# Improved Information Structure Analysis of Scientific Documents Through Discourse and Lexical Constraints

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

**Roi Reichart**
University of Cambridge, UK
rr439@cam.ac.uk

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

## Abstract

Inferring the information structure of scientific documents is useful for many downstream applications. Existing feature-based machine learning approaches to this task require substantial training data and suffer from limited performance. Our idea is to guide feature-based models with declarative domain knowledge encoded as posterior distribution constraints. We explore a rich set of discourse and lexical constraints which we incorporate through the Generalized Expectation (GE) criterion. Our constrained model improves the performance of existing fully and lightly supervised models. Even a fully unsupervised version of this model outperforms lightly supervised feature-based models, showing that our approach can be useful even when no labeled data is available.

## 1 Introduction

Techniques that enable automatic analysis of the information structure of scientific articles can help scientists identify information of interest in the growing volume of scientific literature. For example, classification of sentences according to argumentative zones (AZ) – an information structure scheme that is applicable across scientific domains (Teufel et al., 2009) – can support information retrieval, information extraction and summarization (Teufel and Moens, 2002; Tbahriti et al., 2006; Ruch et al., 2007; Liakata et al., 2012; Contractor et al., 2012).

Previous work on sentence-based classification of scientific literature according to categories of information structure has mostly used feature-based machine learning, such as Support Vector Machines (SVM) and Conditional Random Fields (CRF) (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)). Unfortunately, the performance of these methods is rather limited, as indicated e.g. by the relatively low numbers reported by Liakata et al. (2012) in biochemistry and chemistry with per-class F-scores ranging from .18 to .76.

We propose a novel approach to this task in which traditional feature-based models are augmented with explicit declarative expert and domain knowledge, and apply it to sentence-based AZ. We explore two sources of declarative knowledge for our task - *discourse* and *lexical*. One way to utilize discourse knowledge is to guide the model predictions by encoding a desired predicted class (i.e. information category) distribution in a given position in the document. Consider, for example, sentence (1) from the first paragraph of the *Discussion* section in a paper:

(1) *In time, this will prove to be most suitable for detailed analysis of the role of these hormones in mammary cancer development.*

Although the future tense and cue phrases such as "*in time*" can indicate that authors are discussing future work (i.e. the "Future work" class in the AZ scheme), in this case they refer to their own contribution (i.e. the "Conclusion" class in AZ). As most authors discuss their own contribution in the beginning of the *Discussion* section and future directions in the end, encoding the desired class distribution as a function of the position in this section can guide the model to the right decision.

Likewise, lexical knowledge can guide the model through predicted class distributions for sentences that contain specific vocabulary. Consider, for example, sentence (2):

(2) *The values calculated for lungs include the presumed DNA adduct of BA and might thus be slightly overestimated.*

The verb "*calculated*" usually indicates the "Method" class, but, when accompanied by the modal verb "*might*", it is more likely to imply that authors are interpreting their own results (i.e. the "Conclusion" class in AZ). This can be explicitly encoded in the model through a target distribution for sentences containing certain modal verbs.

Recent work has shown that explicit declaration of domain and expert knowledge can be highly useful for structured NLP tasks such as parsing, POS tagging and information extraction (Chang et al., 2007; Mann and McCallum, 2008; Ganchev et al., 2010). These works have encoded expert knowledge through constraints, with different frameworks differing in the type of constraints and the inference and learning algorithms used. We build on the Generalized Expectation (GE) framework (Mann and McCallum, 2007) which encodes expert knowledge through a preference (i.e. soft) constraints for parameter settings for which the predicted label distribution matches a target distribution.

In order to integrate domain knowledge with a features-based model, we develop a simple taxonomy of constraints (i.e. desired class distributions) and employ a top-down classification algorithm on top of a Maximum Entropy Model augmented with GE constraints. This algorithm enables us to break the multi-class prediction into a pipeline of consecutive, simpler predictions which can be better assisted by the encoded knowledge.

We experiment in the biological domain with the eight-category AZ scheme (Table 1) adapted from (Mizuta et al., 2006) and described in (Contractor et al., 2012). The results show that our constrained model substantially outperforms a baseline unconstrained Maximum Entropy Model. While this type of constrained models have previously improved the feature-based model performance mostly in the weakly supervised and domain adaptation scenarios (e.g. (Mann and McCallum, 2007; Mann and McCallum, 2008; Ganchev et al., 2010)), we demonstrate substantial gains both when the Maximum Entropy Model is fully trained and when its training data is sparse. This demonstrates the importance of expert knowledge for our task and supports our modeling decision that combines feature-based methods with domain knowledge encoded via constraints.

Table 1: The AZ categories included in the categorization scheme of this paper.

| 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 |

## 2 Previous work

**Information structure analysis** The information structure of scientific documents (e.g. journal articles, abstracts, essays) can be analyzed in terms of patterns of topics, functions or relations observed in multi-sentence scientific text. Computational approaches have mainly focused on analysis based on argumentative zones (Teufel and Moens, 2002; Mizuta et al., 2006; Hachey and Grover, 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).

