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

Yufan Guo1, Anna Korhonen1,*, Ilona Silins2 and Ulla Stenius2

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

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 18, 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 suggest 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)

<table>
<thead>
<tr>
<th>Category</th>
<th>Abbreviation</th>
<th>Definition and example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>bkg</td>
<td>The circumstances pertaining to the current work, situation, or its causes, history, etc.</td>
</tr>
<tr>
<td>Objective</td>
<td>obj</td>
<td>A thing aimed at or sought, a target or goal</td>
</tr>
<tr>
<td>Method</td>
<td>meth</td>
<td>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</td>
</tr>
<tr>
<td>Result</td>
<td>res</td>
<td>The effect, consequence, issue or outcome of an experiment; the quantity, formula, etc. obtained by calculation</td>
</tr>
<tr>
<td>Conclusion</td>
<td>con</td>
<td>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</td>
</tr>
<tr>
<td>Related work</td>
<td>rel</td>
<td>A comparison between the current work and the related work</td>
</tr>
<tr>
<td>Future work</td>
<td>fut</td>
<td>The work that needs to be done in the future</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>bkg</th>
<th>obj</th>
<th>meth</th>
<th>res</th>
<th>con</th>
<th>rel</th>
<th>fut</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>368</td>
<td>234</td>
<td>415</td>
<td>89</td>
<td>30</td>
<td>246</td>
<td>174</td>
</tr>
<tr>
<td>Sentence</td>
<td>1429</td>
<td>674</td>
<td>1473</td>
<td>3185</td>
<td>1082</td>
<td>95</td>
<td>47</td>
</tr>
<tr>
<td>Sentence (%)</td>
<td>18</td>
<td>8</td>
<td>18</td>
<td>40</td>
<td>14</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

2 METHODS

2.1 Data

We used in our experiments the recent dataset of Guo et al. (2010), consisting of 1000 texts (7985 sentences and 225785 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 48% 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 (κ=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.
- Class. 60 verb classes appearing in biomedical journal articles.
- Part-of-Speech. The root tag of each verb in the corpus.
- Grammatical Relation – cir, Subject (ncsubj), direct object (dobj), indirect object (indobj) and second object (ob2) relations in the corpus. e.g. (ncsubj observed | difference | obj).
- Subj and Obj. The subjects and objects appearing with any verbs in the corpus (extracted from cir).
- 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. output was used for Word, Bi-gram and Verb features, and the output of sequence labeling for Word, Bi-gram and Verb classes).

The next step is to assign sentences in abstracts to the corresponding categories. The most common approach is to label sentences as positive if a certain expression appears and negative otherwise. The agreement (κ=0.85) is relatively high according to Cohen (1960).

2.2.2 Machine learning methods The next step is to assign sentences in abstracts to zones categories using machine learning. Support vector machines (SVM) and conditional random fields (CRF) have proven 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:

\[
\text{Maximize} \quad \frac{1}{2} \|w\|^2 \quad \text{subject to} \quad y(w \cdot x - b) \geq 1,
\]

where \(x\) is data, \(y\) is label, \(w\) is a normal vector to the hyperplane and \(b\) is the margin. We used Weka software (Hall et al., 2009) employing the SVM algorithm (Platt, 1999) 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:

\[
p(y|x) = \frac{1}{Z(x)} \exp \left( \sum_{i,j} F_j(y|x) \right),
\]

where \(F_j(x, y)\) is a real-valued feature function of the states and the observations; \(\phi\) 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 (ASSVM) 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 learning. 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 ASSVM 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:

\[
\text{Maximize} \quad \frac{1}{2} \|w\|^2
\]

Subject to

\[
y_i^0 \leq w \cdot x_i - b \leq y_i^0,
\]

where \(\pi\) 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 ccpp algorithm (Collobert et al., 2006) for SVM experiments.

