Unsupervised learning of rhetorical structure with un-topic models

Diarmuid Ó Séaghdha
Computer Laboratory
University of Cambridge
Cambridge, UK
do242@cam.ac.uk

Simone Teufel
Computer Laboratory
University of Cambridge
Cambridge, UK
sht25@cam.ac.uk

Abstract

In this paper we investigate whether unsupervised models can be used to induce conventional aspects of rhetorical language in scientific writing. We rely on the intuition that the rhetorical language used in a document is general in nature and independent of the document's topic. We describe a Bayesian latent-variable model that implements this intuition. In two empirical evaluations based on the task of argumentative zoning (AZ), we demonstrate that our generality hypothesis is crucial for distinguishing between rhetorical and topical language and that features provided by our unsupervised model trained on a large corpus can improve the performance of a supervised AZ classifier.

1 Introduction

Scientific writing has many conventions. Some exist at the level of sentence construction, such as a preference for the passive voice or for deverbal nominalisations. Others relate to the high-level organisation of a paper: a typical paper at an NLP conference may be divided into sections covering the introduction, related work, methods, experimental results and conclusion. There are also intermediate levels of convention that use lexical and phrasal items to signal the role played by each part of the text in the argument the authors wish to construct. The theory of *argumentative zoning* (AZ) describes how a scientific article can be analysed in terms of text blocks (or *zones*) that share a rhetorical function (Teufel, 2010). For example: part of the article may consist of background information, another part may describe the aim of the research, other parts may report the authors' own work or compare that work to alternative approaches in the literature. Supervised computational systems can be trained to mark up the AZ structure of a text automatically (see Section 2); the output of such systems has been shown to aid summarisation and human browsing of the scientific literature (Teufel and Moens, 2002; Guo et al., 2011a; Contractor et al., 2012). However, supervised systems require manually annotated training data that must be created anew for each discipline (and language) before they can be deployed, while large quantities of unannotated text are often available. For this reason, there is considerable value in developing unsupervised systems that induce aspects of rhetorical structure from unannotated text.

In this paper we advance a hypothesis about the *generality* of rhetorical language. We propose that the words and linguistic constructs used to express rhetorical function in a scientific paper are independent of the paper's topic. Naturally there will be some variation across research areas and there may be large differences across disciplines, but within a discipline we do not expect that the specific subject of a paper plays a significant role in how the authors construct their argument. For example, the following template could be used to generate an abstract for very many papers in NLP and other fields:

> The problem of _____ has received a lot of attention because of its relevance to _____. CITATION proposed an approach based on the method of _____. In this paper we present a method for _____ that has the following advantages over prior work: _____. We demonstrate the empirical effectiveness of our method by reporting experiments on _____ data, where it outperforms the approach of CITATION by __%.

This leads us to the idea of two-stage "recipes" for scientific papers, whereby the authors start with a framework of boilerplate text that matches the rhetorical argument they wish to make. The authors can then fill in the gaps with the substance of their research contribution.

The two-stage model is of course an idealisation of how scientists construct their papers, but it is useful as an inspiration for a computational model that implements the generality hypothesis. We propose BOILERPLATE-LDA, a generative model that assigns responsibility for generating each word in an abstract to a document-specific topic model or to a rhetorical language model that is not specific to the document. Essentially, we induce argumentative structure from the parts of the text that are not well-explained by the topic model. Hence we describe BOILERPLATE-LDA as an "un-topic model". We evaluate our model in two settings: a clustering evaluation that treats BOILERPLATE-LDA as performing unsupervised argumentative zoning, and a downstream evaluation where the induced structure is not taken as explicitly modelling argumentative zones but is used to provide informative features for a supervised AZ classifier. In both cases, we show that BOILERPLATE-LDA performs well on a very challenging task.

2 Related work

There has been great interest in unsupervised learning among NLP researchers due to the availability of large amounts of unprocessed text through the Web, newswire providers, scientific repositories and other sources in contrast to the onerous requirements of creating task-specific manually annotated data for training supervised analysers. Particularly relevant to our work is the field of topic modelling, where Bayesian latent-variable models are used to induce meaningful generalisations from observations of co-occurrences. Blei et al. (2003) introduced Latent Dirichlet Allocation (LDA) as a model of thematic structure in documents, but subsequent work has adapted the general framework to many different purposes in modelling text as well as other kinds of data. This includes research on modelling aspects of document structure such as topic segmentation, implementing the intuitions that neighbouring blocks of text are coherent in the sense of lexical similarity (Purver et al., 2006; Gruber et al., 2007; Eisenstein and Barzilay, 2008; Du et al., 2013). The model most similar to ours (that we are aware of) is the model of Ritter et al. (2010), which captures dialogue acts and transitions between them in Twitter conversations.