Most existing methods for analyzing scientific text according to information structure use full supervision in the form of thousands of manually annotated sentences (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). Because manual annotation is prohibitively expensive, approaches based on light supervision are now emerging for the task, including those based on active learning and self-training (Guo et al., 2011) and unsupervised methods (Varga et al., 2012; Reichart and Korhonen, 2012). Unfortunately, these approaches do not reach the performance level of fully supervised models, let alone exceed it. Our novel method addresses this problem.

**Declarative knowledge and constraints** Previous work has shown that incorporating declarative constraints into feature-based machine learning models works well in many NLP tasks (Chang et al., 2007; Mann and McCallum, 2008; Druck et al., 2008; Bellare et al., 2009; Ganchev et al., 2010). Such constraints can be used in a semi-supervised or unsupervised fashion. For example, (Mann and McCallum, 2008) shows that using CRF in conjunction with auxiliary constraints on unlabeled data significantly outperforms traditional CRF in information extraction, and (Druck et al., 2008) shows that using declarative constraints alone for unsupervised learning achieves good results in text classification. We show that declarative constraints can be highly useful for the identification of information structure of scientific documents. In contrast with most previous works, we show that such constraints can improve the performance of a fully supervised model. The constraints are particularly helpful for identifying low-frequency information categories, but still yield high performance on high-frequency categories.

## 3 Maximum-Entropy Estimation and Generalized Expectation (GE)

In this section we describe the Generalized Expectation method for declarative knowledge encoding.

**Maximum Entropy (ME)** The idea of Generalized Expectation (Dudík, 2007; Mann and McCallum, 2008; Druck et al., 2008) stems from the principle of maximum entropy (Jaynes, 1957; Pietra and Pietra, 1993) which raises the following constrained optimization problem:

$$\begin{aligned} \max_{p} \quad & H(\cdot) \\ \text{subject to} \quad & E_p[\mathbf{f}(\cdot)] = E_{\tilde{p}}[\mathbf{f}(\cdot)] \\ & p(\cdot) \geq 0 \\ & \sum p(\cdot) = 1, \end{aligned} \tag{1}$$

where $\tilde{p}(\cdot)$ is the empirical distribution, $p(\cdot)$ is a probability distribution in the model and $H(\cdot)$ is the corresponding information entropy, $\mathbf{f}(\cdot)$ is a collection of feature functions, and $E_p[\mathbf{f}(\cdot)]$ and $E_{\tilde{p}}[\mathbf{f}(\cdot)]$ are the expectations of $\mathbf{f}$ with respect to $p(\cdot)$ and $\tilde{p}(\cdot)$. An example of $p(\cdot)$ could be a conditional probability distribution $p(y|x)$, and $H(\cdot)$ could be a conditional entropy $H(y|x)$. The optimal $p(y|x)$ will take on an exponential form:

$$p_\lambda(y|x) = \frac{1}{Z_\lambda} exp(\lambda \cdot \mathbf{f}(x,y)), \tag{2}$$

where $\lambda$ is the Lagrange multipliers in the corresponding unconstrained objective function, and $Z_\lambda$ is the partition function. The dual problem becomes maximizing the conditional log-likelihood of labeled data $\mathcal{L}$ (Berger et al., 1996):

$$\max_\lambda \sum_{(x_i,y_i)\in\mathcal{L}} log(p_\lambda(y_i|x_i)), \tag{3}$$

which is usually known as a Log-linear or Maximum Entropy Model (MaxEnt).

**ME with Generalized Expectation** The objective function and the constraints on **expectations** in (1) can be **generalized** to:

$$\begin{aligned} \max_\lambda & - \sum_x \tilde{p}(x) D(p_\lambda(y|x)||p_0(y|x)) \\ & - g(E_{\tilde{p}(x)}[E_{p_\lambda(y|x)}[\mathbf{f}(x,y)|x]]), \end{aligned} \tag{4}$$

where $D(p_\lambda||p_0)$ is the divergence from $p_\lambda$ to a base distribution $p_0$, and $g(\cdot)$ is a constraint/penalty function that takes empirical evidence $E_{\tilde{p}(x,y)}[\mathbf{f}(x,y)]$ as a reference point (Pietra and Pietra, 1993; Chen et al., 2000; Dudík, 2007). Note that a special case of this is MaxEnt where $p_0$ is set to be a uniform distribution, $D(\cdot)$ to be the KL divergence, and $g(\cdot)$ to be an equality constraint.

