Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes

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

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

Maria Liakata
Aberystwyth University, UK
mal@aber.ac.uk

Ilona Silins
Karolinska Institutet, SWEDEN
Ilona.Silins@ki.se

Lin Sun
University of Cambridge, UK
ls418@cam.ac.uk

Ulla Stenius
Karolinska Institutet, SWEDEN
Ulla.Stenius@ki.se

Abstract

Many practical tasks require accessing specific types of information in scientific literature; e.g. information about the objective, methods, results or conclusions of the study in question. Several schemes have been developed to characterize such information in full journal papers. Yet many tasks focus on abstracts instead. We take three schemes of different type and granularity (those based on section names, argumentative zones and conceptual structure of documents) and investigate their applicability to biomedical abstracts. We show that even for the finest-grained of these schemes, the majority of categories appear in abstracts and can be identified relatively reliably using machine learning. We discuss the impact of our results and the need for subsequent task-based evaluation of the schemes.

1 Introduction

Scientific abstracts tend to be very similar in terms of their information structure. For example, many abstracts provide some background information before defining the precise objective of the study, and the conclusions are typically preceded by the description of the results obtained.

Many readers of scientific abstracts are interested in specific types of information only, e.g. the general background of the study, the methods used in the study, or the results obtained. Accordingly, many text mining tasks focus on the extraction of information from certain parts of abstracts only. Therefore classification of abstracts (or full articles) according to the categories of information structure can support both the manual study of scientific literature as well as its automatic analysis, e.g. information extraction, summarization and information retrieval (Teufel and Moens, 2002; Mizuta et al., 2005; Tbahriti et al., 2006; Ruch et al., 2007).

To date, a number of different schemes and techniques have been proposed for sentence-based classification of scientific literature according to information structure, e.g. (Teufel and Moens, 2002; Mizuta et al., 2005; Lin et al., 2006; Hirohata et al., 2008; Teufel et al., 2009; Shatkay et al., 2008; Liakata et al., 2010). Some of the schemes are coarse-grained and merely classify sentences according to typical section names seen in scientific documents (Lin et al., 2006; Hirohata et al., 2008). Others are finer-grained and based e.g. on argumentative zones (Teufel and Moens, 2002; Mizuta et al., 2005; Teufel et al., 2009), qualitative dimensions (Shatkay et al., 2008) or conceptual structure (Liakata et al., 2010) of documents.

The majority of such schemes have been developed for full scientific journal articles which are richer in information and also considered to be more in need of the definition of information structure (Lin, 2009). However, many practical tasks currently focus on abstracts. As a distilled summary of key information in full articles, abstracts may exhibit an entirely different distribution of scheme categories than full articles. For tasks involving abstracts, it would be useful to know which schemes are applicable to abstracts and which can be automatically identified in them with reasonable accuracy.

In this paper, we will compare the applicability of three different schemes – those based on section names, argumentative zones and conceptual structure of documents – to a collection of biomedical abstracts used for cancer risk assessment (CRA). CRA is an example of a real-world task which could greatly benefit from knowledge about the information structure of abstracts since cancer risk assessors look for a variety of information in them ranging from specific methods to

results concerning different chemicals (Korhonen et al., 2009). We report work on the annotation of CRA abstracts according to each scheme and investigate the schemes in terms of their distribution, mutual overlap, and the success of identifying them automatically using machine learning. Our investigation provides an initial idea of the practical usefulness of the schemes for tasks involving abstracts. We discuss the impact of our results and the further task-based evaluation which we intend to conduct in the context of CRA.

