A Weakly-supervised Approach to Argumentative Zoning of Scientific Documents

Yufan Guo, Anna Korhonen, Thierry Poibeau

A Weakly-supervised Approach to Argumentative Zoning of Scientific Documents

Yufan Guo
Computer Laboratory
University of Cambridge, UK
yg244@cam.ac.uk

Anna Korhonen
Computer Laboratory
University of Cambridge, UK
alk23@cam.ac.uk

Thierry Poibeau
LaTTiCe, UMR8094
CNRS & ENS, France
thierry.poibeau@ens.fr

Abstract

Argumentative Zoning (AZ) – analysis of the argumentative structure of a scientific paper – has proved useful for a number of information access tasks. Current approaches to AZ rely on supervised machine learning (ML). Requiring large amounts of annotated data, these approaches are expensive to develop and port to different domains and tasks. A potential solution to this problem is to use weakly-supervised ML instead. We investigate the performance of four weakly-supervised classifiers on scientific abstract data annotated for multiple AZ classes. Our best classifier based on the combination of active learning and self-training outperforms our best supervised classifier, yielding a high accuracy of 81% when using just 10% of the labeled data. This result suggests that weakly-supervised learning could be employed to improve the practical applicability and portability of AZ across different information access tasks.

1 Introduction

Many practical tasks require accessing specific types of information in scientific literature. For example, a reader of scientific literature may be looking for information about the objective of the study in question, the methods used in the study, the results obtained, or the conclusions drawn by authors. Similarly, many Natural Language Processing (NLP) tasks focus on the extraction of specific types of information in documents only.

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

The best of current approaches have yielded promising results and proved useful for information retrieval, information extraction and summarization tasks (Teufel and Moens, 2002; Mizuta et al., 2006; Tbahriti et al., 2006; Ruch et al., 2007). 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 scientific domains and tasks.

A potential solution to this bottleneck is to develop techniques based on weakly-supervised ML. Relying on a small amount of labeled data and a large pool of unlabeled data, weakly-supervised techniques (e.g. semi-supervision, active learning, co/tri-training, self-training) aim to keep the advantages of fully supervised approaches. They have been applied to a wide range of NLP 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.

To the best of our knowledge, such techniques

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 273–283,
Edinburgh, Scotland, UK, July 27–31, 2011.

have not yet been applied to the analysis of information structure of scientific documents by aforementioned approaches. Recent experiments have demonstrated the usefulness of weakly-supervised learning for classifying discourse relations in scientific texts, e.g. (Hernault et al., 2011). However, focusing on local (rather than global) structure of documents and being much more fine-grained in nature, this related task differs from ours considerably.

In this paper, we investigate the potential of weakly-supervised learning for Argumentative Zoning (AZ) of scientific abstracts. AZ is an approach to information structure which provides an analysis of the rhetorical progression of the scientific argument in a 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). Although the basic scheme is said to be discipline-independent (Teufel et al., 2009), its application to different domains has resulted in various modifications and laborious annotation exercises. This suggests that a weakly-supervised approach would be more practical than a fully supervised one for the real-world application of AZ.

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 on the AZ task: 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).

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. When using just one third of the labeled data, it performs equally well as a fully supervised SVM which uses 100% of the labeled data. Our investigation suggests that weakly-supervised learning could be employed to improve the practical applicability and portability of AZ to different information access tasks.

2 Data

We used in our experiments the recent dataset of (Guo et al., 2010). Guo et al. (2010) provide a corpus of 1000 biomedical abstracts (consisting of 7985 sentences and 225785 words) annotated according to three schemes of information structure – those based on section names (Hirohata et al., 2008), AZ (Mizuta et al., 2006) and Core Scientific Concepts (CoreSC) (Liakata et al., 2010). We focus here on AZ only, because it subsumes all the categories of the simple section name -based scheme, and according to the inter-annotator agreement and ML experiments reported by Guo et al. (2010) it performs better on this data than the fairly fine-grained CoreSC scheme.

