BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 22 2011, pages 3179–3185
doi:10.1093/bioinformatics/btr536

Data and text mining

Advance Access publication September 22, 2011

Weakly supervised learning of information structure of scientific abstracts—is it accurate enough to benefit real-world tasks in biomedicine?

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

[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: Many practical tasks in biomedicine require accessing specific types of information in scientific literature; e.g. information about the methods, results or conclusions of the study in question. Several approaches have been developed to identify such information in scientific journal articles. The best of these have yielded promising results and proved useful for biomedical text mining tasks. However, relying on fully supervised machine learning (ML) and a large body of annotated data, existing approaches are expensive to develop and port to different tasks. A potential solution to this problem is to employ weakly supervised learning instead. In this article, we investigate a weakly supervised approach to identifying information structure according to a scheme called Argumentative Zoning (AZ). We apply four weakly supervised classifiers to biomedical abstracts and evaluate their performance both directly and in a real-life scenario in the context of cancer risk assessment.

Results: Our best weakly supervised classifier (based on the combination of active learning and self-training) performs well on the task, outperforming our best supervised classifier: it yields a high accuracy of 81% when just 10% of the labeled data is used for training. When cancer risk assessors are presented with the resulting annotated abstracts, they find relevant information in them significantly faster than when presented with unannotated abstracts. These results suggest that weakly supervised learning could be used to improve the practical usefulness of information structure for real-life tasks in biomedicine.

Availability: The annotated dataset, classifiers and the user test for cancer risk assessment are available online at http://www.cl.cam.ac.uk/~yg244/11bioinfo.html.

Contact: anna.korhonen@cl.cam.ac.uk

Received on April 21, 2011; revised on August 29, 2011; accepted on September 19, 2011

1 INTRODUCTION

Many practical tasks in biomedicine require accessing specific types of information in scientific literature. For example, a biomedical scientist may be looking for information about the objective of the study in question, the methods used, the results obtained or the conclusions drawn by the authors. Similarly, many biomedical text mining tasks (e.g. information extraction, summarization) focus on the extraction of specific types of information in documents only.

To date, a number of approaches have been proposed for the classification of sentences in scientific literature according to categories of information structure (or discourse, rhetorical, argumentative or conceptual structure, depending on the framework in question). Some of the approaches classify sentences according to typical section names seen in scientific documents (Hirohata *et al.*, 2008; Lin *et al.*, 2006), while others are based e.g. on argumentative zones (Mizuta *et al.*, 2006; Teufel and Moens, 2002; Teufel *et al.*, 2009), qualitative dimensions (Shatkay *et al.*, 2008) or conceptual structure (Liakata *et al.*, 2010) of documents.

The best current approaches have yielded promising results and proved useful for information retrieval, information extraction and summarization tasks (Mizuta *et al.*, 2006; Ruch *et al.*, 2007; Tbahriti *et al.*, 2006; Teufel and Moens, 2002). However, relying on fully supervised machine learning (ML) and a large body of annotated data, existing approaches are expensive to develop and port to different domains and tasks, and thus intractable for use in real-life applications.

A potential solution to this bottleneck is to develop techniques based on weakly supervised ML instead. Making use of a small amount of labeled data and a large pool of unlabeled data, weakly supervised learning (e.g. semi-supervision, active learning, co/tri-training, self-training) aims to keep the advantages of fully supervised approaches. It has been applied to a wide range of natural language processing (NLP) and text mining tasks, including named-entity recognition, question answering, information extraction, text classification and many others (Abney, 2008), yielding performance levels similar or equivalent to those of fully supervised techniques.

In this article, we investigate the potential of weakly supervised learning for Argumentative Zoning (AZ) of biomedical abstracts. AZ provides an analysis of the argumentative structure (i.e. the rhetorical progression of the argument) of a scientific document (Teufel and Moens, 2002). It has been used to analyze scientific texts in various disciplines—including computational linguistics (Teufel and Moens, 2002), law, (Hachey and Grover, 2006), biology (Mizuta *et al.*, 2006) and chemistry (Teufel *et al.*, 2009)—and has proved useful for NLP tasks such as summarization (Teufel and Moens, 2002). However, the application of AZ to different domains has resulted in laborious annotation exercises that suggests that a weakly supervised approach would be more practical for the real-world application of AZ.

*To whom correspondence should be addressed.

