BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 11 2013, pages 1440–1447
doi:10.1093/bioinformatics/btt163

Data and text mining

Advance Access publication April 5, 2013

Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review

Yufan Guo[1,*], Ilona Silins[2], Ulla Stenius[2] and Anna Korhonen[1]

[1]Computer Laboratory, University of Cambridge, Cambridge CB3 0FD, UK and [2]Institute of Environmental Medicine, Karolinska Institutet, Stockholm SE-171 77, Sweden

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Techniques that are capable of automatically analyzing the information structure of scientific articles could be highly useful for improving information access to biomedical literature. However, most existing approaches rely on supervised machine learning (ML) and substantial labeled data that are expensive to develop and apply to different sub-fields of biomedicine. Recent research shows that minimal supervision is sufficient for fairly accurate information structure analysis of biomedical abstracts. However, is it realistic for full articles given their high linguistic and informational complexity? We introduce and release a novel corpus of 50 biomedical articles annotated according to the Argumentative Zoning (AZ) scheme, and investigate active learning with one of the most widely used ML models—Support Vector Machines (SVM)—on this corpus. Additionally, we introduce two novel applications that use AZ to support real-life literature review in biomedicine via question answering and summarization.

Results: We show that active learning with SVM trained on 500 labeled sentences (6% of the corpus) performs surprisingly well with the accuracy of 82%, just 2% lower than fully supervised learning. In our question answering task, biomedical researchers find relevant information significantly faster from AZ-annotated than unannotated articles. In the summarization task, sentences extracted from particular zones are significantly more similar to gold standard summaries than those extracted from particular sections of full articles. These results demonstrate that active learning of full articles' information structure is indeed realistic and the accuracy is high enough to support real-life literature review in biomedicine.

Availability: The annotated corpus, our AZ classifier and the two novel applications are available at http://www.cl.cam.ac.uk/~yg244/12bioinfo.html.

Contact: yg244@cam.ac.uk

Received on November 22, 2012; revised on March 22, 2013; accepted on March 26, 2013

1 INTRODUCTION

Techniques that can help biomedical scientists access information in the rapidly growing body of biomedical literature are much needed. Especially, those capable of automatically identifying specific types of information in literature (e.g. information related to the methods or results used in different studies) can be particularly helpful. We focus on techniques that aim to identify the discourse or information structure of scientific documents (Guo *et al*., 2010; Liakata *et al*., 2010; Mizuta *et al*., 2006; Shatkay *et al*., 2008; Teufel and Moens, 2002; Teufel *et al*., 2009) because such techniques have yielded promising results and have proved to benefit important tasks such as information extraction, summarization and scientific literature review (Guo *et al*., 2011a; Mizuta *et al*., 2006; Ruch *et al*. 2007; Teufel and Moens, 2002). However, most existing approaches use fully supervised machine learning (ML), which requires a large body of annotated data that is difficult and costly to obtain. This limits the applicability of the approaches to different domains and tasks in biomedicine.

A more practical approach would use only minimal supervision for the learning process. A recent experiment shows that active learning using only a small amount of labeled data can identify the information structure of biomedical abstracts reliably (Guo *et al*., 2011b). However, most scientific tasks require access to full-text articles instead. Full articles are considerably more complex linguistically and in terms of information structure than abstracts. For example, their sentences tend to be longer and their structure more complex than in abstracts, including, for example, more frequent use of coordination, negation, passive and anaphora (Cohen *et al*., 2010). The information provided is richer (Cohen *et al*., 2005; Schuemie *et al*., 2004) and the categories of information more numerous and less evenly distributed than in abstracts (see also our analysis in section 2.1). As a result, information structure analysis has proved more challenging for full articles than for abstracts (Mullen *et al*., 2005).

We will investigate whether despite these challenges, weakly supervised learning could be used to identify the information structure of full-text biomedical articles with useful accuracy. We focus on Argumentative Zoning (AZ)—a scheme that describes the rhetorical progression in scientific text (Teufel and Moens, 2002)—because it has been successfully applied to various disciplines (Mizuta *et al*., 2006; Teufel *et al*., 2009; Teufel and Moens, 2002) and tasks (Guo *et al*., 2011a, c; Ruch *et al*., 2007; Teufel, 2005; Teufel and Moens, 2002).

However, as none of the AZ-annotated full-text corpora were publicly available, we developed our own corpus including 50 biomedical articles (consisting of 8171 sentences and 234 619 words). We use active learning—a method that aims to reduce the cost of annotation by iteratively selecting the most informative instances to be labeled—for learning AZ of biomedical articles in our new corpus. Active learning was used for AZ of

*To whom correspondence should be addressed.

biomedical abstracts earlier by Guo *et al.* (2011b) and it has been applied to various text classification tasks (e.g. Brinker, 2006; Esuli and Sebastiani, 2009; Hoi *et al.*, 2006; Lewis and Gale, 1994; Novak *et al.*, 2006; Silva and Ribeiro, 2007), achieving performances close to that of fully supervised learning. In this study, we investigate its performance with one of the most widely used ML models—Support Vector Machines (SVM). In comparison with the related work on biomedical abstracts (Guo *et al.*, 2011b), we deal with the complexity of full articles by performing a more detailed linguistic analysis (from sentence splitting to feature extraction), by making use of features that take context information into account, and by introducing a wider range of active learning strategies (e.g. Query-by-Committee).