Despite the general popularity of unsupervised approaches, rhetorical analysis has generally been treated as a problem for supervised machine learning. Classification-based approaches to argumentative zoning typically use a sequence classifier such as a maximum-entropy Markov model or conditional random field (Teufel and Moens, 2002; Siddharthan and Teufel, 2007; Hirohata et al., 2008; Guo et al., 2010). Guo et al. (2011b) take a semi-supervised approach based on active learning and self-training.

Two unsupervised approaches in the literature are Varga et al. (2012) and Reichart and Korhonen (2012). Varga et al. use a topic model variant called ZONE-LDA that assigns each sentence a latent variable index or "topic" and assumes that the words in the sentence are generated from a distribution particular to the topic; in this situation each topic is assumed to correspond to a distinct argumentative zone. Such a model will have the effect of clustering sentences that share lexical items. Varga et al. also propose a model they call ZONE-LDA-B, in which some common words are assigned to a "background" distribution that is independent of the sentence category; this model performs worse than ZONE-LDA in their evaluation. Reichart and Korhonen take an approach based on Markov random fields. They construct a graphical model in which sentence vertices are connected by potentials weighted according to adjacency and sentence similarity, as well as hand-defined rules about passivisation and sentence location.

The papers cited in the two preceding paragraphs have focused on rhetorical analysis in scientific writing, yet there are many other textual genres where argumentation is conventionalised. For example, Burstein et al. (2003) identify building blocks analogous to AZ zones in the writing of English language learners and demonstrate that a supervised classification approach can be used to mark up their essays. Also in the educational domain, Madnani et al. (2012) train a supervised classifier to detect the "shell" language that learners use to organise the high-level structure of their compositions; this is quite close to our idea of "templates" or "recipes" for scientific papers. Sauper and Barzilay (2009) and Chen et al. (2009) both present models that learn structural conventions in Wikipedia articles without relying on human annotation. Sauper and Barzilay's model induces the typical section structure of Wikipedia articles

about a specific entity type (e.g., *Actors* or *Diseases*) and retrieves web snippets relevant to each section for a target entity, before performing multidocument summarisation to produce a new entry for posting to Wikipedia. Chen et al. take a Bayesian segmentation approach to implicitly learn the topical section structure of articles and use a generalised Mallows model, a distribution over permutations, to identify a canonical ordering for sections.[1] Other forms of general rhetorical analysis include Rhetorical Structure Theory (Mann and Thompson, 1988; Marcu, 2000), which captures local discourse relations between segments of text; RST provides a layer of analysis that is separate and complementary to more global schemes such as argumentative zoning.

3 Intuitions

The performance of unsupervised learning depends on how intuitions about the task are incorporated in the statistical model. Our approach relies on three main intuitions:

Sentence similarity: All else being equal, we expect that lexically similar sentences will have similar purposes. At the same time, lexical similarity alone is not sufficient to capture shared argumentative function: all sentences in a paper about parsing will be similar to each other, while the introductory sentences of a parsing paper and a machine translation paper may share few similar lexical items.

Adjacency: The theory of argumentative zones suggests that sentences with the same rhetorical function will often be grouped together into blocks. Additionally, we expect that authors will follow general conventions about the order of zones, e.g., starting with background and goal statements and progressing to results and conclusions.

Generality: We expect that the language used to convey rhetorical function is independent of the topical content of the paper.

Sentence similarity can be captured using standard lexical similarity measures or through the clustering effects of a topic model. The adjacency assumption can be implemented using a linear-chain sequence model such as a Hidden Markov Model. The ZONE-LDA approach of Varga et al. (2012) relies on sentence similarity alone. Reichart and Korhonen's (2012) model combines sentence similarity and adjacency. To the best of our knowledge, the generality hypothesis has not previously been investigated. The model we describe in Section 4 incorporates all three intuitions in its structure.

4 Models

The model we propose assumes that each word in a sentence is generated either from an LDA-style topic model or from a distribution associated with the rhetorical category assigned to the sentence. The former captures the subject matter of the document; the latter captures conventional language that is independent of the document's subject matter. The sentence categories are generated from a first-order Markov model. The assignment of responsibility for a word is implemented through a so-called "switching variable", a binary-valued latent variable. This is a commonly used mechanism for interpolating language models (Griffiths et al., 2004; Reisinger and Mooney, 2010; Ahmed and Xing, 2010); in many cases, the goal is to assign common words to a "background" distribution that is not considered an object of interest from a topic modelling perspective. In our case it is this non-topical part of the text that is the object of interest.

