Bioinformatics, 31(7), 2015, 1084–1092
doi: 10.1093/bioinformatics/btu758
Advance Access Publication Date: 18 November 2014
Original Paper

OXFORD

Data and text mining

Unsupervised discovery of information structure in biomedical documents

Douwe Kiela[1], Yufan Guo[1], Ulla Stenius[2] and Anna Korhonen[1],*

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

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on May 14, 2014; revised on October 24, 2014; accepted on November 10, 2014

Abstract

Motivation: Information structure (IS) analysis is a text mining technique, which classifies text in biomedical articles into categories that capture different types of information, such as objectives, methods, results and conclusions of research. It is a highly useful technique that can support a range of Biomedical Text Mining tasks and can help readers of biomedical literature find information of interest faster, accelerating the highly time-consuming process of literature review. Several approaches to IS analysis have been presented in the past, with promising results in real-world biomedical tasks. However, all existing approaches, even weakly supervised ones, require several hundreds of hand-annotated training sentences specific to the domain in question. Because biomedicine is subject to considerable domain variation, such annotations are expensive to obtain. This makes the application of IS analysis across biomedical domains difficult. In this article, we investigate an unsupervised approach to IS analysis and evaluate the performance of several unsupervised methods on a large corpus of biomedical abstracts collected from PubMed.

Results: Our best unsupervised algorithm (multilevel-weighted graph clustering algorithm) performs very well on the task, obtaining over 0.70 *F* scores for most IS categories when applied to well-known IS schemes. This level of performance is close to that of lightly supervised IS methods and has proven sufficient to aid a range of practical tasks. Thus, using an unsupervised approach, IS could be applied to support a wide range of tasks across sub-domains of biomedicine. We also demonstrate that unsupervised learning brings novel insights into IS of biomedical literature and discovers information categories that are not present in any of the existing IS schemes.

Availability and Implementation: The annotated corpus and software are available at http://www.cl.cam.ac.uk/∼dk427/bio14info.html.

Contact: alk23@cam.ac.uk

1 Introduction

Recent developments in the areas of natural language processing, machine learning and data mining have led to considerable progress in Biomedical Text Mining (Bio-TM) (Chapman and Cohen, 2009; Harmston *et al*., 2010; McDonald *et al*., 2012; Simpson and Demner-Fushman, 2012), producing highly useful techniques that enable users to retrieve and extract information from scientific text easily and efficiently.

Information structure (IS) analysis is a highly useful TM technique that has attracted increasing attention recently. It aims to classify units of text (typically sentences) into a fixed number of categories, which capture different *types of information* in text (Guo *et al*., 2011b; Webber *et al*., 2011). For instance, IS categories could capture the aim of the study or a type of the study (e.g. animal, human and *in vitro*), the design of experiments (e.g. exposure length,

dose and group size), the results obtained (e.g. end points, positive or negative results), the conclusions drawn and so forth.

To date, several IS schemes have been proposed for the analysis of scientific documents. There are coarse-grained schemes that stem from typical section headings seen in academic journals (Agarwal and Yu, 2009; Sollaci and Pereira, 2004), as well as finer-grained schemes. Among the latter are those based on the rhetorical structure and scientific argumentation of a scientific paper [namely argumentative zones (AZs)] (Guo *et al.*, 2010; Mizuta *et al.*, 2006; Teufel *et al.*, 1999, 2009), the qualitative dimensions or the properties of factual information (e.g. focus, polarity, certainty, evidence and directionality) (Wilbur *et al.*, 2006), the distinction between different types of evidence in empirical studies (e.g. explicit versus implicit claims, correlations, comparisons and observations) (Blake, 2009) and the content and conceptual framework for a scientific investigation (Liakata *et al.*, 2010), among others.

Automatic labeling of biomedical text according to IS has many practical applications. It has been shown to support a variety of Bio-TM tasks (e.g. information retrieval, information extraction and text summarization) (Contractor *et al.*, 2012; Mizuta *et al.*, 2006; Ruch *et al.*, 2007; Tbahriti *et al.*, 2006; Teufel and Moens, 2002) and manual study of biomedical literature. For example, Guo *et al.* (2011a) reported a speed reading test where IS annotations significantly speeded up literature review in cancer risk assessment.

However, the usefulness of IS analysis has been limited by the fact that existing approaches suffer from poor portability. Biomedicine is subject to considerable sub-domain variation at different levels of linguistic description (Lippincott *et al.*, 2011). Application of IS analysis to different sub-domains of biomedicine (e.g. chemistry, molecular biology and cancer research) has required the development of domain-specific training and evaluation data (Guo *et al.*, 2010; Teufel *et al.*, 2009). Created by hand, such data are very expensive to develop and are only available for a handful of sub-domains.

Recent research has investigated reducing the need for annotations via lightly supervised learning (e.g. active learning) of IS. This research has produced useful results—accurate enough to support high-speed literature review (Guo *et al.*, 2011b, 2013). However, such an approach still requires several hundreds of annotated sentences for optimal performance. This is unrealistic as biomedical literature is not only varied in terms of domains but also highly dynamic (Mihăilă *et al.*, 2012).

