# Unsupervised document zone identification using probabilistic graphical models

**Andrea Varga*, Daniel Preoţiuc-Pietro†, Fabio Ciravegna***

*Organisations, Information and Knowledge Research Group,
†Natural Language Processing Research Group,
Department of Computer Science, 211 Portobello, Sheffield, S1 4DP
{a.varga, daniel, f.ciravegna}@dcs.shef.ac.uk

**Abstract**

Document zone identification aims to automatically classify sequences of text-spans (e.g. sentences) within a document into predefined zone categories. Current approaches to document zone identification mostly rely on supervised machine learning methods, which require a large amount of annotated data, which is often difficult and expensive to obtain.

In order to overcome this bottleneck, we propose graphical models based on the popular Latent Dirichlet Allocation (LDA) model. The first model, which we call zoneLDA aims to cluster the sentences into zone classes using only unlabelled data. We also study an extension of zoneLDA called zoneLDAb, which makes distinction between common words and non-common words within the different zone types.

We present results on two different domains: the scientific domain and the technical domain. For the latter one we propose a new document zone classification schema, which has been annotated over a collection of 689 documents, achieving a Kappa score of 85%.

Overall our experiments show promising results for both of the domains, outperforming the baseline model. Furthermore, on the technical domain the performance of the models are comparable to the supervised approach using the same feature sets. We thus believe that graphical models are a promising avenue of research for automatic document zoning.

**Keywords:** document zoning, probabilistic graphical models, unsupervised learning

## 1. Introduction

In many practical tasks there is a need to extract and access certain types of information from a large collection of textual documents. For example in corporate environments, such as manufacturing, there are a huge amount of unstructured historical data generated during the lifecycle of a product. In such environments, engineers wanting to resolve an issue on a particular product are often interested in finding out the cause of the issue, problems encountered on other similar product types and the conclusion drawn after each investigation.

Another example are the biomedical researchers aiming to stay abreast with current research they are typically interested in accessing information from PubMed[1] focusing on particular parts of the articles, such as the method proposed in the study, the results and conclusion obtained.

A major approach in such cases is to employ document zone classification to recognise the information structure of the documents, thus helping to assist information extraction and organisation of factual information from the documents.

The vast majority of the approaches applied for document zoning, rely on supervised machine learning (Liakata et al., 2010; Nawaz et al., 2010; Teufel et al., 2009), which can achieve state-of-the-art performance given that a large amount of annotated data is available. However, gathering these annotations is often time consuming and expensive. Furthermore, in most domains the format and style of the documents can change rapidly, resulting that these approaches could achieve suboptimal results, and thus collecting more training data might be required.

In this paper, we thus investigate the possibility of employing unsupervised approaches for document zoning, which to date has only been scarcely studied (Barzilay and Lee, 2004). We examine a couple of generative models for zone identification, which we call zoneLDA and zoneLDAb models, being extensions of the widely used Dirichlet Allocation (LDA) (Blei et al., 2003) model. Both of the proposed zoneLDA and zoneLDAb models can thus flexibly model the zones categories by ignoring the order in which the sentences occur in the documents. In addition, the zoneLDAb model discovers words which are common to the different zone types, and those are not good predictors for a zone category, and those that are specific to the zone types.

Our study also focused on evaluating our models on different domains, such as the scientific biomedical domain, and a more technical aerospace domain. For the latter one, we also proposed a novel document zone annotation schema.

Our extensive experiments of zoneLDA and zoneLDAb models show that although zoneLDA is equivalent with zoneLDAb without modelling the common words, in the majority of the cases zoneLDA outperforms zoneLDAb. These results are also superior to the baseline LDA model, and are close to the supervised Naive Bayes classifier for the aerospace domain, achieving an F-measure less than 5%.

The remainder of the paper is organised as follows. In Section 2. we present the current state-of-the-art approaches to document zone identification. In Section 3. we formally introduce the task of document zone identification. Section 4. describes our proposed generative models that are modified versions of Latent Dirichlet Allocation model. Next, in Section 5. we describe the datasets used in our experiments and the novel document zone classification schema proposed for the aerospace domain. Section 6. presents our