2 The three schemes

We investigate three different schemes – those based on Section Names (S1), Argumentative Zones (S2) and Core Scientific Concepts (S3):
S1: The first scheme differs from the others in the sense that it is actually designed for abstracts. It is based on section names found in some scientific abstracts. We use the 4-way classification from (Hirohata et al., 2008) where abstracts are divided into objective, method, results and conclusions. Table 1 provides a short description of each category for this and other schemes (see also this table for any category abbreviations used in this paper).
S2: The second scheme is based on Argumentative Zoning (AZ) of documents. The idea of AZ is to follow the knowledge claims made by authors. Teufel and Moens (2002) introduced AZ and applied it to computational linguistics papers. Mizuta et al. (2005) modified the scheme for biology papers. More recently, Teufel et al. (2009) introduced a refined version of AZ and applied it to chemistry papers. As these schemes are too fine-grained for abstracts (some of the categories do not appear in abstracts at all), we adopt a reduced version of AZ which integrates seven categories from (Teufel and Moens, 2002) and (Mizuta et al., 2005) - those which actually appear in abstracts.
S3: The third scheme is concept-driven and ontology-motivated (Liakata et al., 2010). It treats scientific papers as humanly-readable representations of scientific investigations and seeks to retrieve the structure of the investigation from the paper as generic high-level Core Scientific Concepts (CoreSC). The CoreSC is a 3-layer annotation scheme but we only consider the first layer in the current work. The second layer pertains to properties of the categories (e.g. "advantage" vs. "disadvantage" of METH, "new" vs. "old" METH or OBJT). Such level of granularity is rare in abstracts. The 3rd layer involves coreference identification between the same instances of each category, which is also not of concern in abstracts. With eleven categories, S3 is the most fine-grained of our schemes. CoreSC has been previously applied to chemistry papers (Liakata et al., 2010, 2009).

3 Data: cancer risk assessment abstracts

We used as our data the corpus of CRA abstracts described in (Korhonen et al., 2009) which contains MedLine abstracts from different subdomains of biomedicine. The abstracts were selected so that they provide rich information about various scientific data (human, animal and cellular) used for CRA. We selected 1000 abstracts (in random) from this corpus. The resulting data includes 7,985 sentences and 225,785 words in total.

4 Annotation of abstracts

Annotation guidelines. We used the guidelines of Liakata for S3 (Liakata and Soldatova, 2008), and developed the guidelines for S1 and S2 (15 pages each). The guidelines define the unit (a sentence) and the categories of annotation and provide advice for conflict resolution (e.g. which categories to prefer when two or several are possible within the same sentence), as well as examples of annotated abstracts.
Annotation tool. We modified the annotation tool of Korhonen et al. (2009) so that it could be used to annotate abstracts according to the schemes. This tool was originally developed for the annotation of CRA abstracts according to the scientific evidence they contain. The tool works as a Firefox plug-in. Figure 1 shows an example of an abstract annotated according to the three schemes.
Description of annotation. Using the guidelines and the tool, the CRA corpus was annotated according to each of the schemes. The annotation proceeded scheme by scheme, independently, so that annotations of one scheme were not based on any of the other two. One annotator (a computational linguist) annotated all the abstracts according to the three schemes, starting from the coarse-grained S1, then proceeding to S2 and finally to the finest-grained S3. It took 45, 50 and 90 hours in total for S1, S2 and S3, respectively.
The resulting corpus. Table 2 shows the distribution of sentences per scheme category in the resulting corpus.

Table 1: The Three Schemes

S1	Objective	OBJ	The background and the aim of the research
	Method	METH	The way to achieve the goal
	Result	RES	The principle findings
	Conclusion	CON	Analysis, discussion and the main conclusions
S2	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
S3	Hypothesis	HYP	A statement that has not been yet confirmed rather than a factual statement
	Motivation	MOT	The reason for carrying out the investigation
	Background	BKG	Description of generally accepted background knowledge and previous work
	Goal	GOAL	The target state of the investigation where intended discoveries are made
	Object	OBJT	An entity which is a product or main theme of the investigation
	Experiment	EXP	Experiment details
	Model	MOD	A statement about a theoretical model or framework
	Method	METH	The means by which the authors seek to achieve a goal of the investigation
	Observation	OBS	The data/phenomena recorded within an investigation
	Result	RES	Factual statements about the outputs of an investigation
	Conclusion	CON	Statements inferred from observations and results, relating to research hypothesis