The constraint $g(\cdot)$ can be set in a relaxed manner:

$$\sum_k \frac{1}{2\rho_k^2}(E_{\tilde{p}(x)}[E_{p_\lambda(y|x)}[f_k(x,y)|x]] - E_{\tilde{p}(x,y)}[f_k(x,y)])^2,$$

which is the logarithm of a Gaussian distribution centered at the reference values with a diagonal covariance matrix (Pietra and Pietra, 1993), and the dual problem will become a regularized MaxEnt with a Gaussian prior ($\mu_k = 0$, $\sigma_k^2 = \frac{1}{\rho_k^2}$) over the parameters:

$$\max_\lambda \sum_{(x_i,y_i)\in\mathcal{L}} log(p_\lambda(y_i|x_i)) - \sum_k \frac{\lambda_k^2}{2\sigma_k^2} \tag{5}$$

Such a model can be further extended to include expert knowledge or auxiliary constraints on unlabeled data $\mathcal{U}$ (Mann and McCallum, 2008; Druck et al., 2008; Bellare et al., 2009):

$$\begin{aligned} \max_\lambda & \sum_{(x_i,y_i)\in\mathcal{L}} log(p_\lambda(y_i|x_i)) - \sum_k \frac{\lambda_k^2}{2\sigma_k^2} \\ & - \gamma g^*(E_{p_\lambda(y|x)}[\mathbf{f}^*(x,y)]) \end{aligned} \tag{6}$$

where $\mathbf{f}^*(\cdot)$ is a collection of auxiliary feature functions on $\mathcal{U}$, $g^*(\cdot)$ is a constraint function that takes expert/declarative knowledge $E_{p^*(y|x)}[\mathbf{f}^*(x,y)]$ as a reference point, and $\gamma$ is the weight of the auxiliary GE term.

The auxiliary constraint $g^*(\cdot)$ can take on many forms and the one we used in this work is an $L^2$ penalty function (Dudík, 2007). We trained the model with L-BFGS (Nocedal, 1980) in supervised, semi-supervised and unsupervised fashions on labeled and/or unlabeled data, using the Mallet software (McCallum, 2002).

## 4 Incorporating Expert Knowledge into GE constraints

We defined the auxiliary feature functions – the expert knowledge on unlabeled data as[1]:

$$f_k^*(x,y) = \mathbb{1}_{(x_k,y_k)}(x,y), \\ \text{such that } E_{p^*(y|x)}[f_k(x,y)] = p^*(y_k|x_k), \tag{7}$$

where $\mathbb{1}_{(x_k,y_k)}(x,y)$ is an indicator function, and $p^*(y_k|x_k)$ is a conditional probability specified in the form of

$$p^*(y_k|x_k) \in [a_k, b_k] \tag{8}$$

by experts. In particular, we took

$$p^*(y_k|x_k) = \begin{cases} a_k & \text{if } p_\lambda(y_k|x_k) < a \\ b_k & \text{if } p_\lambda(y_k|x_k) > b \\ p_\lambda(y_k|x_k) & \text{if } a \le p_\lambda(y_k|x_k) \le b \end{cases} \tag{9}$$

as the reference point when calculating $g^*(\cdot)$.

We defined two types of constraints: those based on *discourse* properties such as the location of a sentence in a particular section or paragraph, and those based on *lexical* properties such as citations, references to tables and figures, word lists, tenses, and so on. Note that the word lists actually contain both lexical and semantic information.

To make an efficient use of the declarative knowledge we build a taxonomy of information structure categories centered around the distinction between categories that describe the authors' OWN work and those that describe OTHER work (see Section 5). In practice, our model labels every sentence with an AZ category augmented by one of the two categories, OWN or OTHER. In evaluation we consider only the standard AZ categories which are part of the annotation scheme of (Contractor et al., 2012).

Table 2: Discourse and lexical constraints for identifying information categories at different levels of the information structure taxonomy.

**(a) OWN / OTHER**

| | |
|---|---|
| OWN | **Discourse**<br>(1) Target(last part of paragraph) = 1<br>(2) Target(last part of section) = 1<br>**Lexical**<br>(3) Target(tables/figures) $\geq$ 1<br>(4) $\exists$x $\in$ $\{$w$\mid$w$\sim$*we*$\}$ Target(x) = 1<br>$\wedge$ $\forall$y $\in$ $\{$w$\mid$w$\sim$*previous*$\}$ Target(y) = 0<br>(5) $\exists$x $\in$ $\{$w$\mid$w$\sim$*thus*$\}$ Target(x) = 1 |
| OTHER | **Lexical**<br>(6) Target(cite) = 1<br>(7) Target(cite) > 1<br>(8) Backward(cite) = 1<br>$\wedge$ $\exists$x $\in$ $\{$w$\mid$w$\sim$*in_addition*$\}$ Target(x) = 1 |

