Using Argumentative Zones for Extractive Summarization of Scientific Articles

Danish Contractor[1,2] *Yufan Guo*[1] *Anna Korhonen*[1]
(1) Computer Laboratory, University of Cambridge, Cambridge, United Kingdom
(2) IBM Research - India, New Delhi, India
{dc536,yg244,alk23}@cam.ac.uk

ABSTRACT
Information structure, i.e the way speakers construct sentences to present new information in the context of old, can capture rich linguistic information about the discourse structure of scientific documents. Information structure has been found useful for important Natural Language Processing (NLP) tasks, such as information retrieval and extraction. Since scientific articles typically follow a certain discourse structure describing the prior work, problem being solved, methods used, and so forth, it could also be useful for summarization of these articles. In this work we focus on a scheme of information structure called Argumentative Zoning (AZ), and investigate whether its categories could support extractive text summarization in a scientific domain. We develop a summarization system that uses AZ categories (i) as features and (ii) in the final sentence selection process. We evaluate the system directly as well as using task-based evaluation. The results show that AZ can support both full document and customized summarization. We report a statistically significant improvement in summarization performance against a competitive baseline that uses journal section labels instead of AZ information.

TITLE AND ABSTRACT IN MANDARIN

一种根据“论证结构”自动摘录科技文献的方法

信息结构是指作者组织语句陈述信息的方式。信息结构例如科技文献的篇章结构包含丰富的语言信息，有助于解决自然语言处理领域的一些重要问题例如信息检索和信息提取等。科技文献通常使用特定的篇章结构来陈述以往的研究，阐述研究问题以及研究方法等等，这些篇章结构可以被用于文献的自动摘录。本文着眼于一类特定的信息结构 — “论证结构”，研究其是否有助于更好地摘录科技文献。在本文开发的摘录系统中，“论证结构”有两种用途：一是作为特征供机器学习，二是用于最终的语句筛选过程。本文对该系统进行了直接和间接的评测，测试结果显示“论证结构”有助于更好地对全文或指定信息进行摘录。基于“论证结构”的摘录系统显著性优于基于章节标题的摘录系统。

KEYWORDS: discourse, information structure, argumentative zones, summarization, document summarization, information access.

KEYWORDS IN MANDARIN: 篇章，信息结构，论证结构，文本摘要，信息获取.

Proceedings of COLING 2012: Technical Papers, pages 663–678,
COLING 2012, Mumbai, December 2012.

1 Introduction

Information structure is the study of how writers package information into a sentence and convey new information (e.g. new methods, results and conclusions) in the context of old information (e.g. previous or related work) within a document. A number of frameworks capturing different aspects of information structure (including e.g. discourse, rhetorical, argumentative and conceptual) have been proposed, many of which focus on scientific documents (Teufel and Moens, 2002; Shatkay et al., 2008; Teufel et al., 2009; Liakata et al., 2010). Scientific documents are highly structured in nature, and knowledge about their information structure can support important Natural Language Processing (NLP) systems aimed at improving information access to scientific literature.

To date, information structure has proven useful for different information retrieval and extraction tasks, as well as for manual literature review (Tbahriti et al., 2006; Ruch et al., 2007; Guo et al., 2011b). One NLP task which is highly important for the scientific domain and which might similarly benefit from information structure is document summarization. Scientific articles are well structured containing sections such as "Prior work", "Method", "Experiments" etc, and also contain a rich network of citations. While section and citation based features have been exploited extensively in prior summarization works (see section 2), categories of information structure have received little attention. A good summarization system should be able to identify the "key" concepts in an article and generate a summary that has good coverage over the ideas expressed in the article. In the case of scientific documents, information structure can capture rich linguistic characteristics defined by eg. the discourse status of sentences, and therefore could help a summarization system select the right mix of sentences to be used from the document.

(Teufel and Moens, 2002) introduced a scheme of information structure called Argumentative Zoning (AZ) which classifies sentences in scientific text into categories (such as Aim, Background, Own, Contrast and Basis) on the basis of their rhetorical status in scientific discourse. They performed experiments which show that AZ can be used to identify and summarize novel contributions as well as background information in a scientific article. However, they did not investigate integrating AZ in an automatic summarization system.

In this paper we focus on this topic and explore whether knowledge about information structure could be used to support an actual summarization system performing extractive summarization in the scientific domain. Like (Teufel and Moens, 2002), we focus on AZ because this scheme has aided many other NLP tasks and has shown wide applicability across different scientific domains (including e.g. computational linguistics, chemistry, biology). Experimenting on biomedical corpus data, we use the version of AZ adapted for biology by (Mizuta et al., 2006).