Trained on 500 labeled sentences (6% of the corpus) our method yields impressive accuracy of 82% on full articles, just 2% lower than that of fully supervised learning. In addition, we introduce two novel applications where the idea is to use AZ to support real-life literature review in biomedicine. We apply them to the area of cancer risk assessment, and use them for task-based evaluation of our AZ approach. The first application focuses on question answering. We investigate whether biomedical researchers find relevant information faster from AZ-annotated than from unannotated full articles. The results show that the AZ-annotations accelerate the process by 27–30%, making literature review considerably faster. The second application focuses on customized summarization. We use the application to create customized summaries for the conclusions of full articles as many biomedical scientists are particularly interested in these. The results show that sentences extracted from the *Conclusion* zone are significantly more similar to gold-standard summaries, with 4% higher ROUGE-1 *F*-score, than those extracted from the *Discussion* section of articles. In sum, our investigation shows that active learning can yield highly accurate results for AZ of full-text articles and that its performance is high enough to benefit real-life tasks in biomedicine.

We make our AZ-annotated corpus available with this article, together with our novel applications so that they can benefit further research in this area where publicly available resources are scarce.

2 METHODS

2.1 Annotated corpus

We developed a corpus of 50 biomedical articles (consisting of 8171 sentences and 234 619 words) from a set of biomedical journals [e.g. *Carcinogenesis*, *Toxicological Sciences*, *Journal of Biological Chemistry*, among others (http://www.cl.cam.ac.uk/~yg244/12bioinfo/readme.txt)], and annotated them according to the AZ scheme that describes the rhetorical progression of scientific text. The scheme was originally introduced by Teufel and Moens (2002) who applied it to computational linguistics articles. We used the version that Mizuta *et al.* (2006) adapted for biology articles, with minor modifications concerning zone names: Background (BKG), Problem (PROB), Method (METH), Result (RES), Conclusion (CON), Connection (CN), Difference (DIFF) and Future work (FUT). Zones BKG, PROB, METH, RES, CON and FUT refer to the background of the study, the research question, the methods used, the results obtained, the conclusions drawn and the future directions, respectively. CN and DIFF refer to related work that is consistent or inconsistent with authors' work.

Table 1. Distribution of sentences (shown in percentages) in abstracts, full articles and their individual sections in the AZ-annotated corpus

Text	BKG	PROB	METH	RES	CON	CN	DIFF	FUT
Abstract	14.6	10.9	15.8	37.5	21.0	–	–	0.2
Article	16.9	2.8	34.8	17.9	22.3	4.3	0.8	0.2
Introduction	74.8	13.2	5.4	0.6	5.9	0.1	–	–
Methods	0.5	0.2	97.5	1.4	0.2	0.2	0.1	–
Results	4.0	2.1	11.7	68.9	12.1	1.1	0.1	–
Discussion	16.9	1.1	0.7	1.5	63.5	13.3	2.4	0.7

We developed a tool that allows users to open an article in the Firefox browser and to annotate the sentences with AZ categories. A biomedical expert annotated all the 50 articles sentence by sentence so that each sentence was assigned to a single zone (a practice followed by most variations of AZ annotation).

Table 1 shows the distribution of sentences in abstracts, full articles and their individual sections in the annotated corpus. As section names vary from article to article, we grouped similar names before calculating the statistics. For instance, sections *Case presentation*, *Experimental procedures*, *Materials and Methods* were merged into *Methods*. Within full articles, METH is the most frequent category, accounting for 34.8% of the data, followed by CON, RES and BKG, accounting for 22.3, 17.9, and 16.9% of the data, respectively. The four low-frequency categories are FUT, DIFF, PROB and CN, covering 0.2–4.3% of the data each. Full articles include a larger number of zones than abstracts, e.g. CN and DIFF. Our statistics also show that although there is one major zone in each section (e.g. BKG for *Introduction*), 2.5–36.5% of the sentences still belong to other categories, demonstrating that the scheme and the annotations are informative. We measured the inter-annotator agreement between the biomedical expert and a computational linguist on 15 articles. According to Cohen's kappa (Cohen, 1960), annotators are in a good agreement with $\kappa = 0.83$.

2.2 Machine learning for AZ

In academic writing, zones tend to appear in sequential order. For example, BKG is usually followed by PROB, and RES is followed by CON. Therefore a natural approach to AZ would be a sequence model such as Conditional Random Fields, which takes into account transition probabilities. However, in weakly supervised (and in particular active) learning, diversity of the selected labeled data is important. This is difficult to achieve with a sequence model where a selected sequence is an entire article that may include hundreds of sentences but tends to be limited in vocabulary and structures. Moreover, recent work has shown that non-sequence models actually perform better than sequence models in identifying information categories in full articles and short abstracts (Guo *et al.*, 2011b; Liakata *et al.*, 2012). Hence we base our active learning approach on SVM—the most widely used non-sequence model—in this work.

2.2.1 SVM and active learning SVM aims to find the maximum-margin hyperplane that separates the classes:

$$\begin{array}{ll} \min\limits_{w,b,\xi} & \frac{1}{2}||w||^2 + C\sum\limits_{i=1}^{n}\xi_i \\ \text{Subject to} & y_i(w\cdot x_i - b) \ge 1-\xi_i, \quad \xi_i \ge 0, \end{array}$$

where x_i is a data point and y_i is its label, w is a normal vector to the hyperplane, ξ_i is a slack variable that measures the degree of misclassification for x_i and C is the penalty term. The parameters can be learned using the SMO algorithm (Platt, 1999b). We used LIBSVM (Chang and Lin, 2011) with linear kernel for our experiments.