[1] http://www.ncbi.nlm.nih.gov/pubmed/

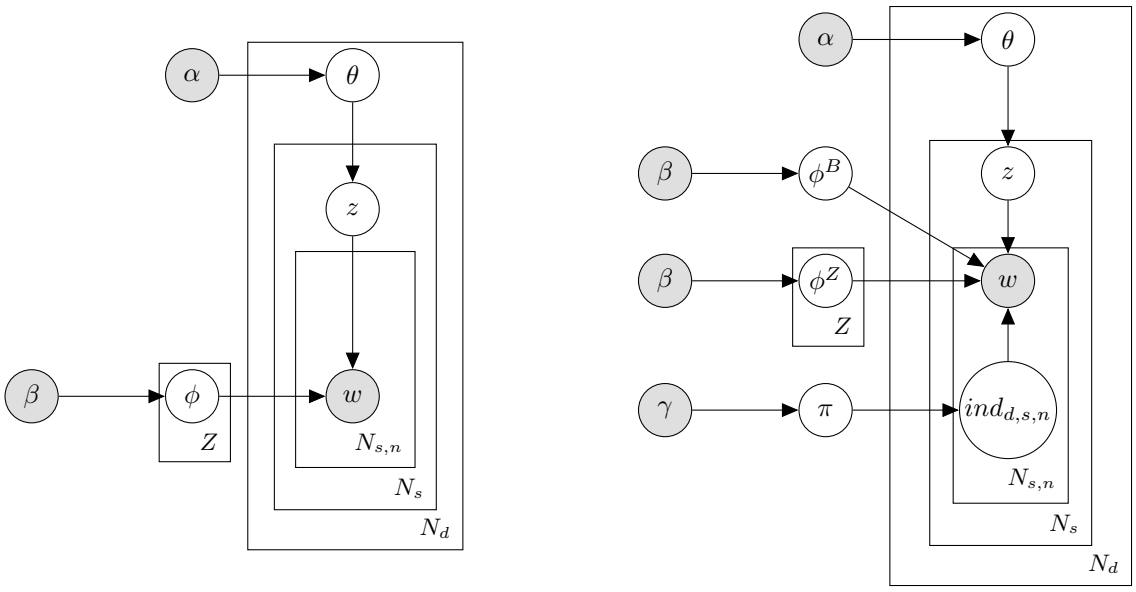


Figure 1: Graphical models of zoneLDA (left) and zoneLDAb models (right). The words $w$ are observed, while the per document zone distributions $\theta$ and per zone word distributions $\phi$ are hidden variables.

obtained experimental results. Conclusions and plans for future work are shown in section 7.

## 2. Related work

Previous work on automatic labelling of document zones mostly employ supervised machine learning, using widely known classifiers such as Naive Bayes (Teufel and Moens, 2002), Hidden Markov Model (Li et al., 2010), Maximum Entropy (Merity et al., 2009), Support Vector Machines (Guo et al., 2011a; McKnight and Srinivasan, 2003) or Conditional Random Fields (Hirohata et al., 2008). Semi-supervised approaches using active learning have only started to gain attention very recently (Guo et al., 2011b).

The majority of approaches have been applied to well-formatted scientific articles in the context of *computational linguistics* (Teufel et al., 2009), *biology* (Mullen et al., 2005; Liakata, 2010; Nawaz et al., 2010; Agarwal and Yu, 2009; Hirohata et al., 2008) or *chemistry* (Liakata, 2010; Teufel et al., 2009), focusing on either the full text or abstract of the articles. The application of document zoning to complex technical domains, such as the aerospace domain has not been studied yet. These domains can pose additional difficulties for a document zoner due to the intrinsic complexity of the language in them (Butters and Ciravegna, 2008) .

Furthermore, there has been only little work on using unsupervised approaches to document zone classification. (Barzilay and Lee, 2004) proposed a Hidden Markov Model (HMM) model to zoning with the states corresponding to topics from the document, and used a state-specific language model to generate the sentences relevant to the topics. Therefore they first applied clustering to compute the similarity between sentences as measured by the cosine metric and then they estimated the parameters for the HMM.

On the other hand, unsupervised approaches based on LDA has been found successful on a variety of different tasks including sentiment analysis (Lin and He, 2009), topic modelling (Zhao et al., 2011) and entity resolution (Dai and Storkey, 2011).

Continuing this success we propose two refined LDA models for the task of document zoning, which in contrast to previous unsupervised approaches are more flexible, in that they don't take into account the order of the sentences in the documents.

In the next section we formally introduce the task of document zoning and then we present our proposed graphical models in Section 4.

## 3. The task of document zoning

We assume a corpus $\mathcal{D}$ consisting of documents $\{D_1 \dots D_{N_d}\}$. Each document $D_i$ in $D$ is a sequence of sentences of $N_s$ which we denote by $s_i = \{s_{i,1}, \dots, s_{i,N_{s_i}}\}$, and each sentence contains a sequence of $N_{s_{i,j},n}$ words (in more general case a sequence of n-grams) $s_{i,j} = \{w_1, \dots, w_{N_{s_{i,j},n}}\}$, where the words are taken from the vocabulary $V$.

The task of document zoning then is to assign to each sentence $s_{i,j}$ in document $D_i$ a zone category $z \in Z$, where for example the distinct zone types $Z \in$ {Introduction, Methods, Results, Abstract, Discussion}, and $|Z| = N_Z = 5$.

## 4. Methodology

We propose two generative models zoneLDA and zoneLDAb for document zoning which are refined versions of the Latent Dirichlet Allocation model.

### 4.1. zoneLDA model

The zoneLDA model as depicted in Figure 1. is based upon the assumption that documents are mixture of zones, where a zone is a probability distribution over words.

Compared to the original LDA model, thus zoneLDA models the documents as mixture of zones and furthermore

it makes the assumption that every word in a sentence has the same zone type assigned.