Inter-annotator agreement. We measured the inter-annotator agreement on 300 abstracts (i.e. a third of the corpus) using three annotators (one linguist, one expert in CRA, and the computational linguist who annotated all the corpus). According to Cohen's Kappa (Cohen, 1960), the inter-annotator agreement for S1, S2, and S3 was $\kappa = 0.84$, $\kappa = 0.85$, and $\kappa = 0.50$, respectively. According to (Landis and Koch, 1977), the agreement 0.81-1.00 is perfect and 0.41-0.60 is moderate. Our results indicate that S1 and S2 are the easiest schemes for the annotators and S3 the most challenging. This is not surprising as S3 is the scheme with the finest granularity. Its reliable identification may require a longer period of training and possibly improved guidelines. Moreover, previous annotation efforts using S3 have used domain experts for annotation (Liakata et al., 2009, 2010). In our case the domain expert and the linguist agreed the most on S3 ($\kappa = 0.60$). For S1 and S2 the best agreement was between the linguist and the computational linguist ($\kappa = 0.87$ and $\kappa = 0.88$, respectively).

Table 2: Distribution of sentences in the scheme-annotated CRA corpus

S1	OBJ	METH	RES	CON								
	61483	39163	89575	35564	Words							
	2145	1396	3203	1241	Sentences							
	27%	17%	40%	16%	Sentences							
S2	BKG	OBJ	METH	RES	CON	REL	FUT					
	36828	23493	41544	89538	30752	2456	1174	Words				
	1429	674	1473	3185	1082	95	47	Sentences				
	18%	8%	18%	40%	14%	1%	1%	Sentences				
S3	HYP	MOT	BKG	GOAL	OBJT	EXP	MOD	METH	OBS	RES	CON	
	2676	4277	28028	10612	15894	22444	1157	17982	17402	75951	29362	Words
	99	172	1088	294	474	805	41	637	744	2582	1049	Sentences
	1%	2%	14%	4%	6%	10%	1%	8%	9%	32%	13%	Sentences

5 Comparison of the schemes in terms of annotations

The three schemes we have used to annotate abstracts were developed independently and have separate guidelines. Thus, even though they seem to have some categories in common (e.g. METH, RES, CON) this does not necessarily guarantee that the latter cover the same information across all three schemes. We therefore wanted to investigate the relation between the schemes and the extent of overlap or complementarity between them.

We used the annotations obtained with each scheme to create three contingency matrices for pairwise comparison. We calculated the chi-squared Pearson statistic, the chi-squared like-

Figure 1: An example of an abstract annotated according to the three schemes

lihood ratio, the contingency coefficient and Cramer's V (Table 3)[1], all of which showed a definite correlation between rows and columns for the pairwise comparison of all three schemes.

However, none of the above measures give an indication of the differential association between schemes, i.e. whether it goes both directions and to what extent. For this reason we calculated the Goodman-Kruskal lambda L statistic (Siegel and Castellan, 1988), which gives us the reduction in error for predicting the categories of one annotation scheme, if we know the categories assigned according to the other. When using the categories of S1 as the independent variables, we obtained a lambda of over 0.72 which suggests a 72% reduction in error in predicting S2 categories and 47% in predicting S3 categories. With S2 categories being the independent variables, we obtained a reduction in error of 88% when predicting S1 and 55% when predicting S3 categories. The lower lambdas for predicting S3 are hardly surprising as S3 has 11 categories as opposed to 4 and 7 for S1 and S2 respectively. S3 on the other hand has strong predictive power in predicting the categories of S1 and S2 with lambdas of 0.86 and 0.84 respectively. In terms of association, S1 and S2 seem to be more strongly associated, followed by S1 and S3 and then S2 and S3.

We were then interested in the correspondence between the actual categories of the three schemes, which is visualized in Figure 2. Looking at the categories of S1, OBJ maps mostly to BKG and OBJ in S2 (with a small percentage in METH and REL). S1 OBJ maps to BKG, GOAL, HYP, MOT and OBJT in S3 (with a small percentage in METH and MOD). S1 METH maps to METH in S2 (with a small percentage in S2 OBJ) while it maps to EXP, METH and MOD in S3 (with a small percentage in GOAL and OBJT). S1 RES covers S2 RES and 40% REL, whereas in S3 it covers RES, OBS and 20% MOD. S1 CON covers S2 CON, FUT, 45% REL and a small percentage of RES. In terms of the S2 vs S3 comparison, S2 BKG maps to S3 BKG, HYP, MOT and a small percentage of OBJT and MOD. S2 CON maps to S3 CON, with a small percentage in RES, OBS and HYP. S2 FUT maps entirely to S3 CON. S2 METH maps to S3 METH, EXP, MOD, 20% OBJT and a small percentage of GOAL. S2 OBJ maps to S3 GOAL and OBJT, with 15% HYP, MOD and MOT and a small percentage in METH. S2 REL spans across S3 CON, RES, MOT and OBJT, albeit in very small percentages. Finally, S2 RES maps to S3 RES and OBS, with 25% in MOD and small percentages in METH, CON, OBJT. Thus, it appears that each category in S1 maps to a couple of categories in S2 and several in S3, which in turn seem to elaborate on the S2 categories.

