Biomedical Informatics Insights

ORIGINAL RESEARCH

Recognizing Scientific Artifacts in Biomedical Literature

Tudor Groza, Hamed Hassanzadeh and Jane Hunter

School of ITEE, University of Queensland, Australia. Corresponding author email: jane@itee.uq.edu.au

Abstract: Today's search engines and digital libraries offer little or no support for discovering those scientific artifacts (hypotheses, supporting/contradicting statements, or findings) that form the core of scientific written communication. Consequently, we currently have no means of identifying central themes within a domain or to detect gaps between accepted knowledge and newly emerging knowledge as a means for tracking the evolution of hypotheses from incipient phases to maturity or decline. We present a hybrid Machine Learning approach using an ensemble of four classifiers, for recognizing scientific artifacts (ie, hypotheses, background, motivation, objectives, and findings) within biomedical research publications, as a precursory step to the general goal of automatically creating argumentative discourse networks that span across multiple publications. The performance achieved by the classifiers ranges from 15.30% to 78.39%, subject to the target class. The set of features used for classification has led to promising results. Furthermore, their use strictly in a local, publication scope, ie, without aggregating corpus-wide statistics, increases the versatility of the ensemble of classifiers and enables its direct applicability without the necessity of re-training.

Keywords: scientific artifacts, conceptualization zones, information extraction

Biomedical Informatics Insights 2013:6 15–27

doi: 10.4137/BII.S11572

This article is available from http://www.la-press.com.

Introduction

Today's search engines and digital libraries offer little or no support for discovering those scientific artifacts (hypotheses, supporting/contradicting statements, findings) that form the core of scientific written communication. Building a view over the supporting or contradicting statements for a particular hypothesis, or summarizing the findings associated with it, is currently extremely difficult and time consuming. For example, consider the hypothesis: "Human apolipoprotein E4 alters the amyloid-β 40:42 ratio and promotes the formation of Cerebral Amyloid Angiopathy." Searching directly for this text in PubMed (currently hosting over 22 million articles) yields only the article that contains this exact hypothesis in its title. No other publications that might discuss it, support it, or contradict it are being returned. Furthermore, from a scientific perspective, it is important to differentiate between the different states the nature of knowledge may take. For example, the statement "aromatic hydrocarbon receptor (AhR) agonists suppress B lymphopoiesis" represents a fact, according to Jensen et al,[1] while in the context of the same article, the statement "two prototypic AhR agonists, 7,12-dimethylbenz [a]anthracene (DMBA) and 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) alter stromal cell cytokine responses" is a hypothesis that requires investigation (the example has been adapted from the ART corpus).[2]

Over the course of the last ten years, most research has focused on mining and analyzing concepts (or named entities) captured within such scientific artifacts and more predominantly on genes, proteins, and their inherent relations.[3–6] Systems like GoPubMed,[7] for example, can find articles relatively easily that contain concepts mentioned in our initial hypothesis example, or in the body of the publications. A simple test on PubMed reveals a high number of hits for the individual concepts: 3,668 hits for Human Apolipoprotein E4; 23,983 hits for amyloid-β; and 1,783 hits for Cerebral Amyloid Angiopathy. However, it is almost impossible to go through all of these hits in order to locate the supporting or contradicting statements. In addition, we currently have no means of comparing such statements to detect gaps between the accepted knowledge and newly emerging knowledge (ie, paradigm shifts),[8] as a means for tracking the evolution of hypotheses from incipient phases to maturity or decline.

Scientific artifacts, spanning within and across multiple publications, provide a rhetorical structure to knowledge and enable the analysis of trends and evolving general patterns. They have been modeled over time via a varied series of rhetorical and argumentation schemes, some of which focus on a rather coarse structure,[9] other on a finer-grained structure with an emphasis on discourse,[2,10] argumentation,[11] or diverse linguistic theories, such as the cognitive theory of discourse representation[12] or the rhetorical structure of text theory.[13] Additionally, they represent the key element in fulfilling the vision of nanopublications[14]—ie, lowering the granularity of the published information to its most atomic form, thus crystallizing the knowledge in the most compact and coherent manner and enabling a richer and more meaningful processing and integration.

Recently, however, research on discourse analysis with a goal of automatically recognizing scientific artifacts carrying a rhetorical role has become more prominent (see for example the outcomes of the ACL 2012 Workshop on Detecting Structure in Scholarly Discourse).[15] As discussed by Liakata et al,[16] there are three directions that have emerged in this area: (i) sentence or zone classification according to a predefined annotation scheme;[16–20] (ii) detection and analysis of speculative language and hedging;[21–23] and (iii) sentence classification according to a multi-dimensional scheme for annotating biological events.[24–26]

In this article, we focus on the first of the directions listed above: the recognition of scientific artifacts within publications based on an existing annotation scheme. Taking into account the local, single publication perspective, we aim to recognize hypotheses, as well as the other scientific artifacts that contextualize and crystallize them, and relate them to other works. More concretely, we target the recognition of five types of statements: (i) hypotheses (HYP)—conjectures on novel ideas/investigations/trends; (ii) motivation (MOT)—statements that provide the context and reasons behind hypotheses; (iii) objectives (OBJ)—propositions that transform hypotheses in measurable goals; (iv) background (BAC)—aspects describing existing information on the topic of the hypothesis; and (v) findings (FIN)—conclusions or observations pertaining to the initial hypothesis. These statements could also act as scaffolding for a structured abstract of the corresponding manuscript.

The process of automatically recognizing scientific artifacts in biomedical publications is particularly challenging. In addition to the inherently complex nature of the task, as the interpretation of what is or is not a hypothesis or a motivational statement is fairly subjective, the domain is very poor in annotated resources. Currently, there is in principle a single openly published corpus of annotated scientific artifacts—the ART corpus,[2,27] which focuses on chemistry and biochemistry. Consequently, in order to achieve our goals, we adapted the CoreSC (Core Scientific Concepts)[10] scheme (used to annotate the ART corpus) for our needs. At the same time, this has also provided us with a common ground for comparing our results to existing research.