The generative process of zoneLDA (as shown in Algorithm 1) can be viewed as a procedure describing how documents are written based on the available zone types $Z$. That is, first the distribution over the mixture of zones ($\theta^d$) is chosen for the document. Then, for each sentence a zone type is randomly selected from the zone distribution, and the corresponding words from that sentence are generated according to the corresponding word-zone distribution ($\phi^z$).

**Algorithm 1** Generative process of zoneLDA. $Z$ denotes the number of zones, $N_d$ denotes the number of documents, $N_s$ denotes the number of sentences, $N_{s,n}$ denotes the number of words in sentence $s$, $\alpha$ refers to a vector for Dirichlet prior for the document zone distributions, $\theta^d$ refers to the document zone distribution for document $d$, $w_{d,s,n}$ denotes the word at the position $n$ of the sentence $s$ in document $d$, $\beta$ refers to the word probability vectors as $Z \times V$ for the Dirichlet prior for each zone

**for all** document $d = \{1, \ldots, N_d\}$ **do**
  **draw** $\theta^d \sim \text{Dir}(\alpha)$
**for all** zone type $z = \{1, \ldots, Z\}$ **do**
  **draw** $\phi^z \sim \text{Dir}(\beta)$
**for all** sentence $z_{d,s}$, where $s \in \{1, \ldots, N_s\}$ **do**
  **draw** a zone class $z_{d,s} \sim \text{Multinomial}(\theta^d)$
**for all** word $w_{d,s,n}$ **do**
  **draw** $w_{d,s,n} \sim \text{Multinomial}(\phi^{z_{d,s}})$
When running the model with the number of zone types($Z$) greater than the number of predefined zone classes, perform k-means clustering with distributions of words as features to obtain $|N_z|$ of zone categories

We used Gibbs sampling for estimating the posterior distribution of the hidden variable $z$ for sentence $i$ in document $d$:

$$P(z_{d,i} = k | z_{-d,i}, w) \propto \frac{n^k_{d,-i,\cdot} + \alpha_k}{n_{d,\cdot,\cdot} + Z\alpha_k} \frac{\sum_{v=1}^{V}(n^k_{d,i,v}) + \beta}{n^k_{\cdot,\cdot,v} + V\beta},$$

where $n^k_{d,-i,\cdot}$ denotes the number of sentences assigned to zone $k$ for document $d$, $n_{d,\cdot,\cdot}$ denotes total number of zone types assigned to document $d$, $n^k_{\cdot,\cdot,v}$ denotes the number of times word $v$ is assigned to zone $k$, $n^k_{d,i,v}$ is the number of times word $v$ from sentence $i$ of document $d$ is assigned to zone $k$.

### 4.2. zoneLDAb model

In this section we present an extended version of the zoneLDA model (shown in Figure 1.), which we call zoneLDAb. As opposed to zoneLDA, zoneLDAb distinguishes between common words or background words (for e.g. "use", "determine", "indicate", "cell") which can appear in multiple zone types and words which are specific to a zone category. This distinction is based on the intuition that words which are related to multiple zone categories are likely to introduce noise (e.g. zone types with incoherent words), and thus those words are not discriminative for a zone category.

As presented in Algorithm 2, the generative process of zoneLDAb differs from that of zoneLDA, in that for each sentence a word distribution is chosen either from the background zone distribution or a selected zone distribution. In zoneLDAb we thus need to infer the zone distribution for each document ($\theta^d$), the word distributions for each zone type ($\theta^z$) and the word distributions for the background words ($\theta^B$). Furthermore, the $\pi$ variable has the role in determining whether a word is a background word or a zone specific word.

**Algorithm 2** Generative process of zoneLDAb. $Z$ denotes the number of zones, $N_d$ denotes the number of documents, $N_{s,n}$ denotes the number of words in sentence $s$, $\alpha$ refers to a vector for Dirichlet prior for the document zone distributions, $\theta^d$ refers to the document zone distribution for document $d$, $w_{d,s,n}$ denotes the word at the position $n$ of the sentence $s$ in document $d$, $\beta$ refers to the word probability vectors as $Z \times V$ for the Dirichlet prior for each zone, $ind_{d,n,s}$ indicates whether a word is a background word or not

**draw** $\phi^B \sim \text{Dir}(\beta)$, $\pi \sim \text{Dir}(\gamma)$
**for all** zone type $z = \{1, \ldots, Z\}$ **do**
  **draw** $\phi^z \sim \text{Dir}(\beta)$
**for all** document $d = \{1, \ldots, N_d\}$ **do**
  **draw** $\theta^d \sim \text{Dir}(\alpha)$
**for all** sentence $z_{d,s}$, where $s \in \{1, \ldots, N_s\}$ **do**
  **draw** a zone class $z_{d,s} \sim \text{Multinomial}(\theta^d)$
**for all** word $w_{d,s,n}$ **do**
  **draw** indicator $ind_{d,s,n} \sim \text{Multinomial}(\pi)$
  **draw** word $w_{d,s,n} \sim \text{Multinomial}(\phi^B)$ if $ind_{d,s,n} = 0$
  and $w_{d,s,n} \sim \text{Multinomial}(\phi^{z_{d,s}})$ if $ind_{d,s,n} = 1$
When running the model with the number of zone types($Z$) greater than the number of predefined zone classes, perform k-means clustering with distributions of words as features to obtain $|N_z|$ of zone categories