Algorithm 1 Pool-based active learning

```
Require: labeled set $L$, unlabeled pool $U$, query strategy $Q(\cdot)$, query batch
  size $B$, stopping criterion $C$
  while $!C$ do
    //training on the current L
    train($L$)
    for $b = 1$ to $B$ do
      //query the most informative instance
      $x^* = \arg\min_{x \in U} \; Q(x)$
      //move the labeled query from $U$ to $L$
      $L = L \cup <x^*, label(x^*)>$
      $U = U - x^*$
    end for
  end while
```

The idea of active learning is to reduce annotation cost by iteratively selecting the most informative instances to be labeled and used as training data for the next iteration. Algorithm 1 shows how pool-based active learning works. The following query strategies were used for SVM-based active learning:

Least confident sampling (Lewis and Gale, 1994) queries the instance whose label the current model is the least confident about:

$$\arg\min_{x \in U} Q(x) = \arg\min_{x \in U} P(y^*|x;\theta).$$

For SVM, a monotonic function (sigmoid function) was used to transform the distance between data points and hyperplanes into posterior probabilities (Platt, 1999a) and, because there were more than two classes, the probabilities were combined by pairwise coupling (Wu *et al.*, 2004).

Margin sampling (Scheffer *et al.*, 2001) is another uncertainty sampling strategy that queries the instance with the smallest margin between the posteriors of the two most likely labelings:

$$\arg\min_{x \in U} Q(x) = \arg\min_{x \in U} P(y1^*|x;\theta) - P(y2^*|x;\theta),$$

Query-by-Bagging (Abe and Mamitsuka, 1998) is a variation of the Query-by-Committee strategy (Seung *et al.*, 1992) for discriminative or non-probabilistic models. It maintains a committee of models that are trained on a portion of the current labeled data but that represent competing hypotheses. The most informative instance is the one about which the committee members disagree the most. Although this method has been shown to work well, no previous work has applied this method to information structure analysis, so we decided to compare it with other query strategies in this work.

2.2.2 Features Each sentence was represented by features that had proved promising in related works (Guo *et al.*, 2011b; Merity *et al.*, 2009; Mullen *et al.*, 2005; Teufel and Moens, 2002):

Section. Normalized section names (Introduction, Methods, Results, Discussion).

Location (i/ii/iii). Each article/section/paragraph was divided into 10 equal parts. Location was defined by the parts where the sentence begins and ends.

Reference (i/ii). The number of citations/referred tables and figures in a sentence (0, 1 or more).

Word. All the words in the corpus (a word feature equals 1 if it occurs in the sentence and 0 if it does not; the rest of the features were defined in a similar way).

Bi-gram. Any combination of two adjacent words in the corpus.

Verb. All the verbs in the corpus.

Verb Class. Sixty verb classes obtained by spectral clustering (Sun and Korhonen, 2009).

Tense and Voice. Tense and voice indicated by the part-of-speech (POS) tag of main verbs and auxiliary verbs, e.g. *have*|*VBZ be*|*VBN* __|*VBN* indicates present perfect tense, passive voice. Previous work such as (Guo *et al.*, 2011c; Liakata *et al.*, 2012) made use of the POS tags of main verbs directly as features. In this work, we take into account not only the main verbs but also the chain of corresponding auxiliary verbs for a more accurate tense and voice analysis.

Grammatical relation (GR). Subject (*ncsubj*), direct object (*dobj*), indirect object (*iobj*) and second object (*obj2*) relations involving verbs, e.g. *(ncsubj observed difference obj)*.

Noun (i/ii). The subjects/objects appearing with any verbs in the corpus (extracted from GRs).

An elaborate sentence splitter and tokenizer was developed to deal with complex biomedical terms and various types of citations in full-text articles. The C&C POS tagger and parser (Curran *et al.*, 2007) trained on biomedical literature were used for extracting syntactic features. We lemmatized the lexical items for all syntactic features using Morpha (Minnen *et al.*, 2001), and removed the words, bi-grams and GRs with fewer than two occurrences.

2.2.3 Context information A well-written scientific article aims for a logical flow of ideas and connections (cohesion) between sentences. Therefore, the context of a sentence can be a good indicator of its information category. For instance, we do not know if the sentence '*The most consistent demographic variable influencing the MN frequency was age, with MN frequency increasing significantly with age (citation)*.' belongs to CN or DIFF before we see the following sentence: '*However, our results indicated that there was no significant increase in MN frequency among older workers compared with younger workers...*'. Also, (Teufel and Moens, 2002) showed that the label of the preceding sentence can be an important feature for AZ with fully supervised learning. However, because the labels of the surrounding sentences are not available in weakly supervised learning, we used the features of both the target and the surrounding sentences for active learning-based AZ.

2.2.4 Evaluation We evaluated the accuracy, precision, recall and *F*-score of SVM with full or light supervision. We also plotted pairwise receiver operating characteristic (ROC) (Landgrebe and Duin, 2007) curves for a more detailed comparison between these methods. All results were averaged across 10-folds using cross-validation to avoid confirmation bias. Each fold was used once as test data and the remaining 9-folds as training (labeled and unlabeled) data.

2.3 Novel AZ-based applications for supporting biomedical literature review

Many real-life tasks in biomedicine require review of scientific literature. We developed two novel applications where the idea is to use AZ to support literature review via question answering and summarization. We used the applications to support the literature-intensive task of cancer risk assessment of chemicals.