Table 1. Categories of AZ appearing in the corpus of Guo *et al.* (2010)

Category	Abbreviation	Definition and example
Background	BKG	The circumstances pertaining to the current work, situation, or its causes, history, etc. *e.g. Concerns about the possible toxic effects of workplace exposures in the synthetic rubber industry have centered on 1,3-butadiene (BD), styrene and dimethyldithiocarbamate (DMDTC).*
Objective	OBJ	A thing aimed at or sought, a target or goal *e.g. The objective of this research was to evaluate techniques for the rapid detection of chromosomal alterations occurring in humans exposed to butadiene.*
Method	METH	A way of doing research, esp. according to a defined and regular plan; a special form of procedure or characteristic set of procedures employed in a field of study as a mode of investigation and inquiry *e.g. The hypoxanthine-guanine phosphoribosyltransferase (HPRT) and thymidine kinase (TK) mutant frequencies (MFs) were measured using a cell cloning assay.*
Result	RES	The effect, consequence, issue or outcome of an experiment; the quantity, formula, etc. obtained by calculation *e.g. Replication past the N3 2'-deoxyuridine adducts was found to be highly mutagenic with an overall mutation yield of approximately 97%.*
Conclusion	CON	A judgment or statement arrived at by any reasoning process; an inference, deduction, induction; a proposition deduced by reasoning from other propositions; the result of a discussion, or examination of a question, final determination, decision, resolution, final arrangement or agreement *e.g. Thus, in terms of mutagenic efficiency, stereochemical configurations of EB and DEB are not likely to play a significant role in the mutagenicity and carcinogenicity of BD.*
Related work	REL	A comparison between the current work and the related work *e.g. These data are much lower compared to previously reported values measured by GC-MS/MS.*
Future work	FUT	The work that needs to be done in the future *e.g. Additional studies are needed to examine the importance of base excision repair (BER) in maintaining genomic integrity, the differential formation of DNA and protein adducts in deficient strains, and the potential for enhanced sensitivity to BD genotoxicity in mice either lacking or deficient in both biotransformation and DNA repair activity.*

Taking two supervised classifiers as a comparison point—Support Vector Machines (SVM) and Conditional Random Fields (CRF)—we investigate the performance of four weakly supervised classifiers for AZ: two based on semi-supervised learning (transductive SVM and semi-supervised CRF) and two on active learning (Active SVM alone and in combination with self-training). We apply these classifiers to AZ-annotated biomedical abstracts in the recent dataset of Guo *et al.* (2010). The results are promising. Our best weakly supervised classifier (Active SVM with self-training) outperforms the best supervised classifier (SVM), yielding high accuracy of 81% when using just 10% of the labeled data for training. When using just one-third of the labeled data, it performs as well as a fully supervised SVM, which uses 100% of the labeled data.

The abstracts in the dataset of Guo *et al.* (2010) were selected on the basis of their suitability for cancer risk assessment (CRA). This enables us to conduct user-based evaluation of the practical usefulness of our approach for the real-world task of CRA. We investigate whether cancer risk assessors find relevant information in abstracts faster when the abstracts are annotated for AZ using our best weakly supervised approach. The results are promising: although manual annotations yield the biggest time savings: 10–13% (compared with the time it takes to examine unannotated abstracts), considerable savings are also obtained with weakly supervised ML annotations: 7–8% (using active SVM with self-training).

In sum, our investigation shows that weakly supervised AZ can be employed to improve the practical applicability and portability of AZ to different information access tasks and that its accuracy is high enough to benefit a real-life task in biomedicine.

Table 2. Distribution of sentences in the AZ-annotated corpus

	BKG	OBJ	METH	RES	CON	REL	FUT
Word	36 828	23 493	41 544	89 538	30 752	2456	1174
Sentence	1429	674	1473	3185	1082	95	47
Sentence (%)	18	8	18	40	14	1	1

2 METHODS

2.1 Data

We used in our experiments the recent dataset of Guo *et al.* (2010), consisting of 1000 CRA abstracts (7985 sentences and 225 785 words) annotated according to AZ. Originally introduced by Teufel and Moens (2002), AZ is a scheme that provides an analysis of the argumentative structure of a document, following the knowledge claims made by the authors. The dataset of Guo *et al.* (2010) has been annotated according to the version of AZ developed for biology papers (Mizuta *et al.*, 2006) (with only minor modifications concerning zone names). Seven categories of this scheme (out of the 10 possible) actually appear in abstracts and in the resulting dataset. These are shown and explained in Table 1. The table also shows one example sentence per category.

Table 2 shows the distribution of sentences per category in the corpus: Result (RES) is by far the most frequent category (accounting for 40% of the corpus), while Background (BKG), Objective (OBJ), Method (METH) and Conclusion (CON) cover 8–18% of the corpus each. Two categories: Related work (REL) and Future work (FUT) are low frequency categories, only covering 1% of the corpus each.

Guo *et al.* (2010) reported the inter-annotator agreement between their three annotators: one linguist, one computational linguist and one domain expert. The agreement ($\kappa = 0.85$) is relatively high according to Cohen (1960).