We develop a simple summarization system for evaluation purposes and use it as a framework when investigating two approaches to integrating both manually and automatically obtained AZ categories into summarization: (i) including them as features in a classification task for selecting sentences that should be part of a summary, along with other features that have traditionally been found useful in such a classification task, and (ii) using them as a selection mechanism for identifying the final set of sentences that should be made part of the summaries.

We evaluate these approaches via two task-based evaluations - extractive summarization of complete articles as well as generation of customized summaries based on user requirements. We also compare their performance against a competitive baseline that makes use of section labels (instead of AZ labels) in biomedical articles. To the best of our knowledge, this is

the first work that compares the use of section labels against AZ in summarization. The results are promising. They demonstrate that AZ can be an effective feature to include in summarization systems and can improve the quality of summaries generated for scientific papers. Both manually and automatically obtained AZ labels prove useful. In the future, the approach could be optimised for integration in state of the art summarization systems as well as used for task-based evaluation and comparison of different automatic AZ lebeling systems.

2 Related work

2.1 Summarization in Scientific Literature

Scientific literature continues to be a major domain for summarization research along with news articles. Different types of summaries can be generated for such documents. For example, one could be interested in automatically generating an abstract-like summary for an article or a group of articles, or one may require a customized summary describing specific types of information (e.g. experiments or results) in articles only. Further, sentences in such summaries may be a paraphrased representation of the information present in the original documents (abstractive summarization) or may be a subset of those in the original article (extractive summarization).

An important characteristic of scientific articles is the presence of citations. Citations have been used in different summarization systems. Recent work such as that by (Abu-Jbara and Radev, 2011), (Qazvinian et al., 2010) and (Qazvinian and Radev, 2010) make use of citation sentences in other scientific papers to summarize the contributions of a paper. Although it is the ideas of one paper that are being summarized, this approach involves searching for references to the paper in other papers, and extracting sentences from them to build summaries.

Other recent work such as that of (Qazvinian and Radev, 2010) uses Markov Random Fields to detect patterns that create context data (background information) for a paper, while (Mei and Zhai, 2008) use citations as a measure of "impact" in a field and use it for summarization.

Latent Semantic Analysis based methods have also been used for summarization of documents. (Steinberger et al., 2005) use LSA along with anaphora resolution to improve document summarization while (Ozsoy et al., 2010) propose multiple LSA based summarization algorithms in which the sentence selection criteria is modified using the "concept" matrix derived at the end of singular value decomposition (SVD) step.

In this paper we use argumentative zones (AZ) for extractive summarization of scientific papers. Extractive summarization generates summaries by selecting a subset of the sentences from the original document. We use a classifier trained using the author-created abstracts of articles to identify sentences from the full document for a system-generated summary and use clustering to further select sentences. AZ information is used both as a feature in classification as well as a guiding step during clustering.

2.2 Argumentative Zoning and NLP Tasks

Argumentative Zoning (AZ) classifies sentences from scientific text based on their rhetorical status in terms of problem solving (e.g."What are the contributions of the paper?"), intellectual attribution (sentences that describe prior work etc) and relatedness amongst articles. The original AZ scheme of (Teufel and Moens, 2002), applied to the domain of computational linguistics, included five rhetorical zone categories (Aim, Background, Own, Contrast and Basis)

and a fully supervised classifier was trained to classify each sentence in scientific articles in one of these categories. Subsequent work and applications of this scheme to other domains (e.g. chemistry, biology) have resulted in finer-grained AZ classifications.

Most approaches to automatic AZ detection rely on fully supervised machine learning. A high accuracy above 80% have been reported with the best of these approaches. (Guo et al., 2011a) has developed an approach based on active learning which performs (as its best) as well as fully supervised approaches but requires only a small amount of labeled data.

Most work on information structure has been evaluated directly on manually annotated data sets. Previous task-based evaluations, mostly conducted on AZ, include information retrieval and extraction tasks, along with literature review in biomedicine (Tbahriti et al., 2006) (Ruch et al., 2007), (Guo et al., 2011b).

(Teufel and Moens, 2002) reported experiments which suggest that AZ should also be helpful for automatic summarization. They used zones to identify and "summarize" new contributions in scientific papers. Sentences from the Aim, Contrast and Basis zones were used to highlight new contributions of a paper. When including information about "background work" in a summary, directly using sentences labeled with the Background zone reduced precision. They therefore trained a classifier based on annotated data that identifies sentences from the Background zone for a short "summarized" version of the document. Although this work suggests that AZ could be useful for summarization, it does not develop or employ an actual summarization system.