2.3.1 Question answering The information that is interesting to risk assessment can be as general as *Is it a study of humans, animals or cells?* or as specific as *Do they refer to any supporting studies?* Although section names (e.g. *Methods*) can be an indicator of information of interest, many sections are rich in information (e.g. in the *Discussion* section, references to related work are usually mixed with the interpretations of new results), and relevant information may be difficult to find. Guo *et al.* (2011a, c) have shown that it is easier for biomedical researchers to find relevant information in AZ-annotated than unannotated abstracts. In this study, we investigated whether AZ annotations can speed up the process of reviewing full-text articles.

We developed an application where scientists can define and fill in a questionnaire relevant for their literature review, and which highlights, for each question, the information in a scientific article, which is most likely to answer the question according to our AZ annotation. We used this application to do task-based evaluation of our weakly supervised AZ approach in the context of cancer risk assessment.

First, experts in risk assessment provided us with a set of questions relevant for their task (e.g. *Are gene/protein changes studied?*) and either a yes/no or multiple choice answer for them (see Table 2). Next, we developed an interface and on-line questionnaire to record the time it takes for experts to answer each question with and without using AZ annotations. Each question came up one at a time with relevant zones (when available) highlighted in different colors (see Table 2 and Fig. 1). An expert who designed the questions and who was familiar with the annotation scheme determined the relevant zones for each question, and we visualized this mapping in the questionnaires. Two experts (A and B) participated in the test. They were asked to complete 15 questionnaires: (i) 5 using original articles, (ii) 5 that had human-annotated zones highlighted and (iii) 5 that had zones identified by our active learning approach highlighted. The length of the articles was approximately the same across different groups for a fair comparison. We compared the time it took for experts to answer the questions in (i–iii) and evaluated whether the results were statistically significant according to Mann–Whitney *U*-test (Mann and Whitney, 1947; Wilcoxon, 1945). We also evaluated the agreement between experts for quality control.

Table 2. Questions used in the question answering experiment and the corresponding highlighted zones

Question	Zone
Q1 Is the motivation behind the study discussed? (e.g. too little knowledge, data are conflicting. . .) y/n	PROB
Q2 What is the main type of study the article focuses on? Cell experiment/animal/human	METH
Q3 Is a (lowest) dose that gives effect mentioned? (e.g. LOEL) y/n	RES
Q4 Is a dose-response/effect studied? y/n	RES
Q5 Are risk estimates mentioned? (e.g. relative risk, odds ratio) y/n	RES
Q6 Are morphological changes studied? y/n	RES
Q7 Are gene/protein changes studied? y/n	RES
Q8 Are clinical endpoints studied? y/n	RES
Q9 According to the authors, is the outcome of the study expected, unexpected or neither/neutral?	CON
Q10 Do the authors refer to any supporting studies? y/n	CN
Q11 Do the authors refer to any conflicting studies? y/n	DIFF
Q12 Do the authors discuss possible directions of future work? y/n	FUT

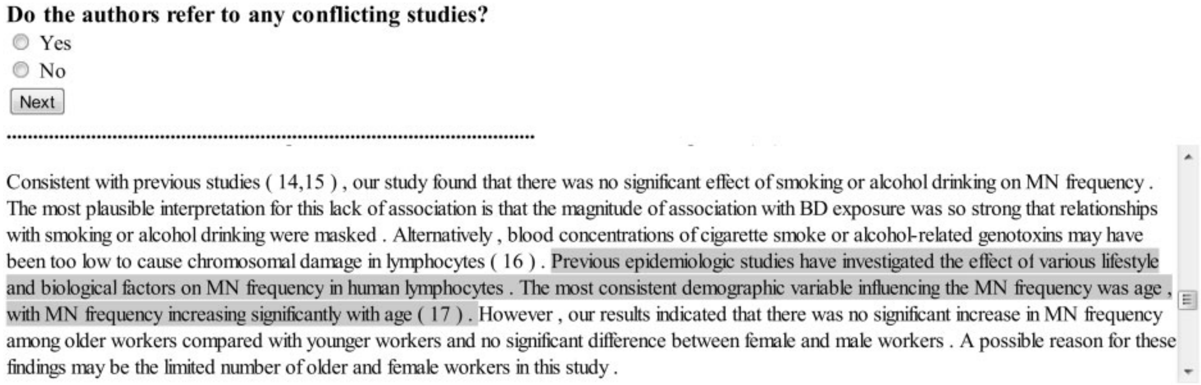
2.3.2 Customized summarization Automatic summarization, which gives a brief statement of the key points of a document could be highly valuable in supporting literature review. Summaries can either be generic, capturing the key idea of a document, or customized, collecting information related to a particular interest of the user (Berger and Mittal, 2000; Mani and Bloedorn, 1998). Most scientific articles include a generic abstract written by authors, so it is easy to get a general idea of the key points of an article. However, when readers need a summary of a particular type of information in an article (e.g. the methods, results or conclusions), they will have to read through the entire article or at least selected sections.