The dependencies between variables in our full BOILERPLATE-LDA model are shown by the plate diagram in Figure 1. The corresponding "generative story" is as follows:

[1] It would be interesting to swap in Chen et al.'s generalised Mallows model for the HMM-style ordering model in BOILERPLATE-LDA. The former has the advantage of capturing non-local ordering effects, while the latter has the advantage of not assuming a single canonical ordering.

for topic $t \in \{1 \dots |T|\}$ **do**
  (Draw a distribution over words)
  $\boldsymbol{\Phi}_t \sim Dirichlet(\boldsymbol{\beta})$
end for
for zone $z \in \{1 \dots |Z|\}$ **do**
  (Draw a distribution over words)
  $\boldsymbol{\Psi}_z \sim Dirichlet(\boldsymbol{\gamma})$
  (Draw a transition distribution)
  $\boldsymbol{\Lambda}_z \sim Dirichlet(\boldsymbol{\lambda})$
end for
(Draw the switch distribution)
$\Sigma \sim Beta(\sigma_0, \sigma_1)$
for doc $d \in \{1 \dots |D|\}$ **do**
  (Draw a distribution over topics)
  $\boldsymbol{\theta}_d \sim Dirichlet(\boldsymbol{\alpha})$
  for sentence $s \in Sentences(d)$ **do**
    $z_s \sim Multinomial(\boldsymbol{\Lambda}_{z_{s-1}})$
    for word $i \in Words(s)$ **do**
      (Draw a switch indicator)
      $b_i = Beta(\Sigma)$
      if $b_i = 0$ **then**
        (Draw a word from the zone-word distribution)
        $w_i \sim Multinomial(\boldsymbol{\Psi}_{z_s})$
      else
        (Draw a topic)
        $t_i \sim Multinomial(\boldsymbol{\theta}_d)$
        (Draw a word from the topic-word distribution)
        $w_i \sim Multinomial(\boldsymbol{\Phi}_{t_i})$
      end if
    end for
  end for
end for

We train the model using Gibbs sampling. Due to Dirichlet-multinomial and beta-Bernoulli conjugacy it is relatively straightforward to integrate out the multinomial and Bernoulli distribution parameters $\boldsymbol{\theta}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$ and Σ and derive update rules for a collapsed Gibbs sampler. Each iteration of the sampler visits each sentence in the corpus in turn, first sampling the sentence label assignment z_s and then sampling for each word in the sentence the switch indicator b_i and (if $b_i = 1$) the topic assignment t_i. The sentence label update is performed using what Gao and Johnson (2008) call a pointwise collapsed Gibbs sampler. Omitting hyperparameters for clarity, the sampling probabilities can be written as

$$P(z_i = z|\boldsymbol{z}^{-i}, \boldsymbol{w}, \boldsymbol{b}) \propto \frac{f_{z_{i-1}\to z} + \kappa_z}{f_{z_{i-1}} + \sum_{z'} \kappa_{z'}} \frac{f_{z\to z_{i+1}} + I(z = z_{i+1}) + \kappa_{z_{i+1}}}{f_z^{-i} + I(z = z_{i+1}) + \sum_{z'} \kappa_{z'}} \prod_{v \in V} \frac{\Gamma(f^{-i}_{zv,b=0} + f_{s_i v,b=0} + \gamma)}{\Gamma(f_z^{-i} + f_{s_i} + \gamma|V|)} \tag{1}$$

where $f_{z->z'}$ is the transition frequency from zone z to zone z', f_z is the number of sentences assigned zone z; $I(z = z_{i+1})$ has value 1 if the two zone assignments are equal and 0 otherwise; V is the vocabulary of word types; $f_{zv,b=0}$ is the number of words of type z that appear in sentences assigned zone z and whose corresponding switch variable has value 0; $f_{s_i v,b=0}$ is the number of words of type v that appear in sentence s_i and whose corresponding switch variable has value 0; the superscript $^{-i}$ indicates that the frequency is calculated over all sentences except s_i. We introduce observed start and end state variables z_s and z_e to handle the boundaries at the beginning and end of each document.