**(b) PROB / METH / RES / CON / FUT**

| | |
|---|---|
| PROB | **Discourse**<br>(1) Target(last part in section) = 1<br>**Lexical**<br>(2) $\exists$x $\in$ $\{$w$\mid$w$\sim$*aim*$\}$ Target(x) = 1<br>(3) $\exists$x $\in$ $\{$w$\mid$w$\sim$*question*$\}$ Target(x) = 1<br>(4) $\exists$x $\in$ $\{$w$\mid$w$\sim$*investigate*$\}$ Target(x) = 1 |
| METH | **Lexical**<br>(5) $\exists$x $\in$ $\{$w$\mid$w$\sim\{$*use,method*$\}\}$ Target(x) = 1 |
| RES | **Lexical**<br>(6) Target(*tables/figures*) $\geq$ 1<br>(7) $\exists$x $\in$ $\{$w$\mid$w$\sim$*observe*$\}$ Target(x) = 1 |
| CON | **Lexical**<br>(8) Target(*cite*) $\geq$ 1<br>(9) $\exists$x $\in$ $\{$w$\mid$w$\sim$*conclude*$\}$ Target(x) = 1<br>(10) $\exists$x $\in$ $\{$w$\mid$w$\sim\{$*suggest, thus, because, likely*$\}\}$<br>Target(x) = 1 |
| FUT | **Discourse**<br>(11) Target(first part in section) = 1<br>(12) Target(last part in section) = 1<br>$\wedge$ $\exists$x $\in$ $\{$w$\mid$w$\sim\{$*will,need,future*$\}\}$ Target(x) = 1<br>**Lexical**<br>(13) $\exists$x $\{$w$\mid$w$\sim$*will,future*$\}$ Target(x) = 1<br>(14) Target(present continuous tense) = 1 |

**(c) BKG / CN / DIFF**

| | |
|---|---|
| BKG | **Discourse**<br>(1) Target(first part in paragraph) = 1<br>(2) Target(first part in section) = 1<br>**Lexical**<br>(3) $\exists$x $\in$ $\{$w$\mid$w$\sim$*we*$\}$ Target(x) = 1<br>$\wedge$ $\forall$y $\in$ $\{$w$\mid$w$\sim$*previous*$\}$ Target(y) = 0<br>(4) Forward(*cite*) = 1<br>$\wedge$ $\forall$x $\in$ $\{$w$\mid$w$\sim\{$*consistent,inconsistent,than*$\}\}$<br>(Target(x) = 0 $\wedge$ Forward(x) = 0) |
| CN | **Lexical**<br>(5) $\exists$x $\in$ $\{$w$\mid$w$\sim$*consistent*$\}$ Target(x) = 1<br>(6) $\exists$x $\in$ $\{$w$\mid$w$\sim$*consistent*$\}$ Forward(x) = 1 |
| DIFF | **Lexical**<br>(7) $\exists$x $\in$ $\{$w$\mid$w$\sim$*inconsistent*$\}$ Target(x) = 1<br>(8) $\exists$x $\in$ $\{$w$\mid$w$\sim$*inconsistent*$\}$ Forward(x) = 1<br>(9) $\exists$x $\in$ $\{$w$\mid$w$\sim\{$*inconsistent,than,however*$\}\}$<br>Forward(x) = 1 $\wedge$ $\exists$y $\in$ $\{$w$\mid$w$\sim$*we*$\}$ Forward(y) = 1<br>$\wedge$ $\forall$z $\in$ $\{$w$\mid$w$\sim$*previous*$\}\}$ Forward(z) = 0 |

[1] Accordingly, $E_{p_\lambda(y|x)}[f_k(x,y)] = p_\lambda(y_k|x_k)$

Table 3: The lexical sets used as properties in the constraints.

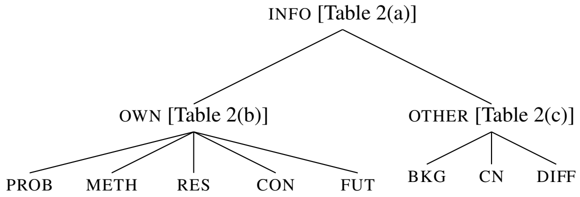
| Cue | Synonyms |
|---|---|
| we | our, present_study |
| previous | previously, recent, recently |
| thus | therefore |
| aim | objective, goal, purpose |
| question | hypothesis, ? |
| investigate | explore, study, test, examine, evaluate, assess, determine, characterize, analyze, report, present |
| use | employ |
| method | algorithm, assay |
| observe | see, find, show |
| conclude | conclusion, summarize, summary |
| suggest | illustrate, demonstrate, imply, indicate, confirm, reflect, support, prove, reveal |
| because | result_from, attribute_to |
| likely | probable, probably, possible, possibly, may, could |
| need | remain |
| future | further |
| consistent | match, agree, support, in_line, in_agreement, similar, same, analogous |
| inconsistent | conflicting, conflict, contrast, contrary, differ, different, difference |
| than | compare |
| however | other_hand, although, though, but |

The constraints in Table 2(a) refer to the top level of this taxonomy: distinction between the authors' own work and the work of others, and the constraints in Tables 2(b)-(c) refer to the bottom level of the taxonomy: distinction between AZ categories related to the authors' own work (Table 2(b)) and other's work (Table 2(c)).

The first and second columns in each table refer to the $y$ and $x$ variables in Equation (8), respectively. The functions Target($\cdot$), Forward($\cdot$) and Backward($\cdot$) refer to the property value for the target, next and preceding sentence, respectively. If their value is 1 then the property holds for the respective sentence, if it is 0, the property does not hold. In some cases the value of such functions can be greater than 1, meaning that the property appears multiple times in the sentence. Terms of the form $\{\mathrm{w}|\mathrm{w}\sim\{w_i\}\}$ refer to any word/bi-grams that have the same sense as $w_i$, where the actual word set we use with every example word in Table 2 is described in Table 3.