Similarly, we run Gibbs sampling for estimating the posterior distribution of the hidden variable $z$ for sentence $i$ in document $d$:

$$P(z_{d,i} = k | z_{-d,i}, w, ind)) \propto \frac{n^k_{d,-i,\cdot} + \alpha_k}{n_{d,\cdot,\cdot} + Z\alpha_k}$$

$$\times \frac{\Gamma(n^k_{\cdot,\cdot,\cdot} + V\beta)}{\Gamma(n^k_{\cdot,\cdot,\cdot} + n^k_{d,i,\cdot} + V\beta)} \prod_{v=1}^{V} \frac{\Gamma(n^k_{\cdot,\cdot,v} + n^k_{d,i,v} + \beta)}{\Gamma(n^k_{\cdot,\cdot,v} + \beta)},$$

where $n^k_{d,-i,\cdot}$ denotes the number of sentences assigned to zone $k$ for document $d$, $n_{d,\cdot,\cdot}$ denotes total number of zone types assigned to document $d$, $n^k_{\cdot,\cdot,\cdot}$ denotes the number of times any word is assigned to zone $k$, $n^k_{d,i,\cdot}$ is the number of times any word from sentence $i$ of document $d$ is assigned to zone $k$, $n^k_{\cdot,\cdot,v}$ denotes the number of times word $v$ is assigned to zone $k$, $n^k_{d,i,v}$ is the number of times word $v$ from sentence $i$ of document $d$ is assigned to zone $k$.

## 5. Corpora

In order to evaluate our models we conducted experiments on corpora belonging to two very different domains: the scientific domain and the technical domain. Corpus statistics are presented in Table 1.

### 5.1. Scientific domain

For the scientific domain, we have built a corpus consisting of biomedical journal articles crawled from the PubMed system. We selected 1,106 articles from the PLoS Pathogens journal[2] published between January 2006 until June 2011. [3] All the documents had to contain all the five zone categories of the widely used IMRAD (Introduction-Method-Results-Abstract-Discussion) (Agarwal and Yu, 2009) classification schema. The articles that didn't contain at least one of the these zones were discarded.

In the data pre-processing phase we removed all text that was contained in the other zones of the document (such as "References", "Supporting Information", "Synopsis", etc.) as they are not the focus of our task. We also eliminated zone names as they would give away valuable information, figures, the text in tables and captions.

In order to reduce data sparsity we removed all numbers, words made out of special characters, citations, references, applied Porter stemming and discarded sentences which contained less than 5 words and words that occured in less then 10 documents. We also removed all stopwords and one-character words. Thus, the resulting corpus will contain only stemmed content words that are not very document specific, a typical procedure when training topic models. We used a Python script[4] to perform these steps and to annotate each sentence with one of the IMRAD zone categories.

The average length of the zones in the corpus is presented in Table 2.

### 5.2. Technical domain and proposed document zone annotation schema

Our second dataset consists of two corpora in the aerospace domain comprising of 689 textual reports, which were collected as part of the SAMULET research project funded by Rolls-Royce and the Technology Strategy Board, in which two of the authors are involved. Due to privacy reasons these corpora have restricted access and we will refer to them as Corpus A and Corpus B. These technical reports are unstructured and semi-structured PDF document containing a mixture of natural language sentences, images and tables. They were written at different stages of an investigation process, which is typically initiated by a customer raising a request regarding an issue on a particular engine.

On these corpora we first conducted a corpus analysis to understand and identify the information they share in common and the possible zone categories they contain. Our analysis revealed that there are *six zone types* common in the reports (see Table 4), as they all follow a *problem-solving* perspective of an investigation. For example these reports typically contain the *Metadata* zone which introduces the main entities (engine, component) under investigation, then they continue with a *Problem description* zone which describes the problems which occurred on these entities; next the *Instructions zone* provides typical instructions regarding what procedure should be taken, finally the *Decision* contains the decision taken after investigation. In addition, two other zones were found in these reports: the *Acknowledgement* zone, which is the formal part of the document, and the *Attachment* zone, which includes further evidence taken during investigation (e.g. in forms of images, emails, faxes). Although the proposed zone categories share some commonalities with existing classification schemas (e.g. IMRAD), the *Instructions* and *Acknowledgement* zones are new in our schema. The instructions are typically split into tasks and subtasks, and may consider reference to some manuals (e.g. the engine manuals). The *Acknowledgement* zone acknowledges the conclusions drawn from the report, containing the signatures of the responsible agents.

These two corpora were then annotated by two independent annotators (the first and the third author of the paper), achieving an inter annotator Kappa agreement of 85%. The average length of the zones in the corpora is presented in Table 3.