AZ is a scheme which provides an analysis of the rhetorical progression of the scientific argument, following the knowledge claims made by authors. (Teufel and Moens, 2002) introduced AZ and applied it first to computational linguistics papers. (Hachey and Grover, 2006) applied the scheme later to legal texts and (Mizuta et al., 2006) modified it for biology papers. More recently, (Teufel et al., 2009) introduced a refined version of AZ and applied it to chemistry papers.

The biomedical 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 corpus. These are shown and explained in Table 1. For example, the Method zone (METH) is for sentences which describe *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*.

An example of a biomedical abstract annotated according to AZ is shown in Figure 1, with different zones highlighted in different colors. For example, the RES zone is highlighted in lemon green.

Table 2 shows the distribution of sentences per scheme category in the corpus: Results (RES) is by far the most frequent zone (accounting for 40% of the corpus), while Background (BKG), Objective (OBJ), Method (METH) and Conclusion (CON) cover

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

Category	Abbr.	Definition
Background	BKG	The circumstances pertaining to the current work, situation, or its causes, history, etc.
Objective	OBJ	A thing aimed at or sought, a target or goal
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
Result	RES	The effect, consequence, issue or outcome of an experiment; the quantity, formula, etc. obtained by calculation
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
Related work	REL	A comparison between the current work and the related work
Future work	FUT	The work that needs to be done in the future

Figure 1: An example of an annotated abstract

Butadiene (BD) metabolism shows gender, species and concentration dependency, making the extrapolation of animal results to humans complex. BD is metabolized mainly by cytochrome P450 2E1 to three epoxides, 1,2-epoxy-3-butene (EB), 1,2;3,4-diepoxybutane (DEB) and 1,2-epoxy-butanediol (EB-diol). For accurate risk assessment it is important to elucidate species differences in the internal formation of the individual epoxides in order to assign the relative risks associated with their different mutagenic potencies. Analysis of N-terminal globin adducts is a common approach for monitoring the internal formation of BD ***Background***es. Our long term strategy is to develop an LC-MS/MS method for simultaneous detection of all three BD hemoglobin adducts. This approach is modeled after the recently reported immunoaffinity LC-MS/MS method for the cyclic N,N-(2,3-dihydroxy-1,4-butadyil)-valine (pyr-Val, derived from DEB). We report herein the analysis of the EB-derived 2-hydroxyl-3-butenyl-valine pep ***Objective***l). The procedure utilizes trypsin hydrolysis of globin and immunoaffinity (IA) purification of alkylated heptapeptides. Quantitation is based on LC-MS/MS monitoring of the transition from the singly charged molecular ion of HB-Val (1-7) to the a(1) fragment. Human HB-Val (1-11) was synthesized and used for antibody production. As internal standard, the labeled rat-[(13)C(5)(15)N]-Val (1-11) was prepared through direct alkylation of the corresponding peptide with EB. Standards were characterized and quantified by LC-MS/MS and LC-UV. The method was validated with different amounts of human HB-Val standard. The recovery was >75% and coefficient of variation <25%. The LOQ was set to 100 fmol/injection. For a proof of principal experiment, globin samples from male and female rats exposed to 1000 ppm BD for 90 days w***Method***ed. The amounts of HB-Val present were 268.2+/-56 and 350+/-70 pmol/g (mean+/-S.D.) for males and females, respectively. No HB-Val was detected i***Result***ls. These data are much lower compared to previously reported values measu***Related work***MS. The difference may be due higher specificity of the LC-MS/MS method to the N-terminal peptide from the alpha-chain versus derivatization of both alpha- and beta-chain by Edman degradation, and possible instability of HB-Val adducts during long term storage (about 10 years) betw***Conclusion***es. These differences will be resolved by examining recently collected samples, using the same internal standard for parallel analysis by GC-MS/M***Future work***MS. Based on our experience with pyr-Val adduct assay we anticipate that this assay will be suitable for evaluation of HB-Val in multiple species.

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

	BKG	OBJ	METH	RES	CON	REL	FUT
Word	36828	23493	41544	89538	30752	2456	1174
Sentence	1429	674	1473	3185	1082	95	47
Sentence	18%	8%	18%	40%	14%	1%	1%

8-18% of the corpus each. Two categories are very low in frequency, only covering 1% of the corpus each: Related work (REL) and Future work (FUT).