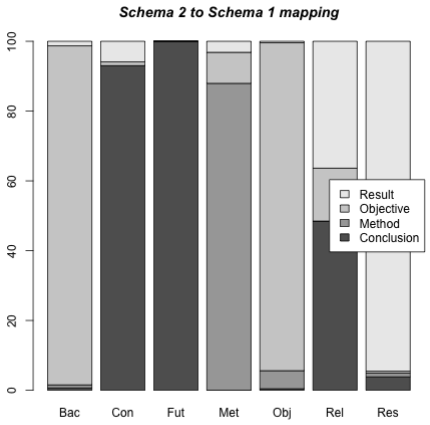
Based on the above analysis of the categories, it is reasonable to assume a subsumption relation between the categories of the type S1 $>$ S2 $>$ S3, with REL cutting across several of the S3 categories and FUT branching off S3 CON. This is an interesting and exciting outcome given that the three different schemes have such a different origin.

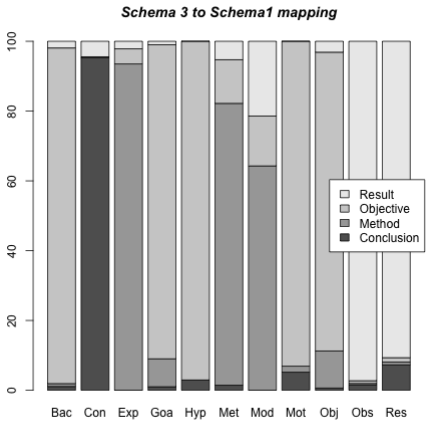
[1]These are association measures for r x c tables. We used the implementation in the vcd package of R (http://www.r-project.org/).

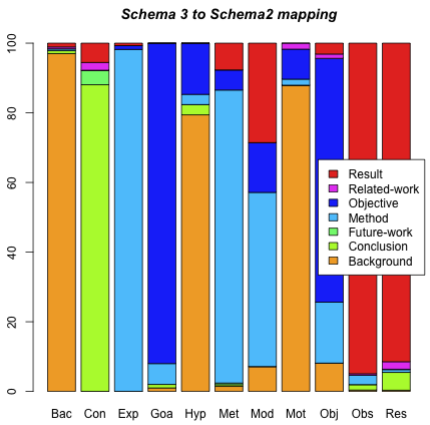
Table 3: Association measures between schemes S1, S2, S3

	S1 vs S2			S1 vs S3			S2 vs S3		
	X^2	**df**	P	X^2	**df**	P	X^2	**df**	P
Likelihood Ratio	5577.1	18	0	5363.6	30	0	6293.4	60	0
Pearson	6613.0	18	0	6371.0	30	0	8554.7	60	0
Contingency Coeff	0.842			0.837			0.871		
Cramer's V	0.901			0.885			0.725		

Figure 2: Pairwise interpretation of categories of one scheme in terms of the categories of the other.







6 Automatic identification of information structure

6.1 Features

The first step in automatic identification of information structure is feature extraction. We chose a number of general purpose features suitable for all the three schemes. With the exception of our novel verb class feature, the features are similar to those employed in related works, e.g. (Teufel and Moens, 2002; Mullen et al., 2005; Hirohata et al., 2008):

History. There are typical patterns in the information structure, e.g. RES tends to be followed by CON rather than by BKG. Therefore, we used the category assigned to the previous sentence as a feature.

Location. Categories tend to appear in typical positions in a document, e.g. BKG occurs often in the beginning and CON at the end of the abstract. We divided each abstract into ten equal parts (1-10), measured by the number of words, and defined the location (of a sentence) feature by the parts where the sentence begins and ends.