The text mining field, within or outside the biomedical scope, consists of a wealth of algorithms and methods, which can usually be classified into two main categories: rule-based methods and (statistical) Machine Learning methods. Rule-based approaches achieve satisfactory results, in particular in Bio-NER (Biomedical Named Entity Recognition) tasks such as gene or protein mentioning.[28,29] They rely on dictionaries, thesauri, and manually crafted rules to perform exact or partial matching. Unfortunately, such methods are not appropriate for recognizing scientific artifacts due to the ambiguous and complex nature of their structure. One could probably envision a method that combines several shallow and deep neuro-linguistic programming (NLP) techniques to produce a series of cascaded transducers. However, the ratio between the amount of manual work required and the flexibility of the end results is not favorable.

On the other hand, Machine Learning (ML) techniques have proved to perform well, both in Bio-NER tasks, as well as in the recognition of scientific artefacts.[16,19,25,30] They are fairly robust and versatile, and capable of detecting patterns that are hard to encode in rules. The main drawback of the ML methods is the necessity of training data, which should contain, in principle, a fair distribution of examples for each of the target classes.

Lately, the focus has shifted towards hybrid methods, either by exploiting the best aspects of both the above-mentioned types of techniques, eg, by using rules to bootstrap the ML classification process,[31,32] or by aggregating several ML techniques into cascaded classifiers.[3] The latter has showed promising results in Bio-NER contexts. Consequently, we have followed this direction and designed our recognition process as a sentence-based classification via an ensemble of four classifiers. The finer-grained annotation level required to capture the content and conceptual structure of a scientific article[16] has motivated our choice of sentence-based classification. This article aims to bring the following contributions, envisioned to support other researchers working on the topic, as well as to enable the development of automated mechanisms for building argumentative discourse networks or for tracking the evolution of scientific artifacts: (i) we propose, develop and evaluate a hybrid Machine Learning ensemble, as opposed to the existing research that makes use of a single classification technique; and (ii) we use classification features built strictly from a local, publication perspective, as opposed to corpus-wide statistics used within all the other approaches. This last aspect can make the difference between a model biased towards the domain/corpus used for training and one that makes use of more generic elements and hence displays an increased versatility. Our experimental results show that such a model achieves accuracy comparable to the state of the art, even without relying on corpus-based features.

Methods

Data

As mentioned in the previous section, our goal is to recognize and classify five types of sentences. To create training and test data for classification, we have adapted the ART corpus to serve our specific goals. The ART corpus[2,27] consists of 256 articles from chemistry and biochemistry, annotated according to the CoreSC scheme at sentence level. The CoreSC annotation scheme[10] defines 11 types of general scientific concepts that can be found in publications, some of which may have attributes attached to them. These attributes denote the difference between the aspects related to the publication under scrutiny and those pertaining to pre-existing work (ie, New vs. Old). In addition, they may also signal certain positive or negative elements related to the current or previous work (ie, Advantage vs. Disadvantage).

Table 1 provides an overview of the CoreSC types and their mapping to our categories. There are a few notes worth mentioning here. Firstly, we have merged the original GOA and OBJ categories under a single Objective (OBJ) class as they both refer to aspects

Table 1. The CoreSC annotation scheme[10] and its adaptation to our goals.

Category	Description	Re-purposed category
Hypothesis (HYP)	A statement that needs to be confirmed by experiments and data	Hypothesis (HYP)
Motivation (MOT)	The reasons supporting the investigation	Motivation (MOT)
Background (BAC)	Accepted background knowledge and previous work	Background (BAC)
Goal (GOA)	The target state of the investigation	Objective (OBJ)
Object (OBJ)	The main theme or product of the investigation	Objective (OBJ)
Method-New (MET)	Means by which the investigation is carried out and the goal is planned to be achieved	Out of scope (O)
Method-Old (MET)	A method proposed in previous works	Background (BAC)
Experiment (EXP)	An experimental method	Out of scope (O)
Model (MOD)	A description of the model or framework used in the investigation	Out of scope (O)
Observation (OBS)	A statement describing data or phenomena encountered during the investigation	Finding (FIN)
Result (RES)	A factual statement about the outcome of the investigation	Finding (FIN)
Conclusion (CON)	A statement that connects observations and results to the initial hypothesis	Finding (FIN)

Notes: The left column presents the original annotation scheme used in the ART corpus, while the right column shows the transformations we have applied to re-purpose this scheme to our goals.

that take the hypothesis from a conjecture to a measurable goal. Secondly, we have also merged the original OBS, RES, and CON into a single Finding (FIN) category. Finally, it can be observed that we make a distinction between Method-New and Method-Old. The CoreSC scheme provides this option, however previous classification methods that have used the ART corpus, such as Liakata et al,[16] merge the two under a single Method category. While the method proposed by a particular paper (ie, Method-New) is not of interest to us, we are interested in capturing previous work (ie, Method-Old) as this is also part of the background knowledge. All the other categories have been marked as outside the scope of our study.

Table 2 lists the statistics of the original corpus together with the new statistics that emerged from re-purposing the original categories into our new classes. Overall, the repurposing has an obvious beneficial effect on those categories resulting from a merge. For example, GOA + OBJ now cover 4.36% rather that 1.45% and 2.90% respectively. However, from a distribution perspective, the corpus is now heavily skewed, in particular, towards the FIN category, which covers 43.71% of the total number of sentences.