Semi-supervised CRF (SSCRF) can be implemented by entropy regularization (Jiao et al., 2006), which extends the objective function on Labeled data \(\sum_y \log(p(y|x,\theta))\) with an additional term \(\sum_y \log(p(y|x,\theta))\) 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 mt. results in terms of accuracy, precision, recall and F-score. We also evaluated 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 5% being manually labeled). The results were then averaged. As randomly selected labeled data were used for SVM, CRF, TSV and SCRF, 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 CRA 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>
<thead>
<tr>
<th>Question</th>
<th>Zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 Do the authors discuss previous or related research on the topic? y/n</td>
<td>BKG, REL</td>
</tr>
<tr>
<td>Q2 Do the authors describe the aim of the research? y/n</td>
<td>OBJ, MTH</td>
</tr>
<tr>
<td>Q3 What is the main type of study the abstract focuses on?</td>
<td>ANIMAL STUDY/HUMAN STUDY/VITRO STUDY</td>
</tr>
<tr>
<td>Q4 Is exposure length mentioned? y/n</td>
<td>MTH</td>
</tr>
<tr>
<td>Q5 Is dose mentioned? y/n</td>
<td>MTH</td>
</tr>
<tr>
<td>Q6 Is group size mentioned? y/n</td>
<td>MTH</td>
</tr>
<tr>
<td>Q7 How many endpoints are mentioned? 0/1/more</td>
<td>RES</td>
</tr>
<tr>
<td>Q8 Are the results positive? y/n/unclear</td>
<td>RES</td>
</tr>
<tr>
<td>Q9 Is the outcome of the study expected/neutral/unexpected?</td>
<td>CON</td>
</tr>
<tr>
<td>Q10 Do the authors mention a need for future research on the topic? y/n</td>
<td>PUT</td>
</tr>
</tbody>
</table>

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, diethylthiurasine, 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. ASSVM performs nearly as well as the best performing supervised method, SVM, with an accuracy of 80% and macro F-score of 0.74.

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.77</td>
<td>0.73</td>
</tr>
<tr>
<td>CRF</td>
<td>0.70</td>
<td>0.64</td>
</tr>
<tr>
<td>TSV</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td>SCRF</td>
<td>0.65</td>
<td>0.48</td>
</tr>
<tr>
<td>ASSVM</td>
<td>0.80</td>
<td>0.75</td>
</tr>
<tr>
<td>SVM</td>
<td>0.77</td>
<td>0.74</td>
</tr>
<tr>
<td>CRF</td>
<td>0.71</td>
<td>0.48</td>
</tr>
<tr>
<td>TSV</td>
<td>0.71</td>
<td>0.65</td>
</tr>
<tr>
<td>SCRF</td>
<td>0.71</td>
<td>0.48</td>
</tr>
<tr>
<td>ASSVM</td>
<td>0.81</td>
<td>0.76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.77</td>
<td>0.73</td>
</tr>
<tr>
<td>CRF</td>
<td>0.70</td>
<td>0.64</td>
</tr>
<tr>
<td>TSV</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td>ASSVM</td>
<td>0.80</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Background Objective Method Result Conclusion Related work Future work

Effects of letrozol granulocyte (LGE) on hepatocarcinogenesis were examined in male Fischer 344 rats, administered diethylthurasine (DEN) at three weekly intraperitoneal doses of 100 mg/kg body weight and partially hepatectomized at the end of week 5. LGE was given at dietary concentrations of 0, 0.2, 0.6 or 1.8% from the end of week 4 for 10 weeks. All rats were sacrificed at the end of week 11. LGE reduced the number of putatively preneoplastic, glutathione S-transferase placental form-positive lesions and the level of oxidative hepatoocyte nuclear DNA injury, as assessed in terms of 8-hydroxydeoxyguanosine production. In contrast, LGE did not affect the size of the preneoplastic lesions, hepatoocyte proliferative activity, activities of phase II enzymes or hepatoocyte extranuclear oxidative injury. These results suggest inhibitory effects of LGE on the early phase hepatocarcinogenesis in rats after initiation with DEN.

END

According to the authors, is the outcome of the study expected, unexpected, or neither/neutral?