Related work by (Farzindar and Lapalme, 2004) used "thematic" structures (rather than AZ) (Introduction, Context, Judicial Analysis and Conclusion) in law judgments to generate summaries. They identified "cue" strings for each of these themes and used verb classes to filter out citation sentences. To summarize the text they used a heuristic function based on position of paragraphs in a document, position of paragraph in a thematic segment, tf-idf distribution and cue words specific to each theme. The summary lengths were controlled by using the distribution of themes in the abstract to select a proportionate number of sentences from each theme. However, scientific articles differ from law judgments because the documents are highly structured and contain "sections" which are defined by the authors.

In this paper we investigate whether AZ could be used to support a summarization system in the scientific domain. We focus on biomedicine and experiment with the AZ scheme developed for biology by (Mizuta et al., 2006). We employ manual AZ annotations in the main experiments (in order to investigate the direct impact of AZ on summarization) but also report experiments where automatic annotations from the weakly supervised AZ labeling system of (Guo et al., 2011a) are used. We integrate the AZ labels into a summarization system as features and also use them to aid the final sentence selection process in summarization. We perform both direct and task-based evaluation which shows that both methods can support automatic summarization.

3 Method

Most scientific papers contain an abstract which provides a short description of the work presented in the paper. The abstracts are created by the authors and can be regarded as summaries of papers. Using such abstracts as the gold-standard we describe a method for generating summaries. The summaries can vary in length (i.e. be more elaborate or concise than the original abstracts) which is useful in the scientific domain where users (e.g. scientists) have highly varied summarization needs.

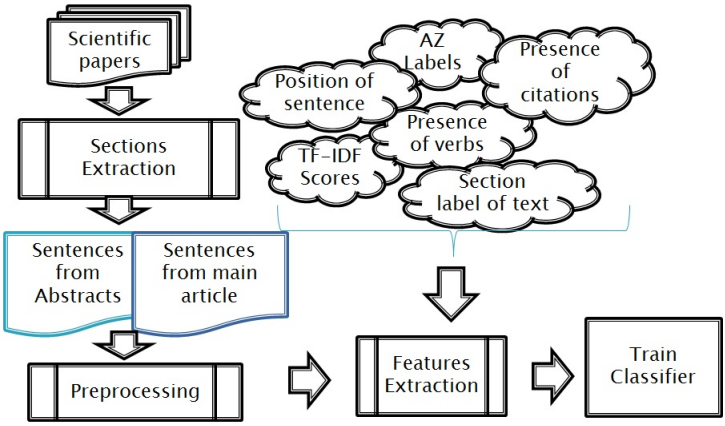

Figure 1: Schematic diagram showing the training phase of the summarization system.

Our method has two main stages: classification and sentence clustering. The classifier creates an initial candidate set of sentences for the summary, and the sentence clusterer identifies groups of similar sentences in this set which are then used to create the final summary. The clustering step, employed by many summarization systems, removes redundancy from the candidate pool and thus improves the quality of the summary.

Figure 1 gives an overview of the training phase of the summarization system. Sentences from the training articles are pre-processed as described in section 4.1. and annotated with the section labels (i.e. the section of the article to which the sentence belongs) and zone labels. They also undergo stop word removal and lemmatization. After pre-processing, the feature vector representations of the sentences are created and used to train a classifier using the Weka tool kit[1].

After training, the system can accept documents for summarization. An article is pre-processed and its feature vector representation is created as during training. A parameter specified by a user controls the compression ratio by adjusting the classifier threshold as well the number of clusters used. After classification, the positively labeled instances are filtered using a sentence clusterer, and a final summary is generated. Figure 2 shows an overview of the execution phase of the summarization system.

The next section describes the actual methods and features used for classification. Section 3.2 gives details about the clustering stage and section 3.3 describes how the parameter specifying the compression ratio is used to adjust the length of the summaries.

3.1 Sub-component for classification

Let A be the set of sentences in the abstracts of papers, and let D be the sentences in the main sections of the papers. Using the set of sentences in A and D, we trained a classifier that learns

[1]http://www.cs.waikato.ac.nz/ml/weka

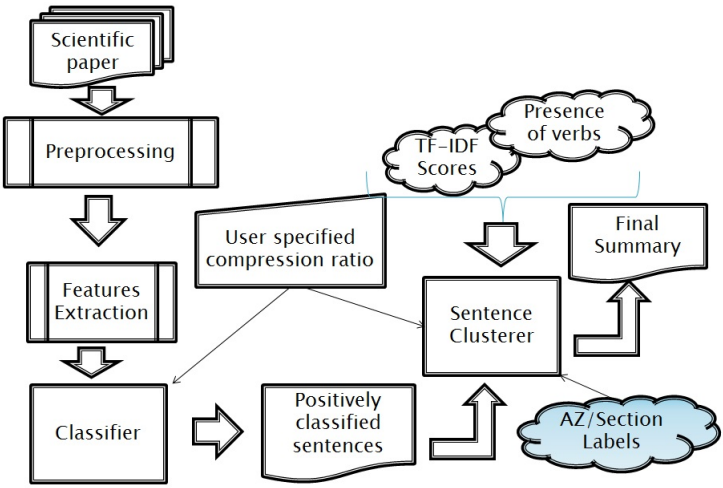

Figure 2: Schematic diagram showing the execution phase of the summarization system.