2.2 Automatic classification

2.2.1 Features and feature extraction The first step in automatic classification is to select a set of features that may indicate AZ categories in abstracts. Following Guo *et al.* (2010), we implemented a set of features that have proved successful in related works, e.g. (Hirohata *et al.*, 2008; Lin *et al.*, 2006; Mullen *et al.*, 2005; Teufel and Moens, 2002):

- Location. The parts where a sentence begins and ends. Each abstract was divided into 10 parts (1–10, measured by the number of words).
- Word. All the words in the corpus (The value of the Word features equals 1 if a certain word occurs in the sentence and 0 if not. The same applies to the following features.).
- Bi-gram. Any combination of two adjacent words in the corpus.
- Verb. All the verbs in the corpus.
- Verb Class. 60 verb classes appearing in biomedical journal articles.
- Part-of-Speech – POS. The POS tag of each verb in the corpus.
- Grammatical Relation – GR. Subject (*ncsubj*), direct object (*dobj*), indirect object (*iobj*) and second object (*obj2*) relations in the corpus. e.g. *(ncsubj observed_14 difference_5 obj)*.
- Subj and Obj. The subjects and objects appearing with any verbs in the corpus (extracted from GRS).
- Voice. The voice of verbs (active or passive) in the corpus.

These features were extracted from the corpus using a number of tools. A tokenizer was used to detect sentence boundaries and to perform basic tokenization (in extreme cases, processing complex biomedical terms e.g. *2-amino-3,8-diethylimidazo[4,5-f]quinoxaline*). The C&C tools (Curran *et al.*, 2007) were used for POS tagging, lemmatization and parsing. The lemma output was used for Word, Bi-gram and Verb features, and the GR output for GR, Subj, Obj and Voice features. The 'obj' marker in a subject relation indicates passive voice [e.g. *(ncsubj observed_14 difference_5 obj)*]. Verb classes were obtained automatically using unsupervised spectral clustering (Sun and Korhonen, 2009). To reduce data sparsity, we lemmatized the lexical items for all the features, and removed words and GRS with <2 occurrences and bi-grams with <5 occurrences.

2.2.2 Machine learning methods The next step is to assign sentences in abstracts to zone categories using machine learning. Support vector machines (SVM) and conditional random fields (CRF) have proved the best performing fully supervised methods in recent related works e.g. (Guo *et al.*, 2010; Hirohata *et al.*, 2008; Mullen *et al.*, 2005; Teufel and Moens, 2002). We therefore implemented these methods as well as weakly supervised variations of them: active SVM with and without self-training, transductive SVM and semi-supervised CRF.

Supervised methods: SVM aims to find the maximum-margin hyperplane, which has the largest distance to the nearest data points of any class. The problem is defined as:

$$\text{Maximize } \frac{2}{|w|}(\text{in } w,b) \text{ subject to } y(w\cdot x-b)\geq 1,$$

where x is data, y is its label, w is a normal vector to the hyperplane and $\frac{2}{|w|}$ is the margin. We used Weka software (Hall *et al.*, 2009) employing the SMO algorithm (Platt, 1999b) with linear kernel for SVM experiments.

CRF is an undirected graphical model that defines a probability distribution over the hidden states (e.g. label sequences) given the observations. The probability of a label sequence y given an observation sequence x can be written as:

$$p(y|x,\theta)=\frac{1}{Z(x)}exp\left(\sum_j \theta_j F_j(y,x)\right),$$

where $F_j(y,x)$ is a real-valued feature function of the states and the observations; θ_j is the weight of F_j, and $Z(x)$ is a normalization factor. We used Mallet software (http://mallet.cs.umass.edu) employing the L-BFGS algorithm (Nocedal, 1980) for CRF experiments.

Weakly supervised methods: Active SVM (ASVM) starts with a small amount of labeled data, and iteratively chooses a certain amount of unlabeled data, about which the classifier is least certain, to be manually labeled (the labels can be restored from the fully annotated corpus) for the next round of learnig. We used an uncertainty sampling query strategy (Lewis and Gale, 1994). In particular, we compared the posterior probabilities of the best estimate given each unlabeled instance, and chose the instances with the lowest probabilities to be labeled for later use. The probabilities can be calculated by fitting a Sigmoid after the standard SVM (Platt, 1999a) and, in the multi-class case, combined using a pairwise coupling algorithm (Hastie and Tibshirani, 1998). We used the -M flag in Weka for computing the posterior probabilities.

Active SVM with self-training (ASSVM) is an extension of ASVM where each round of learning has two steps:

(i) Active learning

  (a) Train a new classifier on all the labeled examples.

  (b) Apply the current classifier to each unlabeled example.

  (c) Find n examples about which the classifier is least certain to be manually labeled.

(ii) Self-training

  (a) Train a new classifier on both labeled and unlabeled/machine-labeled data using the estimates from step (i)(b).

  (b) Test the current classifier on test data.

Transductive SVM (TSVM) is an extension of SVM that aims to:

$$\text{Maximize } \frac{2}{|w|}(\text{in } w,b,y^{(u)}),$$

$$\text{Subject to}$$

$$y^{(l)}(w\cdot x^{(l)}-b)\geq 1,$$

$$y^{(u)}(w\cdot x^{(u)}-b)\geq 1,$$

$$y^{(u)}\in\{-1,1\},$$

where $x^{(u)}$ is unlabeled data and $y^{(u)}$ the estimate of its label. The idea is to find a prediction on unlabeled data such that the decision boundary has the maximum margin on both the labeled and the unlabeled (now labeled) data. The latter guides the decision boundary away from dense regions. We used UniverSVM software (http://3t.kyb.tuebingen.mpg.de/bs/people/fabee/universvm.html) employing the CCCP algorithm (Collobert *et al.*, 2006) for TSVM experiments.