Classifiers

Our method relies on an ensemble of four classifiers, trained and tested on the above-described data via nine-fold cross validation. Four divergent classifiers have been trained, each using a different package. Two of the classifiers were represented by

Table 2. The coverage of the re-purposed classes from the ART corpus.[2]

Category	No. sentences	Coverage	Re-purposed category	No. sentences	Coverage
HYP	780	1.95%	HYP	780	1.95%
MOT	541	1.35%	MOT	541	1.35%
BAC	7,606	19.05%	BAC	10,229	25.62%
MET (old)	2,623	6.57%			
GOA	582	1.45%	OBJ	1,743	4.36%
OBJ	1,161	2.90%			
MET (new)	1,658	4.15%	O	9,172	22.97%
EXP	3,858	9.66%			
MOD	3,656	9.15%			
OBS	5,410	13.55%	FIN	17,450	43.71%
RES	8,404	21.05%			
CON	3,636	9.10%			

Notes: Naturally, by merging some of the initial types, our re-purposed classes gained more weight in the overall corpus distribution.

Conditional Random Fields[33] chunkers, trained using the MALLET[34] and the CRF++ (http://crfpp.google-code.com/) packages. Both are freely available and were used to train forward parsing chunkers. MALLET was trained without feature induction and with dense weights, while the CRF++ chunker was trained with the hyperparameter set to 3.5. The ensemble has been completed with two Support Vector Machines[35] classifiers provided by the YamCha package.[36] Both are multi-class classifiers trained using a second-degree polynomial kernel. The difference between the two was the training method: one was trained using the one vs. one method, while the other using the one vs. all method.

To combine the results from all four classifiers, we used different aggregation strategies. In addition to individual classification, we also experimented with:

- Set operations—results of the individual classifiers have been treated as sets, which have then been combined using direct or composite operations. Direct operations refer to combinations of pairs of single classification results (eg, YamCha1vs1 $\cup$ MALLET), while composite operations refer to combinations of direct operations (eg, (CRF++ $\cup$ YamCha1vs1) $\cap$ (MALLET $\cup$ YamCha1vsAll)). In both cases, we have used union and intersection as basic set operators.
- Simple majority voting—results from the individual classifiers have been treated as votes for a particular class. These votes were then counted and the winning result was established via a simple majority. The veto option was used when a tie occurred or when the classifiers were in complete disagreement. In these cases, the veto owner established the final result.

Classification features

We used five types of features to build the classifiers, detailed below.

Structure-based features place the sentence under scrutiny in the overall picture provided by the linear structure of the publication. Four such features were used:

- Section-based publication placement—the number of the top-level section that contains the sentence. The abstract was considered to be section 0.
- Subsection-based section placement—the number of the subsection within a top-level section that contains the sentence.
- Relative section placement—the paragraphs within a section have been split into four parts: (1) the first paragraph; (2) the last paragraph; (3) the first half of the remaining paragraphs except the first and the last; and (4) the second half of the remaining paragraphs. Certain types of sentences tend to be present more within the first or last paragraphs (usually MOT/OBJ and FIN), while others, which describe the main ideas of the section, are part of the middle paragraphs.
- Relative paragraph placement—similar to the relative section placement, however seen at the paragraph level. The split was done between the first two sentences, the last to sentences and the rest (first half and second half).

The first feature is similar to the SectionId feature used by Liakata et al.[16]

Linguistic features, focuses on the linguistic aspects that characterize the sentence's tokens. Some features take discrete values that encode counts, such as no, 1, 2, 3+, while others are signaling flags, they take yes/no values.

- Part of speech tags: adjectives, adverbs, coordinating conjunctions, and pronouns.
- Verb information: type, tense, voice, negation.
- Verb classes: the ten verb classes adopted from Guo et al.[37]
- Hedging: dictionary-based flag denoting the presence of hedging (eg, seem, may, sometimes, possibly) adapted from Medlock et al.[22] and Szarvas et al.[23]
- Rhetorical relations: flags denoting the presence of 13 rhetorical relations adapted from the Rhetorical Structure of Text theory.[38] These are: antithesis, conjunction, circumstance, concession, condition, cause, consequence, elaboration, evidence, means, preparation, purpose, and restatement. Their actual presence is signaled by a series of cue-phrases compiled by Marcu.[39]

Other miscellaneous features take into account elements usually present within scientific publications:

- Figures/tables/citations: the presence of any of these elements in the sentence

- Topics: We have performed topic modeling using Latent Dirichlet Allocation (LDA)[40] on each publication to compile the five most probable unigram and n-gram topics.

Distribution features capture the distribution of a certain feature at paragraph level and then encode, via discrete values, the coverage of that feature in the sentence under scrutiny. We experimented with four intervals: 0%–25%, 25%–50%, 50%–75%, and 75%–100%. The actual features considered were: (i) rhetorical relations, (ii) verb classes, (iii) topics, (iv) adverbs, (v) pronouns, and (vi) citations. Sentence context (with variable window) includes the features of the neighboring sentences. The size of the window specifies the number of adjacent sentences to be considered. We experimented with sizes 1 to 5.

One important aspect of the features we used for classification is that they are not dependent on the corpus. Independently of the role or target, all features compute their values strictly in the context of the publication. This is a major difference between our approach and all the other methods described in the literature. For example, a third of the features used by Liakata et al[16] take into account measures compiled at corpus level (some similar to the concept of "information content").[41] For example, topical n-grams compiled at corpus level (instead of publication level as we did), grammatical triples compiled again on the entire corpus, or section headings usually encountered in biomedical publications. According to both the single feature classification, as well as the leave-one-out (LOOF) tests, these features had a decisive role in achieving a good efficiency. However, it remains unclear whether the same model could be applied in another domain, without re-training it, or altering some of the features to fit the domain.

Results

As previously mentioned, all experiments detailed below were carried out on the re-purposed ART corpus. We performed nine-fold cross-validation with stratification and averaged the results. In the following sections, we discuss the results achieved by the individual classifiers, as well as those achieved via different aggregation techniques.

Individual classifier results

Table 3 lists the results achieved by the individual classifiers. We can observe that, with the exception of

Table 3. Experimental results of the individual classifiers.