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

3 Automatic identification of AZ

3.1 Features and feature extraction

Guo et al. (2010) used a variety of features in their fully supervised ML experiments on different schemes of information structure. Since their feature types cover the best performing feature types in earlier works e.g. (Teufel and Moens, 2002; Lin et al., 2006; Mullen et al., 2005; Hirohata et al., 2008; Merity et al., 2009) we re-implemented and used them in our experiment[1]. However, being aware of the fact that some of these features may not be optimal for weakly-supervised learning (i.e. when learning from smaller data), we evaluate their performance and suitability for the task later in section 4.3.

- **Location**. Zones tend to appear in typical positions in abstracts. Each abstract was there-

[1]The only exception is the history feature which was left out because it cannot be applied to all of our methods

fore divided into ten parts (1-10, measured by the number of words), and the location was defined by the parts where the sentence begins and ends.

- **Word**. All the words in the corpus.
- **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)*. The value of this feature equals 1 if it occurs in a particular sentence (and 0 if not).
- **Subj and Obj**. The subjects and objects appearing with any verbs in the corpus (extracted from above 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 the boundaries of sentences and to separate punctuation from adjacent words e.g. in complex biomedical terms such as *2-amino-3,8-diethylimidazo[4,5-f]quinoxaline*. The C&C tools (Curran et al., 2007) trained on biomedical literature were employed for POS tagging, lemmatization and parsing. The lemma output was used for creating Word, Bi-gram and Verb features. The GR output was used for creating the GR, Subj, Obj and Voice features. The "obj" marker in a subject relation indicates passive voice (e.g. *(ncsubj observed_14 difference_5 obj)*). The verb classes were acquired automatically from the corpus using the unsupervised spectral clustering method of (Sun and Korhonen, 2009). To control the number of features we lemmatized the lexical items for all the features, and removed the words and GRs with fewer than 2 occurrences and bi-grams with fewer than 5 occurrences.

3.2 Machine learning methods

Support Vector Machines (SVM) and Conditional Random Fields (CRF) have proved the best performing fully supervised methods in most recent works on information structure, e.g. (Teufel and Moens, 2002; Mullen et al., 2005; Hirohata et al., 2008; Guo et al., 2010). 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.

3.2.1 Supervised methods

SVM constructs hyperplanes in a multidimensional space to separate data points of different classes. Good separation is achieved by the hyperplane that has the largest distance from the nearest data points of any class. The hyperplane has the form $w \cdot x - b = 0$, where w is its normal vector. We want to maximize the distance from the hyperplane to the data points, or the distance between two parallel hyperplanes each of which separates the data. The parallel hyperplanes can be written as: $w \cdot x - b = 1$ and $w \cdot x - b = -1$, and the distance between them is $\frac{2}{|w|}$. The problem reduces to:

Minimize $|w|$ (in w, b)

Subject to

$w \cdot x - b \geq 1$ for x of one class,

$w \cdot x - b \leq -1$ for x of the other,

which can be solved by using the SMO algorithm (Platt, 1999b). We used Weka software (Hall et al., 2009) (employing its linear kernel) for SVM experiments.

CRF is an undirected graphical model which 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:

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

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. The θ parameters can be learned using the L-BFGS algorithm (Nocedal, 1980). We used Mallet software (McCallum, 2002) for CRF experiments.

3.2.2 Weakly-supervised methods

Active SVM (ASVM) starts with a small amount of labeled data, and iteratively chooses a proportion of

unlabeled data for which SVM has less confidence to be labeled (the labels can be restored from the original corpus) and used in the next round of learning, i.e. active learning. Query strategies based on the structure of SVM are frequently employed (Tong and Koller, 2001; Novak et al., 2006). For example, it is often assumed that the data points close to the separating hyperplane are those that the SVM is uncertain about. Unlike these methods, our learning algorithm compares the posterior probabilities of the best estimate given each unlabeled instance, and queries those with the lowest probabilities for the next round of learning. The probabilities can be obtained by fitting a Sigmoid after the standard SVM (Platt, 1999a), and combined using a pairwise coupling algorithm (Hastie and Tibshirani, 1998) in the multi-class case. We used the SVM linear kernel in Weka for classification, and the -M flag in Weka for calculating the posterior probabilities.

