Journal of Bioinformatics and Computational Biology

Recognising Discourse Causality Triggers in the Biomedical Domain

Claudiu Mihăilă and Sophia Ananiadou
The National Centre for Text Mining
School of Computer Science, The University of Manchester,
131 Princess Street, Manchester, M1 7DN, United Kingdom
{*claudiu.mihaila, sophia.ananiadou*}*@manchester.ac.uk*

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Current domain-specific information extraction systems represent an important resource for biomedical researchers, who need to process vast amounts of knowledge in a short time. Automatic discourse causality recognition can further reduce their workload by suggesting possible causal connections and aiding in the curation of pathway models. We describe here an approach to the automatic identification of discourse causality triggers in the biomedical domain using machine learning. We create several baselines and experiment with and compare various parameter settings for three algorithms, i.e., Conditional Random Fields (CRF), Support Vector Machines (SVM) and Random Forests (RF). We also evaluate the impact of lexical, syntactic and semantic features on each of the algorithms, showing that semantics improves the performance in all cases. We test our comprehensive feature set on two corpora containing gold standard annotations of causal relations, and demonstrate the need for more gold standard data. The best performance of 79.35% F-score is achieved by CRFs when using all three feature types.

Keywords: discourse analysis; causality; text mining; biomedical text mining.

1. Introduction

It has become increasingly important to be able to provide automated, efficient and accurate means of retrieving and extracting user-oriented biomedical knowledge, considering the ever-increasing amount of knowledge published daily in the form of research articles.[1,5] Based on this need, biomedical text mining has seen significant recent advancements in the last few years,[39] including named entity recognition,[7] coreference resolution[2,34] and relation[19,28] and event extraction.[21,20] Additionally, biomedical text mining tools have been included in specifically-designed frameworks and systems, in which biomedical researchers can easily build workflows to extract information, such as Argo[30] and U-Compare,[11,14] or create and curate pathways and link them to the literature, such as PathText.[12] Using biomedical text mining technology, text can now be enriched via the addition of semantic metadata and thus can support tasks such as analysing molecular pathways[32] and semantic searching.[22]

However, most of the undertaken research is restricted to punctual facts that are expressed in at most one sentence if not one clause. More complex tasks, such as question answering and automatic summarisation, require the extraction of information that spans across several sentences, together with the recognition of relations that exist across sentence boundaries, in order to achieve high levels of performance. These relations, such as causal, temporal and conditional, which characterise how facts in text are related, create a coherent sequence of clauses and sentences, known as *discourse*. Thus, the relations that connect facts in a logical manner are known as *discourse relations*. These help readers to infer deeper, more complex knowledge about the facts mentioned in the discourse. These relations can be either explicit or implicit, depending whether or not they are expressed in text using overt *discourse connectives* (also known as *triggers*). Take, for instance, the case in example (1), where the trigger *Therefore* signals a justification between the two sentences: because "a normal response to mild acid pH from PmrB requires both a periplasmic histidine and several glutamic acid residues", the authors believe that the "regulation of PmrB activity could involve protonation of some amino acids".

(1) In the case of PmrB, a normal response to mild acid pH requires not only a periplasmic histidine but also several glutamic acid residues.
Therefore, regulation of PmrB activity may involve protonation of one or more of these amino acids.

Thus, by identifying these types of causal relations, search engines become able to automatically discover possibly novel relations between biomedical entities, processes and events or between experimental evidence and associated conclusions. This can happen especially if the mechanism is applied to a collection of articles, some of which might be overlooked by humans. However, phrases acting as causal triggers in certain contexts may not denote causality in all cases. Therefore, a dictionary-based approach is likely to produce a very high number of false positives. In this article, we describe the first supervised machine-learning approaches to the automatic identification of triggers that actually denote causality. We show that by adding a deep semantic layer of information, the performance can increase significantly, and that more gold standard data is much needed for better results.