We developed an application that enables automatic creation of customized summaries of different types of information in a scientific article. We used this application to create summaries of the conclusions of articles that are always the most interesting to scientists. Although we could run summarization on the *Discussion* section of an article to create a customized summary for conclusions, this section typically provides also other information, e.g. background and comparison with other works (see Table 1). Our idea was, therefore, to investigate whether AZ labels (i.e. CON for conclusions) are more informative than section names for customized summarization. First, we asked an expert in cancer risk assessment to summarize the conclusions of each article in the corpus. We then used the Microsoft AutoSummarize (http://www.microsoft.com/education/en-us/teachers/how-to/Pages/autosummarize-document.aspx) system to select 10% of sentences from (i) the CON zone and (ii) the *Discussion* section of each article, and evaluated the two types of summaries against those written by the expert in terms of ROUGE scores (Lin, 2004). ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a measure of how many words/word pairs/word sequences are shared between the automatic summary and the gold standard summary.

3 RESULTS

3.1 Machine learning

Table 3 shows the accuracy and *F*-score results for our baseline (which maps section names to zones directly, not relying on ML or annotated data) and those for fully and weakly supervised learning. Fully supervised learning uses all the labeled training data (~7350 sentences), whereas random selection and


Do the authors refer to any conflicting studies?

Consistent with previous studies (14,15) , our study found that there was no significant effect of smoking or alcohol drinking on MN frequency . The most plausible interpretation for this lack of association is that the magnitude of association with BD exposure was so strong that relationships with smoking or alcohol drinking were masked . Alternatively , blood concentrations of cigarette smoke or alcohol-related genotoxins may have been too low to cause chromosomal damage in lymphocytes (16) . Previous epidemiologic studies have investigated the effect of various lifestyle and biological factors on MN frequency in human lymphocytes . The most consistent demographic variable influencing the MN frequency was age , with MN frequency increasing significantly with age (17) . However , our results indicated that there was no significant increase in MN frequency among older workers compared with younger workers and no significant difference between female and male workers . A possible reason for these findings may be the limited number of older and female workers in this study .

Fig. 1. A questionnaire used in the question answering experiment. Questions came up one at a time with the relevant zones highlighted in different colors

Table 3. Accuracy and *F*-scores for individual zones

Method	Acc	*F*-score							
		BKG	PROB	METH	RES	CON	CN	DIFF	FUT
Baseline	0.77	0.69	–	0.93	0.80	0.71	–	–	–
Full supervision	0.84	0.80	0.40	0.95	0.88	0.80	0.50	–	0.11
Random selection	0.79	0.75	0.20	0.93	0.82	0.75	0.23	–	–
Active learning	0.82	0.77	0.42	0.94	0.86	0.77	0.36	–	–

The baseline did not use any labeled data. Fully supervised learning used ~7350 labeled sentences, and random selection and active learning used 500. We report results for the best-performing active learning method: least confident sampling.

active learning use only 500 labeled sentences, selected randomly or according to a particular query strategy. Because the performances of the three query strategies are similar to one another, Table 3 shows the results for least confident sampling only.

Given the distribution of sentences in each section (see Table 1), it is not surprising that the baseline performs so well, with an accuracy of 0.77 and *F*-score ranging from 0.69 to 0.93 for the four major zones (BKG, METH, RES and CON). However, we show that ML can identify also minor categories in each section that enables a more detailed information structure analysis. As shown in Table 3, fully supervised learning outperforms the baseline significantly with 7% higher accuracy and so does active learning with 5% higher accuracy. Notably, the accuracy of active learning is just 2% lower than that of fully supervised learning, which needs substantial labeled data for training.

The highest *F*-score is observed for the METH zone, which makes sense because 97.5% of the *Methods* section belongs to this zone. Although CON is the second largest zone in the corpus, its *F*-scores are 7–9% lower than those for RES, probably because the *Discussion* section is usually a more complex mixture of CON and other zones than the *Results* section. The low-frequency categories such as DIFF and FUT are not identified owing to the lack of training data, but unlike the baseline method, ML does find two more categories, CN and PROB. We further looked into zones where the achievement of active learning is the most significant (e.g. PROB) or trivial (e.g. CON), as illustrated by the pairwise ROC curves in Figure 2. We can see that in both cases, active learning performs as well as full supervision, whereas random selection hardly identifies a low-frequency category.

Table 4 shows the accuracy of weakly supervised learning when 50–500 labeled sentences are used. Active learning has a clear advantage over random selection as the amount of labeled data increases. When 250–500 sentences are labeled, it performs significantly better than random selection with 2–3% higher accuracy [with $P<0.001$ in McNemar's test (McNemar, 1947)]. The three query strategies perform similarly. Least confident sampling performs slightly better than the other two with 1% higher accuracy when 500 sentences are labeled.

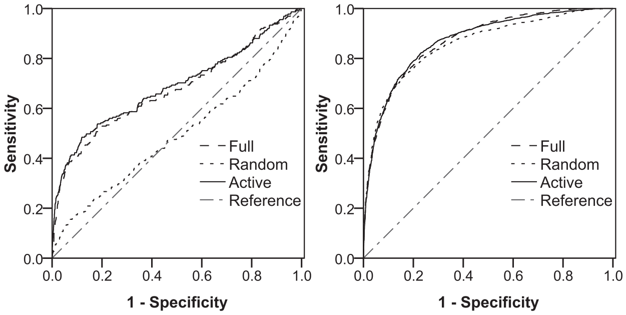


Fig. 2. ROC curves of full supervision, random selection and active learning for PROB versus BKG (left) and CON versus BKG (right). BKG is the most frequent category coming along with PROB/CON in *Introduction*/*Discussion*