For example, take constraints (1) and (4) in Table 2(a). The former is a standard discourse constraint that refers to the probability that the target sentence describes the authors' own work given that it appears in the last of the ten parts in the paragraph. The latter is a standard lexical constraint that refers to the probability that a sentence presents other people's work given that it contains any words in {*we, our, present_study*} and that it doesn't contain any words in {*previous, previously, recent, recently*}. Our constraint set further includes constraints that combine both types of information. For example, constraint (12) in Table 2(b) refers to the probability that a sentence discusses future work given that it appears in the last of the ten parts of the section (discourse) and that it contains at least one word in {*will, future, further, need, remain*} (lexical).

Figure 1: The constraint taxonomy for top-down modeling.



## 5 Top-Down Model

An interesting property of our task and domain is that the available expert knowledge does not directly support the distinctions between AZ categories, but it does provide valuable indirect guidance. For example, the number of citations in a sentence is only useful for separating the authors' work from other people's work, but not for further fine grained distinctions between zone categories. Moreover, those constraints that are useful for making fine grained distinctions between AZ categories are usually useful only for a particular subset of the categories only. For example, all the constraints in Table 2(b) are conditioned on the assumption that the sentence describes the authors' own work.

To make the best use of the domain knowledge, we developed a simple constraint taxonomy, and apply a top-down classification approach which utilizes it. The taxonomy is presented in Figure 1. For classification we trained three MaxEnt models augmented with GE constraints: one for distinguishing between OWN and OTHER[2], one for distinguishing between the AZ categories under the OWN auxiliary category and one for distinguishing between the AZ categories under the OTHER auxiliary category. At test time we first apply the first classifier and based on its prediction we apply either the classifier that distinguishes between OWN categories or the one that distinguishes between OTHER categories.

[2] For the training of this model, each training data AZ category is mapped to its respective auxiliary class.

## 6 Experiments

**Data** We used the full paper corpus used by Contractor et al. (2012) which contains 8171 sentences from 50 biomedical journal articles. The corpus is annotated according to the AZ scheme described in Table 1. AZ describes the logical structure, scientific argumentation and intellectual attribution of a scientific paper. It was originally introduced by Teufel and Moens (2002) and applied to computational linguistics papers, and later adapted to other domains such as biology (Mizuta et al., 2006) – which we used in this work – and chemistry (Teufel et al., 2009).

Table 4 shows the AZ class distribution in full articles as well as in individual sections. Since section names vary across scientific articles, we grouped similar sections before calculating the statistics (e.g. *Discussion* and *Conclusions* sections were grouped under *Discussion*). We can see that although there is a major category in each section (e.g. CON in *Discussion*), up to 36.5% of the sentences in each section still belong to other categories.

**Features** We extracted the following features from each sentence and used them in the feature-based classifiers: (1) Discourse features: location in the article/section/paragraph. For this feature each text batch was divided to ten equal size parts and the corresponding feature value identifies the relevant part; (2) Lexical features: number of citations and references to tables and figures (0, 1, or more), word, bi-gram, verb, and verb class (obtained by spectral clustering (Sun and Korhonen, 2009)); (3) Syntactic features: tense and voice (POS tags of main and auxiliary verbs), grammatical relation, subject and object. The lexical and the syntactic features were extracted for the represented sentence as well as for its surrounding sentences. We used the C&C POS tagger and parser (Curran et al., 2007) for extracting the lexical and the syntactic features. Note that all the information encoded into our constraints is also encoded in the features and is thus available to the feature-based model. This enables us to properly evaluate the impact of our modeling decision which augments a feature-based model with constraints.

**Baselines** We compared our model against four baselines, two with full supervision: Support Vector Machines (SVM) and Maximum Entropy Models (MaxEnt), and two with light supervision: Transductive SVM (TSVM) and semi-supervised MaxEnt based on Entropy Regularization (ER) (Vapnik, 1998; Jiao et al., 2006). SVM and MaxEnt have proved successful in information structure analysis (e.g. (Merity et al., 2009; Guo et al., 2011)) but, to the best of our knowledge, their semi-supervised versions have not been used for AZ of full articles.

Table 4: Class distribution (shown in percentages) in articles and their individual sections in the AZ-annotated corpus.

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

Table 5: Performance of baselines on the *Discussion* section.

| | BKG | PROB | METH | RES | CON | CN | DIFF | FUT |
|---|---|---|---|---|---|---|---|---|
| **Full supervision** | | | | | | | | |
| SVM | .56 | 0 | 0 | 0 | .84 | .35 | 0 | 0 |
| MaxEnt | .55 | .08 | 0 | 0 | .84 | .38 | 0 | 0 |
| **Light supervision with 150 labeled sentence** | | | | | | | | |
| SVM | .26 | 0 | 0 | 0 | .80 | .05 | 0 | 0 |
| TSVM | .25 | .04 | .04 | .03 | .33 | 14 | .06 | .02 |
| MaxEnt | .25 | 0 | 0 | 0 | .80 | .10 | 0 | 0 |
| MaxEnt+ER | .23 | 0 | 0 | 0 | .80 | .07 | 0 | 0 |

**Parameter tuning** The boundaries of the reference probabilities ($a_k$ and $b_k$ in Equation (8)) were defined and optimized on the development data which consists of one third of the corpus. We considered six types of boundaries: Fairly High for 1, High for [0.9,1), Medium High for [0.5,0.9), Medium Low for [0.1,0.5), Low for [0,0.1), and Fairly Low for 0.