how to generate a set of sentences $C, C \subseteq D$, which is the candidate set of sentences that are to be made part of a summary.

We require both positive and negative labeled instances for training a classifier. The sentences in the abstract can be considered as positive labeled instances but those in the main text are unlabeled, i.e. they could be positive or negative. The problem of training a classifier using positive and unlabeled data has been studied before. A state-of-the-art method has been described in (Elkan and Noto, 2008), where the classifier is built using positive and unlabeled instances. The model predicts probabilities that differ by a constant factor from the actual conditional probabilities of being positive. Using the constant factor, one can estimate the probabilities of positive and negative instances. We employ this method to train a classifier for selecting the candidate set of sentences (referred to as the "non-traditonal classifier" based method in section 4.3.2).

Instead of using the positive and unlabeled data, one could artificially generate some labeled data and use that for training a classifier. Consider a similarity metric $\omega(S_1, S_2)$ that returns a score indicating the similarity between two sentences S_1 and S_2. Using this similarity metric ω [2] we can identify those sentences in the main text that are most similar to the sentences in the abstract and label them as positive instances, while the others can be labeled as negative instances. A traditional classifier model based on Support Vector Machines was built using these positive and negative instances.

[2]We used two measures – n-gram overlap based similarity between sentences and cosine similarity between sentence feature vectors – and found that the latter gave better results.

3.1.1 Features used for classification

We used various features for representing the training and test samples used by the classifier, including the new AZ feature as well as features that have proved successful in previous related works, e.g. (Abu-Jbara and Radev, 2011).

- Verbs feature: Sentences in the abstract tend to contain many verbs. For example, text fragments like "We showed that", "we found that", "We used", "we proved" are very common in abstracts, often using past tense. Using the StanfordNLP Part-of-speech (POS) tagger (Toutanova et al., 2003), each sentence was tagged and verbs along with their tenses were identified and included in the feature set for that sentence.
- tf-idf values: TF-IDF scores for each sentence were used as features for the sentences. The size of tf-idf features is the size of the vocabulary. Each word in the vocabulary along with its tf-idf value in a sentence were used as features.
- Citation and reference occurrences: Sentences containing citations are frequently found where published related work is discussed. They tend to occur in the background, prior or related work sections. We also keep track of sentences that contain references to figures and tables. These point to a section in the same document, while a citation points to a different document. Each sentence is assigned two boolean features, indicating the presence of a citation and a reference to figures or tables, respectively.
- Argumentative zones: Each sentence is labeled with one of eight AZ categories both manually and automatically using the system of (Guo et al., 2011a).
- Locative features: Sentences tend to have locative characteristics, e.g. most sentences describing prior work occur in the beginning of a paper, while those describing future work tend to occur at the end. Position of sentences has earlier been found to be useful in summarization tasks (Baxendale, 1958) (Conroy and O'leary, 2001).

We experimented with different combinations of these features. Citation and location based features were only used by the traditional classifier, as these features are unavailable when the original abstracts are used directly for training.

3.2 Sub-component for clustering

A well rounded summary should briefly describe the nature of the problem, the work conducted, and the nature of the results obtained, without repeating any information. Once the sentences have been classified, a clustering step is used to remove redundancy from them and to identify similar sentences. Using the section labels to group similar sentences, we applied the k-means clustering (Lloyd, 1982) to detect clusters within each group. By selecting the centroid from each of the k-clusters in each section group, the final set of sentences were identified for a summary.

An alternative to using the section labels for grouping sentences is to make use of AZ. Sentences with the same AZ label can be grouped together, and the clusters can be identified within each group. We experimented with this option as well. The feature vector representation of sentences used for clustering consists of tf-idf weights as well as variables indicating the presence or absence of verbs.

3.3 Controlling the length of summaries

The compression ratio of a summary is defined as the ratio of the number of sentences in the summary generated and the number of sentences in the original article. Using the compression ratio as an input parameter to the system, the length of the summaries can be controlled.