Word. Like many text classification tasks, we employed all the words in the corpus as features.

Bi-gram. We considered each bi-gram (combination of two word features) as a feature.

Verb. Verbs are central to the meaning of sentences, and can vary from one category to another. For example, *experiment* is frequent in METH and *conclude* in CON. Previous works have used the matrix verb of each sentence as a feature. Because the matrix verb is not the only meaningful verb, we used all the verbs instead.

Verb Class. Because individual verbs can result in sparse data problems, we also experimented with a novel feature: verb class (e.g. the class of EXPERIMENT verbs for verbs such as *measure* and *inject*). We obtained 60 classes by clustering verbs appearing in full cancer risk assessment articles using the approach of Sun and Korhonen (2009).

POS. Tense tends to vary from one category to another, e.g. past is common in RES and past partici-

ple in CON. We used the part-of-speech (POS) tag of each verb assigned by the C&C tagger (Curran et al., 2007) as a feature.

GR. Structural information about heads and dependents has proved useful in text classification. We used grammatical relations (GRs) returned by the C&C parser as features. They consist of a named relation, a head and a dependent, and possibly extra parameters depending on the relation involved, e.g. *(dobj investigate mouse)*. We created features for each subject (*ncsubj*), direct object (*dobj*), indirect object (*iobj*) and second object (*obj2*) relation in the corpus.

Subj and Obj. As some GR features may suffer from data sparsity, we collected all the subjects and objects (appearing with any verbs) from GRs and used them as features.

Voice. There may be a correspondence between the active and passive voice and categories (e.g. passive is frequent in METH). We therefore used voice as a feature.

6.2 Methods

We used Naive Bayes (NB) and Support Vector Machines (SVM) for classification. NB is a simple and fast method while SVM has yielded high performance in many text classification tasks.

NB applies Bayes' rule and Maximum Likelihood estimation with strong independence assumptions. It aims to select the class c with maximum probability given the feature set F:

$$\begin{aligned}\arg\max_c P(c|F) &= \arg\max_c \frac{P(c)\cdot P(F|c)}{P(F)} \\ &= \arg\max_c P(c)\cdot P(F|c) \\ &= \arg\max_c P(c)\cdot \textstyle\prod_{f\in F} P(f|c)\end{aligned}$$

SVM constructs hyperplanes in a multidimensional space that separates 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 the normal vector to the hyperplane. 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 the two is $\frac{2}{|w|}$. The problem reduces to:

Minimize $|w|$

Subject to $w \cdot x_i - b \geq 1$ for x_i of one class,
and $w \cdot x_i - b \leq -1$ for x_i of the other.

7 Experimental evaluation

7.1 Preprocessing

We developed a tokenizer to detect the boundaries of sentences and to perform basic tokenisation, such as separating punctuation from adjacent words e.g. in tricky biomedical terms such as *2-amino-3,8-diethylimidazo[4,5-f]quinoxaline*. We used the C&C tools (Curran et al., 2007) for POS tagging, lemmatization and parsing. The lemma output was used for extracting *Word*, *Bi-gram* and *Verb* features. The parser produced GRs for each sentence from which we extracted the *GR*, *Subj*, *Obj* and *Voice* features. We only considered the GRs relating to verbs. The "obj" marker in a subject relation indicates a verb in passive voice (e.g. *(ncsubj observed_14 difference_5 obj)*). To control the number of features we removed the words and GRs with fewer than 2 occurrences and bi-grams with fewer than 5 occurrences, and lemmatized the lexical items for all the features.

7.2 Evaluation methods

We used Weka (Witten, 2008) for the classification, employing its NB and SVM linear kernel. The results were measured in terms of accuracy (the percentage of correctly classified sentences), precision, recall, and F-Measure. We used 10-fold cross validation to avoid the possible bias introduced by relying on any one particular split of the data. The data were randomly divided into ten parts of approximately the same size. Each individual part was retained as test data and the remaining nine parts were used as training data. The process was repeated ten times with each part used once as the test data. The resulting ten estimates were then combined to give a final score. We compare our classifiers against a baseline method based on random sampling of category labels from training data and their assignment to sentences on the basis of their observed distribution.