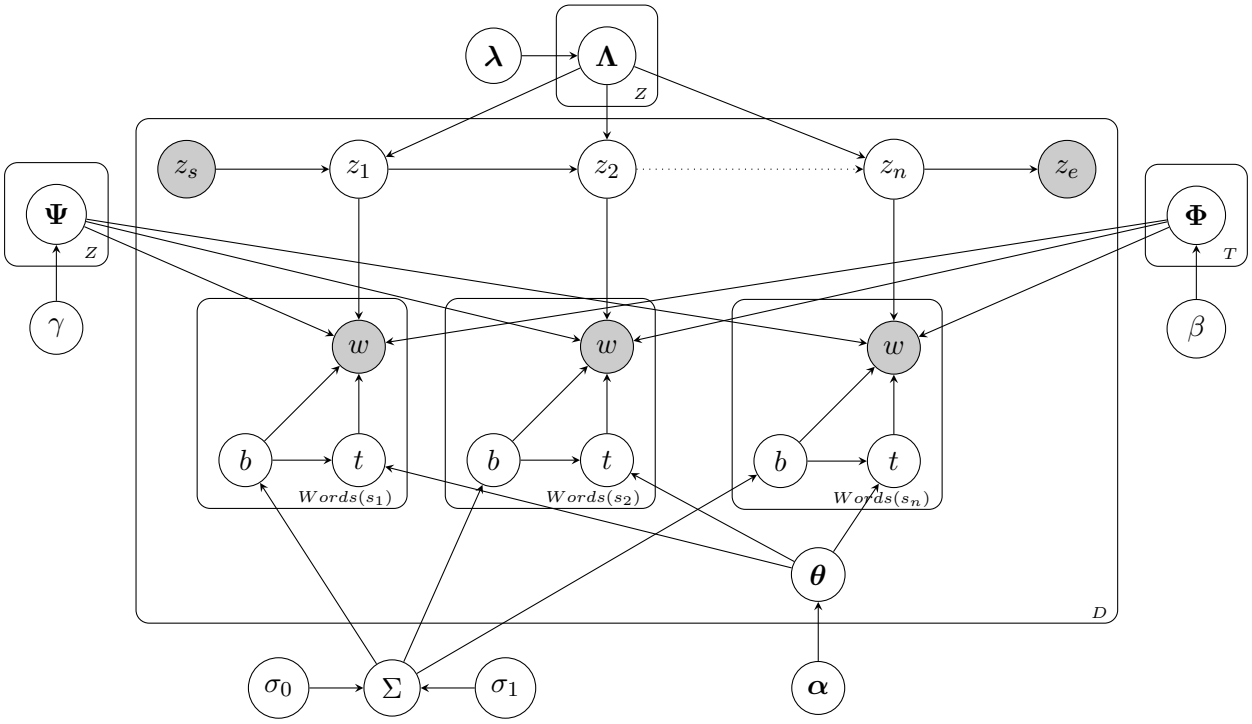

Figure 1: Plate diagram for BOILERPLATE-LDA

The topic and switch variables for each word are sampled in a blocked fashion; the sampling probabilities are similar to the standard LDA updates:

$$
\begin{aligned}
P(b_j = 0, t_j = \emptyset | \mathbf{z}^{-j}, \mathbf{b}^{-j}, \mathbf{t}, \mathbf{w}) &\propto (f^{-j}_{b=0} + \Sigma_0) \frac{f^{-j}_{z_i w_j, b=0} + \gamma}{f_{z_i, b=0} + |V|\gamma} \\
P(b_j = 1, t_j = t | \mathbf{z}^{-j}, \mathbf{b}^{-j}, \mathbf{t}, \mathbf{w}) &\propto (f^{-j}_{b=1} + \Sigma_1) \frac{f^{-j}_{tw_j} + \alpha_z}{f^{-j}_{w_j, b=1} + \sum_{z'} \alpha_{z'}} \frac{f^{-j}_{zw_j} + \beta}{f^{-j}_{z} + |V|\beta} \\
P(b_i = 0, t_i \neq \emptyset | \mathbf{z}^{-j}, \mathbf{b}^{-j}, \mathbf{t}, \mathbf{w}) &= 0 \\
P(b_i = 1, t_i = \emptyset | \mathbf{z}^{-j}, \mathbf{b}^{-j}, \mathbf{t}, \mathbf{w}) &= 0
\end{aligned} \tag{2}
$$

where we use j to index words and i to index sentences; f_{tw_j} is the number of words of type w_j that are assigned topic t; the superscript $^{-j}$ indicates that the frequency is calculated over all words except j.