The compression ratio is used to adjust the classification threshold of the classifier. If the number of sentences is too low, the classification threshold is reduced by a fixed step size and sentences are re-classified. This process is repeated until the the compression ratio is a fixed constant t % more than the actual compression ratio required. This relaxed compression ratio is used so that the clustering stage has enough sentences to choose from.

In clustering, the compression ratio is used to determine the number of sub-clusters to be created within each AZ/section label group. Let the compression ratio be denoted by cr, the length of original document be l and the number of distinct AZ labels/section labels in the classified set be m. Then the number of clusters k is given by :

$$k = ceil(\frac{cr * l}{m}) \quad (1)$$

where $ceil(x)$ is a function that returns the smallest integral value that is greater than or equal to the real number x.

4 Experiments and Results

4.1 Data and Pre-processing

We used a corpus of 50 biomedical articles sourced from a number of journals on cancer which are available online at PubMed[3]. The corpus contains 580 sentences in the abstracts and 7,989 sentences in the main body. The sentences were annotated according to the AZ annotation scheme of (Mizuta et al., 2006). Eight AZ categories[4] appeared in the annotated data[5], including Background, Conclusion, Problem, Connection, Method, Difference, Result and Future work. Inter-annotator agreement between the two annotators (one domain expert and one computational linguist) was high $\kappa = 0.83$ according to Cohen's kappa (Cohen, 1960).

	Sentence Type	Training data	Test data	Validation set
Biomedical corpus	Abstract sentences	308	206	66
	Main article sentences	5046	2943	-

Table 1: Details of data set

For the experiments with automatically obtained AZ labels, we used the weakly-supervised method of (Guo et al., 2011a) to identify the AZ category of each sentence. Based on the active learning and self-training, the method was trained using just 10% of labeled data in a corpus of 1000 biomedical articles. With accuracy of 81% it performs similarly with supervised methods that employ all the labels (Guo et al., 2010).

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

[4]Please see the paper of (Mizuta et al., 2006) for the full details of the annotation scheme and examples of different zone categories.

[5]The data and the source code of the methods described in this paper are available on request.

We split the articles in our corpus into 3 sets for training, testing and validation. The validation set was created from a small set of sentences from abstracts and was used to learn a classifier from positive and unlabeled training samples (Elkan and Noto, 2008).

All sentences were tagged using the Stanford NLP POS tagger (Toutanova et al., 2003) with the Penn treebank tagset[6]. In addition, the articles were lemmatized using the Stanford NLP lemmatizer[7] and stop words were removed using stop-lists available on the Internet.

4.2 Experiments

We evaluated the method on two tasks: full document and customized summarization. In full document summarization, a user-specified compression ratio is used to automatically summarize the contents of the entire article. In customized summarization, the user specifies the length and the focus of the summary to be generated (e.g. a summary of the "methods" described in the paper only).

4.2.1 Evaluation Measures

The ROUGE-N measure (Lin, 2004) is used frequently for evaluation of summarization systems. ROUGE stands for Recall Oriented Understudy for Gisting Evaluation and the ROUGE-N score is calculated by counting the number of overlapping N-grams between a user generated/reference summary and a system summary. ROUGE does not consider the length of the summaries, and therefore, if an entire article is returned, it could get the best ROUGE score as the number of n-gram matches will be high. Therefore, compression ratios (cr), i.e. ratio of the number of sentences in the summary ($|S_{summary}|$) generated and the number of sentences in the original article ($|S_{article}|$) are also frequently used.

$$cr = \frac{|S_{summary}|}{|S_{article}|} \tag{2}$$

We used as the primary evaluation metric the F_1 [8] measure, calculated using the number of overlapping n-grams between the summary generated at different compression ratios and the abstracts created by the authors.

4.3 Full article summarization

In this task, sentences from the abstracts were used to learn how to generate full length summaries of articles. Different combinations of features were used to train classifiers.

4.3.1 Training

Sentences from the abstracts were used to train a non-traditional classifier of (Elkan and Noto, 2008) and to create an artificial set of positive and negative instances for training the traditional classifier. The artificial data set was created by selecting such sentences from the main text that were similar to sentences from the abstract. The similarity criteria was based on the cosine distance between the feature vectors of the sentences. The following section describes the results for both these methods with different combinations of features.