This article investigates whether unsupervised methods could be realistic for IS analysis. Such methods have the distinct advantage of removing the need for any IS annotation. Not only would this allow for easily applying the approach to the many (sub-)domains where annotated data are not readily available, but it could also lead to the discovery of novel, undefined information categories emerging from the data.

We experiment with a large corpus of biomedical abstracts annotated according to two different IS schemes: section names (SN) and AZs. We apply to this corpus two canonical clustering algorithms—spherical k-means (Zhong, 2005) and Expectation Maximization-Gaussian Mixture Model (EM-GMM) (Dempster *et al.*, 1977)—as well as the-state- of-the-art multilevel-weighted graph clustering algorithm (Dhillon *et al.*, 2007), which is an efficient approximation of the very popular spectral clustering algorithm. The latter algorithm has the added advantage that it avoids expensive eigenvector computation by exploiting weighted kernel k-means to locally optimize graph clustering objectives (Dhillon *et al.*, 2004). Using a selection of clustering evaluation metrics, we evaluate these algorithms on the two IS schemes, comparing them against fully supervised and weakly supervised algorithms. Our unsupervised approach performs surprisingly well, especially when multilevel-weighted graph clustering is applied to the SN scheme: we obtain over 0.70 *F* scores for most categories—performance which is close to that of lightly supervised methods and which has proven sufficient to aid practical tasks in biomedicine. We also demonstrate that unsupervised learning brings novel insights into IS of biomedical literature and discovers information categories that are not present in any of the existing IS schemes.

There is only one previous study in unsupervised IS analysis, which focused on detecting SNs (Varga *et al.*, 2012). They used probabilistic graphical models and reported an *F* score of 35%, which is rather low and unlikely to be used in real-world tasks. Our results suggest that if we use more sophisticated unsupervised techniques and better features, IS analysis could realistically be used to benefit a much wider range of tasks in biomedicine.

2 Materials and methods

2.1 Data

We experiment with two well-known and widely used IS schemes. The first is the SN scheme, which is grounded on SNs found in some scientific abstracts. We use the four-way classification from Hirohata *et al.* (2008), where abstracts are divided into Objective-, Method-, Result- and Conclusion-type sentences. The second is argumentative zoning (AZ)—a scheme originally introduced by Teufel and Moens (2002). AZ provides an analysis of the argumentative structure of a document, following the knowledge claims made by authors. We use the version of AZ developed for biology papers (Mizuta *et al.*, 2006), with the same modifications as in Guo *et al.* (2010).

Our experiments use the dataset by Guo *et al.* (2010), which consists of 1000 biomedical abstracts with a total of 7985 sentences, annotated according to both the SN and AZ schemes. Guo *et al.* (2010) reported a high inter-annotator agreement of $\kappa = 0.85$ for their three annotators: one linguist, one computational linguist and one domain expert. These two IS annotation schemes were selected because they are good examples of a coarse-grained scheme that typically relies on section headings (SN) and a more fine-grained scheme based on scientific rhetorical structure (AZ), respectively. The two schemes with their respective categories are listed in Table 1. Table 2 presents the distribution of scheme-annotated sentences in the corpus. Although there is a subsumption relation between the schemes (Guo *et al.*, 2010), SN as a coarser-grained version of AZ is still worth investigating because it is widely used in academic writing most notably in the field of biomedicine (Sollaci and Pereira, 2004).

2.2 Automatic classification

2.2.1 Feature selection

To apply our algorithms, we need to first select a set of features which may indicate which category in our scheme is appropriate. We follow Guo *et al.* (2010) in implementing a set of features that have proved successful in related works (e.g. Teufel and Moens 2002; Guo *et al.* 2011b, 2013; Hirohata *et al.* 2008; Lin *et al.* 2006; Mullen *et al.* 2005):

- **Location**: each abstract is divided into 10 equal parts, measured by the number of words. The location feature is defined by the parts where a 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**: sixty verb classes appearing in biomedical journal articles (e.g. the experiment class includes verbs such as 'measure' and 'inject').

Table 1. The section-based and AZ annotation schemes in the corpus of Guo *et al.* (2010)

Scheme	Category	Abbreviation	Definition and example
Section	Objective	OBJ	The background and the aim of the research
	Method	METH	The way to achieve the goal
	Result	RES	The principal findings
	Conclusion	CON	Analysis, discussion and the main conclusions
AZ	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, especially 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

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

Scheme	OBJ	METH	RES	CON	BKG	REL	FUT
Section	27%	17%	40%	16%	–	–	–
AZ	8%	18%	40%	14%	18%	1%	1%

- **Part-of-Speech (POS)**: the POS tag of each verb [Penn Treebank tagset (Santorini, 1990)], as an indicator of verb tense.
- **Grammatical relation (GR)**: subject, direct object, indirect object and second object relations between verbs and nouns.
- **Subj/Obj**: the subjects/objects appearing with any verbs.
- **Voice**: the voice of verbs.