7.3 Results

Table 4 shows F-measure results when using each individual feature alone, and Table 5 when using all the features but the individual feature in question. In these two tables, we only report the results for SVM which performed considerably better than NB. Although we have results for most scheme categories, the results for some are missing due to the lack of sufficient training data (see Table 2), or due to a small feature set (e.g. *History* alone).

Table 4: F-Measure results when using each individual feature alone

		a	b	c	d	e	f	g	h	i	j	k
S1	OBJ	.39	.83	.71	.69	.52	.45	.45	.45	.54	.39	-
	METH	-	.47	.81	.74	.63	.49	-	.46	.03	.42	.51
	RES	-	.76	.85	.86	.76	.70	.72	.69	.70	.68	.54
	CON	-	.72	.70	.65	.63	.53	.49	.57	.68	.20	-
S2	BKG	.26	.73	.69	.67	.45	.38	.56	.33	.33	.29	-
	OBJ	-	.13	.72	.68	.54	.63	-	.49	.48	.20	-
	METH	-	.50	.81	.72	.64	.47	-	.47	.03	.42	.51
	RES	-	.76	.85	.87	.76	.72	.72	.70	.69	.68	.54
	CON	-	.70	.73	.71	.62	.51	.40	.61	.67	.23	-
	REL	-	-	-	-	-	-	-	-	-	-	-
	FUT	-	-	-	-	-	-	-	-	-	-	-
S3	HYP	-	-	-	-	.67	-	-	-	-	-	-
	MOT	.18	.57	.70	.49	.39	.13	.36	.33	.30	.40	-
	BKG	-	-	.54	.40	.21	-	-	.11	.06	.06	-
	GOAL	-	-	.53	.33	.22	-	.19	.31	-	.25	-
	OBJT	-	-	.73	.63	.60	.10	-	.26	.32	-	-
	EXP	-	.22	.63	.46	.33	.30	-	.31	.07	.44	.25
	MOD	-	-	-	-	-	-	-	-	-	-	-
	METH	-	-	.82	.61	.39	.39	-	.50	-	.37	-
	OBS	-	.59	.75	.71	.63	.56	.56	.54	.48	.52	.47
	RES	-	-	.87	.73	.41	.34	-	.38	.24	.35	-
	CON	-	.74	.68	.65	.65	.50	.48	.49	.55	.21	-

a-k: History, Location, Word, Bi-gram, Verb, Verb Class, POS, GR, Subj, Obj, Voice

Looking at individual features alone, *Word*, *Bi-gram* and *Verb* perform the best for all the schemes, and *History* and *Voice* perform the worst. In fact *History* performs very well on the training data, but for the test data we can only use estimates rather than the actual labels. The *Voice* feature works only for RES and METH for S1 and S2, and for OBS for S3. This feature is probably only meaningful for some of the categories.

When using all but one of the features, S1 and S2 suffer the most from the absence of *Location*, while S3 from the absence of *Word*/*POS*. *Verb Class* on its own performs worse than *Verb*, however when combined with other features it performs better: leave-Verb-out outperforms leave-Verb Class-out.

After comparing the various combinations of features, we found that the best selection of features was *all but the Verb* for all the schemes. Table 6 shows the results for the baseline (BL), and the best results for NB and SVM. NB and SVM perform clearly better than BL for all the schemes. The results for SVM are the best. NB yields the highest performance with S1. Being sensitive to sparse data, it does not perform equally well on S2 and S3 which have a higher number of categories, some of which are low in frequency (see Table 2).

For S1, SVM finds all the four scheme categories with the accuracy of 89%. F-measure is 90 for OBJ, RES and CON and 81 for METH. For S2, the classifier finds six of the seven categories, with the accuracy of 90% and the average F-measure of 91 for the six categories. As with S2, METH has the lowest performance (at 85 F-measure). The one missing category (REL) appears in our abstract data with very low frequency (see Table 2).

For S3, SVM uncovers as many as nine of the 11 categories with accuracy of 81%. Six categories perform well, with F-measure higher than 80. EXP, BKG and GOAL have F-measure of 70, 62 and 62, respectively. Like the missing categories HYP and MOD, GOAL is very low in frequency. The lower performance of the higher frequency EXP and BKG is probably due to low precision in distinguishing between EXP and METH, and BKG and other categories, respectively.