[6] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQP-HTMLDemo/PennTreebankTS.html
[7] http://www-nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/process/Morphology.html
[8] http://en.wikipedia.org/wiki/F1_score

Classification features	Cluster groups created using	Precision (P)	Recall (R)	F1 Measure
Verb, tf–idf	None	0.1966	0.3926	0.2620
	Section label	0.1946	0.4058	0.2630
	AZ Label	**0.2091**	**0.4301**	**0.2814**
Verb,tf–idf, AZ	None	0.1938	0.3766	0.2559
	Section label	0.1861	0.3985	0.2537
	AZ Label	0.2054	0.4287	0.2777

Figure 3: Full Document summarization results using non-traditional classifier and compression ratio of 10 %

4.3.2 Results

Figure 3 shows the performance of the non-traditional classifier methods for summarization when manual AZ labels are used. We varied the compression ratio between 5% and 25% in steps of 5% and studied the performance. We report here the results using the compression ratio of 10% because it produces summary length that corresponds the closest to the length of the actual abstracts. As can be seen, during clustering, the use of AZ labels considerably improves the F_1 scores and also improves both precision and recall (ROUGE-1). The features used during classification are the verb features, the tf-idf features and the AZ label features. Other features (e.g. locative and citation based) were not used in this method as they are unavailable in the sentences from the abstracts and in the training data.

Figure 4 shows the performance of the traditional classifier based summarization, again using manual AZ labels. Here the results are better when clustering employs the AZ labels instead of section labels for grouping sentences. Features based on location of sentences and citations were used as the training data consists of sentences from the main text. The use of citation features along with locative features improves the performance of the summarizer, though the non-traditional classifier outperforms the traditional classifier.

The use of section labels for clustering is a tough baseline to beat, as these sections contain sentences that the authors themselves deemed fit to belong to those sections. For example, sentences belonging to a section called “Result” could be considered to identify sentences relating to “Results”, but the use of sentences belonging to the argumentative zone called “Result” are not confined just to the “Result” section, and the use of the AZ labels shows a significant improvement in performance. We also experimented using clustering without creating cluster groups based on sections or AZ and found that the use of AZ labels improves performance. The improvement in F_1 scores when using AZ during the clustering stage was found to be statistically significant ($p < 0.03$).

Finally, we performed an experiment using automatically detected AZ labels by the method of (Guo et al., 2011a). Using the best feature configurations for the task, this results in a small (2%) drop in F_1 scores when compared to the use of manual labels, but slight improvement when compared to the use of section labels (See Figure 5).

Classification features	Cluster groups created using	Precision (P)	Recall (R)	F1 Measure
Verb, tf–idf	None	0.1984	0.3779	0.2602
	Section label	0.1966	0.3827	0.2597
	AZ Label	0.1994	0.3993	0.2660
Verb,tf–idf, AZ	None	0.1952	0.3871	0.2595
	Section label	0.1936	0.3950	0.2599
	AZ Label	0.1953	0.4215	0.2669
Verb,tf–idf, AZ, Citation	None	0.2063	0.3288	0.2535
	Section label	0.2061	0.3290	0.2535
	AZ Label	0.2146	0.3372	0.2632
Verb, tf–idf, AZ, Citation, Locative	None	0.2066	0.3882	0.2697
	Section label	0.2048	0.3733	0.2644
	AZ Label	**0.2102**	**0.3781**	**0.2707**

Figure 4: Full Document summarization results using traditional classifier and compression ratio of 10 %

Classification features	Cluster groups created using	Precision (P)	Recall (R)	F1 Measure
Verb, tf–idf (Non Traditional Classifier)	None	0.1966	0.3926	0.2620
	Section label	0.1946	0.4058	0.2630
	Weakly Supervised AZ Label	0.2096	0.4010	0.2753
	AZ Label	**0.2091**	**0.4301**	**0.2814**

Figure 5: Performance when using AZ labels automatically generated using the weakly supervised method of (Guo et al., 2011a)

4.4 Customized summarization

In this task, the system generates summaries for some parts of the paper based on user requirements. We evaluated the system on two customized summary tasks that our experts found useful for biomedical literature review: the summarization of "Results" in a paper and the "Discussion" in a paper. The evaluation was done against gold-standard summaries for 50 articles, generated by a human expert (an expert in biomedical research). The expert was asked to generate the customized summaries by selecting sentences from the main body of the articles and to ensure that the summaries generated do not exceed 40 % of the full article length. The gold-standard summaries were also used for training as described in the next section.

4.4.1 Training

The gold-standard summary set was split for training and testing (60-40 % split). The sentences from the gold-standard were used as positive labels for training the classifier, while the negative instances were labeled in two ways.

In the first method, sentences from the main article, which were not part of the gold-standard summary, were labeled as negative instances. Thus, in this case, the whole set of sentences from the article is available for training. In the second method, the AZ labels of the sentences are used to get a reduced set of negative instances. Sentences from the zone best suited to the customized summary task are selected. For example, in the customized summarization task for "Results"' in a paper, the negative instances will only contain sentences which are not part of the gold-standard and which belong to the "Results" zone. This method is referred to as "Zone pre-selection"' in the results section.