	HYP			MOT			BAC			OBJ			FIN		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
MALLET	14.84	1.77	3.16	15.74	0.52	0.98	59.41	46.60	51.86	49.39	12.24	19.41	65.23	86.21	74.20
CRF++	18.87	8.38	10.95	20.47	2.59	4.54	60.93	59.21	**59.88**	52.93	29.15	**37.40**	71.78	82.98	**76.93**
YamCha1vs1	19.70	11.31	**13.71**	17.12	9.07	**11.72**	54.47	60.96	57.35	43.44	25.84	32.30	73.20	77.91	75.43
YamCha1vsAll	12.08	6.05	7.79	12.69	7.63	9.45	55.93	57.43	56.50	34.44	26.01	29.41	71.29	78.11	74.49

Notes: Bold numbers denote the best F1 score achieved for the particular class. We can observe that YamCha1vs1 outperforms the other classifiers in the first two classes, while CRF++ achieves the best results in the latter three.

one category, MOT, CRF++ and YamCha1vs1 have a similar performance, the average difference between them being approximately 3%. CRF++ performs better on the classes that are better represented in the corpus (BAC, OBJ, and FIN), while YamCha1vs1 performs better on the more problematic classes (HYP and MOT). The largest difference in performance—of almost 7%—is present in the MOT class, which has the lowest coverage in the corpus. Similar to CRF++, the other two classifiers—MALLET and YamCha1vsAll—have a poor performance on the first two classes, and approach to the best score in the classes well represented in the corpus.

The precision of CRF++ should be noted. It is consistent throughout all classes and is either the best or the second best (at a very small difference), independently on whether CRF++ achieves the best score in that particular class. For example, in the MOT class, CRF++ has a 20% precision, although the final F1 score is almost the lowest (4%). This leads to the conclusion that CRF++ is poor at discriminating classes that don't have a good coverage in the corpus. However, when it does find those classes, there are very high chances for these to be correctly classified.

Set operations

In order to improve the individual classification results, we attempted to aggregate them via different direct or combined set operations. Direct set operations represent the aggregation of pairs of single classifiers, eg, CRF++ ∪ MALLET, or YamCha1vs1 ∩ CRF++, while combined set operations aggregate pairs of direct set operations, eg, (CRF++ ∪ MALLET) ∩ (YamCha1vs1 ∪ YamCha1vsAll). In both cases, we used union and intersection as atomic operators.

A series of results of direct set operations are listed in Table 4. As expected, the best scores were achieved by sets that included at least one of the two best individual classifiers, ie, CRF++ and YamCha1vs1. In practice, the results follow the same trend as in the individual classification. In the first two categories, HYP and MOT, the union of the best scoring individual classifiers (CRF++ and YamCha1vs1, and YamCha1vs1 and YamCha1vsAll,) achieves the highest scores (15.30% and 13.44% respectively). Similarly, in the OBJ and FIN, the union of CRF++ and YamCha1vs1 performs the best—39.42% and

Table 4. Experimental results of the direct set operations (see list below).

	HYP			MOT			BAC			OBJ			FIN		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DS_OP1	17.14	15.12	**15.30**	17.53	10.54	13.05	51.46	70.88	59.47	43.91	36.05	**39.42**	68.74	89.63	**77.76**
DS_OP2	14.48	12.39	13.59	16.74	9.60	12.06	49.92	66.67	56.89	42.54	29.58	34.76	64.28	92.64	75.86
DS_OP3	14.48	12.39	12.75	14.50	9.82	11.63	52.66	69.68	**59.85**	37.44	38.83	37.93	67.27	90.65	77.18
DS_OP4	14.80	13.75	13.72	13.98	13.19	**13.44**	50.72	66.27	57.29	35.02	38.40	36.47	68.93	86.02	76.48

Notes: Bold numbers denoted the best F1 score achieved. As expected, the best results have been achieved by those set operations that included the two best individual classifiers—CRF++ and YamCha1vs1. Direct set operations: **DS_OP1**: CRF++ ∪ YamCha1vs1; **DS_OP2**: MALLET ∪ YamCha1vs1; **DS_OP3**: CRF++ ∪ YamCha1vsAll; **DS_OP4**: YamCha1vs1 ∪ YamCha1vsAll.

77.76% respectively. The only exception from this pattern can be found in the BAC category, where this union achieves the second best score, at a minimal difference of 0.38%, behind the union of CRF++ and YamCha1vsAll.

Union, as an operator, has a positive impact on the recall and a negative impact on precision—more classes being found dilutes the correctness of the classification. This can be clearly seen in the above described results: (i) the first two categories feature a consistent 4% increase in recall, associated with a decrease in precision of 2% to 4%; and (ii) in the last three categories this is accentuated, as the increase in recall is of almost 10%, achieving a maximum of 92.65% recall in the union of MALLET and YamCha1vs1. Intersection, on the other hand, has the opposite effect: it improves precision (only the correct scores are retained from both classifiers) at the expense of recall. In our particular case, this has led to extremely low scores of under 1% F1 in the first two categories and moderate, yet lower, scores in the last three categories (not listed here). The only notable results are with respect to precision, which peaked at 30% in HYP, 74% in BAC, and 78% in FIN.

The results for the second type of set operations are listed in Table 5. Again, making the distinction between the two groups of classes, we can observe that the poor performance of MALLET is directly reflected in the joint set results in the first two classes, while in the last three classes—where the classifiers have a similar performance—the results reflect a good complementarity between the approaches. It should be noted that we attempted other different set combinations, eg, by using intersection first and then union, or by combining results in a more serial manner such as combining two approaches then intersecting or adding the results to a third one, and so on. However, none of them worked better than the ones listed in Table 5, and thus, we have not included them in the discussion.