Active SVM with self-training (ASSVM) is an extension of ASVM where each round of training has two steps: (i) training on the labeled, and testing on the unlabeled data, and querying; (ii) training on both labeled and unlabeled/machine-labeled data by using the estimates from step (i). The idea of ASSVM is to make the best use of the labeled data, and to make the most use of the unlabeled data.

Transductive SVM (TSVM) is an extension of SVM which takes advantage of both labeled and unlabeled data (Vapnik, 1998). Similar to SVM, the problem is defined as:

Minimize $|w|$ (in $w, b, y^{(u)}$)
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 problem can be solved by using the CCCP algorithm (Collobert et al., 2006). We used UniverSVM software (Sinz, 2011) for TSVM experiments.

Semi-supervised CRF (SSCRF) can be implemented with entropy regularization (ER). It 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 (Jiao et al., 2006; Mann and Mccallum, 2007). We used Mallet software (McCallum, 2002) for SSCRF experiments.

4 Experimental evaluation

4.1 Evaluation methods

We evaluated the ML results in terms of accuracy, precision, recall, and F-measure against manual AZ annotations in the corpus:

$$acc = \frac{no.\ of\ correctly\ classified\ sentences}{total\ no.\ of sentences\ in\ the\ corpus}$$

$$p = \frac{no.\ of\ sentences\ correctly\ identified\ as\ Class_i}{total\ no.\ of\ sentences\ identified\ as\ Class_i}$$

$$r = \frac{no.\ of\ sentences\ correctly\ identified\ as\ Class_i}{total\ no.\ of\ sentences\ in\ Class_i}$$

$$f = \frac{2*p*r}{p+r}$$

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. More specifically, the data was randomly assigned to ten folds of roughly the same size. Each fold was used once as test data and the remaining nine folds as training data. The results were then averaged.

Following (Dietterich, 1998), we used McNemar's test (McNemar, 1947) to measure the statistical significance between the results of different ML methods. The chosen significance level was .05.

4.2 Results

Table 3 shows the results for the four weakly-supervised and two supervised methods when 10% of the training data (i.e. ∼700 sentences) has been labeled. We can see that ASSVM is the best performing method with an accuracy of 81% and the macro

Table 3: Results when using 10% of the labeled data

	Acc.	F-score							
		MF	BKG	OBJ	METH	RES	CON	REL	FUT
SVM	.77	.74	.84	.68	.71	.82	.64	-	-
CRF	.70	.65	.75	.46	.48	.78	.76	-	-
ASVM	.80	.75	.88	.56	.68	.87	.78	.33	
ASSVM	.81	.76	.86	.56	.76	.88	.76	-	-
TSVM	.76	.73	.84	.61	.71	.79	.71	-	-
SSCRF	.73	.67	.76	.48	.52	.81	.78	-	-

MF: Macro F-score of the five high frequency categories: BKG, OBJ, METH, RES, CON.

Figure 2: Learning curve for different methods when using 0-100% of the labeled data

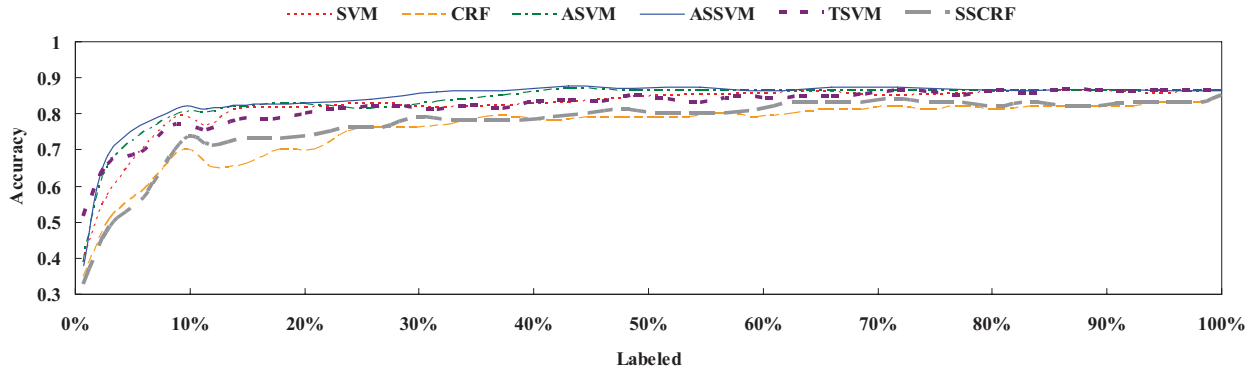
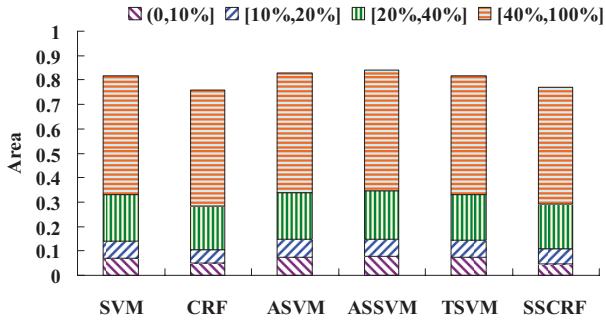