These features were extracted from the corpus using a number of tools, including a tokenizer that detects sentence boundaries and performs basic tokenization (designed specifically for handling complex biomedical terms, e.g. 2-amino-3,8-diethylimidazo[4,5-f]quinoxaline), the C&C tools (Curran *et al.*, 2007) for POS tagging and parsing (the output was used for extracting aforementioned lexical and syntactic features) and the Morpha lemmatizer (http://svn.ask.it.usyd.edu.au/trac/candc) for lemmatizing the lexical items for all the features. Verb classes were obtained automatically using unsupervised spectral clustering (Sun and Korhonen, 2009). To reduce data sparsity, we removed words and GRs with fewer than two occurrences and bi-grams with fewer than five occurrences.

The combination of these features leads to the large number of 26 459 components per feature vector. We do not include the feature vectors of surrounding sentences, as some previous studies have done. The feature vectors are binary and tend to be sparse.

2.2.2 Unsupervised algorithms

We use two canonical clustering algorithms in the form of k-means and EM-GMM and one state-of-the-art algorithm that is an efficient approximation of the very popular spectral clustering algorithm in the form of multilevel-weighted graph clustering. We have selected three algorithms that can be argued to be particularly well suited for the task at hand: they have been shown to perform well on text data and are efficient enough to compute with such large vectors (which is not the case, e.g. spectral clustering).

Spherical k-means The k-means algorithm (MacQueen, 1967) is very well known. Here we apply a version called *spherical* k-means (Zhong, 2005), which differs from traditional k-means in having a different objective function: instead of minimizing Euclidean distance, we maximize cosine similarity, which is defined as follows:

$$cosine(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1||v_2|} \tag{1}$$

where v_1 and v_2 are feature vectors. The objective function thus becomes:

$$\underset{S}{\operatorname{argmax}} \sum_{i=1}^{k} \sum_{x_j \in S_i} cosine(\mu_i, x_j) \tag{2}$$

where S are the clusters, k is the number of clusters, x_j is a data point and μ_i is a centroid. It has been found that cosine similarity is a much better metric on large amounts of sparse data, which tends to be the case for text mining features, than Euclidean distance (Dhillon and Modha, 2001).

EM-GMM The EM algorithm (Dempster *et al.*, 1977) has a long history and is also well known in the unsupervised learning literature. In a probabilistic clustering scenario, we assume that the underlying distribution is a GMM and employ the EM algorithm to obtain the maximum (log-)likelihood estimate of the parameters:

$$\ln p(\mathbf{x}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{i=1}^{K} \pi_i \mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i) \right\} \tag{3}$$

where $\mathbf{x}$ is a D-dimensional vector, $\pi_i, i = 1, \ldots, K$ are the mixture weights and $\mathcal{N}$ defines the component Gaussian densities. This approach allows us to capture uncertainty in cluster assignments. Final assignments are then computed according to the highest probability for each of the data points.

Multilevel-Weighted Graph Clustering Two clustering methods that have recently gained attraction (especially for data that is not linearly separable) are kernel k-means and spectral clustering. It has been found that they have equivalent objective functions, i.e. a general weighted kernel k-means objective is mathematically equivalent to a weighted graph clustering objective (Dhillon *et al.*, 2007). This fact can be exploited by a multilevel algorithm that directly optimizes various weighted graph clustering objectives, such as the popular ratio cut, normalized cut and ratio association criteria (Dhillon *et al.*,

2005). The resultant algorithm is very fast: it eliminates the need for any eigenvector computation for graph clustering problems, which can be prohibitive for very large graphs, as is the case with our dataset. We used this algorithm as an *approximation* of spectral clustering (Dhillon *et al.*, 2007), for which the objective function is:

$$\max_{\tilde{Y}} \text{trace}(\tilde{Y}^T W^{-1/2} A W^{-1/2} \tilde{Y})$$

where A is an adjacency (*similarity*) matrix, W is a diagonal matrix of the weight/degree of each cluster and $\tilde{Y}$ is an orthonormal matrix that indicates the cluster membership and that is proportional to $W^{1/2}$. The problem can be solved using an efficient local search algorithm with good scalability.

We used an open source implementation of the algorithm called Graclus (http://www.cs.utexas.edu/users/dml/Software/graclus.html) for multilevel-weighted graph clustering. For the other two algorithms, we used MATLAB implementations.