In the pre-processing stage we converted all the PDF document into plain text[5] and thus removed all the formatting information and figures. Similarly to the biomedical corpus, we removed all the numbers and stop words. Furthermore, due to the diverse format of the documents, consisting of tables and natural language text, we considered as smallest unit of classification the lines of the documents, as opposed to sentences.

## 6. Experiments

In this section we discuss the results obtained in our experiments using the two graphical models proposed in Section 4. For evaluating the accuracy of the clusters generated by the zoneLDA and zoneLDAb models we used a baseline LDA model, a supervised machine learning algorithm, and the random baseline. The baseline LDA model, employs the original LDA model proposed in (Blei et al., 2003) for discovering the topics in a document, and based on the discovered topics, at inference time, assigns to each sentence the most likely topic among the words within that sentence. For all the graphical models we run Gibbs sampling for 10,000 number of iterations and a burn-in of 500. The supervised classifier employed was the Naive Bayes classifier which was trained with the same-bag-of-words features. We used the implementation of Naive Bayes available in the Mallet toolkit [6].

### 6.1. Dataset Preparation

We evaluated the proposed models on two very different domains: a publicly available scientific domain and a tech-

[2] http://www.ncbi.nlm.nih.gov/pmc/journals/349/

[3] The IDs of the Plos journals articles used in our experiments can be found at `www.dcs.shef.ac.uk/~daniel/plos.ids`

[4] The Python script is available at `www.dcs.shef.ac.uk/~daniel/pubmed.py`

[5] We used Apache PDFBox library available at `http://pdfbox.apache.org/` for converting the Pdf documents into plain text format.

[6] `http://mallet.cs.umass.edu/`

| Domain | Corpus name | Number of documents | Average number of sentences per document | Average number of distinct words |
|---|---|---|---|---|
| scientific domain | PLOS journal articles | 1,106 | $241 \pm 59$ | $966 \pm 174$ |
| technical domain | Corpus A | 317 | $226 \pm 295$ | $329 \pm 241$ |
| | Corpus B | 372 | $394 \pm 405$ | $518 \pm 273$ |

Table 1: Corpus statistics of documents in the scientific and technical domains.

| Abstract | Introduction | Methods | Results | Discussion |
|---|---|---|---|---|
| 11 | 27 | 64 | 88 | 53 |

Table 2: Average length for each IMRAD zone in the PLOS journal corpus

| Corpus name | Metadata | Problem description | Decision | Instructions | Acknowledgement | Attachment |
|---|---|---|---|---|---|---|
| Corpus A | 36 | 3 | 22 | 76 | 17 | 35 |
| Corpus B | 30 | 49 | 13 | 107 | 13 | 25 |

Table 3: Average length for each proposed zone category in the technical corpora

| Zone Category | Description |
|---|---|
| 1. *Metadata* | contains general information about the report: the title of the report; the entities (for e.g. engine, component) under investigation; and other entities (e.g. agents) participating in the investigation |
| 2. *Problem description* | aims to describe the problem encountered on a specific entity (e.g. engine, component) |
| 3. *Decision* | summarises the decision taken after investigation. (for e.g. the conclusion drawn) |
| 4. *Instructions* | describes the general procedure to follow in a certain situation (for e.g. a given problem) |
| 5. *Acknowledgement* | denotes the formal part of the document, consisting of the details of the agents (e.g. name) and their signature |
| 6. *Attachment* | contains further evidence attached to the investigation (mostly pictures, email, faxes) |

Table 4: Proposed document zone annotation schema in the aerospace domain

nical domain with restricted availability. The scientific domain data consists of 1,106 biomedical articles downloaded from PubMed.

After stemming and removing the stop words the size of the vocabulary of this corpus was 46,698. We furthermore reduced the size of the vocabulary to avoid sparsity and decrease the time required for training our graphical models. We discarded all the words which occurred in less than 10 documents, and in more than 70% of the corpus, resulting in a reduced vocabulary of 6,843 words. In addition, we conducted experiments keeping only the top 1,000 most frequent words in the corpus; and another experiment in which we only removed words which occurred in less than 10 documents, but did not find any significant improvement.

Our technical domain data consists of two aerospace corpora. After stemming and removing the stop words the size of the vocabulary of Corpus A was 4,153. After removing words which occur in less than 5 documents the vocabulary became 1,023. For the Corpus B the initial vocabulary has been reduced from 2,964 to 1,340 after discarding words which occur in less than 5 documents.

### 6.2. Evaluation metric

In the case of scientific domain we split the original corpus into 60% training, 10% development, and 30% testing, and we averaged the results over 5 independent runs. For the case of technical domain, we split the original CorpusA and CorpusB into 45% training, 10% development and 45% testing and we averaged the results over 5 independent runs. In each of the case we compared the performance of the unsupervised graphical models, baseline LDA model and the supervised classifier over the same held-out test data.