Figure 3: Area under learning curves at different intervals



F-score of .76 (the macro F-score is calculated for the 5 scheme categories which are found by all the methods). ASVM performs nearly as well, with an accuracy of 80% and F-score of .75. Both methods outperform supervised SVM with a statistically significant difference ($p < .001$).

TSVM is the lowest performing SVM-based method. Yielding an accuracy of 76% and F-score of .73 its performance is lower than that of the supervised SVM. However, it does outperform both CRF-based methods. SSCRF performs better than CRF with 3% higher accuracy and .02 higher F-score. The difference in accuracy is statistically significant ($p < .001$).

Only one method (ASVM) identifies six out of the seven possible categories. Other methods identify five categories. The 1-2 missing categories are very low in frequency (accounting for 1% of the corpus data each, see table 2). Looking at the results for other categories, they seem to reflect the amount of corpus data available for each category (Table 2), with RES (Results) being the highest and OBJ (Objective) the lowest performing category with most methods. Interestingly, the only method that performs relatively well on OBJ is the supervised SVM.

The best method ASSVM outperforms other methods most clearly on METH (Method) category. Although METH is a high frequency category (accounting for 18% of the corpus data) other methods tend to confuse it with OBJ, presumably because a single sentence may contain elements of both (e.g. scientists may describe some of their method when describing the objective of the study).

Figure 2 shows the learning curve of different methods (in terms of accuracy) when the percentage of the labeled data (in the training set) ranges from 0 to 100%. ASSVM outperforms other methods, reaching its best performance of 88% accuracy when using ∼40% of the labeled data. Indeed when using 33% of the labeled data, it performs already equally well as fully-supervised SVM using 100% of the labeled data. The advantage of ASSVM over ASVM (the second best method) is clear especially when 20-40% of the labeled data is used. SVM and TSVM tend to perform quite similarly with each other when more than 25% of the labeled data is used, but when less data is available, SVM performs better. Looking at the CRF-based methods, SSCRF outperforms CRF in particular when 10-25% of the labeled data is used. However, neither of them reaches the performance level of SVM-based methods.

Figure 3 shows the area under the learning curves (by the trapezoidal rule) at different intervals, which gives a reasonable approximation to the overall performance of different methods. The area under ASSVM is the largest at each of the four intervals, with a value of .08 at (0,10%], .07 at [10%,20%],

.20 at [20%, 40%] and .50 at [40%,100%]. The difference between supervised and weakly-supervised methods is more significant at (0, 20%] than at [20%,100%].

4.3 Further analysis of the features

As explained in section 3.1, we employed in our experiments a collection of features which had performed well in previous supervised AZ experiments. We conducted further analysis to investigate which of these features are the most (and the least) 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 are shown in Table 4.