Semi-supervised CRF (SSCRF) can be implemented by entropy regularization (Jiao *et al.*, 2006), which extends the objective function on Labeled data $\sum_L \log p(y^{(l)}|x^{(l)},\theta)$ with an additional term $\sum_U \sum_Y p(y|x^{(u)},\theta)\log p(y|x^{(u)},\theta)$ to minimize the conditional entropy of the model's predictions on Unlabeled data. We used Mallet software for SSCRF experiments.

2.2.3 Evaluation methods We evaluated the ML results in terms of accuracy, precision ($\frac{\text{true positive}}{\text{true positive+false positive}}$), recall ($\frac{\text{true positive}}{\text{true positive+false negative}}$) and F-score ($\frac{2*\text{precision}*\text{recall}}{\text{precision+recall}}$) against manual annotations. We used 10-fold cross-validation for all the methods to avoid the possible bias introduced by relying on any particular split of the data. The data were randomly assigned to 10-folds of roughly the same size. Each fold was used once

as test data and the remaining nine folds as training data (with x% being manually labeled). The results were then averaged. As randomly selected labeled data were used for SVM, CRF, TSVM and SSCRF, the results for these methods were averaged from five runs. Following Dietterich (1998), we used McNemar's test (McNemar, 1947) to measure the statistical significance of the differences between the results of supervised and weakly supervised learning. The chosen significance level was 0.05.

2.3 User test in the context of CRA

A major time-consuming component of chemical cancer risk assessment (CRA) is the review and analysis of existing scientific literature on the chemical in question. MEDLINE (http://www.nlm.nih.gov/databases/databases_medline.html) abstracts are typically used as a starting point in this work. Risk assessors (e.g. toxicologists, biologists) read the abstracts of interest, looking for various information in them (e.g. about the methods, results and conclusions of the study in question) (Korhonen *et al.*, 2009). One way to speed up this work is to annotate the abstracts with categories of information structure so that the information of interest can be found faster. Guo *et al.* (2011) investigated this idea first and showed that time savings can be obtained in literature review when abstracts are annotated either manually or automatically (using fully supervised ML) according to different information structure schemes.

We evaluated our weakly supervised approach to AZ in a similar way but re-designed the evaluation of Guo *et al.* (2011) so that it is better controlled and covers a wider range of information. Cancer risk assessors working in Karolinska Institutet (Stockholm, Sweden) provided us with a list of 10 questions considered when studying abstracts for CRA purposes.

Table 3. Questions and highlighted zones

Question	Zone
Q1 Do the authors discuss previous or related research on the topic? y/n	BKG REL
Q2 Do the authors describe the aim of the research? y/n	OBJ
Q3 What is the main type of study the abstract focuses on? animal study/human study/in vitro study	METH
Q4 Is exposure length mentioned? y/n	METH
Q5 Is dose mentioned? y/n	METH
Q6 Is group size mentioned? y/n	METH
Q7 How many endpoints are mentioned? 0/1/more	RES
Q8 Are the results positive? y/n/unclear	RES
Q9 Is the outcome of the study expected/unexpected/neutral?	CON
Q10 Do the authors mention a need for future research on the topic? y/n	FUT

We turned any open-ended questions (e.g. Author's conclusions?) into more controlled ones (e.g. Is the outcome of the study expected, unexpected, or neither/neutral?) so that each question has either a yes/no or multiple choice answer (Table 3). We then designed an online questionnaire where each question–answer pair is displayed to an expert on a separate page and the zone(s) most relevant for answering the question are highlighted with colors as to attract expert's attention (Fig. 1).

Two experts participated in the test: one professor level expert (A) with a long experience in CRA (over 25 years) and one more junior expert (B) with a PhD in toxicology and over 5 years of experience in CRA. Each expert was presented with the same set of 200 abstracts focusing on four chemicals (butadiene, diethylnitrosamine, diethylstilbestrol and phenobarbital): (i) 50 unannotated, (ii) 50 manually annotated, (iii) 50 ASSVM-annotated and (iv) 50 randomly annotated abstracts (i.e. annotated so that sentences were assigned to zones on the basis of their observed distribution in the training data). We compared the time it took for experts to answer the questions when presented with abstracts in (i)–(iv), and examined whether the differences are statistically significant (significance level of 0.05, Mann–Whitney U Test (Mann and Whitney, 1947; Wilcoxon, 1945)). In addition, we evaluated the impact of (i)–(iv) on the quality of experts' answers by examining inter-expert agreement.