Voting

The last aggregation technique used was a simple majority voting. Results are listed in Table 6, where the first column represents the veto owner in case of a tie or complete disagreement. We can clearly draw a mapping between these results and those of the individual classifiers, with the remark that voting has, in

Table 5. Experimental results of the paired set operations (see list below)—best F1 scores are marked in bold.

	HYP			MOT			BAC			OBJ			FIN		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PS_OP1	23.17	4.56	7.34	25.66	1.31	2.45	63.81	55.64	59.21	58.52	25.16	**34.99**	72.07	83.42	77.28
PS_OP2	26.82	6.14	**9.59**	22.99	4.41	**7.22**	59.31	60.90	**59.95**	53.80	23.90	32.96	70.72	87.37	78.13
PS_OP3	21.49	4.24	6.93	21.54	3.69	6.18	58.93	60.47	59.53	52.94	21.95	30.80	71.16	86.89	**78.20**

Notes: Here we observe the effect induced by the best individual classifiers on the overall pair efficiency: they complement the less efficient classifiers to achieve the best results (eg, PS_OP2 in HYP, MOT and BAC). Paired set operations: **PS_OP1**: (MALLET ∪ CRF++) ∩ (YamCha1vs1 ∪ YamCha1vsAll); **PS_OP2**: (MALLET ∪ YamCha1vs1) ∩ (CRF++ ∪ YamCha1vsAll); **PS_OP3**: (MALLET ∪ YamCha1vsAll) ∩ (CRF++ ∪ YamCha1vs1).

Table 6. Experimental results of the voting mechanism.

	HYP			MOT			BAC			OBJ			FIN		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
MALLET	22.63	3.13	5.33	5.92	0.40	0.74	62.77	56.36	59.15	55.14	22.72	32.01	70.26	87.60	77.93
CRF++	21.01	5.06	7.87	12.89	1.31	2.36	62.40	58.06	**59.97**	57.12	26.09	**35.67**	71.64	86.64	**78.39**
YamCha1vs1	23.31	6.14	**9.53**	21.55	4.57	**7.37**	59.11	60.08	59.44	54.51	24.28	33.40	72.29	83.85	77.60
YamCha1vsAll	16.41	3.78	5.99	20.06	4.59	7.34	59.55	59.66	59.44	53.42	24.24	33.20	71.63	84.55	77.51

Notes: We can observe that they follow the same pattern as in the case of the individual classification: the classifiers that have achieved the best results in the individual classification perform well also as veto holders in the voting mechanism.

general, a positive effect over precision and a moderately negative effect over recall (these effects are milder than those present in set operations). Similar to the results listed in Table 3, the best scores in the first two categories were achieved with YamCha1vs1 as veto owner (9.53% and 7.37%, respectively) and in the last three categories with CRF++ as veto owner (59.97%, 35.67% and 78.39%). Overall, the voting mechanism displays a behavior very similar to the paired set operations.

Discussion

In order to get a better understanding of the role and importance of the features used for classification, we have performed two additional experiments using CRF++ as a single classifier. In the first experiment, we have trained CRF++ with each individual feature from the overall best model. In the second experiment, we have trained it using a leave-one-out setting, ie, from the best CRF++ model we have left out one feature at a time. Both experiments used a nine-fold cross validation with stratification.

Table 7 lists the F1 scores achieved in the one-feature setting. It can be clearly observed that in the context of the classes that are poorly represented in the corpus, no feature was able to perform a correct classification. The only exceptions are the context features (listed with a "*cf_*" prefix in the table) that have performed surprisingly homogeneously in the OBJ category. On the other hand, in the remaining two classes, BAC and FIN, we are able to identify a series of discriminative features. More concretely, we can observe that some of the structural and miscellaneous features have performed extremely well: *f_citation* (51.85% and 64.24%), *f_citaton_distro* and *f_paperplace* (51.56% and 68.06%) as well as the linguistic features focused on verbs—*f_vbclasses* (10.76% and 63.11%) and *f_verbs* (7.71% and 64.09%). Some of these results are particularly interesting because they account for more than 80% of the final result—eg, *f_citation* in BAC achieves 51.85% of the maximum 59.88% achieved by the entire model, or in FIN, where it scores 64.24% of the maximum 76.95% F1. The above observations lead to two conclusions: (i) citations and the location in the overall paper structure are key elements in distinguishing background and findings sentences, and (ii) background and findings sentences are characterized by fairly uniform verb patterns.

Table 7. F1 scores for one feature classification using the best CRF++ model.

Feature	HYP	MOT	BAC	OBJ	FIN
f_adjectives	0.00	0.00	0.00	0.00	61.05
f_cc	0.00	0.00	0.00	0.00	61.05
f_figs	0.00	0.00	0.00	0.00	61.05
f_pronouns	0.00	0.00	0.00	0.00	61.05
f_relsectplace	0.00	0.00	0.00	0.00	61.05
f_sectionplace	0.00	0.00	0.00	0.00	60.98
f_topics_distro	0.00	0.00	0.00	0.00	61.26
f_verbs	0.00	0.00	**7.71**	0.00	**64.09**
f_adverbs	0.00	0.00	0.00	0.00	61.05
f_citation	0.00	0.00	**51.84**	0.00	**64.27**
f_hedging	0.00	0.00	0.00	0.00	61.05
f_pronouns_distro	0.00	0.00	0.00	0.00	61.05
f_rhetrel	0.00	0.00	0.02	0.00	61.05
f_tables	0.00	0.00	0.00	0.00	61.05
f_vbclasses	0.00	0.00	**10.76**	0.00	**63.11**
f_adverbs_distro	0.00	0.00	0.00	0.00	61.33
f_citation_distro	0.00	0.00	**51.84**	0.00	**64.27**
f_paperplace	0.00	0.00	**51.56**	0.00	**68.06**
f_relparplace	0.00	0.00	0.00	0.00	61.05
f_rhetrel_distro	0.00	0.00	1.37	0.00	61.40
f_topic	0.00	0.00	0.00	0.00	61.05
f_vbclasses_distro	0.00	0.00	**21.13**	0.00	**66.00**
cf_adverbs	0.00	0.00	0.00	13.78	60.87
cf_hedging	0.00	0.00	0.00	13.78	61.17
cf_pronouns	0.00	0.00	0.00	13.78	61.17
cf_rhetrel	0.00	0.00	0.32	13.78	61.09
cf_vbclasses	0.00	0.00	**12.35**	13.78	**62.35**
cf_verbs	0.00	0.00	4.62	13.78	**63.33**

Notes: Bold numbers denote the most interesting F1 scores achieved by diverse features. We can observe that only the well represented classes in the corpus have associated successful F1 scores.