For evaluating the clusters generated by zoneLDA and

zoneLDAb models we employed pairwise clustering evaluation metric implemented in Mallet, which taking the gold standard into account for each pair of sentences computes the false positives and false negatives in order to decide whether the pair should be in the same cluster or not:

$$\text{Prec}_{\text{pair}} = \frac{|\text{clustered sentence pairs which should be clustered}|}{|\text{sentence pairs which are clustered}|}$$

$$\text{Rec}_{\text{pair}} = \frac{|\text{clustered sentence pairs which should be clustered}|}{|\text{sentence pairs which should be clustered}|}$$

$$\text{F1}_{\text{pair}} = \frac{2 \times \text{Prec}_{\text{pair}} \times \text{Rec}_{\text{pair}}}{\text{Prec}_{\text{pair}} + \text{Rec}_{\text{pair}}}$$

### 6.3. Hyper-parameter setting

We evaluated both zoneLDA and zoneLDAb models with different values for their parameters. We first varied the number of word distributions, considering $Z \in \{5, 50, 100\}$ for the biomedical domain and $Z \in \{7, 50, 100\}$ for the aerospace domain. In addition for the different number of zone types we experimented with asymmetric and symmetric values for $\alpha$ . We thus investigated six different values for $\alpha$.

In the first setting we chose a symmetric Dirichlet prior with $\alpha = 0.1$, which discovers zone types which are sparse. In the second case we chose a symmetric Dirichlet prior with $\alpha = 1$. In the third case we chose a symmetric Dirichlet prior with $\alpha = 10$, which discovers zone types which are dense.

We furthermore experimented with assymetric Dirichlet priors, where we initialised the $\alpha_i, i \in \{1, \ldots, Z\}$ values for the different zone categories based on the development set. As such, in the fourth setting, we set the $\alpha_i, i \in \{1, \ldots, Z\}$, so that $\sum_{i \in \{1,\ldots,N_Z\}} \alpha_i = 0.1$. In the fifth case we set the $\alpha_i, i \in \{1, \ldots, Z\}$, so that $\sum_{i \in \{1,\ldots,N_Z\}} \alpha_i = 1$. And in the sixth case we set $\alpha_i, i \in \{1, \ldots, Z\}$, so that $\sum_{i \in \{1,\ldots,N_Z\}} \alpha_i = 10$. We set the value for the $\beta$ parameter to $0.01$.

In our experiments we will refer to the first case with "/1", to the second case with "/2", to the third case with "/3", to the fourth case with "/4", to the fifth case with "/5" and to the sixth case with "/6".

### 6.4. Results

Figure 2 shows the results obtained in terms of F1-measure over the biomedical corpus. As we can observe, the accuracy of the zoneLDA model slightly increases with the number of zone types learned. The zoneLDA model achieved an F1-measure over 30% for 50 and 100, having the best F1-measure of $35,22\%$ for 50 topics with an asymmetric Dirichlet prior (case "50/6" in Figure 2). In contrast, when we look at the results of the zoneLDAb model, the improvement obtained with different number of zone types is less significant. The best F1-measure of $32.08\%$ being achieved with 5 zone types and an asymmetric Dirichlet prior (case "5/6" in Figure 2). Compared to the baseline LDA model we can also notice, that when we have a small number of zone types (e.g. 5), the baseline LDA model outperforms both zoneLDA and zoneLDAb models, but as

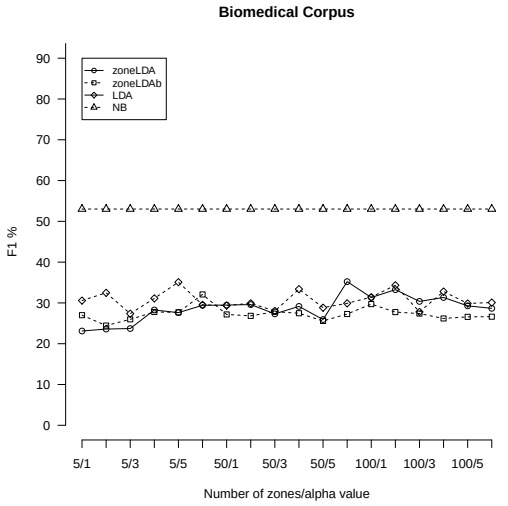


Figure 2: The performance of zoneLDA and zoneLDAb model over the biomedical corpus.

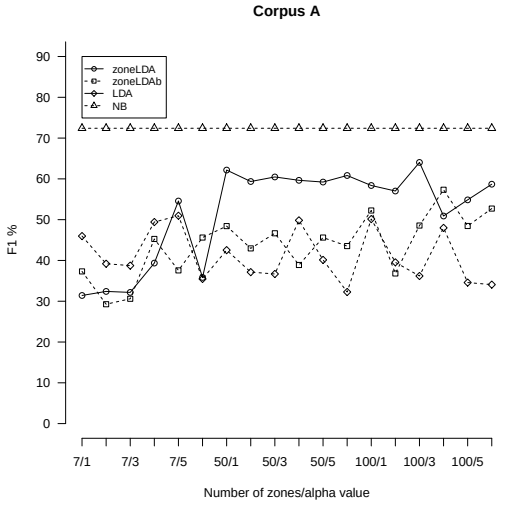


Figure 3: The performance of zoneLDA and zoneLDAb model over CorpusA.

the number of zone types increases the performance of the zoneLDA model becomes superior in most of the cases.