Table 4. Accuracy of weakly supervised learning with 50–500 labeled sentences

Method	50	100	150	200	250	300	400	500
Random selection	0.73	0.76	0.77	0.78	0.78	0.78	0.79	0.79
Active learning								
Least confident	0.73	0.76	0.77	0.79	0.80	0.80	0.81	0.82
Margin	0.73	0.76	0.78	0.79	0.80	0.80	0.81	0.81
Query-by-bagging	0.73	0.77	0.78	0.78	0.79	0.79	0.80	0.81

3.2 Question answering and customized summarization

Table 5 shows the time it took experts A and B to answer questions using (i) unannotated articles, (ii) articles that highlight zones annotated by humans and (iii) articles that highlight zones learned by SVM through active learning, as well as the percentage of time savings using AZ annotations. Because questions 3–8 are all related to the RES zone, we sum up the time it took to answer those questions. Looking at the total time spent on all the questions, experts found the information of interest significantly faster from AZ-annotated articles (with *P* ranging from 0.002 to 0.019) than from unannotated ones. They used 36–42% less time when examining manually annotated articles (ii) and 27–30% when examining automatically annotated articles (iii). The experts were in a good agreement and >89% of their answers were the same. Looking at the results for individual questions, AZ annotations are particularly useful for answering questions related to PROB, CON and CN (Q1, Q9, Q10), but not those related to METH or FUT (Q2, Q12) because the METH zone overlaps with the *Methods* section to a large extent, and the FUT zone is rarely seen in the corpus.

Table 6 shows examples of customized summaries focusing on the conclusions of an article, where the zone-based summary was extracted from the CON zone (learned through active learning) and the section-based summary from the *Discussion* section. Comparing the two auto summaries against the gold standard, we can see that the zone-based summary is considerably more accurate: it shares three sentences (in bold) with the gold standard, whereas the section-based summary only shares one (the rest contains background information or related work). Table 7 shows averaged ROUGE scores for zone-based and section-based summaries. Zone-based summaries are significantly more similar to the gold standard with 4–5% higher *F*-scores and 8–10% higher precision than section-based summaries. The recall scores for zone- and section-based summaries are similar, suggesting that both schemes can retrieve equal amount of relevant information. The differences in precision and *F*-score indicate that zone-based summaries are more compact and precise than section-based summaries.

Table 5. The time (measured in seconds) it took experts A and B to answer questions for (i) unannotated articles, (ii) articles with highlighted zones annotated by human and (iii) articles with highlighted zones annotated through active learning

Article	Q1	Q2	Q3–Q8	Q9	Q10	Q11	Q12	Total
A								
(i)	35.8	21.4	125.4	44.2	21.6	26.8	18.6	293.8
(ii)	16.4	21.4	82.2	29.6	11.8	15.2	12.2	188.8
(i)–(ii)	54%	0%	34%	33%	45%	43%	34%	36%
(iii)	20.4	15.4	99	33.4	16.8	11.6	10.2	206.8
(i)–(iii)	43%	28%	21%	24%	22%	57%	45%	30%
B								
(i)	69.6	12	124	157	27.4	33.4	16.8	440.2
(ii)	11.6	13	106	50.2	21	30.2	22.4	254.4
(i)–(ii)	83%	−8%	15%	68%	23%	10%	−33%	42%
(iii)	40.8	13.2	95.8	118.4	20.2	15.4	16.2	320
(i)–(iii)	41%	−10%	23%	25%	26%	54%	4%	27%

(i)–(ii) and (i)–(iii) refer to the percentage of time savings when using AZ annotations.

4 DISCUSSION AND CONCLUSIONS

We have introduced an AZ-annotated corpus of 50 full biomedical articles, twice as big as the one developed by Mizuta *et al.* (2006) for biology articles, and the first publicly available AZ-annotated corpus of scientific articles. Using this corpus, we have investigated, for the first time, whether a weakly supervised approach is realistic for information structure analysis

Table 7. Averaged ROUGE scores for customized summaries: ROUGE-1 for word co-occurrence, ROUGE-2 for bi-gram co-occurrence (any pair of adjacent words) and ROUGE-SU4 for skip-bi-gram co-occurrence (any pair of words in the same order as they appear in sentences)

ROUGE	Zone-based			Section-based		
	Recall	Precision	*F*-score	Recall	Precision	*F*-score
ROUGE-1	0.49	0.60	0.51	0.50	0.50	0.47
ROUGE-2	0.32	0.37	0.33	0.30	0.29	0.28
ROUGE-SU4	0.33	0.39	0.34	0.32	0.31	0.29

