features
1	{do} be {done} {doing} {be_done} {have_been_done} induce {may_do} {to_do} show have {have_done} increase {did} suggest indicate report cause include inhibit find observe involve associate activate demonstrate result use lead play {could_do} know {do_do} form contribute {can_do} {would_do} promote reduce
2	{were_done} {done} {doing} {did} use be describe contain perform incubate {do} determine analyze follow add isolate purchase wash accord {to_do} treat collect remove prepare obtain measure store stain centrifuge transfer detect purify assess supplement carry dissolve plate receive kill
3	{did} {done} be {doing} {were_done} [tab_fig] {do} show increase observe compare {to_do} expose use have find {did_do} treat {be_done} report follow drink reduce result administer decrease determine measure include evaluate affect detect induce indicate associate provide reveal suggest occur

Table 7: Topics and key features extracted by the article-level model (including modal, tense and voice marked in curly brackets, reference to tables or figures marked in square brackets, and verbs in the base form)

Topic	Features
1	[no_cite] {did} (we) {done} {do} {doing} use {were_done} (present) {to_do} investigate be (mammary) determine provide (our) treat compare examine
2	{did} {done} [cite] {doing} {were_done} be expose find [no_cite] drink increase report (recent) (previous) administer {do} contain evaluate (early)
3	{do} [cite] be {done} [no_cite] {doing} {be_done} {have_been_done} induce {have_done} (it) show {may_do} have {to_do} include increase (their) associate

Table 8: Topics and key features extracted by the section-specific topic model of the *Introduction* section (including citations marked in square brackets, pronouns and the follow-up adjective modifiers marked in parentheses, modal, tense and voice marked in curly brackets, and verbs in their base form)

Topic	Features
1	(we) [no_cite] (our) higher (mammary) as because (first) significant possible high (early) (positive) most
2	[cite] present (present) (previous) similar different (its) although consistent furthermore greater due most whereas
3	[no_cite] not also (it) but however more (their) both therefore only thus significant lower

Table 9: Topics and key features extracted by the section-specific topic model of the *Discussion* section (including citations marked in square brackets, pronouns and the follow-up adjective modifiers marked in parentheses, and conjunctions, adjectives and adverbs)

TopicGE and TopicGC based summaries outperform the other systems, even the one that uses gold standard information structure categorization. A potential explanation for the better performance of our models compared to ExpertGE is that the relative strength of our models is in identifying the major category of each section while ExpertGE is better at identifying low or medium frequency categories.

Qualitative Analysis We next provide a qualitative analysis of the topics induced by our topic models — the article-level model as well as the section-level models — and their key features. Note that both our models, TopicGE and TopicGC, assume that the induced topics provide a good approximation of the information structure categories and build their constraints (expert knowledge) from these topics accordingly. Below we examine this assumption.

Table 7 presents the topics and key features obtained from global topic modeling applied to full articles. The table reveals a strong correlation between present/future tense and topic 1, and between past tense and topics 2 and 3 (Modal, Tense and Voice features). The table further demonstrates that topics 1 and 3 are linked to verbs that describe research findings, such as "show" and "demonstrate" in topic 1, and "report" and "indicate" in topic 3, whereas topic 2 seems related to verbs that describe methods and experiments such as "use" and "prepare". The feature corresponding to tables and figures [tab_fig] is only seen in topic 3. Based on these observations, topics 1, 2 and 3 seem to be related to AZ categories CON, METH and RES respectively.

Tables 8 and 9 present the topics and the key features obtained from the section-specific topic mod-

eling for the *Introduction* and *Discussion* sections. Due to space limitations we cannot provide a detailed analysis of the information included in these tables, but it is easy to see that they provide evidence for the correlation between topics in the section specific models and AZ categories. Table 8 demonstrates that for the *Introduction* section topic 1 correlates with the author's work and topics 2 and 3 with previous work. Table 9 shows that for the *Discussion* section topics 1 and 3 well correlate with the AZ CON category and topic 2 with the BKG, CN and DIFF categories. Our analysis therefore demonstrates that the induced topics are well aligned with the actual categories of the AZ classification scheme or with distinctions (e.g. the author's own work vs. works of others) that are very relevant for this scheme. Note that we have not seeded our models with word-lists and the induced topics are therefore purely data-driven.

6 Discussion

We presented a new framework for automatic induction of declarative knowledge and applied it to constraint-based modeling of the information structure analysis of scientific documents. Our main contribution is a topic-model based method for unsupervised acquisition of lexical, syntactic and discourse knowledge guided by the notion of topics and their key features. We demonstrated that the induced topics and key features can be used with two different unsupervised learning methods – a constrained unsupervised generalized expectation model and a graph clustering formulation. Our results show that this novel framework rivals more supervised alternatives. Our work therefore contributes to the important challenge of automatically inducing declarative knowledge that can reduce the dependence of ML algorithms on manually annotated data.