4.4.2 Results

For the customized summarization task, the summary lengths are controlled based on the number of sentences desired instead of the compression ratio. Estimating a compression ratio based on the full length of the article is not suitable because the task focuses on summarizing a "part" of the article (only the "results" for example), and not the entire article. The "part" of the article that is summarized is based on the "type" of information the user is interested in, and is not by itself, explicitly demarcated in the original article.

In this task, therefore, if x is the number of sentences required in the summary, the the number of clusters are given by:

$$n = ceil(\frac{x}{m}) \tag{3}$$

When zone pre-selection is employed, the clustering stage was found to be less useful for the "Results" and "Discussion" summarization tasks, because the classification stage itself was able to identify a good set of sentences to be used for the summary. The problem of redundant information is reduced, because the summaries are generated of some "parts" of the document and not the entire document, reducing chances of redundancy.

It must also be noted that this behaviour may not be universally true for all types of customized summaries. For example, summarization of the "Background" work in a paper, which usually contains more information and tend to be larger sections, may contain redundant information and the classifier may not be as selective due to the increase in length of the text being summarized.

Customized Summary Type	Features	Precision	Recall	F1
Summarize Results	Verbs, tf–idf	0.1649	0.4766	0.2451
	Verbs,tf–idf,AZ	0.1613	0.4906	0.2428
	Verbs,tf–idf, Zone pre–selection	**0.3312**	**0.6253**	**0.4330**
Summarize Discussion	Verbs, tf–idf	0.1178	0.3318	0.1739
	Verbs,tf–idf,AZ	0.1285	0.3354	0.1858
	Verbs,tf–idf, Zone pre–selection	**0.1838**	**0.4959**	**0.2682**

Figure 6: Customized Summarization results for summary length of 15 sentences

Customized Summary Type	Zone Pre–selection	Precision	Recall	F1
Summarize Results	Weakly Supervised AZ	0.3257	0.6145	0.4257
	Manually labeled AZ	0.3312	0.6253	0.4330
Summarize Discussion	Weakly Supervised AZ	0.1867	0.4882	0.2702
	Manually labeled AZ	0.1838	0.4959	0.2682

Figure 7: Comparison of customized summarization results using weakly supervised and manually labeled AZ

Figure 6 presents results for the two customized summarization tasks. The use of zone pre-selection improves the results considerably against the baseline system which does not make use of zone pre-selection. There is a 76 % improvement in F_1 scores in the "Results" task and an improvement of 54 % in the "Discussion" task. It can also be seen that the use of AZ as a feature for classification does not cause significant change in performance. This was also noticed in the full document summarization task, where AZ were found to be most effective during the clustering stage and not during the classification stage.

Figure 7 compares the performance of the best performing customized summarization configuration (Zone pre-selection) using weakly supervised AZ labels. An analysis of the errors made by the weakly supervised automated AZ labeling method showed that the "Results" and "Conclusion" AZ labels account for 9% and 20% of the errors respectively. The "Conclusion" AZ category is one where the annotators to have most disagreement, partly because many sentences include elements of both discussion and some other zone (e.g. methods or results), yet annotators are asked to assign each sentence to one category only. Nevertheless, our experiments shows that the performance of the summarization system when using automatically generated AZ labels is comparable to that of a system using manually labeled AZ labels.

The results presented in this section are promising, showing that AZ can yield improvements in both full document and customized summarization tasks in biomedicine.

Conclusion and Future Work

Most work on the information structure of scientific literature has been evaluated directly against manually labeled data. Task-based evaluation has mainly concentrated on information retrieval and extraction tasks. We have investigated whether AZ could be used to benefit summarization of scientific articles. Although previous work had suggested that AZ can improve summarization, no experiment had been conducted using a full AZ scheme and a real summarization system.

We developed a simple summarization system that uses a classifier to identify a set of candidate sentences, and uses clustering along with AZ labels to reduce redundancy in the summaries generated. The system is capable of creating full document summaries of different length and information density as well as customized summaries based on user requirements. Both types of summaries can be helpful for users in the scientific domain.

We evaluated the summarization performance on both full document and customized summarization and reported statistically significant improvement in performance scores when using AZ labels. The system outperforms a strong baseline method that uses section labels instead of the AZ labels. The improvement of approximately 7 % in F_1 scores in the full document summarization and an improvement of 54-76 % in customized summarization clearly shows that AZ can benefit automatic summarization.