Table 6. Examples of customized summaries focusing on the conclusions

Gold standard summary	Zone-based summary	Section-based summary[a]
We found an elevated risk of colorectal cancer associated with high levels of mono-ortho PCBs 28 and 118. In our population, high body burdens of mono-ortho PCBs 28 and 118 were associated with an elevated risk of colorectal cancer. Nevertheless, higher serum levels of PCB-28 were associated with increased risk of colon cancer in this study. In our population, p,p'-DDE also increased risk for tumors with wild-type K-ras but not when this oncogene was mutated. In conclusion, these results suggest that exposure to mono-ortho PCBs is associated with an increased risk of colorectal cancer.	To overcome these limitations, we studied a population not occupationally exposed using serum OC levels as exposure markers. **We found an elevated risk of colorectal cancer associated with high levels of mono-ortho PCBs 28 and 118. In our population, high body burdens of mono-ortho PCBs 28 and 118 were associated with an elevated risk of colorectal cancer.** These OCs are among the most toxic of the PCBs, together with non-ortho PCBs. The role of other OCs in colorectal cancer risk may be more complex. **In conclusion, these results suggest that exposure to mono-ortho PCBs is associated with an increased risk of colorectal cancer**.	OCs have previously been associated with increased risk of colorectal cancers in studies of occupationally exposed individuals (Acquavella *et al.*, 1996; Soliman *et al.*, 1997; Wilkinson *et al.*, 1997). **We found an elevated risk of colorectal cancer associated with high levels of mono-ortho PCBs 28 and 118**. Many studies that also used plasma OC concentrations as exposure markers have reported mixed results for breast cancer (Calle *et al.*, 2002) and non-Hodgkin lymphoma (Cantor *et al.*, 2003; De Roos *et al.*, 2003; Rothman *et al.* 1997) but increased risk of pancreatic cancer (Hoppin *et al.*, 2000; Porta *et al.*, 1999; Slebos *et al.*, 2000). Some studies on breast cancer have also reported increased risk being limited to mono-ortho PCBs (Aronson *et al.*, 2000; Demers *et al.*, 2002; Lucena *et al.*, 2001).

[a]References cited in this column are from an example in our dataset.

of full scientific articles. We have shown that an approach based on active learning performs well with an accuracy of 82% when using 500 labeled sentences. Nearly as good as fully supervised learning (accuracy 84%), our results are promising, especially considering the high linguistic and informational complexity of full articles. We have also introduced two novel task-based evaluations of AZ that involve doing literature review via question answering and customized summarization. In our question answering experiment, researchers find relevant information from AZ-annotated articles 27–42% faster than from unannotated articles, regardless of whether manual or automatic annotations are used. In the customized summarization experiment, sentences extracted from a particular zone are significantly more similar to gold standard summaries, with 8–10% higher precision, than those extracted from a particular section. Both experiments show that active learning-based AZ can support biomedical literature review.

Recent studies have shown promising results on weakly supervised learning of information structure of biomedical abstracts (Guo *et al*., 2011a, c). We have focused on full-text articles whose information structure is considerably more complex than that of abstracts. Despite the challenges, we have reported promising results in our direct and task-based evaluations.

In the future, we intend to improve this work in several directions. First, it may be useful to develop a more fine-grained scheme for analyzing the information structure of full-text articles. Refining the existing scheme will offer an opportunity for a more detailed analysis of the article's information structure, which is likely to benefit also real-life applications such as literature review. With respect to the weakly supervised learning, there has been a lot of recent work on integrating declarative knowledge/constraints into standard ML models for improved performance, such as Generalized Expectation (Mann and McCallum, 2010) and Posterior Regularization (Bellare *et al*., 2009). The power of these methods has been demonstrated for a variety of natural language processing tasks, in particular for text classification (Druck *et al*., 2008). Thus, we plan to investigate these methods for AZ on full articles and evaluate their usefulness in real-world applications. We are also interested in developing domain adaptation technologies for porting of AZ across the sub-fields of biomedicine and other areas of science.

Funding: The work reported in this article was funded by the Royal Society (UK), Swedish Research Council, EPSRC (UK) grant EP/G051070/1 and EU grant 7FP-ITC-248064. Y.G. was funded by the Cambridge International Scholarship. I.S. was funded by the Swedish Governmental Agency for Innovation System.

Conflict of Interest: none declared.

REFERENCES

Abe,N. and Mamitsuka,H. (1998) Query learning strategies using boosting and bagging. In: *Proceedings of the Fifteenth International Conference on Machine Learning*. San Francisco, CA, USA, pp. 1–9.

Bellare,K. *et al.* (2009) Alternating projections for learning with expectation constraints. In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. UAI '09 AUAI Press, Arlington, Virginia, USA, pp. 43–50.

Berger,A. and Mittal,V.O. (2000) Query-relevant summarization using FAQS. In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Hong Kong, pp. 294–301.

Brinker,K. (2006) On active learning in multi-label classification. In:Spiliopoulou,M., Kruse,R., Borgelt,C., Nürnberger,A. and Gaul,W. (eds) *From Data and Information Analysis to Knowledge Engineering*. Springer-Verlag, Berlin/Heidelberg, pp. 206–213.

Chang,C.-C. and Lin,C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 1–27.

Cohen,J. (1960) A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, **20**, 37–46.

Cohen,K.B. *et al.* (2005) Corpus design for biomedical natural language processing. In: *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*. Association for Computational Linguistics, Detroit, Michigan, pp. 38–45.

Cohen,K.B. *et al.* (2010) The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, **11**, 492.

Curran,J.R. *et al.* (2007) Linguistically motivated large-scale nlp with c&c and boxer. In: *Proceedings of the ACL 2007 Demonstrations Session*. Association for Computational Linguistics, Prague, Czech Republic, pp. 33–36.

Druck,G. *et al.* (2008) Learning from labeled features using generalized expectation criteria. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Singapore, Singapore, pp. 595–602.

Esuli,A. and Sebastiani,F. (2009) Active learning strategies for multi-label text classification. In: *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval*. Springer-Verlag, Toulouse, France, pp. 102–113.

Guo,Y. *et al.* (2010) Identifying the information structure of scientific abstracts: an investigation of three different schemes. In: *Proceedings of BioNLP*. Association for Computational Linguistics, Uppsala, Sweden, pp. 99–107.

Guo,Y. *et al.* (2011a) A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics*, **12**, 69.