5 Experiments

5.1 Data

For evaluation, we use a collection of abstracts compiled by Guo et al. (2010). These abstracts had originally been collected in the context of semi-automated cancer risk assessment by searching PubMed for abstracts mentioning one or more of a list of chemicals known to have carcinogenic properties (Korhonen et al., 2009). Guo et al. annotated abstracts for five of these chemicals using an AZ scheme with seven categories: *Background*, *Objective*, *Method*, *Result*, *Conclusion*, *Related work* and *Future work*.[2] In order to test whether our models can also perform over a large, heterogeneous dataset, we also used a collection of 129,595 abstracts taken from a collection of open-access journal articles. Preprocessing involved sentence splitting, tokenisation and part-of-speech tagging using the Stanford CoreNLP toolkit[3] and the removal of all tokens containing non-alphanumeric characters, all tokens of character length one

[2]The annotated dataset has been made available at `http://www.cl.cam.ac.uk/~yg244/abstract_az.html`.
[3]`http://nlp.stanford.edu/software/corenlp.shtml`

and a small set of stop words.[4] This left a training corpus of 16,841,280 tokens.

5.2 Clustering Evaluation

5.2.1 Evaluation

Our first quantitative evaluation investigates whether the zones induced by BOILERPLATE-LDA correspond to the argumentative zones identified by human theorists. We treat this as a clustering task with the gold standard provided by Guo et al.'s (2010) dataset. The clustering evaluation measures we use are the Adjusted Rand Index (Hubert and Arabie, 1985) and Adjusted Mutual Information (Vinh et al., 2010); both measures are normalised to have a maximum value of 1 and are adjusted for chance so that the expected score given to a random clustering is 0. This second property makes them conservative in comparison to other evaluation measures. We report results with the number of zones $|Z| \in \{10, 20, 50\}$ and number of topics $|T| \in \{10, 20, 50, 100\}$; for each combination of settings we report the average evaluation score attained by three independent runs of the learning algorithm.

5.2.2 Models

For our evaluation, we test the following models:

BOILERPLATE-LDA: Our full model, as described in Section 4.

BOILERPLATE-LDA-MULT: A simplified model where the Markov dependencies between zone assignments are replaced by a flat multinomial; the probability of a zone is independent of the adjacent sentences.

BOILERPLATE-LDA-NOTOPICS: A simplified model where all words in a sentence are generated from the zone distribution $\boldsymbol{\Psi}_{z_s}$; this is almost identical to Varga et al.'s (2012) ZONE-LDA model.

K-MEANS: A standard k-means clustering model run until convergence. The features for each sentence consist of tf-idf-transformed lexical frequencies, part-of-speech tags and a location feature computed by dividing the abstract into 5 bins.

The BOILERPLATE-LDA models are all trained for 1000 iterations of Gibbs sampling. The Dirichlet hyperparameters are re-estimated every 10 iterations; the topic hyperparameters $\boldsymbol{\alpha}$ are optimised using a fixed-point iteration to maximise the log-evidence (Minka, 2003; Wallach, 2008), while the other hyperparameters are sampled using Hamiltonian Monte Carlo (Neal, 2010). K-MEANS was run until convergence.

5.2.3 Results

Figure 2 gives an illustration of the zone representation induced at the end of one run of BOILERPLATE-LDA with the settings $|Z| = 10$, $|T| = 100$. Firstly, we list the most probable words for each zone (2a). While the model may not find a perfect match for the gold-standard inventory of argumentative zones, we can see that some induced zones describe standard methodology (8,9), others describe results and implications (1,3,7) and others describe motivations (2,5,6). Inspection of the transition matrix (2b) confirms our expectation that self-transitions have the highest probability; we also observed that the zones most frequently transitioned to from the start state are the motivational zones and the zones most frequently transitioned from to the end state are the results/implications zones. The example abstracts in Figure 3 illustrate how BOILERPLATE-LDA can be used to mark up the text of an abstract as "boilerplate" or "non-boilerplate" based on the values of the switch variables b_i.

[4] The part-of-speech tags are not used by BOILERPLATE-LDA but they are used as features for other models.

1	results, suggest, our, data, study, role, findings, we, between, indicate, important, studies
2	study, we, using, used, investigated, determine, present, between, investigate, analysis, aim
3	increased, significantly, levels, showed, found, observed, significant, after, compared, higher
4	two, sequence, we, found, region, sequences, we, three, identified, between, different, analysis
5	use, more, studies, study, used, however, important, health, most, treatment, clinical, potential
6	role, important, known, studies, however, shown, including, involved, mechanisms, cell
7	case, we, patient, report, rare, most, common, reported, presented, disease, associated, cause
8	CI, significantly, respectively, significant, between, group, mean, higher, compared, more, found
9	study, years, using, two, patients, included, total, group, three, data, after, used, collected, age
10	we, data, analysis, used, using, new, approach, based, method, information, developed, more