3 RESULTS

3.1 Automatic classification

Table 4 shows the results for the four weakly supervised and two supervised methods when using 10% of the labeled data (i.e. ∼700 sentences). ASSVM is the best performing method, with an accuracy of 81% and macro F-score of 0.76. ASVM performs nearly as

Table 4. Results when using 10% of the labeled data

	Acc	F-score							
		MF	BKG	OBJ	METH	RES	CON	REL	FUT
SVM	0.77	0.73	0.79	0.60	0.70	0.84	0.69	–	–
CRF	0.70	0.64	0.74	0.52	0.46	0.77	0.73	–	–
ASVM	0.80	0.75	0.88	0.56	0.68	0.87	0.78	0.33	
ASSVM	0.81	0.76	0.86	0.56	0.76	0.88	0.76	–	–
TSVM	0.76	0.72	0.82	0.57	0.69	0.82	0.72	0.08	–
SSCRF	0.71	0.65	0.78	0.50	0.48	0.77	0.73	–	–

MF: Macro F-score calculated for the five high frequency categories: BKG, OBJ, METH, RES, CON which are found by all the methods.

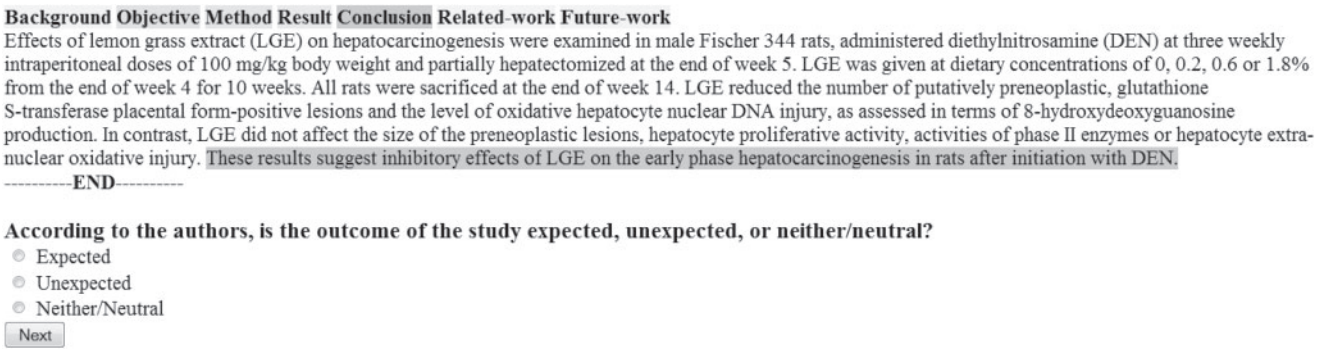

Fig. 1. An example of the questionnaire.

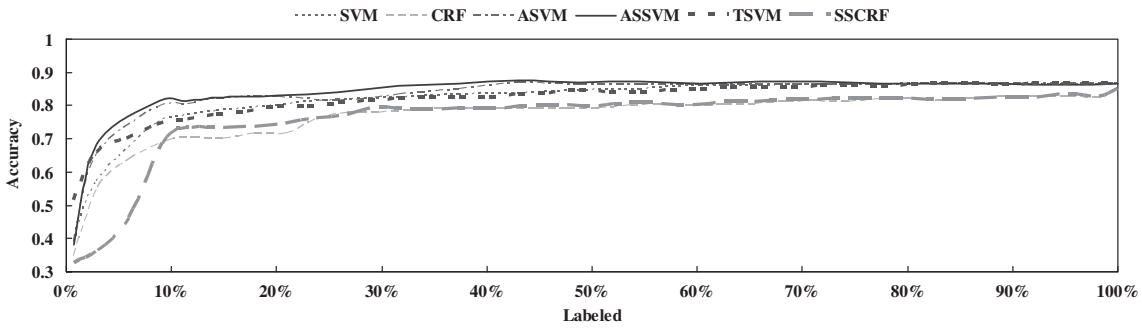

Fig. 2. Learning curve of different methods when using 0–100% of the labeled data.

well, with an accuracy of 80% and *F*-score of 0.75. Both methods outperform supervised SVM with a statistically significant difference ($P < 0.001$). TSVM is the lowest performing SVM-based method: its performance is lower than that of the supervised SVM. Yet, it outperforms both CRF-based methods. SSCRF performs slightly better than CRF with 1% higher accuracy and 0.01 higher *F*-score. Only two methods (ASVM, TSVM) find six out of the seven possible zone categories. Other methods find five categories. The 1–2 missing categories are low frequency categories, accounting for 1% of the corpus data each (Table 2). The results for other categories also seem to reflect the amount of corpus data available per category (Table 2), with RES being the highest and OBJ the lowest performing category with most methods.

Figure 2 shows the learning curve of different methods (in terms of accuracy) when using 0–100% of the labeled data. ASSVM outperforms other methods, reaching its best performance of 88% accuracy when using ∼40% of the labeled data. It outperforms ASVM (the second best method) in particular when 20–40% of the labeled data is used. When using 33% of the labeled data, it performs already as well as fully supervised SVM (i.e. using 100% of the labeled data). SVM and TSVM tend to perform quite similarly with each other when >10% of the labeled data are used, but when less data are available, TSVM performs better. Looking at the CRF-based methods, SSCRF outperforms CRF in particular when 10–25% of the labeled data are used. However, neither of them reaches the performance level of SVM-based methods, which is in line with the results of fully supervised CRF and SVM in Guo *et al.* (2011).