2.2.3 Evaluation

Evaluation of clustering results is difficult. In the case of supervised or weakly supervised learning, evaluation is relatively easy: we have a one-to-one mapping of predicted labels to expected labels. In the case of unsupervised learning, however, we have a set of cluster assignments on the one hand and a set of expected labels on the other. The evaluation is thus complicated by the lack of an adequate mapping between the two sides, for how to decide whether a given cluster assignment is the 'right' cluster for an expected label in the gold standard. To mitigate this problem, we elected to evaluate using a variety of metrics and to choose the best clustering based on the combination of these metrics:

V measure The V measure is calculated on the basis of two criteria: homogeneity and completeness (Rosenberg and Hirschberg, 2007). To satisfy the homogeneity criterion, a clustering must assign only those data points that are members of a single class to a single cluster. That is, a perfectly homogeneous clustering has an injective mapping $C \rightarrow K$, where C is the set of classes and K is the set of clusters. To satisfy the completeness criterion, a clustering must assign all of those data points that are members of a single class to a single cluster. That is, a perfectly complete clustering has an injective mapping $K \rightarrow C$. Homogeneity is defined as:

$$h(n) = \begin{cases} 1 & H(C,K) = 0 \\ 1 - \dfrac{H(C\,|\,K)}{H(C)} & \text{else} \end{cases} \quad (4)$$

where

$$H(C\,|\,K) = -\sum_{k=1}^{|K|}\sum_{c=1}^{|C|}\frac{a_{ck}}{N}\log\frac{a_{ck}}{\sum_{c=1}^{|C|} a_{ck}} \quad (5)$$

a_{ck} is the number of data points that are members of class c and cluster k and

$$H(C) = -\sum_{c=1}^{|C|}\frac{\sum_{k=1}^{|K|}}{n}\log\frac{\sum_{k=1}^{|K}}{n} \quad (6)$$

Completeness is then defined as the 'reciprocal' of homogeneity:

$$c(n) = \begin{cases} 1 & H(K,C) = 0 \\ 1 - \dfrac{H(K\,|\,C)}{H(K)} & else \end{cases} \quad (7)$$

The V measure is simply the harmonic mean between homogeneity and completeness.

Table 3. Results for both schemes using our evaluation metrics

Algorithm	k	*V* measure	Purity	*F* measure
SN				
Baseline	4	0.00	0.38	0.43
Spherical k-means	9	0.29	0.68	0.50
EM-GMM	5	0.09	0.46	0.43
Multilevel-weighted graph	6	0.31	0.69	0.65
AZ				
Baseline	7	0.00	0.40	0.38
Spherical k-means	9	0.31	0.64	0.51
EM-GMM	5	0.09	0.43	0.40
Multilevel-weighted graph	6	0.30	0.62	0.58

Purity Purity is a simple and transparent evaluation measure that is often used for clustering (Amigó *et al.*, 2009). To compute it, each cluster is assigned to the class that is most frequent in the cluster and then the accuracy of this assignment is measured by counting the number of correctly assigned classes and divided by the total number of data points:

$$Purity(C,K) = \frac{1}{N}\sum_{j=1}^{K}\max_{C_i \in C}(|K_j \cap C_i|) \quad (8)$$

F *measure* The F measure is traditionally defined as the weighted harmonic mean between precision and recall (van Rijsbergen, 1974). In our case, we do not have a one-to-one mapping between predicted and expected labels, so instead we take the weighted mean of the highest combined precision and recall per class-cluster mapping, which we called the macro-F (Amigó et al., 2009):

$$F_{macro}(C,K) = \sum_{i=1}^{C}\frac{|C_i|}{N}\max_{K_j \in K} F(C_i,K_j) \quad (9)$$

where

$$F(C_i,K_j) = \frac{2 \times Recall(C_i,K_j) \times Precision(C_i,K_j)}{Recall(C_i,K_j) + Precision(C_i,K_j)} \quad (10)$$

As the F measure is probably the best known amongst our evaluation metrics, especially outside of the unsupervised learning community, we will focus on the macro-F measure. When we report F scores for individual class-cluster mappings, we will call these micro-F scores to distinguish them.

3 Results

3.1 Overall performance

Table 3 presents the results for our three unsupervised algorithms, compared with a baseline that maximizes homogeneity by assigning all sentences to the same cluster, which is a much better baseline than chance. Experiments were done with $2 \leq k \leq 20$. We report the results for the highest performing k per algorithm. With k set to the number of categories (which would allow for a true 1-1 mapping), we observe the same performance per algorithm as described below.

For the SN scheme, multilevel-weighted graph clustering performs best, with a macro F measure of 0.65, which is very high for unsupervised learning. Spherical k-means comes second, judging by V measure and Purity but is not as accurate when measured by macro-F. EM-GMM does not do particularly well, which we suspect is due to sparsity issues. For AZs, we observe a similar pattern when we look at the macro-F score: multilevel-weighted graph clustering

Table 4. Micro-*F* score per category for best performing in both schemes

Scheme	OBJ	METH	RES	CON	BKG	REL	FUT
SN	0.70	0.48	0.74	0.70	—	—	—
AZ	—	0.46	0.75	0.73	0.59	—	—

performs best, followed by spherical k-means, followed by EM-GMM.

3.2 Class-based analysis

To investigate the clustering results more closely, we look into the dominant class per cluster. Through a *winner takes all* approach (i.e. each cluster is labeled with its dominant class), we can calculate micro-*F* scores for each of the individual classes. Table 4 presents the *F* scores for the best performing algorithm with the optimal *k* for the two schemes.