Our main focus was on manual AZ annotations because we wanted to investigate the direct impact and the upper bound of AZ on summarization. However, also our pilot experiments using automatic AZ annotations show improvement in summarization performance. Future work could use our method as a framework for task-based evaluation of AZ labeling systems.

In this initial investigation on the topic, we kept the summarization framework intentionally simple for evaluation purposes. Future work could optimise the use of AZ for state-of-the-art summarization systems and also explore further ways of integrating AZ in the task. For example, our experiments show that zones are useful for building better clusters. Instead of employing clustering to reduce redundancy, one could investigate the use of diversity ranking algorithms. Once the sentences have been grouped based on zone labels, as described in section 3.2, diversity ranking algorithms, e.g. (Radlinski et al., 2008), could be used to obtain a ranked list of topically or information "diverse" sentences, from which the summary could be built.

Alternatively, instead of using a classification and clustering based approach, sentences from the main article could be selected based on a diversity ranking algorithm and then the final summary could be built using the distribution of zones in the abstracts or gold standard summaries as a "summary template".

Although we focused on AZ due to its good applicability to different scientific domains, success in previous task-based evaluations, and the availability of a weakly-supervised AZ detection method which enables easy porting between NLP tasks, it would be interesting to investigate and compare the usefulness of other schemes of information structure for summarization.

Acknowledgement

The first author's studies at the University of Cambridge were partially funded by the Cambridge Overseas Trust. We would also like to thank Dr. Ilona Silins, University of Cambridge for creating the gold-standard summaries for the customized summarization task. The work in this paper was also funded by the Royal Society (UK), EPSRC (UK) grant EP/G051070/1 and EU grant 7FP-ITC-248064.

References

Abu-Jbara, A. and Radev, D. (2011). Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 500–509, Stroudsburg, PA, USA. Association for Computational Linguistics.

Baxendale, P. B. (1958). Machine-made index for technical literature: an experiment. *IBM J. Res. Dev.*, 2(4):354–361.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Conroy, J. M. and O'leary, D. P. (2001). Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 406–407, New York, NY, USA. ACM.

Elkan, C. and Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 213–220, New York, NY, USA. ACM.

Farzindar, A. and Lapalme, G. (2004). Letsum, a text summarization system in law field. In *THE FACE OF TEXT conference (Computer Assisted Text Analysis in the Humanities)*, pages 27–36, McMaster University, Hamilton, Ontario, Canada.

Guo, Y., Korhonen, A., Liakata, M., Karolinska, I. S., Sun, L., and Stenius, U. (2010). Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, BioNLP '10, pages 99–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guo, Y., Korhonen, A., and Poibeau, T. (2011a). A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 273–283, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Liakata, M., Teufel, S., Siddharthan, A., and Batchelor, C. R. (2010). Corpora for the conceptualisation and zoning of scientific papers. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *LREC*. European Language Resources Association.

Lin, C. Y. (2004). Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough? In *Proceedings of the NTCIR Workshop 4*.

Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, pages 129–136.

Mei, Q. and Zhai, C. (2008). Generating impact-based summaries for scientific literature. In *Proceedings of ACL-08: HLT*, pages 816–824, Columbus, Ohio. Association for Computational Linguistics.

Mizuta, Y., Korhonen, A., Mullen, T., and Collier, N. (2006). Zone analysis in biology articles as a basis for information extraction. *I. J. Medical Informatics*, pages 468–487.

Ozsoy, M. G., Cicekli, I., and Alpaslan, F. N. (2010). Text summarization of turkish texts using latent semantic analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 869–876, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qazvinian, V. and Radev, D. R. (2010). Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 555–564, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qazvinian, V., Radev, D. R., and Özgür, A. (2010). Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 895–903, Stroudsburg, PA, USA. Association for Computational Linguistics.

Radlinski, F., Kleinberg, R., and Joachims, T. (2008). Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 784–791, New York, NY, USA. ACM.

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

Shatkay, H., Pan, F., Rzhetsky, A., and Wilbur, W. J. (2008). Multi-dimensional classification of biomedical text. *Bioinformatics*, 24(18):2086–2093.

Steinberger, J., Kabadjov, M. A., and Poesio, M. (2005). Improving lsa-based summarization with anaphora resolution. In *In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 1–8.

Tbahriti, I., Chichester, C., Lisacek, F., and Ruch, P. (2006). Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the medline digital library. *I. J. Medical Informatics*, 75(6):488–495.

Teufel, S. and Moens, M. (2002). Summarizing scientific articles - experiments with relevance and rhetorical status. *Computational Linguistics*, 28:2002.

Teufel, S., Siddharthan, A., and Batchelor, C. (2009). Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1493–1502, Stroudsburg, PA, USA. Association for Computational Linguistics.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.