Table 8. F1 scores for "leave one feature out" classification using the best CRF++ model.

Feature	HYP	MOT	BAC	OBJ	FIN
f_adjectives	10.81	3.16	59.55	37.14	76.60
f_cc	9.65	3.27	59.67	37.11	76.57
f_figs	9.80	2.97	59.13	**37.58**	76.25
f_pronouns	**11.59**	4.06	59.62	37.39	76.52
f_relsectplace	8.20	2.67	59.36	34.99	76.10
f_sectionplace	10.73	4.09	59.64	37.21	76.46
f_topics_distro	7.94	1.94	59.77	35.35	76.58
f_verbs	9.61	2.88	59.51	34.13	76.02
f_adverbs	10.36	3.54	59.61	36.64	76.61
f_citation	10.39	**4.96**	59.68	36.92	76.63
f_hedging	6.91	3.51	59.68	36.03	76.62
f_pronouns_distro	10.63	3.50	59.85	**37.82**	76.68
f_rhetrel	**11.12**	4.34	**59.93**	36.86	76.66
f_tables	10.79	**5.03**	59.50	36.98	76.23
f_vbclasses	10.03	4.30	59.72	37.40	76.64
f_adverbs_distro	9.80	**4.59**	59.63	**37.54**	76.68
f_citation_distro	10.36	3.89	58.90	36.26	76.25
f_paperplace	9.32	3.48	56.94	35.71	74.45
f_relparplace	10.45	3.64	59.53	35.33	76.26
f_rhetrel_distro	9.10	3.23	59.97	36.45	76.63
f_topic	9.64	**4.61**	59.35	37.17	76.30
f_vbclasses_distro	10.03	4.25	59.54	36.10	76.36
cf_adverbs	9.53	**4.94**	59.60	**37.78**	76.74
cf_hedging	9.68	4.39	59.75	36.89	76.77
cf_pronouns	10.51	4.10	59.81	36.42	76.62
cf_rhetrel	10.11	**4.83**	59.87	36.75	76.69
cf_vbclasses	10.92	4.43	59.70	37.27	76.42
cf_verbs	10.29	2.61	59.77	**37.58**	76.35

Notes: Similar to the one feature classification, bold numbers denote the most interesting F1 scores achieved, this time, by leaving the corresponding feature out from the classification model. Interestingly, in some cases the F1 score is higher than in the case of the overall F1 score achieved by the final model.

The second experiment complements the first by showing the impact of leaving a particular feature out of the model. Table 8 lists the F1 scores achieved by models that leave out the feature present in the first column. In principle, it seems that the large majority of features have a small negative impact on the performance of the model. However, due to the fact that we are dealing with several classes and we had to reach a compromise at the model level, discarding some of the features actually leads to a positive effect on the performance. Examples of this phenomenon are emphasized in Table 8 with bold font. For example, leaving out *f_pronouns* or *f_rhetrel* in the HYP category leads to an increased F1 of 11.59% and 11.12%, respectively; the best F1 score achieved by CRF++ in this category is 10.95%. Similarly, leaving out *f_tables* or the context feature *cf_rhetrel* in the MOT category increases the F1 score to 5.03% and 4.83% respectively (from an initial 4.54%). The other categories contain similar examples, however, in some cases the increase in F1 is marginal, ie, 0.05% in BAC (by leaving out *f_rhetrel*) and up to 0.42% in OBJ (by leaving out *f_pronouns_distro*).

We have continued the analysis of the classifiers behavior by looking at the confusion matrix according to the best CRF++ model (see Table 9). The HYP class is confused in the vast majority of cases with the FIN class because the language used to describe findings is fairly similar to the one used to state hypotheses. The FIN class is, in general, problematic as it aggregates three types of statements (see Table 2—observations, results, and conclusions), each of which has similarities with the other classes. Consequently, with the exception of the MOT class, this class accounts for most of the class-to-class confusions (see BAC or OBJ). A final remark can be noted about the O class,

Table 9. Classification confusion matrix based on the best CRF++ model.

	HYP	MOT	BAC	OBJ	FIN	O
HYP	**249**	1	83	4	398	45
MOT	3	**14**	396	19	90	19
BAC	39	31	**6,208**	123	2,409	1,420
OBJ	0	5	368	**510**	572	288
FIN	58	0	1,399	160	**14,665**	1,168

Note: Bold numbers denote correctly classified instances.

those sentences that are outside the scope of our study. We can observe that, when compared to the confusions raised by other classes, the O class has a very small impact. The only noted exception is the BAC class, and this is due to the fact that we have split the original ART MET category into MET-New and MET-Old and used only one of them within the BAC class. Naturally, the choice of features and the coverage of each of these classes in the overall corpus are responsible for the above listed results.

The experiments presented in these sections lead to a series of conclusions. Firstly, hybrid classification methods depend heavily on the individual performance of the underlying classifiers used for aggregation. Subject to their configuration, such ensembles of classifiers are able to exploit the diversity and consistency among the individual elements to reach a final decision. Usually, this decision is better than using a single classifier. This can also be observed in our case (Table 10). The hybrid methods have outperformed the individual classifiers, with the direct set operations performing constantly better. Nevertheless, the efficiency and applicability of such hybrid methods requires consideration on a per use-case basis. Secondly, our experiments have showed that CRF++ performs fairly consistent when

Table 10. Comparative overview of the F1 scores achieved by the different techniques.