The micro-*F* scores per class indicate that we are able to obtain very high scores for some of our classes, which are comparable to the weakly supervised scores obtained by Guo *et al.* (2011b). With the exception of the Method class, the SN scheme *F* scores are all over 0.70. Although all available classes are identified for the SN scheme, in the AZ scheme, only four out of the seven available classes are ever dominant in a cluster. However, we have found that increasing *k* allows the other classes to become dominant in at least one of the clusters as well, although identifying Future work and Related work is still a bit challenging, which is not surprising as there are relatively few sentences with these labels.

3.3 Feature-based analysis

As explained in Section 2.2.1, we made use of a collection of features, which have been shown to perform well in previous supervised and lightly supervised experiments. We conducted further analysis to investigate which of these feature types are the most (and the least) useful for unsupervised learning. We took the best-performing algorithm and conducted a leave-one-out analysis of the features. Table 5 presents the *F* scores per excluded feature type.

We can clearly observe that the most important feature types are location, word and voice. It makes sense that word is so important as it is the basic unit of a text corpus. Location and voice are good indicators of certain information categories as explained in Section 2.2.1. Bi-grams seem to be the least useful probably due to the problem of sparse data. These findings are generally in line with those reported in Guo *et al.* (2011a, b) on the same dataset, except for the word feature that plays a less important role in fully or lightly supervised IS analysis. A possible explanation could be that features such as verb, subj and obj are actually subsets of the word feature that may compensate for the absence of word features when provided with highly informative annotations.

We can now combine the leave-one-out analysis with the dominant class-based analysis to see which features are more or less important for each of the class labels. To save space, we only report results for the SN scheme, on which we have obtained the highest scores. Table 6 lists the results. Interestingly, we observe that the importance of the location and word features depends on the class: location seems to be particularly important for the Objective class, whereas word features are very important for the Method class. The verb feature type is very important for the Objective class.

Table 5. Macro *F* score per excluded feature type (LC, location; WO, word; BI, bi-gram; VE, verb; VC, verb class; SU, subject; OB, object; VO, voice) for both schemes. The lower the score, the higher the impact of the feature type

Scheme	All	LC	WO	BI	VE	VC	POS	GR	SU	OB	VO
SN	0.65	0.45	0.44	0.58	0.53	0.53	0.51	0.50	0.50	0.51	0.46
AZ	0.58	0.42	0.43	0.52	0.49	0.48	0.52	0.46	0.47	0.52	0.41

Table 6. Micro *F* score per excluded feature type per class for the SN scheme. The lower the score, the higher the impact of the feature type for the given class

Class	Conclusion	Method	Objective	Result
Location	0.57	0.46	0.38	0.60
Word	—	0.38	0.58	0.58
Bi-gram	0.55	0.42	0.60	0.67
Verb	0.52	0.44	0.37	0.71
Verb Class	—	0.44	0.54	0.68
POS	0.53	0.43	0.43	0.65
GRs	0.61	—	0.60	0.71
Subj	0.59	—	0.52	0.68
Obj	0.53	0.43	0.43	0.65
Voice	0.55	—	0.60	0.66
All	0.70	0.48	0.70	0.74

Table 7. Confusion matrix for best performing. Rows are the gold standard classes and columns are the winner-takes-all class assignments

	Conclusion	Method	Objective	Result
Conclusion	817	3	114	442
Method	5	474	177	614
Objective	85	126	1419	600
Result	57	103	118	2734

4 Discussion

4.1 Error analysis

In this section, we examine the errors that the algorithm makes more closely. In the case of AZs, it is unsurprising that we have difficulty with the Future-work and Related-work classes for the obvious reason that they are relatively infrequent in the corpus (as presented in Table 2). Furthermore, they are arguably very fine-grained classes that could be said to overlap with the Background class present in AZs.

The more interesting question concerns the Method class: why do we get scores above 0.70 for all other classes but a much lower score for this class? Table 7 presents the confusion matrix for the best-performing algorithm. We can clearly observe that the Method class is confused disproportionately often with the Result class. In fact, out of the 1300 sentences that should have been labeled Method, almost half of them have been labeled Result. This problem is not unique to our approach: the same phenomenon can be observed in weakly supervised and even in supervised learning. Table 8 presents some example method sentences and how they were classified. One might say that the errors are understandable, considering the misleading key words (e.g. 'detected') or the various numerical values that are frequently seen in the Result class.

Table 8. Example Method sentences and how they were classified in the winner-takes-all assignments

Cluster	Dominant label	Gold standard label	Example sentence
1	RES	METH	The metabolites were determined by HPLC.
1	RES	METH	These four reaction products were used as analytical standards for kinetic studies of the reaction of valinamide with BMO at physiological pH (7.4) and temperature (37°C).
1	RES	METH	Adducts detected *in vivo* were identified by comparison with the products formed from the reaction of the individual epoxides with 2'-deoxyguanosine (dG).
2	RES	METH	B6C3F1 lacI transgenic mice were exposed to air or to 62.5, 625 or 1250 ppm BD for 4 weeks (6 h/day, 5 days/week) and euthanized 14 days after the last exposure.
2	RES	METH	Groups of control and exposed animals ($n = 4$–12/group) were necropsied at multiple time points after exposure and the T-cell cloning assay was used to measure Hprt mutant frequencies in lymphocytes isolated from spleen.
5	METH	METH	F344 rats were given a single i.p. injection of diethylnitrosamine (200 mg/kg body weight) and subjected to two-thirds partial hepatectomy at week 3.
5	METH	METH	Female Wistar rats were administered NDEA (200 ppm) through drinking water (5 days/week) for 4 weeks.
5	METH	METH	Six-week-old male F344 rats were given a single dose of diethylnitrosamine (DEN, 200 mg/kg b.w., i.p.).