- Expected
- Unexpected
- Neither/Neutral

Fig. 1. An example of the questionnaire.
well, with an accuracy of 80% and F-score of 0.75. Both methods outperform supervised SVM with a statistically significant difference ($P < 0.001$). T SVM is the lowest performing SVM-based method: its performance is lower than that of the supervised SVM. Yet, it outperforms both CRF-based methods. S S C R F performs slightly better than CRF with 1% higher accuracy and 0.01 higher F-score. Only two methods (ASSVM, T SVM) find six out of the seven possible zone categories. Other methods find five categories. The I–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 SVM (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 T SVM tend to perform quite similarly with each other when >10% of the labeled data are used, but when less data are available, T SVM performs better. Looking at the CRF-based methods, S S C R F 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.

### 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)/(iii) compared with (i). TIME stands for the sample mean, and $\text{SAVE}$ 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 5. Leave-one-feature-out results for ASSVM with 10% of labeled data

<table>
<thead>
<tr>
<th>Feature</th>
<th>ACC</th>
<th>MF</th>
<th>BKG</th>
<th>OBJ</th>
<th>METH</th>
<th>RES</th>
<th>CON</th>
<th>REL</th>
<th>PUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>0.73</td>
<td>0.67</td>
<td>0.67</td>
<td>0.55</td>
<td>0.62</td>
<td>0.85</td>
<td>0.65</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Word</td>
<td>0.80</td>
<td>0.78</td>
<td>0.78</td>
<td>0.70</td>
<td>0.74</td>
<td>0.85</td>
<td>0.72</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Bigram</td>
<td>0.81</td>
<td>0.75</td>
<td>0.83</td>
<td>0.57</td>
<td>0.71</td>
<td>0.87</td>
<td>0.78</td>
<td>0.33</td>
<td>–</td>
</tr>
<tr>
<td>Verb</td>
<td>0.81</td>
<td>0.79</td>
<td>0.84</td>
<td>0.77</td>
<td>0.73</td>
<td>0.87</td>
<td>0.75</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>VC</td>
<td>0.79</td>
<td>0.75</td>
<td>0.86</td>
<td>0.62</td>
<td>0.72</td>
<td>0.84</td>
<td>0.70</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>POS</td>
<td>0.74</td>
<td>0.70</td>
<td>0.66</td>
<td>0.65</td>
<td>0.66</td>
<td>0.82</td>
<td>0.73</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>GR</td>
<td>0.79</td>
<td>0.75</td>
<td>0.83</td>
<td>0.67</td>
<td>0.69</td>
<td>0.84</td>
<td>0.72</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Subj</td>
<td>0.80</td>
<td>0.76</td>
<td>0.87</td>
<td>0.65</td>
<td>0.73</td>
<td>0.85</td>
<td>0.72</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Obj</td>
<td>0.80</td>
<td>0.78</td>
<td>0.84</td>
<td>0.75</td>
<td>0.70</td>
<td>0.85</td>
<td>0.75</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Voice</td>
<td>0.78</td>
<td>0.75</td>
<td>0.88</td>
<td>0.70</td>
<td>0.71</td>
<td>0.83</td>
<td>0.62</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$\Phi$</td>
<td>0.81</td>
<td>0.76</td>
<td>0.86</td>
<td>0.56</td>
<td>0.76</td>
<td>0.88</td>
<td>0.76</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Φ: Employing all the features.
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 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 21 14.7 25 8.6 19 5.2 0 5.5 9 12.8 5 7.4 4 9.5 10 5 39 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 12.7 3.8
   (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 10.5 5 5.2 6 4.9 15 11.8 5 6 29 9.9 22 4.0 0 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