**Evaluation** We evaluated the precision, recall and F-score for each category, using a standard ten-fold cross-validation scheme. The models were tested on each of the ten folds and trained on the rest of them, and the results were averaged across the ten folds.

## 7 Results

We report results at two levels of granularity. We first provide detailed results for the *Discussion* section which should be, as is clearly evident from Table 4, the most difficult section for AZ prediction as only 63.5% of its sentences take its most dominant class (CON). As we show below, this is also where our constrained model is most effective. We then show the advantages of our model for other sections.

**Results for the *Discussion* section** To get a bet-

Table 6: *Discussion* section performance of MaxEnt, MaxEnt+GE and a MaxEnt+GE model that does not include our top-down classification scheme. Results are presented for different amounts of labeled training data. The MaxEnt+GE (Top-down) model outperforms the MaxEnt in 44 out of 48 cases, and MaxEnt+GE (Flat) in 39 out of 48 cases.

| | **MaxEnt** | | | | | | **MaxEnt + GE (Top-down)** | | | | | | **MaxEnt + GE (Flat)** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 50 | 100 | 150 | 500 | 1000 | Full | 50 | 100 | 150 | 500 | 1000 | Full | 50 | 100 | 150 | 500 | 1000 | Full |
| BKG | .10 | .26 | .25 | .44 | .48 | .55 | .49 | .49 | .48 | .52 | .55 | .57 | .35 | .37 | .37 | .46 | .51 | .53 |
| PROB | 0 | 0 | 0 | 0 | 0 | 0 | .38 | .16 | .29 | .13 | .30 | .41 | .38 | .23 | .19 | .39 | .38 | .33 |
| METH | 0 | 0 | 0 | 0 | 0 | 0 | .17 | .22 | .37 | .35 | .50 | .39 | .16 | .17 | .21 | .24 | .32 | .29 |
| RES | 0 | 0 | 0 | 0 | 0 | 0 | .18 | .24 | .58 | 0 | 0 | .46 | .13 | .05 | .21 | .31 | .25 | .34 |
| CON | .79 | .80 | .80 | .83 | .83 | .84 | .77 | .78 | .82 | .83 | .84 | .84 | .63 | .66 | .68 | .74 | .78 | .78 |
| CN | .02 | .04 | .10 | .24 | .34 | .38 | .29 | .31 | .33 | .35 | .40 | .39 | .21 | .21 | .24 | .26 | .30 | .32 |
| DIFF | 0 | 0 | 0 | 0 | 0 | 0 | .26 | .25 | .25 | .19 | .24 | .21 | .14 | .16 | .15 | .14 | .18 | .17 |
| FUT | 0 | 0 | 0 | 0 | 0 | 0 | .35 | .38 | .31 | .25 | .35 | .31 | .36 | .36 | .39 | .33 | .25 | .37 |

Figure 2: Performance of the MaxEnt and MaxEnt+GE models on the *Introduction* (left), *Methods* (middle) and *Results* (right) sections. The MaxEnt+GE model is superior.

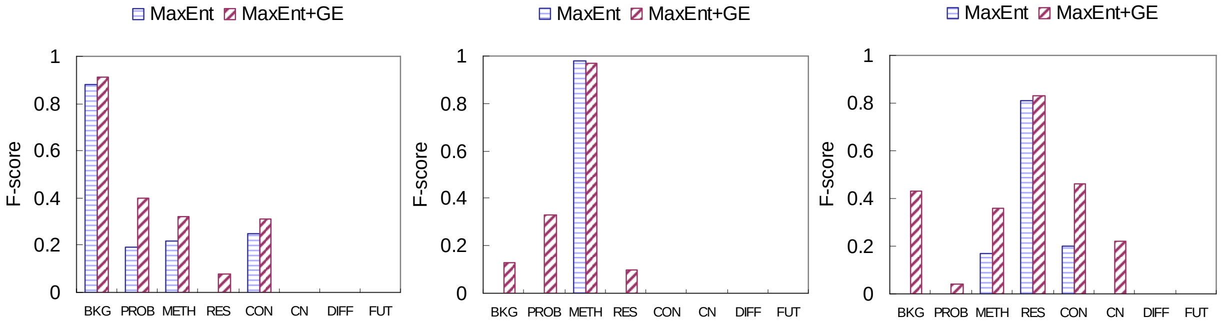


Table 7: *Discussion* section performance of the MaxEnt, MaxEnt+GE and unsupervised GE models when the former two are trained with 150 labeled sentences. Unsupervised GE outperforms the standard MaxEnt model for all categories except for CON – the major category of the section. The result pattern for the other sections are very similar.