Table 9. Examples of CON misclassified as RES and OBJ misclassified as METH

Cluster	Dominant label	Gold standard label	Example sentence
6	RES	CON	Treatment with DMDTC significantly increased the protein carbonyl contents of hepatic microsomes compared with that of controls, a finding that may be related to DMDTC's activity as a prooxidant.
5	METH	OBJ	Potential modifying effects of epoprostenol sodium administration on liver carcinogenesis were investigated in male F344/DuCrj rats initially treated with N-nitrosodiethylamine (DEN).

A possible solution to this problem would be to assign higher weights to verb features that are more crucial to the meaning of a sentence and that can possibly make a better distinction between the Method and Result sentences.

Table 9 lists other common errors such as CON being misclassified as RES and OBJ being misclassified as METH. These errors often occur in complex sentences where different clauses/constituents carry different types of information. Under the current schemes, a unique information category is assigned to a full sentence according to the distribution or the priority of information contained in the sentence. However, in the case of machine learning, a computer may be confused by the contradictory evidence in a complex sentence (e.g. 'increased' for RES versus 'may be related' for CON, 'investigated' for OBJ versus 'treated' for METH). A potential solution would be to split the text into smaller units than sentences for more accurate IS analysis.

4.2 Algorithm performance

The best performing algorithm is undoubtedly multilevel-weighted graph clustering. It achieves a macro-*F* score of .65, which is very high for an unsupervised approach. In comparison, the approach by Guo *et al.* (2011b), which was weakly supervised but still required a large proportion of the data (thousands of labeled sentences) for optimal performance, achieved a macro-*F* score of 0.81. Multilevel-weighted graph clustering is preferable to the other unsupervised algorithms not only because of accuracy but also because it is faster than any of the others [unlike EM or k-means that works on a random initialization, the multilevel algorithm ensures a good initial clustering at each level and its convergence rate is better (Dhillon *et al.*, 2007)].

It is perhaps slightly surprising that EM-GMM does not perform very well. We speculate that this is due to sparsity issues: it has been commented in the past that EM-GMM cannot deal very well with sparseness (Neal and Hinton, 1999). Our data are particularly sparse, because all features are binary. Running EM-GMM on a subset of the features, such as location, voice and POS-tag feature types, yields better performance but is still not comparable to that of multilevel-weighted graph clustering.

4.3 Discovering novel IS categories

As mentioned above, unsupervised learning techniques may also yield novel insight into the appropriate information categories for biomedical literature. Interestingly, one of the clusters appears to be very 'noisy' compared with the others, which might indicate that the clustering finds an information category that is not present in our labeling scheme. Cluster 1, as we have called it, contains 889 sentences, of which 225 are Objective (25%), 350 are Result (39%), 204 are Method (23%) and 83 are Conclusion (9%). Using the winner-takes-all approach, we would assign this cluster the dominant label Result. However, these percentages show that this is not really a Result cluster at all.

To further examine the properties of this cluster, we looked at the sentences closest to the cluster centroid. Most sentences describe experimental data, either as reporting the outcome of an experiment (Result), the goal or rationale of an experiment (Objective) or the methods used in an experiment (Method). Some example sentences are listed in Table 10. This raises the distinct possibility that

Table 10. Sentences closest to the centroid in the 'Experiment' cluster

Gold standard label	Sentence
RES	In addition, the average Hprt MF in mice exposed to 3 ppm BD [1.54 ± 0.82 (SD) x 10(−6)] was significantly increased by 1.6-fold over the average control value of 0.96 ± 0.51 (SD)x10(−6) ($P = 0.004$).
METH	We then compared Hprt mutant frequencies (MFs) among these groups.
OBJ	The higher levels of these two DNA-reactive metabolites in mice compared with rats probably contribute to the species differences in carcinogenic effects of BD between mice and rats.
CON	These results suggest that EB causes mutation primarily by base substitution and that the spectrum of these mutations closely resembles that of BD.

performance would be improved by adding an 'Experiment' label to the SN scheme—it seems that the clustering algorithm prefers clustering these types of sentences together instead of with their respective gold standard label categories.