Table 4: Leaving one feature out results for ASSVM when using 10% of the labeled data

	Acc.	F-score							
		MF	BKG	OBJ	METH	RES	CON	REL	FUT
Location	.73	.67	.67	.55	.62	.85	.65	-	-
Word	.80	.78	.87	.70	.74	.85	.72	-	-
Bigram	.81	.75	.83	.57	.71	.87	.78	.33	-
Verb	.81	.79	.84	.77	.73	.87	.75	-	-
VC	.79	.75	.86	.62	.72	.84	.70	-	-
POS	.74	.70	.66	.65	.66	.82	.73	-	-
GR	.79	.75	.83	.67	.69	.84	.72	-	-
Subj	.80	.76	.87	.65	.73	.85	.72	-	-
Obj	.80	.78	.84	.75	.70	.85	.75	-	-
Voice	.78	.75	.88	.70	.71	.83	.62	-	-
Φ	.81	.76	.86	.56	.76	.88	.76	-	-

MF: Macro F-score of the five high frequency categories: BKG, OBJ, METH, RES, CON.
Φ: Employing all the features.

We can see that the Location feature is by far the most useful feature for ASSVM. The performance drops 8% in accuracy and .09 in F-score in the absence of this feature. Location is particularly important for BKG (which nearly always appears in the same location: in the beginning of an abstract) and is highly useful for METH and CON as well. Removing POS has almost equally strong effect, in particular on BKG and METH, suggesting that verb tense is particularly useful for distinguishing these categories.

Also Voice, Verb class and GR contribute to general performance, especially to accuracy. Voice is particularly important for CON, which differs from other categories in the sense that it is marked by frequent usage of active voice. Verb class is helpful for METH, RES and CON while GR is helpful for all high frequency categories.

Among the least helpful features are those which suffer from sparse data problems, including e.g. Word, Bi-gram, and Verb. They perform particularly badly when applied to low frequency zones. However, this is not the case when using fully-supervised methods (i.e. 100% of the labeled data), suggesting that a good performance in fully supervised experiments does not necessarily translate into a good performance in weakly-supervised experiments, and that careful feature analysis and selection is important when aiming to optimize the performance when learning from sparse data.

5 Discussion

In our experiments, the majority of weakly-supervised methods outperformed their corresponding supervised methods when using just 10% of the labeled data. The SVM-based methods performed better than the CRF-based ones (regardless of whether they were weakly or fully supervised). Guo et al. (2010) made a similar discovery when comparing fully supervised versions of SVM and CRF.

Our best performing weakly-supervised methods were those based on active learning. Making a good use of both labeled and unlabeled data, active learning combined with self-training (ASSVM) proved to be the most useful method. Given 10% of the labeled data, ASSVM obtained an accuracy of 81% and F-score of .76, outperforming the best supervised method SVM with a statistically significant difference. It reached its top performance (88% accuracy) when using 40% of the labeled data, and performed equally well as fully supervised SVM (i.e. 100% of the labeled data) when using just one third of the labeled data.

This result is in line with the results of many other text classification works where active learning (alone or in combination with other techniques such as self-training) has proved similarly useful, e.g. (Lewis and Gale, 1994; Tong and Koller, 2002; Brinker, 2006; Novak et al., 2006; Esuli and Sebastiani, 2009; Yang et al., 2009).

While active learning iteratively explores the unknown aspects of the unlabeled data, semi-supervised learning attempts to make the best use

of what it already knows about the data. In our experiments, semi-supervised methods (TSVM and SSCRF) did not perform equally well as active learning – TSVM even produced a lower accuracy than SVM with the same amount of labeled data – although these methods have gained success in related works.

We therefore looked into related works using TSVM, e.g. (Chapelle and Zien, 2005), and discovered that our dataset is much higher in dimensionality than those employed in many other works. High dimensional data is more sensitive, and therefore fine-tuning with unlabeled data may cause a big deviation. We also looked into related works using SSCRF, in particular the work of (Jiao et al., 2006) who used the same SSCRF as the one we used in our experiments. Jiao et al. (2006) employed a much larger data set than we did – one including 5448 labeled instances (in 3 classes) and 5210-25145 unlabeled instances. Given more labeled and unlabeled data per class we might be able to obtain better performance using SSCRF also on our task. However, given the high cost of obtaining labeled data methods not needing it are preferable.