(a) Most probable words for each zone

From \ To	Start	1	2	3	4	5	6	7	8	9	10	End
Start	0.00	0.00	0.10	0.01	0.08	0.24	0.36	0.10	0.00	0.03	0.08	0.00
1	0.00	0.37	0.01	0.04	0.01	0.03	0.01	0.00	0.00	0.00	0.01	0.50
2	0.00	0.02	0.26	0.25	0.07	0.01	0.02	0.00	0.05	0.29	0.01	0.00
3	0.00	0.27	0.02	0.59	0.02	0.02	0.02	0.00	0.02	0.00	0.01	0.04
4	0.00	0.12	0.02	0.09	0.62	0.00	0.03	0.00	0.01	0.00	0.05	0.05
5	0.00	0.01	0.10	0.00	0.00	0.55	0.03	0.02	0.00	0.06	0.04	0.18
6	0.00	0.06	0.21	0.05	0.05	0.05	0.50	0.01	0.00	0.01	0.05	0.02
7	0.00	0.02	0.02	0.01	0.01	0.15	0.04	0.63	0.01	0.02	0.00	0.08
8	0.00	0.09	0.01	0.14	0.01	0.07	0.00	0.02	0.61	0.04	0.01	0.01
9	0.00	0.01	0.11	0.05	0.01	0.05	0.00	0.02	0.20	0.54	0.01	0.00
10	0.00	0.05	0.02	0.01	0.07	0.02	0.01	0.00	0.01	0.01	0.68	0.12
End	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(b) Zone transition probabilities between adjacent sentences

Figure 2: Zones induced by one run of BOILERPLATE-LDA ($|Z| = 10$, $|T| = 100$)

VASP: A Volumetric Analysis of Surface Properties Yields Insights into Protein-Ligand Binding Specificity

Many algorithms that compare protein structures can reveal similarities that suggest related biological functions, even at great evolutionary distances. Proteins with related function often exhibit differences in binding specificity, but few algorithms identify structural variations that effect specificity. To address this problem, we describe the Volumetric Analysis of Surface Properties (VASP), a novel volumetric analysis tool for the comparison of binding sites in aligned protein structures. VASP uses solid volumes to represent protein shape and the shape of surface cavities, clefts and tunnels that are defined with other methods. Our approach, inspired by techniques from constructive solid geometry, enables the isolation of volumetrically conserved and variable regions within three dimensionally superposed volumes. We applied VASP to compute a comparative volumetric analysis of the ligand binding sites formed by members of the steroidogenic acute regulatory protein (StAR)-related lipid transfer (START) domains and the serine proteases. Within both families, VASP isolated individual amino acids that create structural differences between ligand binding cavities that are known to influence differences in binding specificity. Also, VASP isolated cavity subregions that differ between ligand binding cavities which are essential for differences in binding specificity. As such, VASP should prove a valuable tool in the studyof protein-ligand binding specificity.

A new usage of functionalized oligodeoxynucleotide probe for site-specific modification of a guanine base within RNA

Site-specific modification of RNA is of great significance to investigate RNA structure, function and dynamics. Recently, we reported a new method for sequence- and cytosine-selective chemical modification of RNA based on the functional group transfer reaction of the 1-phenyl-2-methylydene-1,3-diketone unit of the 6-thioguanosine base incorporated in the oligodeoxynucleotide probe. In this study, we describe that the functionality transfer rate is greatly enhanced and the selectivity is shifted to the guanine base when the reaction is performed under alkaline conditions. Detailed investigation indicated that the 2-amino group of the enolate form of rG is the reactant of the functionality transfer reaction. As a potential application of this efficient functionality transfer reaction, a pyrene group as a relatively large fluorescent group was successfully transferred to the target guanine base of RNA with a high guanine and site selectivity. This functionality transfer reaction with high efficiency and high site-selectivity would provide a new opportunity as a unique tool for the study of RNA.

Figure 3: Examples of abstracts marked up for boilerplate (underlined) and non-boilerplate (faded text) by BOILERPLATE-LDA

Model	$\|T\|$	$\|Z\| = 10$		$\|Z\| = 20$		$\|Z\| = 50$	
		ARI	NMI	ARI	NMI	ARI	NMI
BOILERPLATE-LDA	10	0.19	0.15	0.09	0.09	0.04	0.07
	20	0.20	0.16	0.03	0.10	0.03	0.08
	50	0.26	0.21	0.18	0.16	0.05	0.10
	100	**0.32**	**0.28**	0.20	0.20	0.07	0.14
BOILERPLATE-LDA-MULT	10	0.13	0.11	0.08	0.08	0.04	0.06
	20	0.10	0.13	0.04	0.09	0.03	0.07
	50	0.21	0.16	0.13	0.14	0.06	0.10
	100	0.18	0.16	0.14	0.14	0.07	0.11
BOILERPLATE-LDA-NOTOPICS	0	0.00	0.02	0.04	0.05	0.06	0.05
K-MEANS	0	0.05	0.05	0.03	0.06	0.03	0.04

Table 1: Results of the clustering evaluation. $|Z|$ is the number of zones; $|T|$ is the number of topics.