Table 5: F-Measure results using all the features and all but one of the features

		ALL	A	B	C	D	E	F	G	H	I	J	K
S1	OBJ	.90	.89	.87	.92	.90	.90	.91	.91	.91	.92	.91	.88
	METH	.80	.81	.80	.80	.79	.81	.79	.80	.80	.80	.81	.81
	RES	.88	.90	.88	.90	.88	.90	.88	.88	.88	.89	.89	.90
	CON	.86	.85	.82	.87	.88	.90	.90	.88	.89	.88	.88	.90
S2	BKG	.91	.94	.90	.90	.93	.94	.94	.91	.93	.94	.92	.94
	OBJ	.72	.78	.84	.78	.83	.88	.84	.81	.83	.84	.78	.83
	METH	.81	.83	.80	.81	.80	.85	.80	.78	.81	.81	.82	.83
	RES	.88	.90	.88	.89	.88	.91	.89	.89	.90	.90	.90	.89
	CON	.84	.83	.77	.83	.86	.88	.86	.87	.88	.89	.88	.81
	REL	-	-	-	-	-	-	-	-	-	-	-	-
	FUT	-	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
S3	HYP	-	-	-	-	-	-	-	-	-	-	-	-
	MOT	.82	.84	.80	.76	.82	.82	.83	.78	.83	.83	.83	.83
	BKG	.59	.60	.60	.54	.67	.62	.62	.59	.61	.61	.62	.61
	GOAL	.62	.67	.67	.62	.71	.62	.67	.43	.67	.67	.67	.62
	OBJT	.88	.85	.83	.74	.83	.85	.83	.74	.83	.83	.83	.85
	EXP	.72	.68	.72	.53	.65	.70	.72	.73	.74	.74	.72	.68
	MOD	-	-	-	-	-	-	-	-	-	-	-	-
	METH	.87	.86	.87	.66	.85	.89	.87	.88	.86	.86	.87	.86
	OBS	.82	.81	.84	.72	.80	.82	.81	.80	.82	.82	.81	.81
	RES	.87	.87	.88	.74	.87	.86	.87	.86	.87	.87	.87	.88
	CON	.88	.88	.82	.88	.83	.87	.87	.84	.87	.88	.87	.86

A-K: History, Location, Word, Bi-gram, Verb, Verb Class, POS, GR, Subj, Obj, Voice

We have 1.0 for FUT in S2 probably because the size of the training data is just right, and the model doesn't overfit the data. We make this assumption because we have 1.0 for almost all the categories in the training data, but only for FUT on the test data.

Table 6: Baseline and best NB and SVM results

		Acc.	**F-Measure**						
S1			OBJ	METH	RES	CON			
	BL	.29	.23	.23	.39	.18			
	NB	.82	.85	.75	.85	.71			
	SVM	.89	.90	.81	.90	.90			
		Acc.	**F-Measure**						
S2			BKG	OBJ	METH	RES	CON	REL	FUT
	BL	.25	.13	.08	.22	.40	.13	-	-
	NB	.76	.79	.25	.70	.83	.66	-	-
	SVM	.90	.94	.88	.85	.91	.88	-	1.0
		Acc.	**F-Measure**						
S3			HYP	MOT	BKG	GOAL	OBJT	EXP	MOD
	BL	.15	-	.10	.06	.04	.06	.11	-
	NB	.53	-	.56	-	-	-	.30	-
	SVM	.81	-	.82	.62	.62	.85	.70	-

		METH	OBS	RES	CON		
	BL	.13	.24	.15	.17		
	NB	.32	.61	.59	.62		
	SVM	.89	.82	.86	.87		

8 Discussion and conclusions

The results from our corpus annotation (see Table 2) show that for the coarse-grained S1, all the four categories appear frequently in biomedical abstracts (this is not surprising because S1 was actually designed for abstracts). All of them can be identified using machine learning. For S2 and S3, the majority of categories appear in abstracts with high enough frequency that we can conclude that also these two schemes are applicable to abstracts. For S2 we identified six categories using machine learning, and for S3 as many as nine, indicating that automatic identification of the schemes in abstracts is realistic.