To investigate which features are the most useful for weakly supervised learning, we took our best performing method ASSVM and conducted leave-one-out analysis of the features with 10% of the labeled data. The results in Table 5 show that Location is by far the most useful feature, in particular for BKG, METH and CON. The overall performance drops 8% in accuracy and 0.09 in *F*-score when removing this feature. Removing POS has almost equally strong effect, in particular on BKG and METH. Also Voice, Verb class and GR contribute to general performance. Among the least helpful features are those which suffer from sparse data problems, e.g. Word, Bi-gram and Verb). They perform particularly badly when applied to low frequency zones.

Table 5. Leave-one feature-out results for ASSVM with 10% of labeled data

	Acc.	*F*-score							
		MF	BKG	OBJ	METH	RES	CON	REL	FUT
Location	0.73	0.67	0.67	0.55	0.62	0.85	0.65	–	–
Word	0.80	0.78	0.87	0.70	0.74	0.85	0.72	–	–
Bigram	0.81	0.75	0.83	0.57	0.71	0.87	0.78	0.33	–
Verb	0.81	0.79	0.84	0.77	0.73	0.87	0.75	–	–
VC	0.79	0.75	0.86	0.62	0.72	0.84	0.70	–	–
POS	0.74	0.70	0.66	0.65	0.66	0.82	0.73	–	–
GR	0.79	0.75	0.83	0.67	0.69	0.84	0.72	–	–
Subj	0.80	0.76	0.87	0.65	0.73	0.85	0.72	–	–
Obj	0.80	0.78	0.84	0.75	0.70	0.85	0.75	–	–
Voice	0.78	0.75	0.88	0.70	0.71	0.83	0.62	–	–
Φ	0.81	0.76	0.86	0.56	0.76	0.88	0.76	–	–

Φ: Employing all the features.

3.2 User test

Table 6 shows the time (measured in seconds) it took for experts A and B to answer questions (individual and total) when presented with (i) unannotated, (ii) manually annotated, (iii) ASSVM annotated and (iv) randomly annotated abstracts (see Section 2 for details of the experts and abstract groups), along with the percentage of time savings obtained when using annotations (ii)–(iv) (compared with (i)). TIME stands for the sample mean, and SAVE for the percentage of time savings. Table 7 shows the statistical significance (*P*-values, Mann–Whitney U Test) of the differences between the results for different abstract groups [e.g. (i) v.(ii)]. Looking at the overall figures (i.e. Total), both manual (ii) and ASSVM (iii) annotations help users find relevant information significantly faster than plain text abstracts (i): the percentage of time savings ranges between 7% and 13%, and the corresponding *P*-values ranges between < 0.001 and 0.027. Although manual annotations save more time than ASSVM annotations (13% versus 7% for A, and 10% versus 8% for B), ASSVM annotations are surprisingly useful. Random annotations (iv) have a negative effect: both experts spend more time examining (iv) than (i) abstracts: 6% for A and 19% for B.

Looking at the results for individual questions, (ii) and (iii) are more helpful for answering broader questions (e.g. Q9 Is the outcome of the study expected/unexpected/neutral?) than more specific questions (e.g. Q4 Is exposure length mentioned?). Although (ii) is more helpful than (iii) for most questions, the majority of differences are not statistically significant, showing that ASSVM annotations are almost equally useful as manual annotations. ASSVM annotations have a negative effect on Q4 and Q5. Q4 and Q5 focus on METH which is a higher frequency (accounting for 18% of the corpus) but less predictable (0.76 *F*-score for ASSVM) category.

Table 6. Time savings

		Q1		Q2		Q3		Q4		Q5		Q6		Q7		Q8		Q9		Q10		Total	
		TIME	SAVE (%)	TIME	SAVE (%)	TIME	SAVE (%)	TIME	SAVE (%)	TIME	SAVE (%)	TIME	SAVE (%)	TIME	SAVE (%)	TIME	SAVE (%)	TIME	SAVE (%)	TIME	SAVE (%)	TIME	SAVE (%)
A	(i)	14.1		6.3		11.8		7.2		5.2		6.1		12.5		7.7		8.7		3.4		83.0	
	(ii)	11.9	16	5.7	9	9.1	23	7.2	0	4.7	9	5.6	8	10.6	15	7.2	7	7.2	17	3.1	9	72.2	13
	(iii)	12.0	15	6.0	4	9.8	16	8.5	−17	5.9	−12	5.1	17	11.5	9	7.5	3	7.6	12	3.2	7	77.1	7
	(iv)	12.6	11	7.5	−19	14.7	−25	8.6	−19	5.2	0	5.5	9	12.8	−2	7.4	4	9.5	−10	3.9	−12	87.8	−6
B	(i)	10.1		9.8		8.8		9.6		4.9		5.7		12.4		9.2		12.7		3.8		87.1	
	(ii)	8.7	14	9.3	5	7.3	17	10.0	−5	4.9	0	5.0	12	12.1	3	7.3	21	9.6	24	3.9	−3	78.3	10
	(iii)	9.0	11	10.0	−2	8.8	1	10.0	−5	5.2	−6	4.9	15	11.8	5	6.9	26	9.9	22	4.0	−6	80.5	8
	(iv)	12.2	−21	12.4	−27	12.7	−44	11.2	−17	5.5	−12	4.9	14	16.0	−29	7.9	15	15.6	−24	5.0	−30	103.5	−19