To delve even further into cluster properties, we looked at words that frequently occur in cluster sentences. Specifically, we are interested in words that occur disproportionately frequently in a particular cluster. Hence, we use a weighting scheme similar to, e.g. TF-IDF (Jones, 1972), where we treat each cluster as a document and weight the words based on the following formula, where N_c^w is the number of sentences containing word w in cluster c, N_c is the number of sentences in a cluster and c' is the set of clusters excluding cluster c (i.e. all clusters except for the current one):

$$weight(w, c) = \frac{N_c^w \times N_{c'}}{N_c \times N_{c'}^w} \quad (11)$$

This weighting scheme allows us to gauge the relative importance of words, compared with the sentences in the cluster and sentences in the other clusters. An alternative is to use the same scheme as in Equation (11) but to upweight common words by multiplying with the number of clusters that the word occurs in. To illustrate how well this approach can work in giving insight into cluster properties, consider some of the top words for the Conclusion and Experiment clusters for both of these weighting schemes (frequency based and common word based) in Table 11.

Using an unsupervised approach to discover novel or optimal IS categories can be particularly useful when applying IS analysis to new tasks, domains and datasets (e.g. full text biomedical articles) where we can expect a higher degree of IS variation.

4.4 Applicability

With respect to the usefulness of unsupervised IS analysis, a natural question is whether the results are good enough to support a real-world task in biomedicine. There are two issues that need to be addressed separately: first, how to obtain class labels and second, whether the performance would still be good enough to be useful. The former question is a side effect of the approach being unsupervised: we only have cluster assignments, without the class labels. We propose that this can be solved by presenting the user with one or more sentences that are closest to each of the centroids. The label that the user picks for those sentences is given to all the sentences in the same cluster. As illustrated in Table 12, a user should be able to choose the appropriate label based on the sentences nearest to the centroid, especially when more than one sentences are presented. The latter question is beyond the scope of this first paper on unsupervised IS—however, we expect good results since previous research on weakly supervised IS showed that performances very similar to those obtained for our best categories (around or above 0.70 F) were sufficient to significantly improve the efficiency of literature review in a real-life user test (Guo *et al.*, 2011b, 2013).

Table 11. Weighted contribution of words in clusters 1 (Experiment) and 3 (Conclusion)

Cluster 3 (Conclusion)		Cluster 1 (Experiment)	
Freq-weight	Common-weight	Freq-weight	Common-weight
Add	Suggest	Exhaled	hprt
Understanding	Useful	Differed	Frequency
Implication	Indicate	Passive	Exposed
Determinant	May	Bioactivated	Genotype
Need	Result	Globin	Spleen
Improve	Demonstrate	Linearly	Inhalation
Causative	Suggests	Cross-sectional	Mutant
Counteract	Together	Leukemia	bdo2
Conjunction	Finding	Collection	ppb
Considering	Taken	Matched	laci

Table 12. Sentences closest to each of the clusters' centroids

Cluster	Dominant label	Sentence closest to centroid
1	RES	*In vitro* activation of p-cresol with horseradish peroxidase produced six DNA adducts with a relative adduct level of 8.03 ± 0.43 x 10(−7).
2	RES	There was no significant effect from individual GSTM1, GSTT1 or mEH genotypes in workers exposed to <150 p.p.b.
3	OBJ	1,3-Butadiene (BD), which is used to make styrene-butadiene rubber, is a potent carcinogen in mice and a probable carcinogen, associated with leukemia, in humans.
4	CON	These results suggest that the alteration of the NMPs may be involved in DEN-induced hepatocarcinogenesis.
5	METH	Two weeks after a single dose of DEN (200 mg/kg, intraperitoneally), rats were given annatto extract at dietary levels of 0, 0.03, 0.1 and 0.3% or phenobarbital sodium at 0.05% as a positive control for 6 weeks.
6	RES	The total number of thyroid-follicular cells was not significantly increased by MEI treatment.

5 Conclusion

To summarize, in this article, we investigate canonical and state-of-the-art clustering algorithms for IS analysis of biomedical literature. This is the first study of unsupervised IS analysis that uses sophisticated unsupervised learning techniques that may be

applicable in real-world tasks. We obtain high performance especially for identifying objective, result and conclusion sentences. We further demonstrate that unsupervised learning brings new insights into IS analysis and allows for the possibility of discovering novel information categories.

We have demonstrated the potential of clustering and have shown that clustering can detect valuable new information in data. In the future, we would like to pursue unsupervised learning further and learn feature representations as well, rather than pre-defining them [e.g. similar to Socher *et al.* (2011, 2013)]. Other areas we want to explore are probabilistic clustering algorithms such as LDA-style topic models (Blei *et al.*, 2003) and clustering algorithms specifically designed for high-dimensional data (Kriegel *et al.*, 2009).

Funding

This work was supported by the Royal Society (UK), the Swedish Research Council, FAS (Sweden) and an Engineering and Physical Sciences Research Council (EPSRC) doctoral training grant (to D.K.).

Conflict of Interest: none declared.

References

Agarwal,S. and Yu,H. (2009) Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, **25**, 3174–3180.

Amigó,E. *et al.* (2009) A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, **12**, 461–486.

Blake,C. (2009) Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *J. Biomed. Inform.*, **43**, 173–189.

Blei,D.M. *et al.* (2003) Latent dirichlet allocation. *J. Machine Learn. Res.*, **3**, 993–1022.