The results of the clustering evaluation are presented in Table 1. Clearly, this is a challenging task; the BOILERPLATE-LDA-NOTOPICS and K-MEANS models, which do not filter out topic-specific vocabulary, perform little better than chance in terms of identifying argumentative zones (recall that for the ARI and AMI measures, zero means "not greater than expected by chance" rather than "no correlation at all"). BOILERPLATE-LDA-MULT performs better than those models though not as well as the full BOILERPLATE-LDA model, indicating that sequential structure is important for inducing rhetorical regularities. In general, the best results are attained with low settings of $|Z|$ and high settings of $|T|$; this seems to create the "bottleneck" effect needed to focus the model on purely rhetorical information. The highest scores (ARI = 0.32, AMI = 0.28) are attained by BOILERPLATE-LDA with the settings $|Z| = 10$, $|T| = 100$.

5.3 Supervised Evaluation

5.3.1 Evaluation

A second evaluation of BOILERPLATE-LDA's usefulness is to test whether it can yield features that improve the performance of a supervised argumentative zoning system. It is possible for an unsupervised model to induce structure that does not map exactly onto a pre-existing set of labels but still captures valuable information about the underlying phenomenon that can be of use to a supervised classifier when combined with other information sources. To this end, we train and evaluate supervised models on the same dataset of Guo et al. (2010) that we used for the clustering evaluation. We perform 10-fold cross-validation and report Accuracy (proportion of sentences labelled correctly) as well as macro-averaged Precision, Recall and F-Score. To measure statistical significance we use two-tailed paired t-tests, following Dietterich (1998).[5]

5.3.2 Models

We use two supervised sequence classification algorithms for training models:

LR: A logistic regression classifier with a "history" feature encoding the previous sentence's label, trained with L_1 regularisation, using the implementation in LibLinear.[6]

CRF: A first-order conditional random field classifier, trained with L_1 regularisation, using the implementation in Mallet.[7]

In both cases, the predicted labelling for a test document is given by the most probable (Viterbi) sequence according to the trained model. We use the following feature sets:

[5] In order to address concerns about the suitability of the t-tests under non-normality, we replicated the tests using Wilcoxon's signed-ranks test as recommended by Demšar (2006); the results were identical.

[6] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

[7] http://mallet.cs.umass.edu/

	LR				CRF			
Model	Acc	P	R	F	Acc	P	R	F
BASELINE	0.83	0.71	0.70	0.70	0.85	**0.75**	0.64	0.67
+BOILERPLATE-LDA	**0.84**	**0.72**	**0.71**	**0.71**	**0.86**	0.74	**0.65**	**0.68**
+LDA-BAG (50)	0.83	0.69	0.68	0.68	0.84	0.73	0.62	0.64
+LDA-BAG (100)	0.83	0.69	0.69	0.69	0.84	0.72	0.64	0.66
+LDA-MAX (50)	0.83	0.71	0.69	0.69	0.85	0.72	0.64	0.66
+LDA-MAX (100)	**0.84**	0.71	0.69	0.70	0.85	0.74	0.63	0.66

Table 2: Results of the supervised evaluation

BASELINE: Our baseline set of features is a standard set for supervised argumentative zoning: all unigrams and bigrams in the sentence, all part-of-speech tags in the sentence and a location feature computed by dividing the abstract into 5 bins.

+BOILERPLATE-LDA: The baseline model with additional features corresponding to the zone index assigned by BOILERPLATE-LDA to the sentence. We set $|Z| = 10$, $|T| = 100$ since that setting performed best in the clustering evaluation. As before, we use the output of three independently learned sampling chains, giving each sentence three zone features; the classifier should learn which chains are better than others during training.

+LDA-BAG: The baseline model with additional features derived from standard Latent Dirichlet Allocation models trained on the same corpus as BOILERPLATE-LDA. As LDA assigns a topic to each word in a sentence, we add all topics assigned to all words in the sentence as additional features. As above, we use the output of three sampling chains. We report results for models with 50 topics and 100 topics.

+LDA-MAX: The baseline model with additional features derived from LDA models. Here each model assigns each sentence the single topic assigned to the greatest number of words in the sentence (ties are broken randomly).