Our analysis in section 5 showed that there is a subsumption relation between the categories of the schemes. S2 and S3 provide finer-grained information about the information structure of abstracts than S1, even with their 2-3 low frequency (or missing) categories. They can be useful for practical tasks requiring such information. For example, considering S3, there may be tasks where one needs to distinguish between EXP, MOD and METH, between HYP, MOT and GOAL, or between OBS and RES.

Ultimately, the optimal scheme will depend on the level of detail required by the application at hand. Therefore, in the future, we plan to conduct task-based evaluation of the schemes in the context of CRA and to evaluate the usefulness of S1-S3 for tasks cancer risk assessors perform on abstracts (Korhonen et al., 2009). Now that we have annotated the CRA corpus for S1-S3 and have a machine learning approach available, we are in an excellent position to conduct this evaluation.

A key question for real-world tasks is the level of machine learning performance required. We plan to investigate this in the context of our task-based evaluation. Although we employed fairly standard text classification methodology in our experiments, we obtained high performance for S1 and S2. Due to the higher number of categories (and less training data for each of them), the overall performance was not equally impressive for S3 (although still quite high at 81% accuracy).

Hirohata et al. (2008) have showed that the amount of training data can have a big impact on our task. They used c. 50,000 Medline abstracts annotated (by the authors of the Medline abstracts) as training data for S1. When using a small set of standard text classification features and Conditional Random Fields (CRF) (Lafferty et al., 2001) for classification, they obtained 95.5% per-sentence accuracy on 1000 abstracts. However, when only 1000 abstracts were used for training the accuracy was considerably worse; their reported per-abstract accuracy dropped from 68.8% to less than 50%. Although it would be difficult to obtain similarly huge training data for S2 and S3, this result suggests that one key to improved performance is larger training data, and this is what we plan to explore especially for S3.

In addition we plan to improve our method. We showed that our schemes partly overlap and that similar features and methods tend to perform the best / worst for each of the schemes. It is therefore unlikely that considerable scheme specific tuning will be necessary. However, we plan to develop our features further and to make better use of the sequential nature of information structure. Currently this is only represented as the History feature, which provides a narrow window view to the category of the previous sentence. Also we plan to compare SVM against methods such as CRF and Maximum Entropy which have proved successful in recent related works (Hirohata et al., 2008; Merity et al., 2009). The resulting models will be evaluated both directly and in the context of CRA to provide an indication of their practical usefulness for real-world tasks.

Acknowledgments

The work reported in this paper was funded by the Royal Society (UK), the Swedish Research Council, FAS (Sweden), and JISC (UK) which is funding the SAPIENT Automation project. YG was funded by the Cambridge International Scholarship.

References

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

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*, pages 33–36.

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

A. Korhonen, L. Sun, I. Silins, and U. Stenius. 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.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditionl random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

M. Liakata and L.N. Soldatova. 2008. Guidelines for the annotation of general scientific concepts. *Aberystwyth University, JISC Project Report http://ie-repository.jisc.ac.uk/88/*.

M. Liakata, Claire Q, and L.N. Soldatova. 2009. Semantic annotation of papers: Interface & enrichment tool (sapient). In *Proceedings of BioNLP-09*, pages 193–200, Boulder, Colorado.

M. Liakata, S. Teufel, A. Siddharthan, and C. Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. *To appear in the 7th International Conference on Language Resources and Evaluation*.

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*, pages 65–72, New York, USA.

J. Lin. 2009. Is searching full text more effective than searching abstracts? *BMC Bioinformatics*, 10:46.

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*, pages 19–26. Association for Computational Linguistics.

Y. Mizuta, A. Korhonen, T. Mullen, and N. Collier. 2005. 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*.

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*, 7:52–58.

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*, 76:195–200.

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*, 18:2086–2093.

S. Siegel and N. J. Jr. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Berkeley, CA, 2nd edition.

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*, 75:488–495.

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 *Proc. of EMNLP*.

I. H. Witten, 2008. *Data mining: practical machine learning tools and techniques with Java Implementations*. http://www.cs.waikato.ac.nz/ml/weka/.