	HYP	MOT	BAC	OBJ	FIN
Individual	13.71	11.72	59.88	37.40	76.93
Direct set operations	**15.30**	**13.44**	59.85	**39.42**	77.76
Paired set operations	9.59	7.22	59.95	34.99	78.20
Voting	9.53	7.37	**59.97**	35.67	**78.39**

Notes: Overall, the proposed hybrid methods perform the best, the most consistent aggregation technique being the direct set operations.

the coverage of the target class is reasonable, in addition to achieving excellent precision results in diverse aggregation schemes. Consequently, together with YamCha1vs1, this should always be considered as a foundation for any ensemble.

Related work

As mentioned in the Introduction, the literature consists of several other approaches that have a similar goal. However, in the context of our research, the most relevant work is that of Liakata et al.[16] Consequently, within all our experiments we have tried to follow similar settings, in order to make the two approaches comparable. While it is impossible to compare them directly due to the difference in target classes, we have performed identical experiments as described in the previous section on the 11 classes of interest for Liakata et al (Table 11). Overall, our model has been outperformed on all classes. However, in most cases, with 2 or 3 exceptions, the difference between the originally reported score and the best performance achieved by one of our models (usually the direct set operations) was of only 2%–3%. These are positive results if we consider the difference in the types of features used for classification.

Our models were built using features only in the local context of a publication, while the features contributing mainly to the results of Liakata et al and of the other similar approaches have used information compiled at the corpus level, such as n-grams, grammatical triples, or section titles. We believe this may affect the versatility of the classifier and would require re-training when applied to a different biomedical domain in order to achieve the same performance. Conversely, our classifiers are not biased towards the domain of the training corpus and hence may be directly applied in a different domain without retraining and probably without detracting from the initial accuracy.

In principle, our conjectures throughout the article and conclusions discussed above rely on the assumption that the underlying target domain, while not necessarily a particular one, does fit into the general biomedical area. Biomedical publications, as in the case of any other community, have a particular structure and share a specific language, aspects that are exploited by the models we have trained. This structure and language are most probably different

Table 11. Comparative overview of the classification results on the 11 classes proposed by Liakata et al.[16]

	BAC	CON	EXP	GOA	MET	MOT	OBS	RES	MOD	OBJ	HYP
Liakata et al	62	45	76	28	30	20	51	51	53	34	19
Individual	57	41	73	19	22	11	46	45	34	30	16
Direct set operations	56	**43**	73	20	**26**	14	**48**	**47**	37	**31**	**18**
Paired set operations	57	40	**74**	16	19	8	42	43	37	24	13
Voting	57	40	73	17	20	8	44	43	36	27	17

Notes: Bold numbers denote F1 scores close to the ones obtained by Liakata. Overall, our model performs fairly well, with a few exceptions; the decrease in efficiency being explained by the increased versatility of our classification model.

in, for example, the case of computer science publications, and thus a straightforward generalization of the trained models are not possible. A comprehensive study would be required to enable a clear understanding of the advantages and disadvantages between generalization and efficiency in the process of recognizing scientific artifacts. As part of our future work, we intend to perform such a study, however by starting from the comparison of different sub-domains of the biomedical field.

Conclusions

In this article we have presented a hybrid Machine Learning approach for extracting particular sentences of interest from scientific publications. Our focus has been on hypotheses and their context, ie, motivation, background, objectives, and findings, and represents the first step towards our general goal of extracting and consolidating argumentative discourse networks that span across multiple publications. The approach consists of an ensemble of classifiers trained to perform sentence-based classification.

We have used several aggregation techniques, including direct and paired set operations and simple majority voting. Experimental results have shown the supremacy of hybrid methods over individual classifiers. Our future work will focus on clustering similar hypotheses denoting a generic, abstract theme and finding their associated supporting and contradicting arguments.

Acknowledgements

The author acknowledges use of computing resources from the NeCTAR research cloud (http://www.nectar.org.au). NeCTAR is an Australian Government project conducted as part of the Super Science initiative and financed by the Education Investment Fund. We are also very grateful to Dr. Maria Liakata for providing unrestricted access to the ART corpus and its associated tools.

Author Contributions

Conceived and designed the experiments: TG, HH. Analyzed the data: TG, HH, JH. Wrote the first draft of the manuscript: TG. Contributed to the writing of the manuscript: HH, JH. Agree with manuscript results and conclusions: TG, HH, JH. Jointly developed the structure and arguments for the paper: TG, HH.

Funding

This research has been funded by the Australian Research Council (ARC) under the Discovery Early Career Researcher Award (DECRA)—DE120100508.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