Guo,Y. *et al.* (2011b) A weakly-supervised approach to argumentative zoning of scientific documents. In:*Proceedingsof the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, United Kingdom, pp. 273–283.

Guo,Y. *et al.* (2011c) Weakly supervised learning of information structure of scientific abstracts–is it accurate enough to benefit real-world tasks in biomedicine? *Bioinformatics*, **27**, 3179–3185.

Hoi,S.C.H. *et al.* (2006) Large-scale text categorization by batch mode active learning. In:*Proceedingsof the 15th international conference on World Wide Web*. ACM, Edinburgh, Scotland, pp. 633–642.

Landgrebe,T.C.W. and Duin,R.P.W. (2007) Approximating the multiclass ROC by pairwise analysis. *Pattern Recogn. Lett.*, **28**, 1747–1758.

Lewis,D.D. and Gale,W.A. (1994) A sequential algorithm for training text classifiers. In:*Proceedingsof the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Springer-Verlag New York, Inc., Dublin, Ireland, pp. 3–12.

Liakata,M. *et al.* (2010) Corpora for the conceptualisation and zoning of scientific papers. In:*Proceedingsof LREC'10*. European Language Resources Association (ELRA), Valletta, Malta.

Liakata,M. *et al.* (2012) Automatic recognition of conceptualisation zones in scientific articles and two life science applications. *Bioinformatics*, **28**, 991–1000.

Lin,C.-Y. (2004) ROUGE: a package for automatic evaluation of summaries. In: *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Association for Computational Linguistics, Barcelona, Spain, pp. 74–81.

Mani,I. and Bloedorn,E. (1998) Machine learning of generic and user-focused summarization. In: *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*. American Association for Artificial Intelligence, Madison, Wisconsin, USA, pp. 820–826.

Mann,G.S. and McCallum,A. (2010) Generalized expectation criteria for semi-supervised learning with weakly labeled data. *J. Mach. Learn. Res.*, **11**, 955–984.

Mann,H.B. and Whitney,D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, **18**, 50–60.

McNemar,Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**, 153–157.

Merity,S., Murphy,T. and Curran,J.R. (2009) Accurate argumentative zoning with maximum entropy models. In: *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*. Association for Computational Linguistics, Suntec, Singapore, pp. 19–26.

Minnen,G. *et al*. (2001) Applied morphological processing of English. *Nat. Lang. Eng*., **7**, 207–223.

Mizuta,Y. *et al*. (2006) Zone analysis in biology articles as a basis for information extraction. *Int. J. Med. Inform*, **75**, 468–487.

Mullen,T. *et al*. (2005) A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *SIGKDD Explor. Newsl*., **7**, 52–58.

Novak,B., Mladeni,Č.D. and Grobelnik,M. (2006) Text classification with active learning. In: Spiliopoulou,M., Kruse,R., Borgelt,C., Nürnberger,A. and Gaul,W. (eds) *From Data and Information Analysis to Knowledge Engineering*. Springer-Verlag, Berlin/Heidelberg, pp. 398–405.

Platt,J.C. (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Alexander,J.S., Peter,B., Bernhard,S. and Dale,S. (eds) *Advances in Large Margin Classiers*. MIT Press, Cambridge, MA, USA, pp. 61–74.

Platt,J.C. (1999b) Using analytic QP and sparseness to speed training of support vector machines. In: *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*. MIT Press, Cambridge, MA, USA, pp. 557–563.

Ruch,P. *et al*. (2007) Using argumentation to extract key sentences from biomedical abstracts. *Int. J. Med. Inform*., **76**, 195–200.

Scheffer,T., Decomain,C. and Wrobel,S. (2001) Active hidden Markov models for information extraction. In: *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*. Springer-Verlag, London, UK, pp. 309–318.

Schuemie,M.J. *et al*. (2004) Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, **20**, 2597–2604.

Seung,H.S., Opper,M. and Sompolinsky,H. (1992) Query by committee. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. ACM, Pittsburgh, Pennsylvania, USA, pp. 287–294.

Shatkay,H. *et al*. (2008) Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, **24**, 2086–2093.

Silva,C. and Ribeiro,B. (2007) Combining active learning and relevance vector machines for text classification. In: *Proceedings of the Sixth International Conference on Machine Learning and Applications*. IEEE Computer Society, Washington, DC, USA, pp. 130–135.

Sun,L. and Korhonen,A. (2009) Improving verb clustering with automatically acquired selectional preference. In: *Proceedings of EMNLP*. Association for Computational Linguistics, Singapore, pp. 638–647.

Teufel,S. (2005) Argumentative Zoning for improved citation indexing. In: Shanahan,J.G., Qu,Yan and Wiebe,Janyce (eds) *Computing Attitude and Affect in Text: Theory and Applications*. Springer, Dordrecht, The Netherlands, pp. 159–170.

Teufel,S. and Moens,M. (2002) Summarizing scientific articles: experiments with relevance and rhetorical status. *Comput. Linguist*., **28**, 409–445.

Teufel,S., Siddharthan,A. and Batchelor,C. (2009) Towards domain-independent argumentative zoning: evidence from chemistry and computational linguistics. In: *Proceedings of EMNLP*. Association for Computational Linguistics, Singapore, pp. 1493–1502.

Wilcoxon,F. (1945) Individual comparisons by ranking methods. *Biometrics Bull*., **1**, 80–83.

Wu,T.-F. *et al*. (2004) Probability estimates for multi-class classification by pairwise coupling. *J. Mach. Learn. Res*., **5**, 975–1005.