Table 7. Significance of the results in the previous table

		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Total
A	(i) v. (ii)	**0.035**	0.837	0.063	0.975	0.397	0.421	**0.032**	0.296	**0.015**	0.285	**<0.001***
	(i) v. (iii)	0.083	0.924	0.405	0.162	0.221	0.075	0.154	0.550	0.139	0.413	**0.005**
	(i) v. (iv)	0.200	0.135	**0.005**	**0.018**	0.864	0.248	0.872	0.232	0.315	0.781	0.159
	(ii) v. (iii)	0.570	0.633	0.235	0.141	0.052	0.242	0.326	0.530	0.321	0.851	**0.041**
B	(i) v. (ii)	0.122	0.923	0.180	0.666	0.986	0.149	0.901	**0.006**	**0.002**	0.781	**0.005**
	(i) v. (iii)	0.266	0.321	0.565	0.381	0.338	0.070	0.532	**0.018**	**0.005**	0.786	**0.027**
	(i) v. (iv)	**0.024**	**0.008**	**0.003**	0.106	0.385	**0.027**	**0.008**	0.193	**0.009**	0.077	**<0.001***
	(ii) v. (iii)	0.682	0.188	**0.050**	0.729	0.535	0.693	0.477	0.341	0.795	0.667	0.619

As we mentioned in Section 2.3: 'we compared the time it took for experts to answer the questions when presented with abstracts in (i)–(iv), and examined whether the differences are statistically significant [significance level of 0.05, Mann-Whitney U Test (Mann and Whitney, 1947; Wilcoxon, 1945)]. Values in bold are less than 0.05, indicating that the differences are statically significant.
*After rounding, this value is 0.00

Table 8. Quality of answers (inter-expert agreement)

	Q1	Q2	Q3a	Q3b	Q3c	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Total
(i)	0.63	0.51	0.78	0.90	0.80	0.86	0.90	0.92	0.86	0.76	0.92	0.90	0.81
(ii)	0.70	0.66	0.92	0.96	0.92	0.82	0.92	0.90	0.72	0.76	0.88	0.88	0.84
(iii)	0.82	0.68	0.96	0.98	0.90	0.86	0.86	0.90	0.74	0.86	0.90	0.88	0.86
(iv)	0.66	0.54	0.90	0.92	0.92	0.74	0.88	0.90	0.82	0.82	0.84	0.90	0.82

Since Q3 is a multiple-choice question, we report the inter-expert agreement for each option: Q3a,b,c.

Table 8 shows the joint probability of users' agreement on the answers. Annotations (ii), (iii) and (iv) do not affect the users' agreement a lot: 0.82–0.86 and 0.81 with and without annotations. Interestingly, experts tend to agree the most on the answers when using ASSVM annotated abstracts. This demonstrates that automatic annotation does not affect the quality of the answers.

4 DISCUSSION AND CONCLUSIONS

Our results show that weakly supervised ML can be used for the identification of information structure in biomedical abstracts. In our experiments, the majority of weakly supervised methods: ASSVM, ASVM and SSCRF outperformed their corresponding supervised methods: SVM and CRF. ASSVM/ASVM selects the most difficult instances (or the instances distinct from the existing labeled data) to be manually labeled and then used for the next round of learning, offering a wider coverage of the possible inputs than SVM. SSCRF extends CRF by taking into account the conditional entropy of the model's predictions on unlabeled data (favoring peaked, confident predictions) so that the decision boundary is moved into the sparse regions of input space.

The best performing weakly supervised methods were those based on active learning. When using 10% of the labeled data, active learning combined with self-training (ASSVM) outperformed the best supervised method SVM with a statistically significant difference. ASSVM reached its top performance (88% accuracy) when using 40% of the labeled data, and performed equally well as fully supervised SVM when using just one-third of the labeled data. This result is in line with the results of other text classification works where active learning has proved similarly useful, e.g. Esuli and Sebastiani (2009); Lewis and Gale (1994). In addition, we have demonstrated that the accuracy of our best weakly supervised method (ASSVM) is high enough to benefit a real-life task in biomedicine: cancer risk assessors find relevant information in abstracts significantly faster (7–8%) when the abstracts are annotated using ASSVM (as opposed to being unannotated). In sum, our research shows that application of AZ-style approaches to real-world biomedical tasks can be realistic as only a limited amount of labeled data is needed for it.