References

1. Jensen BA, Leeman RJ, Schlezinger JJ, Sherr DH. Aryl hydrocarbon receptor (AhR) agonists suppress interleukin-6 expression by bone marrow stromal cells: an immunotoxicology study. *Environ Health*. 2003;2(1):16.
2. Liakata M, Soldatova L. The ART Corpus. *Tech Rep JISC Project Report*, *Aberystwyth University*. UK; 2009.
3. Li L, Fan W, Huang D, Dang Y, Sun J. Boosting performance of gene mention tagging system by hybrid methods. *J Biomed Inform*. 2012;45(1):156–64.
4. Saha S, Sarkar S, Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition. *J Biomed Inform*. 2009;42(5): 905–11.
5. Torii M, Hu Z, Wu C, Liu H. BioTagger-GM: a gene/protein name recognition system. *J Am Med Inform Assoc*. 2009;16(2):247–55.
6. Yang Z, Lin H, Li Y. Exploiting the contextual cues for bio-entity name recognition in biomedical literature. *J Biomed Inform*. 2008;41(4):580–7.
7. Dietze H, Alexopoulou D, Alvers MR, et al. GoPubMed: Exploring PubMed with Ontological Background Knowledge. *Bioinform Syst Biol*. 2009;V: 385–99.
8. Lisacek F, Chichester C, Kaplan A, Sandor A. Discovering paradigm shift patterns in biomedical abstracts: application to neurodegenerative diseases. In *Proc*. *of the 1st International Symposium on Semantic Mining in Biomedicine (SMBM)*. European Bioinformatics Institute, Hinxton, UK; 2005.
9. de Waard A. A pragmatic structure for research articles. In *Proc*. *of the 2nd International Conference on Pragmatic Web*, New York, NY, USA; 2007:83–9.
10. Soldatova L, Liakata M. An ontology methodology and CISP—The proposed core information about scientific papers. *Tech Rep JISC Project Report*, *Aberystwyth University*. UK; 2007.
11. Ciccarese P, Wu E, Wong G, et al. The SWAN biomedical discourse ontology. *J Biomed Inform*. 2008;41(5):739–51.
12. Mancini C, Shum SB. Modeling discourse in contested domains: a semiotic and cognitive framework. *Internat J Human-Comp Stud*. 2006;64(11): 1154–71.
13. Groza T, Handschuh S, Decker S. Capturing rhetoric and argumentation aspects with scientific publications. *J Data Semantics*. 2011;15:1–36.
14. Groth P, Gibson A, Velterop J. The Anatomy of a Nanopublication. *Inf Serv*. 2010;30:51–6.
15. Ananiadou S, van den Bosch A, Sandor A, Shatkay H, de Waard A. Proceedings of the ACL 2012 Workshop on Detecting Structure in Scholarly Discourse (DSSD 2012), Jeju, Korea, Jul 2012. http://www.nactem.ac.uk/dssd (last retrieved 17/01/2013).
16. Liakata M, Saha S, Dobnik S, Batchelor C, Rebholz-Schuhmann D. Automatic recognition of conceptualisation zones in scientific articles and two life science applications. *Bioinformatics*. 2012;28(7):991–1000.
17. Teufel S, Siddharthan A, Batchelor C. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proc of the 2009 Conference on Empirical Methods in Natural Language*, Volume 3, Stroudsburg, PA. USA; 2009:1493–502.
18. Teufel S. The structure of scientific articles: applications to citation indexing and summarization. *CSLI Studies in Computational Linguistics*, Chicago University Press; 2010.
19. Mizuta Y, Korhonen A, Mullen T, Collier N. Zone analysis in biology articles as a basis for information extraction. *Int J Med Inform*. 2006;75(6): 468–87.
20. Teufel S. Argumentative Zoning: Information Extraction from Scientific Text. PhD thesis, *School of Cognitive Science*, *University of Edinburgh*, Edinburgh, UK; 2000.
21. Kilicoglu H, Bergler S. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics.* 2008;9 Suppl 11:S10.
22. Medlock B, Briscoe T. Weakly supervised learning for hedge classification in scientific literature. In *Proc of the 45th Annual Meeting of the ACL*, Prague, Czech Republic. 2007:23–30.
23. Szarvas G. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proc of the ACL-08*: *HLT*, Columbus, Ohio, USA; 2008:281–9.
24. Wilbur WJ, Rzhetsky A, Shatkay H. New directions in biomedical text annotations: definitions, guidelines and corpus construction. *BMC Bioinformatics.* 2006;7:356.
25. Shatkay H, Pan F, Rzhetsky A, Wilbur WJ. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*. 2008;24(18):2086–93.
26. Thompson P, Nawaz R, McNaught J, Ananiadou S. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*. 2011;12:393.
27. Liakata M, Teufel S, Siddharthan A, Batchelor C. Corpora for the conceptualisation and zoning of scientific papers. In *Proc of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta; 2010:2054–61.
28. Hanisch D, Fundel K, Mevissen H-T, Zimmer R, Fluck J. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*. 2005; 6(Suppl 1):S14.
29. Yang Z, Lin H, Li Y. Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature. *Comput Biol Chem*. 2008;32(4):287–91.
30. Li L, Zhou R, Huang D, Liao W. Integrating divergent models for gene mention tagging. In *Proc of the International Conference on Natural Language Processing and Knowledge Engineering*. Dalian, China; 2009:1–7.
31. Pedersen T. Rule-based and lightly supervised methods to predict emotions in suicide notes. *Biomedical Informatics Insights.* 2012;5(Suppl 1): 185–93.
32. McCart JA, Finch DK, Jarman J, et al. Using ensemble models to classify the sentiment expressed in suicide notes. *Biomed Inform Insights*. 2012: 5(Suppl 1):185–93.
33. Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc of the International Conference on Machine Learning (ICML 2001*). Francisco, CA, USA; 2001:282–9.
34. McCallum AK. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu 2002 (last retrieved 17/01/2013).
35. Vapnik V. *The Nature of Statistical Learning Theory*. Springer-Verlag; 1995.
36. Kudoh T, Matsumoto Y. Use of support vector learning for chunk identification. In *Proc of CoNLL and ALL 2000*, Lisbon, Portugal; 2000: 142–4.
37. Guo Y, Korhonen A, Liakata M, Silins I, Sun L, Stenius U. Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proc of the 2010 Workshop on Biomedical Natural Language Processing*, Uppsala, Sweden; 2010:99–107.
38. Mann WC, Thompson SA. Rhetorical structure theory: a theory of text organization. *Tech*. *Rep*. *RS-87-190*, *Information Science Institute*. 1987.
39. Marcu D. The theory and practice of discourse parsing and summarization. *Bradford Books.* 2000.
40. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Machine Learning Res*. 2003;3(4–5):993–1022.
41. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In *Proc of the 14th International Joint Conference on Artificial Intelligence*. Montreal, Canada; 1995:448–53.