The next natural step in this research is generalizing our framework and make it applicable to more applications, domains and machine learning models. We are currently investigating a number of ideas which will hopefully lead to better natural language learning with reduced human supervision.

References

Sam Anzaroot, Alexandre Passos, David Belanger, and Andrew McCallum. 2014. Learning soft linear constraints with application to citation field extraction. In *ACL*, pages 593–602.

Kedar Bellare, Gregory Druck, and Andrew McCallum. 2009. Alternating projections for learning with expectation constraints. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 43–50.

Catherine Blake. 2009. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, 43(2):173–189.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2007. Guiding semi-supervision with constraint-driven learning. In *ACL*, pages 280–287.

Danish Contractor, Yufan Guo, and Anna Korhonen. 2012. Using argumentative zones for extractive summarization of scientific articles. In *COLING*, pages 663–678.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 33–36.

Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. 2007. Weighted graph cuts without eigenvectors: A multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1944–1957.

Gregory Druck, Gideon Mann, and Andrew McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 595–602.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.

Sharon Goldwater and Tom Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *ACL*, pages 744–751.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins Karolinska, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *BioNLP*, pages 99–107.

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

Yufan Guo, Anna Korhonen, Ilona Silins, and Ulla Stenius. 2011b. Weakly-supervised learning of information structure of scientific abstracts–is it accurate enough to benefit real-world tasks in biomedicine? *Bioinformatics*, 27(22):3179–3185.

Yufan Guo, Roi Reichart, and Anna Korhonen. 2013a. Improved information structure analysis of scientific documents through discourse and lexical constraints. In *NAACL HLT*, pages 928–937.

Yufan Guo, Ilona Silins, Ulla Stenius, and Anna Korhonen. 2013b. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review. *Bioinformatics*, 29(11):1440–1447.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *IJCNLP*, pages 381–388.

Douwe Kiela, Yufan Guo, Ulla Stenius, and Anna Korhonen. 2014. Unsupervised discovery of information structure in biomedical documents. *Bioinformatics*, page btu758.

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *ACL*, pages 478–485.

Joel Lang and Mirella Lapata. 2014. Similarity-driven semantic role induction via graph partitioning. *Computational Linguistics*, 40(3):633–669.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for conceptualization and zoning of scientific papers. In *LREC*, pages 2054–2061.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *BioNLP*, pages 65–72.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Gideon S Mann and Andrew McCallum. 2007. Simple, robust, scalable semi-supervised learning via expectation regularization. In *ICML*, pages 593–600.

Gideon S. Mann and Andrew McCallum. 2008. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *ACL*, pages 870–878.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *ACL*, pages 795–804.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

David McClosky and Christopher D. Manning. 2012. Learning constraints for consistent timeline extraction. In *EMNLP-CoNLL*, pages 873–882.

Microsoft. 2007. AutoSummarize: Automatically summarize a document. https://support.office.com/en-us/article/Automatically-summarize-a-document-b43f20ae-ec4b-41cc-b40a-753eed6d7424.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of english. *Natural Language Engineering*, 7(3):207–223.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics on Natural Language Processing in Biomedicine and Its Applications*, 75(6):468–487.

Diarmuid Ó Séaghdha and Simone Teufel. 2014. Unsupervised learning of rhetorical structure with un-topic models. In *Proceedings of COLING 2014: Technical Papers*, pages 2–13.

Roi Reichart and Regina Barzilay. 2012. Multi-event extraction guided by global constraints. In *NAACL HLT*, pages 70–79.

Roi Reichart and Anna Korhonen. 2012. Document and corpus level inference for unsupervised and transductive learning of information structure of scientific documents. In *COLING*, pages 995–1006.

Roi Reichart and Ari Rappoport. 2009. The nvi clustering evaluation measure. In *CoNLL*, pages 165–173.

Laura Rimell and Stephen Clark. 2009. Porting a lexicalized-grammar parser to the biomedical domain. *Journal of Biomedical Informatics*, 42(5):852–865.

Patrick Ruch, Clia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbhler, Paul Fabry, Julien Gobeill, Violaine Pillet, Dietrich Rebholz-Schuhmann, Christian Lovis, and Anne-Lise Veuthey. 2007. Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics*, 76(2-3):195–200.

Alexander Rush, Roi Reichart, Michael Collins, and Amir Globerson. 2012. Improved parsing and pos tagging using inter-sentence consistency constraints. In *EMNLP-CoNLL*, pages 1434–1444.

Hagit Shatkay, Fengxia Pan, Andrey Rzhetsky, and W John Wilbur. 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093.

Imad Tbahriti, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. 2006. Using argumentation to retrieve articles with similar citations. *International Journal of Medical Informatics*, 75(6):488–495.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *EMNLP*, pages 1493–1502.

Andrea Varga, Daniel Preotiuc-Pietro, and Fabio Ciravegna. 2012. Unsupervised document zone identification using probabilistic graphical models. In *LREC*, pages 1610–1617.

Bonnie Webber, Markus Egg, and Valia Kordoni. 2011. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.