2. Related work

A large amount of work related to discourse parsing and discourse relation identification exists in the general domain, where researchers have not only identified discourse connectives, but also developed end-to-end discourse parsers. Most work is based on the Penn Discourse Treebank (PDTB),[26] a corpus of lexically-grounded annotations of discourse relations.

Some researchers have tackled the problem of identifying discourse connectives, but without determining the discourse relation, as a disambiguation task.[25] Using almost exclusively syntactic features related to the trigger, they achieve an F-score

of around 95%.

Basing on the above work, other researchers introduced new features and manage to slightly improve the overall performance.[16] They included features related to the immediate context of the discourse trigger, such as the previous and next words, their part-of-speech and syntactic interaction with the trigger itself. Also, they added as a feature the entire path from the connective to the root of the parse tree.

A further two approaches consider the syntactic constituency and dependency structure of the context of the trigger.[37] Features include the path from the trigger to the syntactic root, syntactic context features and conjunctive features in the case of the syntactic approach, whilst the dependency approach relies on features such as immediately neighbouring words and their part-of-speech, parents and siblings of the connective and clause detection.

Another small increase in F-score, with just under 1% over Ref. 25 and even less over Ref. 37 is reached by combining certain aspects of the surface level and syntactic feature sets of these respective works.[9]

Until now, comparatively little work has been carried out on causal discourse relations in the biomedical domain, although causal associations between biological entities, events and processes are central to most claims of interest.[13] The equivalent of the PDTB for the biomedical domain is the BioDRB corpus,[27] containing 16 types of discourse relations, e.g., temporal, causal and conditional. A slightly larger corpus is BioCause,[18] containing manually annotated causal discourse relations in full-text open-access journal articles from the infectious diseases domain.

To the best of our knowledge, there is no previous work identifying discourse causal relations in the biomedical domain. Using the BioDRB corpus as data, some researchers have explored the identification of discourse connectives.[31] They do not distinguish, however, between the types of discourse relations and identify them as discourse markers in general. Using mostly a set of orthographic features, they obtain the best F-score of 75.7% using CRF, with SVM reaching only 65.7%. These results were obtained by using only syntactic features, as semantic features were shown to lower the performance. Also, they prove that there exist differences in discourse triggers between the biomedical and general domains by training a model on the BioDRB and evaluating it against PDTB and vice-versa.

The same conclusions were reached in another study,[9] which manages to improve these results by around 3%. They notice that the automatic named entity recognition performed by ABNER[35] lowers the overall performance, due to its use of orthographic features, which thus become duplicated in the feature vector.

3. Methodology

In this section, we describe our data and the features of causal triggers. We also explain our evaluation methodology.

3.1. *Data*

The data for the experiments comes from the BioCause and BioDRB corpora. BioCause is a collection of 19 open-access full-text journal articles pertaining to the biomedical subdomain of infectious diseases, manually annotated with 850 causal relationships. Two types of spans of text are marked in the text, namely causal triggers and causal arguments. Each causal relation is composed of three text-bound annotations: a trigger, a cause or evidence argument and an effect argument. Some causal relations have implicit triggers, so these are excluded from the current research.

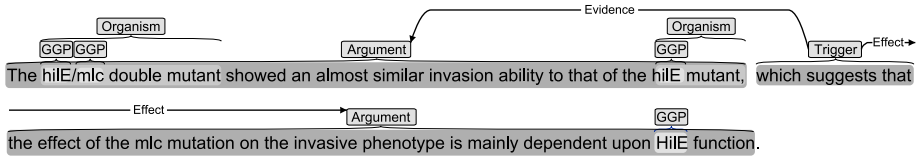

Fig. 1. Causal relation in the BioCause.

Figure 1 shows an example of discourse causality from BioCause, marking the causal trigger and the two arguments with their respective relation. Named entities are also marked in this example.

BioCause contains 381 unique explicit triggers, each being used, on average, only 2.10 times. The number decreases to 347 unique triggers when they are lemmatised, corresponding to an average usage of 2.30 times per trigger. Both count settings demonstrate the diversity of causality-triggering phrases that are used in the biomedical domain.