Chapman,W. and Cohen,K.B. (2009) Current issues in biomedical text mining and natural language processing. *J. Biomed. Inform.*, **5**, 757–759.

Contractor,D. *et al.* (2012) Using argumentative zones for extractive summarization of scientific articles. In: *Proceedings of the International Conference on Computational Linguistics (COLING)*, ACL, Mumbai, India, pp. 663–678.

Curran,J.R. *et al.* (2007) Linguistically motivated large-scale nlp with c&c and boxer. In: *ACL '07 Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions,* ACL, Prague, Czech Republic, pp. 33–36.

Dempster,A. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B*, **39**, 1–38.

Dhillon,I.S. and Modha,D.M. (2001) Concept decompositions for large sparse text data using clustering. *Machine Learn.*, **42**, 143–175.

Dhillon,I. *et al.* (2004) Kernel k-means, spectral clustering and normalized cuts. In: *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Seattle, WA, pp. 551–556.

Dhillon,I. *et al.* (2005) A fast kernel-based multilevel algorithm for graph clustering. In: *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Chicago, IL, pp. 629–634.

Dhillon,I. *et al.* (2007) Weighted graph cuts without eigenvectors: a multilevel approach. *IEEE Trans. Pattern Anal. Machine Intell.*, **29**, 1944–1957.

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

Guo,Y. *et al.* (2011a) A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics.*, **69**. http://www.biomedcentral.com/1471-2105/12/69.

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

Guo,Y. *et al.* (2013) Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review. *Bioinformatics.*, **29**, 1440–1447.

Harmston,N. *et al.* (2010) What the papers say: text mining for genomics and systems biology. *Hum. Genomics*, **5**, 17–29.

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

Jones,K. (1972) A statistical interpretation of term specificity and its application in retrieval. *J. Doc.*, **28**, 11–21.

Kriegel,H.-P. *et al.* (2009) Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowledge Discov. Data*, **3**, 1–58.

Liakata,M. *et al.* (2010) Corpora for the conceptualisation and zoning of scientific papers. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association, Valletta, Malta, pp. 2054–2061.

Lin,J. *et al.* (2006) Generative content models for structural analysis of medical abstracts. In: *HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*. ACL, New York, NY, pp. 65–72.

Lippincott,T. *et al.* (2011) Exploring subdomain variation in biomedical language. *BMC Bioinformatics*, **12**, 212.

MacQueen,J.B. (1967) Some methods for classification and analysis of multivariate observations. In: Le Cam,L.M. and Neyman,J. (eds) *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, pp. 281–297.

McDonald,D. *et al.* (2012) Value and benefits of text mining. *Technical report* **811**, JISC.

Mihăilă,C. *et al.* (2012) Analysing entity type variation across biomedical subdomains. In: Ananiadou,S. *et al.* (eds.) *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012)*.

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

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

Neal,R. and Hinton,G. (1999) A view of the EM algorithm that justifies incremental, sparse, and other variants. In: M.I.,Jordan, (ed.) *Learning in Graphical Models*. MIT Press, Cambridge, MA. pp. 355–368.

Rosenberg,A. and Hirschberg,J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, ACL, Prague, Czech Republic, pp. 410–420.

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

Santorini,B. (1990) *Part-of-speech tagging guidelines for the penn treebank project (3rd revision). Technical report MS-CIS-90-47*, University of Pennsylvania.

Simpson,M. and Demner-Fushman,D. (2012) Biomedical text mining: a survey of recent progress. In: Aggarwal,C.C. and Zhai,C.X. (eds) *Mining Text Data*. Springer Science+Business Media LLC, New York, Philadelphia, pp. 465–517.

Socher,R. *et al.* (2011) Parsing natural scenes and natural language with recursive neural networks. In: *The 28th International Conference on Machine Learning (ICML)*, Bellevue, WA, pp. 129–136.

Socher,R. *et al.* (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Seattle, WA, pp. 1631–1642.

Sollaci,L.B. and Pereira,M.G. (2004) The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *J. Med. Libr. Assoc.*, **92**, 364–367.

Sun,L. and Korhonen,A. (2009) Improving verb clustering with automatically acquired selectional preference. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),* ACl, Suntec, Singapore, pp. 638–647.

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

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

Teufel,S. *et al*. (1999) An annotation scheme for discourse-level argumentation in research articles. In: *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, ACL, Bergen, Norway, pp. 110–117.

Teufel,S. *et al*. (2009) Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),* ACL, Suntec, Singapore, pp. 1493–1502.

van Rijsbergen,C. (1974) Foundation of evaluation. *J. Doc.*, **30**, 365–373.

Varga,A. *et al*. (2012) Unsupervised document zone identification using probabilistic graphical models. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, European Language Resources Association, Istanbul, Turkey, pp. 1610–1617.

Webber,B. *et al*. (2011) Discourse structure and language technology. *Nat. Lang. Eng.*, **18**, 437–490.

Wilbur,W.J. *et al*. (2006) New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, **7**, 356.

Zhong,S. (2005) Efficient online spherical k-means clustering. In: *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN 2005)*, Montreal, Canada. pp. 3180–3185.