| | **MaxEnt** | | | **MaxEnt + GE** | | | **Unsup GE** | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | P | R | F |
| BKG | .38 | .19 | .25 | .49 | .48 | **.48** | .49 | .44 | .46 |
| PROB | 0 | 0 | 0 | .38 | .23 | .29 | .28 | .38 | **.32** |
| METH | 0 | 0 | 0 | .29 | .50 | **.37** | .08 | .56 | .14 |
| RES | 0 | 0 | 0 | .68 | .51 | **.58** | .08 | .51 | .14 |
| CON | .69 | .96 | .80 | .81 | .84 | **.82** | .74 | .69 | .71 |
| CN | .35 | .06 | .10 | .39 | .29 | **.33** | .40 | .13 | .20 |
| DIFF | 0 | 0 | 0 | .21 | .30 | **.25** | .12 | .13 | .12 |
| FUT | 0 | 0 | 0 | .24 | .44 | .31 | .26 | .61 | **.36** |

Table 8: Analysis of the impact of the different constraint types for the lightly supervised and the fully supervised cases. Results are presented for the *Discussion* section. Using only the lexical constraints is generally preferable in the fully supervised case. Combining the different constraint types is preferable for the lightly supervised case.

| | **Discourse** | | **Lexical** | | **Discourse+Lexical** | |
|---|---|---|---|---|---|---|
| | 150 | Full | 150 | Full | 150 | Full |
| BKG | .29 | .55 | .46 | **.58** | **.48** | .57 |
| PROB | 0 | 0 | **.37** | .40 | .29 | **.41** |
| METH | 0 | .11 | .29 | .35 | **.37** | **.39** |
| RES | 0 | .06 | .32 | **.47** | **.58** | .46 |
| CON | .81 | .84 | .80 | **.84** | **.82** | **.84** |
| CN | .12 | .34 | **.35** | **.42** | .33 | .39 |
| DIFF | 0 | 0 | .21 | **.21** | **.25** | **.21** |
| FUT | 0 | 0 | 0 | .29 | **.31** | **.31** |

ter understanding of the nature of the challenge we face, Table 5 shows the F-scores of fully- and semi-supervised SVM and MaxEnt on the *Discussion* section. The dominant zone category CON, which accounts for 63.5% of the section sentences, has the highest F-scores for all methods and scenarios. Most of the methods also identify the second and the third most frequent zones BKG and CN, but with relatively lower F-scores. Other low-frequency categories can hardly be identified by any of the methods regardless of the amount of labeled data available for training. Note that the compared models perform quite similarly. We therefore use the MaxEnt model, which is most naturally augmented with GE constraints, as the baseline unconstrained model.

When adding the GE constraints we observe a substantial performance gain, in both the fully and the lightly supervised cases, especially for the low-frequency categories. Table 6 presents the F-scores of MaxEnt with and without GE constraints ("MaxEnt+GE (Top-down)" and "MaxEnt") in the light and full supervision scenarios. Incorporating GE into MaxEnt results in a substantial F-score improvement for all AZ categories except for the major category CON for which the performance is kept very similar. In total, MaxEnt+GE (Top-down) is better in 44 out of the 48 cases presented in the table. Importantly, the constrained model provides substantial improvements for both the relatively high-frequency classes (BKG and CN which together label 30.2% of the sentences) and for the low-frequency classes (which together label 6.4% of the sentences).

The table also clearly demonstrates the impact of our tree-based top-down classification scheme, by comparing the Top-down version of MaxEnt+GE to the standard "Flat" version. In 39 out of 48 cases, the Top-down model performs better. In some cases, especially for high-frequency categories and when the amount of training data increases, unconstrained MaxEnt even outperforms the flat MaxEnt+GE model. The results presented in the rest of the paper for the MaxEnt+GE model therefore refer to its Top-down version.

**All sections** We next turn to the performance of our model on the three other sections. Our experiments show that augmenting the MaxEnt model with domain knowledge constraints improves performance for all the categories (either low or high frequency), except the major section category, and keep the performance for the latter on the same level. Figure 2 demonstrates this pattern for the lightly supervised case with 150 training sentences but the same pattern applies to all other amounts of training data, including the fully supervised case. Naturally, we cannot demonstrate all these cases due to space limitations. The result patterns are very similar to those presented above for the *Discussion* section.

**Unsupervised GE** We next explore the quality of the domain knowledge constraints when used in isolation from a feature-based model. The objective function of this model is identical to Equation (6) except that the first (likelihood) term is omitted. Our experiments reveal that this unsupervised GE model outperforms standard MaxEnt for all the categories except the major category of the section, when up to 150 training sentences are used. Table 7 demonstrates this for the *Discussion* section. This pattern holds for the other scientific article sections. Even when more than 150 labeled sentences are used, the unsupervised model better detects the low frequency categories (i.e. those that label less than 10% of the sentences) for all sections. These results provide strong evidence for the usefulness of our constraints even when they are used with no labeled data.

**Model component analysis** We finally analyze the impact of the different types of constraints on the performance of our model. Table 8 presents the *Discussion* section performance of the constrained model with only one or the full set of constraints. Interestingly, when the feature-based model is fully trained the application of the lexical constraints alone results in a very similar performance to the application of the full set of lexical and discourse constraints. It is only in the lightly supervised case where the full set of constraints is required and results in the best performing model.