When looking at the errors made by the zoneLDA model, we noticed that the most difficult zone to identify was the Abstract zone, for which the zoneLDA model achieved an F1-measure of $1\%$. The second most difficult zone type was the Introduction zone, for which the performance was $8\%$, next for the Discussion zone type the F1-measure was $21,4\%$, for the Methods zone was $30.00\%$ and for the Results zone was $66.8\%$. Similar trends can be observed for the zoneLDAb model. The Abstract zone type still being the most difficult zone type to be discovered with an F1-measure of $2\%$. Then for the Introduction the zoneLDAb model achieved an F1-measure of $13.1\%$, next for the Discussion zone type it achieved an F1-measure of $25,1\%$. The best two performances were achieved for the Methods zone

| Method | Plos journal articles | | | Corpus A | | | Corpus B | | |
|---|---|---|---|---|---|---|---|---|---|
| | Prec | Recall | F1 | Prec | Recall | F1 | Prec | Recall | F1 |
| Random baseline | $20,00\%$ | $20,00\%$ | $20,00\%$ | $14,28\%$ | $14,28\%$ | $14,28\%$ | $14,28\%$ | $14,28\%$ | $14,28\%$ |
| LDA(baseline) | $27,30\%$ | $49,10\%$ | $35,08\%$ | $42,90\%$ | $33,55\%$ | $37,65\%$ | $26,70\%$ | $42,50\%$ | $32,79\%$ |
| zoneLDA | $29,50\%$ | $43,70\%$ | $35,22\%$ | $49,25\%$ | $91,50\%$ | $64,03\%$ | $32,75\%$ | $84,80\%$ | $47,25\%$ |
| zoneLDAb | $26,63\%$ | $40,33\%$ | $32,08\%$ | $49,35\%$ | $68,40\%$ | $57,34\%$ | $33,30\%$ | $75,95\%$ | $46,29\%$ |
| NaiveBayes | $52,40\%$ | $51,23\%$ | $51,8\%$ | $78,07\%$ | $67,53\%$ | $72,41\%$ | $68,35\%$ | $40,99\%$ | $51,23\%$ |

Table 5: The best results obtained on scientific and technical domains

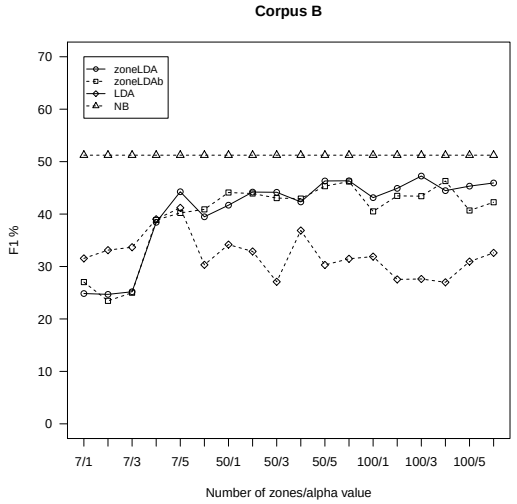


Figure 4: The performance of zoneLDA and zoneLDAb model over CorpusB.

type, an F1-measure of $38,9\%$; and for the Results zone type an F1-measure of $41,5\%$. These results are as well in light with the results obtained with the supervised Naive Bayes classifier. Namely, the worst F1-measure of $1\%$ was achieved on the Abstract zone type, followed by the Introduction zone type with an F1-measure of $43.32\%$, then for the remaining of the zone categories the classifier achieved an F1-measure of over $50\%$. For the Discussion zone type being $54.94\%$, for the Results zone type was $70.43\%$, and for the Methods zone type achieving $85.8\%$.

These results furthermore show, that for both zoneLDA, zoneLDAb models and the supervised Naive Bayes classifier, the best performances were achieved for the long zone categories, which contain the most number of sentences. This is because in case short zone categories such as Abstract uses words from other zones that are not discriminative of this zone type.

When examining the results obtained by the baseline LDA model, however, the results look different. The worst results were obtained for the Discussion zone, an F1-measure of $2\%$, for the Methods zone type, an F1-measure of $3.4\%$, and for the Abstract zone type, an F1-measure of $5.6\%$. Then for the Introduction the baseline model achieved an F1-measure of $34.8\%$, and for the Results zone type an F1-measure of $43.2\%$.

Figure 3 shows the results obtained in terms of F1-measure over the aerospace Corpus A. As we can see, the zoneLDA model is more sensitive to the number of zone types learned. The best F1-measure of $64.03\%$ was achieved with 100 zone categories using a symmetric Dirichlet prior (case "100/3" in Figure 3). This result is also close to the supervised Naive Bayes classifier, which achieved and F1-measure of $72,41\%$. Similar trends can be seen for the zoneLDAb model, which improves its performance with the number of zone types learned, achieving the best F1-measure of $57.33\%$ using 100 zone categories with an asymmetric Dirichlet prior (case "100/4" in Figure 3). Compared to the baseline LDA model, we can also notice that both zoneLDA and zoneLDAb models perform consistently better when having a large number of word distributions and clustering.