To the best of our knowledge, no previous work has been done on weakly supervised learning of textual information structure according to the family of schemes we have focused on Guo *et al.* (2011); Hirohata *et al.* (2008); Liakata *et al.* (2010); Lin *et al.* (2006); Mizuta *et al.* (2006); Shatkay *et al.* (2008). Previous works on these schemes have been fully supervised in nature. In addition, although some works have been evaluated in the context of text mining tasks (e.g. information extraction, summarization), the only previous work which has reported user-centered evaluation in the context of a real-life biomedical task is that of Guo *et al.* (2011).

In the future, we plan to improve and extend this work in several directions. Semi-supervised learning (TSVM and SSCRF) did not perform equally well as active learning in our experiments, although it has proved successful in related works e.g. (Jiao *et al.*, 2006). We suspect that this is due to the high dimensionality and sparseness of our labeled dataset. Given the high cost of obtaining labeled data, methods not needing it are preferable. We plan to thus experiment

with more sophisticated active learning algorithms, e.g. margin sampling (Scheffer *et al.*, 2001), query-by-committee (QBC) (Seung *et al.*, 1992) and SVM simple margin (Tong and Koller, 2001). Combinations of other weakly supervised methods, e.g. EM+active learning (McCallum and Nigam, 1998) and co-training+EM+active learning (Muslea *et al.*, 2002) would also be worth investigating. In addition, we plan to replace the SVM-based model with other models e.g. Logistic Regression, which outperforms SVM in active learning as reported in (Hoi *et al.*, 2006). CRF-based active learning might be a good option too.

The work presented in this article has focused on the AZ scheme. In the future, we plan to investigate the usefulness of weakly supervised learning for identifying information structure according to other popular schemes, e.g. (Hirohata *et al.*, 2008; Liakata *et al.*, 2010; Lin *et al.*, 2006; Shatkay *et al.*, 2008) and not only in scientific abstracts but also in full journal papers, which typically exemplify a larger set of scheme categories. Focusing on full journal papers will also enable further user-based evaluation. For example, although abstracts are used as a typical starting point in CRA, subsequent steps of CRA focus on information in full articles. These more challenging steps may benefit from AZ (and other type of) annotations to a greater degree.

Funding: Royal Society (UK); Swedish Research Council; FAS (Sweden); Cambridge International Scholarship (to Y.G.) and EPSRC (EP/G051070/1 UK).

Conflict of Interest: none declared.

REFERENCES

Abney,S. (2008) *Semi-Supervised Learning for Computational Linguistics*. Chapman & Hall / CRC.

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

Collobert,R. *et al.* (2006) Trading convexity for scalability. In *Proceedings of the 23rd International Conference on Machine Learning*. ACM Press, pp. 201–208.

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

Dietterich,T.G. (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, **10**, 1895–1923.

Esuli,A. and Sebastiani,F. (2009) Active learning strategies for multi-label text classification. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*. Springer, Berlin/Heidelberg, 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*. ACL, pp. 99–107.

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

Hachey,B. and Grover,C. (2006) Extractive summarisation of legal texts. *Artif. Intell. Law*, **14**, 305–345.

Hall,M. *et al.* (2009) The weka data mining software: an update. *SIGKDD Explor. Newsl.*, **11**, 10–18.

Hastie,T. and Tibshirani,R. (1998) Classification by pairwise coupling. *Ann. Stat.*, **26**, 451–471.

Hirohata,K. *et al.* (2008) Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of 3rd International Joint Conference on Natural Language Processing*. ACL, pp. 381–388.

Hoi,S.C.H. *et al.* (2006) Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th International Conference on World Wide Web*. ACM, pp. 633–642.

Jiao,F. *et al.* (2006) Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *COLING/ACL*. ACL, pp. 209–216.

Korhonen,A. *et al.* (2009) The first step in the development of text mining technology for cancer risk assessment: identifying and organizing scientific evidence in risk assessment literature. *BMC Bioinformatics*, **10**, 303.

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

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

Lin,J. *et al.* (2006) Generative content models for structural analysis of medical abstracts. In *Proceedings of BioNLP-06*. pp. 65–72.

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.

McCallum,A. and Nigam,K. (1998) Employing em and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., pp. 350–358.

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

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

Mullen,T. *et al.* (2005) A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *Nat. Lang. Process. Text Min.*, **7**, 52–58.

Muslea,I. *et al.* (2002) Active + semi-supervised learning = robust multi-view learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., pp. 435–442.

Nocedal,J. (1980) Updating Quasi-Newton matrices with limited storage. *Math. Comput.*, **35**, 773–782.

Platt,J.C. (1999a) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classiers*. MIT Press, 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, 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. *et al.* (2001) Active hidden Markov models for information extraction. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*. Springer, pp. 309–318.

Seung,H.S. *et al.* (1992) Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. ACM, 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.

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

Tbahriti,I. *et al.* (2006) Using argumentation to retrieve articles with similar citations. *Int. J. Med. Inform.*, **75**, 488–495.

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

Teufel,S. *et al.* (2009) Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP*. pp. 1493–1502.

Tong,S. and Koller,D. (2001) Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, **2**, 45–66.

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