5.3.3 Results

Results for the supervised evaluation are presented in Table 2. +BOILERPLATE-LDA is the only augmented feature set that consistently gives an improvement over the baseline features. The improvements in accuracy are statistically significant ($p < 0.01$). In every case but one (which is not statistically significant), the LDA models fail to improve on the baseline in either accuracy or F-Score, showing that the latent structure induced by BOILERPLATE-LDA captures aspects of rhetorical language that are not captured by topical word clustering.

6 Conclusion

We consider the work presented in this paper to be a first step towards the ambitious goal of inducing latent descriptions of the templates used by scientists and writers in other fields. We have shown how our hypothesis about the generality of rhetorical language allows the construction of models that can separate out topical and rhetorical language use. One focus for future work will be to enrich the model structure; an approach based on adaptor grammars (Johnson et al., 2006) could be used to break the reductive unigram assumption in BOILERPLATE-LDA and identify multiword collocations that carry rhetorical information. Another focus will be to broaden our understanding of how unsupervised rhetorical models trained on large corpora can improve the robustness of supervised systems. For example, we have observed that lexicalised AZ classifiers trained on texts from one scientific domain will often perform poorly on texts from another domain; unsupervised models have the potential to induce relevant lexical commonalities across domains.

Acknowledgements

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D11PC20153. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Amr Ahmed and Eric P. Xing. 2010. Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of EMNLP-10*, Cambridge, MA.

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.

Harr Chen, S.R.K. Branavan, Regina Barzilay, and David R. Karger. 2009. Global models of document structure using latent permutations. In *Proceedings of NAACL-09*, Boulder, CO.

Danish Contractor, Yufan Guo, and Anna Korhonen. 2012. Using argumentative zones for extractive summarization of scientific articles. In *Proceedings of COLING-12*, Mumbai, India.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Lan Du, Wray Buntine, and Mark Johnson. 2013. Topic segmentation with a structured topic model. In *Proceedings of NAACL-13*, Atlanta, GA.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of EMNLP-08*, Honolulu, HI.

Jianfeng Gao and Mark Johnson. 2008. A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers. In *Proceedings of EMNLP-08*, Honolulu, HI.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2004. Integrating topics and syntax. In *Proceedings of NIPS-04*, Vancouver, BC.

Amit Gruber, Michal Rosen-Zvi, and Yair Weiss. 2007. Hidden topic Markov models. In *Proceedings of AISTATS-07*, San Juan, Puerto Rico.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: An investigation of three different schemes. In *Proceedings of BioNLP-10*, Uppsala, Sweden.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Johan Högberg, and Ulla Stenius. 2011a. A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics*, 12:69.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011b. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of EMNLP-11*, Edinburgh, UK.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of IJCNLP-08*, Hyderabad, India.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Proceedings of NIPS-06*, Vancouver, BC.

Anna Korhonen, Ilona Silins, Lin Sun, and Ulla Stenius. 2009. The first step in the development of text mining technology for cancer risk assessment: identifying and organizing scientific evidence in risk assessment literature. *BMC Bioinformatics*, 10:303.

Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of NAACL-12*, Montreal, QC.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

Thomas P. Minka. 2003. Estimating a Dirichlet distribution. Available at `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/`.

Radford M. Neal. 2010. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC Press, Boca Raton, FL.

Matthew Purver, Konrad Körding, Tom Griffiths, and Josh Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of COLING-ACL-06*, Sydney, Australia.

Roi Reichart and Anna Korhonen. 2012. Document and corpus level inference for unsupervised and transductive learning of information structure of scientific documents. In *Proceedings of COLING-12*, Mumbai, India.

Joseph Reisinger and Raymond Mooney. 2010. A mixture model with sharing for lexical semantics. In *Proceedings of EMNLP-10*, Cambridge, MA.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Proceedings of NAACL-HLT-10*, Los Angeles, CA.

Christina Sauper and Regina Barzilay. 2009. Automatically generating Wikipedia articles: A structure-aware approach. In *Proceedings of ACL-IJCNLP-09*, Singapore.

Advaith Siddharthan and Simone Teufel. 2007. Whose idea was this, and why does it matter? Attributing scientific work to citations. In *Proceedings of NAACL-07*, Rochester, NY.

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

Simone Teufel. 2010. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. CSLI Publications, Stanford, CA.

Andrea Varga, Daniel Preoţiuc-Pietro, and Fabio Ciravegna. 2012. Unsupervised document zone identification using probabilistic graphical models. In *Proceedings of LREC-12*, Istanbul, Turkey.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854.

Hanna M. Wallach. 2008. *Structured topic models for language*. Ph.D. thesis, University of Cambridge.