## 8 Conclusions and Future Work

We have explored the application of posterior discourse and lexical constraints for the analysis of the information structure of scientific documents. Our results are strong. Our constrained model outperforms standard feature-based models by a large margin in both the fully and the lightly supervised cases. Even an unsupervised model based on these constraints provides substantial gains over feature-based models for most AZ categories.

We provide a detailed analysis of the results which reveals a number of interesting properties of our model which may be useful for future research. First, the constrained model significantly outperforms its unconstrained counterpart for low-medium frequency categories while keeping the performance on the major section category very similar to that of the baseline model. Improved modeling of the major category is one direction for future research. Second, our full constraint set is most beneficial in the lightly supervised case while the lexical constraints alone yield equally good performance in the fully supervised case. This calls for better understanding of the role of discourse constraints for our task as well as for the design of additional constraints that can enhance the model performance either in combination with the existing constraints or when separately applied to the task. Finally, we demonstrated that our top-down tree classification scheme provides a substantial portion of our model's impact. A clear direction for future research is the design of more fine-grained constraint taxonomies which can enable efficient usage of other constraint types and can result in further improvements in performance.

## References

Kedar Bellare, Gregory Druck, and Andrew McCallum. 2009. Alternating projections for learning with expectation constraints. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 43–50, Arlington, Virginia, United States. AUAI Press.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71.

Catherine Blake. 2009. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *J Biomed Inform*, 43(2):173–89.

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.

M.W. Chang, L. Ratinovc, and D. Roth. 2007. Guiding semi-supervision with constraint-driven learning. In *ACL*.

Stanley F. Chen, Ronald Rosenfeld, and Associate Member. 2000. A survey of smoothing techniques for me models. *IEEE Transactions on Speech and Audio Processing*, 8:37–50.

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

J. R. Curran, S. Clark, and J. Bos. 2007. Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the ACL 2007 Demonstrations Session*, pages 33–36.

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.

Miroslav Dudík. 2007. *Maximum entropy density estimation and modeling geographic distributions of species*. Ph.D. thesis.

K. Ganchev, J. Graca, J. Gillenwater, and B. Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*.

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 *Proceedings of BioNLP*, pages 99–107.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 273–283.

Ben Hachey and Claire Grover. 2006. Extractive summarisation of legal texts. *Artif. Intell. Law*, 14:305–345.

K. Hirohata, N. Okazaki, S. Ananiadou, and M. Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of 3rd International Joint Conference on Natural Language Processing*, pages 381–388.

E. T. Jaynes. 1957. Information Theory and Statistical Mechanics. *Physical Review Online Archive (Prola)*, 106(4):620–630.

F. Jiao, S. Wang, C. Lee, R. Greiner, and D. Schuurmans. 2006. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *COLING/ACL*, pages 209–216.

M. Liakata, S. Teufel, A. Siddharthan, and C. Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of LREC'10*.

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

J. Lin, D. Karakos, D. Demner-Fushman, and S. Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *Proceedings of BioNLP-06*, pages 65–72.

G. Mann and A. McCallum. 2007. Simple, robust, scalable semi-supervised learning via expectation regularization. In *ICML*.

G. Mann and A. McCallum. 2008. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *ACL*.

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

A. K. McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

S. Merity, T. Murphy, and J. R. Curran. 2009. Accurate argumentative zoning with maximum entropy models. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 19–26.

Y. Mizuta, A. Korhonen, T. Mullen, and N. 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.

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

S. Della Pietra and V. Della Pietra. 1993. Statistical modeling by me. Technical report, IBM.

Roi Reichart and Anna Korhonen. 2012. Document and corpus level inference for unsupervised and transductive learning of information structure of scientic documents. In *Proceedings of COLING 2012*.

P. Ruch, C. Boyer, C. Chichester, I. Tbahriti, A. Geissbuhler, P. Fabry, J. Gobeill, V. Pillet, D. Rebholz-Schuhmann, C. Lovis, and A. L. Veuthey. 2007. Using argumentation to extract key sentences from biomedical abstracts. *Int J Med Inform*, 76(2-3):195–200.

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

L. Sun and A. Korhonen. 2009. Improving verb clustering with automatically acquired selectional preference. In *Proceedings of EMNLP*, pages 638–647.

I. Tbahriti, C. Chichester, Frederique Lisacek, and P. Ruch. 2006. Using argumentation to retrieve articles with similar citations. *Int J Med Inform*, 75(6):488–495.

S. Teufel and M. Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28:409–445.

S. Teufel, A. Siddharthan, and C. Batchelor. 2009. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *EMNLP*.

V. N. Vapnik. 1998. *Statistical learning theory*. Wiley, New York.

Andrea Varga, Daniel Preotiuc-Pietro, and Fabio Ciravegna. 2012. Unsupervised document zone identification using probabilistic graphical models. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

B. Webber, M. Egg, and V. Kordoni. 2011. Discourse structure and language technology. *Natural Language Engineering*, 18:437–490.