6 Conclusions and future work

Our experiments show that weakly-supervised learning can be used to identify AZ in scientific documents with good accuracy when only a limited amount of labeled data is available. This is helpful thinking of the real-world application and porting of the approach to different tasks and domains. To the best of our knowledge, no previous work has been done on weakly-supervised learning of information structure according to schemes of the type we have focused on (Teufel and Moens, 2002; Mizuta et al., 2006; Lin et al., 2006; Hirohata et al., 2008; Shatkay et al., 2008; Liakata et al., 2010).

Recently, some work has been done on the related task of classification of discourse relations in scientific texts: (Hernault et al., 2011) used structural learning (Ando and Zhang, 2005) for this task. They obtained 30-60% accuracy on the RST Discourse Treebank (including 41 relation types) when using 100-10000 labeled and 100000 unlabeled instances. The accuracy was 20-60% when using the labeled data only. However, although related, the task of discourse relation classification differs substantially from our task in that it focuses on local discourse relations while our task focuses on the global structure of the scientific document.

In the future, we plan to improve and extend this work in several directions. First, the approach to active learning could be improved in various ways. The query strategy we employed (uncertainty sampling) is a relatively straightforward method which only considers the best estimate for each unlabeled instance, disregarding other estimates that may contain useful information. In the future, we plan to experiment with more sophisticated strategies, e.g. the margin sampling algorithm by (Scheffer et al., 2001) and the query-by-committee (QBC) algorithm by (Seung et al., 1992). In addition, there are algorithms designed for reducing the redundancy in queries which may be worth investigating (Hoi et al., 2006).

Also, (Hoi et al., 2006) shows that Logistic Regression (LR) outperforms SVM when used with active learning, yielding higher F-score on the Reuters-21578 data set (binary classification, 10,788 documents in total, 100 of them labeled). It would be interesting to explore whether supervised methods other than SVM are optimal for active learning when applied to our task.

Secondly, we plan to investigate other semi-supervised methods, for example, the Expectation-Maximization (EM) algorithm. (Lanquillon, 2000) has shown that EM SVM performs better than supervised and transductive SVM on a text classification task when applied to the dataset of 20 Newsgroups (20 classes, 4000 documents for testing, 10000 unlabeled ones), yielding up to $\sim$10% higher accuracy when 200-5000 labeled documents are used for training.

In addition, other combinations of weakly-supervised methods might be worth looking into, such as EM+active learning (McCallum and Nigam, 1998) and co-training+EM+active learning (Muslea et al., 2002), which have proved promising in related text classification works.

Besides looking for optimal ML strategies, we plan to look for optimal features for the task. Our feature analysis showed that not all the features which had proved promising in fully supervised experiments were equally promising when applied to weakly-supervised learning from smaller data. We

plan to look into ways of reducing the sparse data problem in features, e.g. by classifying not only verbs but also other word classes into semantically-motivated categories.

One the key motivations for developing a weakly-supervised approach is to facilitate easy porting of schemes such as AZ to new tasks and domains. Recent research shows that active learning in a target domain can leverage information from a different but related (source) domain (Rai et al., 2010). Making use of existing annotated datasets in biology, chemistry, computational linguistics and law (Teufel and Moens, 2002; Mizuta et al., 2006; Hachey and Grover, 2006; Teufel et al., 2009) we will explore optimal ways of combining weakly-supervised learning with domain-adaptation.

The work presented in this paper has focused on the abstracts annotated according to the AZ scheme. In the future, we plan to investigate the usefulness of weakly-supervised learning for identifying other schemes of information structure, e.g. (Lin et al., 2006; Hirohata et al., 2008; Shatkay et al., 2008; Liakata et al., 2010), and not only in scientific abstracts but also in full journal papers which typically exemplify a larger set of scheme categories.

Finally, an important avenue of future research is to evaluate the usefulness of weakly-supervised identification of information structure for NLP tasks such as summarization and information extraction (Tbahriti et al., 2006; Ruch et al., 2007), and for practical tasks such as manual review of scientific papers for research purposes (Guo et al., 2010).

Acknowledgments

The work reported in this paper was funded by the Royal Society (UK). YG was funded by the Cambridge International Scholarship.

References

Steven Abney. 2008. *Semi-supervised learning for computational linguistics*. Chapman & Hall / CRC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853.