Figure 4 shows the results obtained in terms of F1-measure over the aerospace Corpus B. The performance of the zoneLDA model slightly increases with the number of zone types learned. The best F1-measure of $47,25\%$ was achieved using 100 zone categories with an asymmetric Dirichlet prior (case "100/3" in Figure 4), which is only $5\%$ less than the performance of the supervised Naive Bayes classifier. The performance of the zoneLDAb model is also very similar, being over $40\%$ for all the different cases when 50 or 100 zone categories are used, the best values of $46,29\%$ (case "100/4" in Figure 4) was obtained for 100 topics with an asymmetric Dirichlet prior. Compared to the baseline LDA model, we can also notice that both zoneLDA and zoneLDAb models perform consistently better when having a large number of word distributions and clustering.

In summary, our results show that in general the performance of the zoneLDA and zoneLDAb models increases with the number of word distributions learned. For the majority of the cases, when the number of word distributions is more than 50, the zoneLDA model consistently outperforms the zoneLDAb model. This is because the background words discovered by the zoneLDAb model seem to contain zoning information.

When we compare the zoneLDA and zoneLDAb models with the baseline LDA model we can furthermore notice that having the number of word distributions relatively small (e.g. equal to the number of predefined zone classes: 5 for the scientific corpus, and 7 for the technical corpora), the baseline LDA model outperforms both models. These results are not surprising, because in such cases, both zoneLDA and zoneLDAb models discover coherent topics, rather than zone types. On the other hand, when the number of word distributions learned increases these mod-

els exhibit a significant improvement over the LDA model. In this case, the discovered word distributions are less sensitive to topic information, allowing the zone information to be discovered. Moreover, on the technical domain, the performance of the unsupervised zoneLDA and zoneLDAb models are comparable with the supervised Naive Bayes algorithm.

## 7. Conclusions

In this paper we introduced zoneLDA and zoneLDAb models for unsupervised document zoning, which cluster the sentences in a document into predefined zone categories. Both of the models ignore the order of the sentences in the documents, and the second zoneLDAb model makes a distinction between content and background words. In our experiment, the two models improve upon the baseline LDA model which first discovers the topics of the words within the sentences, and then infers the most likely zone for a sentence.

We demonstrated the effectiveness of the models on two very different domains: a scientific biomedical domain and on two technical aerospace corpora. For the latter domain, we also proposed a novel document zone annotation schema, and our unsupervised graphical models achieved performance comparable with the supervised Naive Bayes classifier, with less than $5\%$ deficit in F1-measure.

Our future work will consist in evaluating these models on corpora from other domains and in constructing more elaborate graphical models for document zoning.

## 8. References


Shashank Agarwal and Hong Yu. 2009. Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. *Bioinformatics*.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceeding of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, Volume 3, March.

Jonathan Butters and Fabio Ciravegna. 2008. Using similarity metrics for terminology recognition. In *Proceeding of the International Conference on Language Resources and Evaluation (LREC)*.

Andrew M. Dai and Amos J. Storkey. 2011. The grouped author-topic model for unsupervised entity resolution. In *Proceeding of International Conference on Artificial Neural Networks (ICANN)*.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Johan Hogberg, 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*.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011b. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceeding of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

Kenji Hirohata, Naoaki Okazaki, and Sophia Ananiadou. 2008. Identifying sections in scientific abstracts using conditional random fields. In *International Joint Conference on Natural Language Processing*.

Ying Li, Sharon Lipsky Gorman, and Noemie Elhadad. 2010. Section classification in clinical notes using supervised hidden markov model. In *Proceeding of the ACM International Health Informatics Symposium (IHI)*.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceeding of the International Conference on Language Resources and Evaluation (LREC)*.

Maria Liakata. 2010. Zones of conceptualisation in scientific papers: a window to negative and speculative statements. In *Proceeding of the Workshop on Negation and Speculation in Natural Language Processing (NSNLP)*.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceeding of the Conference on Information and Knowledge Management (CIKM)*.

Larry McKnight and Padmini Srinivasan. 2003. Categorization of sentence types in medical abstracts. *Proceeding of the Annual Symposium of the American Medical Informatics Association (AMIA)*, pages 440–444.

Stephen Merity, Tara Murphy, and James R. Curran. 2009. Accurate argumentative zoning with maximum entropy models. In *Workshop Proceeding of the Text and citation analysis for scholarly digital libraries (NLPIR4DL)*.

Tony Mullen, Yoko Mizuta, and Nigel Collier. 2005. A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *Proceeding of the ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)*.

Raheel Nawaz, Paul Thompson, John McNaught, and Sophia Ananiadou. 2010. Meta-knowledge annotation of bio-events. In *Proceeding of the International Conference on Language Resources and Evaluation (LREC)*.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*.

Simone Teufel, Advaith Siddharthan, and Colin R. Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceeding of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *Proceeding of European Conference on Information Retrieval (ECIR)*.