<table>
<thead>
<tr>
<th>Q1</th>
<th>Q2</th>
<th>Q3a</th>
<th>Q3b</th>
<th>Q3c</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>v. (ii)</td>
<td>0.035</td>
<td>0.877</td>
<td>0.063</td>
<td>0.975</td>
<td>0.397</td>
<td>0.421</td>
<td>0.032</td>
<td>0.236</td>
<td>0.015</td>
<td>0.285</td>
<td>-0.001*</td>
</tr>
<tr>
<td></td>
<td>v. (ii)</td>
<td>0.081</td>
<td>0.924</td>
<td>0.405</td>
<td>0.162</td>
<td>0.221</td>
<td>0.075</td>
<td>0.154</td>
<td>0.550</td>
<td>0.341</td>
<td>0.431</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>v. (i)</td>
<td>0.200</td>
<td>0.135</td>
<td>0.005</td>
<td>0.081</td>
<td>0.864</td>
<td>0.248</td>
<td>0.872</td>
<td>0.322</td>
<td>0.310</td>
<td>0.781</td>
<td>0.159</td>
</tr>
<tr>
<td></td>
<td>v. (ii)</td>
<td>0.570</td>
<td>0.633</td>
<td>0.235</td>
<td>0.140</td>
<td>0.052</td>
<td>0.242</td>
<td>0.326</td>
<td>0.530</td>
<td>0.321</td>
<td>0.851</td>
<td>0.041</td>
</tr>
</tbody>
</table>

B (i) v. (ii) 0.122 | 0.930 | 0.180 | 0.666 | 0.986 | 0.149 | 0.001 | 0.006 | 0.020 | 0.781 | 0.005 |
|     | v. (iii) | 0.260 | 0.321 | 0.568 | 0.318 | 0.380 | 0.070 | 0.552 | 0.018 | 0.005 | 0.786 | 0.027 |
|     | v. (iv) | 0.624 | 0.008 | 0.000 | 0.106 | 0.385 | 0.027 | 0.000 | 0.015 | 0.009 | 0.077 | -0.001* |
|     | v. (ii) | 0.062 | 0.888 | 0.050 | 0.729 | 0.553 | 0.693 | 0.477 | 0.340 | 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 statistically significant.

*After rounding, this value is 0.00

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

<table>
<thead>
<tr>
<th>Q1</th>
<th>Q2</th>
<th>Q3a</th>
<th>Q3b</th>
<th>Q3c</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>0.63</td>
<td>0.51</td>
<td>0.78</td>
<td>0.90</td>
<td>0.80</td>
<td>0.86</td>
<td>0.90</td>
<td>0.92</td>
<td>0.86</td>
<td>0.76</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>ii</td>
<td>0.70</td>
<td>0.66</td>
<td>0.92</td>
<td>0.96</td>
<td>0.92</td>
<td>0.82</td>
<td>0.92</td>
<td>0.90</td>
<td>0.72</td>
<td>0.76</td>
<td>0.88</td>
<td>0.84</td>
</tr>
<tr>
<td>iii</td>
<td>0.82</td>
<td>0.68</td>
<td>0.96</td>
<td>0.98</td>
<td>0.90</td>
<td>0.86</td>
<td>0.86</td>
<td>0.90</td>
<td>0.74</td>
<td>0.86</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td>iv</td>
<td>0.66</td>
<td>0.54</td>
<td>0.90</td>
<td>0.92</td>
<td>0.92</td>
<td>0.74</td>
<td>0.68</td>
<td>0.82</td>
<td>0.82</td>
<td>0.84</td>
<td>0.90</td>
<td>0.82</td>
</tr>
</tbody>
</table>

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

Table 8 shows the joint probability of users’ agreement on the answers. Annotations (ii), (iii) and (iv) do not affect the users’ agreement at 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, ASSVM and SSCRF outperformed their corresponding supervised methods: SVM and CRF. ASSVM/ASSVM 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). 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 when using just 10% of the abstracts.

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...
Semi-Supervised Learning for Computational Linguistics
REFERENCES
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 (McCullum and Ngiam, 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). SVM-based active learning might be
a good option too.

The work presented in this article has focused on the AE 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 category 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 cia, subsequent steps
of cia focus on information in full articles. These more challenging
steps may benefit from AE (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
McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 153-157.