Klaus Brinker. 2006. On active learning in multi-label classification. In *From Data and Information Analysis to Knowledge Engineering*, pages 206–213.

Olivier Chapelle and Alexander Zien. 2005. Semi-supervised classification by low density separation.

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

Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. 2006. Trading convexity for scalability. In *Proceedings of the 23rd international conference on Machine learning*.

J. R. Curran, S. Clark, and J. Bos. 2007. Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the ACL 2007 Demonstrations Session*.

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

Andrea Esuli and Fabrizio Sebastiani. 2009. Active learning strategies for multi-label text classification. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins Karolinska, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*.

Ben Hachey and Claire Grover. 2006. Extractive summarisation of legal texts. *Artif. Intell. Law*, 14:305–345.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18.

T. Hastie and R. Tibshirani. 1998. Classification by pairwise coupling. *Advances in Neural Information Processing Systems*, 10.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2011. Semi-supervised discourse relation classification with structural learning. In *CICLing (1)*.

K. Hirohata, N. Okazaki, S. Ananiadou, and M. Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of 3rd International Joint Conference on Natural Language Processing*.

Steven C. H. Hoi, Rong Jin, and Michael R. Lyu. 2006. Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th international conference on World Wide Web*.

F. Jiao, S. Wang, C. Lee, R. Greiner, and D. Schuurmans. 2006. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *COLING/ACL*.

Carsten Lanquillon. 2000. Learning from labeled and unlabeled documents: A comparative study on semi-supervised text classification. In *Proceedings of the*

4th European Conference on Principles of Data Mining and Knowledge Discovery.

David D. Lewis and William A. Gale. 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*.

M. Liakata, S. Teufel, A. Siddharthan, and C. Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

J. Lin, D. Karakos, D. Demner-Fushman, and S. Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *Proceedings of BioNLP-06*.

G. S. Mann and A. Mccallum. 2007. Efficient computation of entropy gradient for semi-supervised conditional random fields. In *HLT-NAACL*.

Andrew McCallum and Kamal Nigam. 1998. Employing em and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*.

A. K. McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Quinn McNemar. 1947. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157.

S. Merity, T. Murphy, and J. R. Curran. 2009. Accurate argumentative zoning with maximum entropy models. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*.

Y. Mizuta, A. Korhonen, T. Mullen, and N. Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics on Natural Language Processing in Biomedicine and Its Applications*, 75(6):468–487.

T. Mullen, Y. Mizuta, and N. Collier. 2005. A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *Natural language processing and text mining*, 7(1):52–58.

Ion Muslea, Steven Minton, and Craig A. Knoblock. 2002. Active + semi-supervised learning = robust multi-view learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*.

Jorge Nocedal. 1980. Updating Quasi-Newton Matrices with Limited Storage. *Mathematics of Computation*, 35(151):773–782.

Bla Novak, Dunja Mladeni, and Marko Grobelnik. 2006. Text classification with active learning. In *From Data and Information Analysis to Knowledge Engineering*, pages 398–405.

J. C. Platt. 1999a. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classiers*, pages 61–74.

John C. Platt. 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*.

Piyush Rai, Avishek Saha, Hal Daumé, III, and Suresh Venkatasubramanian. 2010. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*.

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

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*.

H. S. Seung, M. Opper, and H. Sompolinsky. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*.

H. Shatkay, F. Pan, A. Rzhetsky, and W. J. Wilbur. 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093.

F. Sinz, 2011. *UniverSVM Support Vector Machine with Large Scale CCCP Functionality*. http://www.kyb.mpg.de/bs/people/fabee/universvm.html.

L. Sun and A. Korhonen. 2009. Improving verb clustering with automatically acquired selectional preference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

I. Tbahriti, C. Chichester, Frederique Lisacek, and P. Ruch. 2006. Using argumentation to retrieve articles with similar citations. *Int J Med Inform*, 75(6):488–495.

S. Teufel and M. Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28:409–445.

S. Teufel, A. Siddharthan, and C. Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP*.

S. Tong and D. Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66.

Simon Tong and Daphne Koller. 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66.

V. N. Vapnik. 1998. *Statistical learning theory*. Wiley, New York.

Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. 2009. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*.