The BioDRB corpus spreads over 24 articles, and the number of purely causal relations annotated in this corpus is 542. There are another 23 relations which are a mixture between causality and one of either background, temporal, conjunction or reinforcement relations. These relations are based on only 45 different trigger types.

3.2. *Features*

Three types of features have been employed in the development of this causality trigger model, i.e., lexical, syntactic and semantic. These features are categorised and described below.

3.2.1. *Lexical features*

The lexical features are built from the actual tokens present in text. Tokenisation is performed by the GENIA tagger[36] using the biomedical model. The first two features represent the token's surface expression and its base form.

Neighbouring tokens have also been considered. We included the token immediately to the left and the one immediately to the right of the current token. This decision is based on two observations. Firstly, in the case of tokens to the left, most triggers are found either at the beginning of the sentence (311 instances) or are preceded by a comma (238 instances). These two left contexts represent 69% of all triggers. Secondly, for the tokens to the right, almost 45% of triggers are followed by a determiner, such as *the*, *a* or *an*, (281 instances) or a comma (71 instances).

3.2.2. *Syntactic features*

The syntax, dependency and predicate argument structure are produced by the Enju parser.[23] Figure 2 depicts a partial lexical parse tree of a sentence which starts with a causal trigger, namely *Our results suggest that*. From the lexical parse trees, several types of features have been generated.

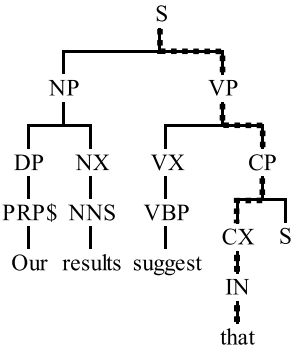

Fig. 2. Partial lexical parse tree of a sentence starting with a causal trigger.

The first two features represent the part-of-speech and syntactic category of a token. For instance, Figure 2 shows that the token *that* has the part-of-speech *IN*. These features are included due to the fact that either many triggers are lexicalised as an adverb or conjunction, or are part of a verb phrase. For the same reason, the syntactical category path from the root of the lexical parse tree to the token is also included. The path also encodes, for each parent constituent, the position of the token in its subtree, i.e., beginning (B), inside (I) or end (E); if the token is the only leaf node of the constituent, this is marked differently, using a C. Thus, the path of *that*, highlighted in the figure, is *I-S/I-VP/B-CP/C-CX*.

Secondly, for each token, we extracted the predicate argument structure and checked whether a relation exists between the token and the previous and following tokens. The values for this feature represent the argument number as allocated by Enju.

Thirdly, the ancestors of each token to the third degree are instantiated as three different features. In the case that such ancestors do not exist (i.e., the root of

the lexical parse tree is less than three nodes away), a "none" value is given. For instance, the token *that* in Figure 2 has as its first three ancestors the constituents marked with *CX*, *CP* and *VP*.

Finally, the lowest common ancestor in the lexical parse tree between the current token and its left neighbour has been included. In the example, the lowest common ancestor for *that* and *suggest* is *VP*.

These last two feature types have been used based on the observation that the lowest common ancestor for all tokens in a causal trigger is S or VP in over 70% of instances. Furthermore, the percentage of cases of triggers with V or ADV as the lowest common ancestor is almost 9% in each case. Also, the average distance to the lowest common ancestor is 3.

3.2.3. *Semantic features*

We have exploited several semantic knowledge sources to identify causal triggers more accurately, as a mapping to concepts and named entities acts as a back-off smoothing, thus increasing performance.

One semantic knowledge source is the BioCause corpus itself. All documents annotated for causality in BioCause had been previously manually annotated with biomedical named entity and event information. This was performed in the context of various shared tasks, such as the BioNLP 2011 Shared Task on Infectious Diseases.[29] We therefore leverage this existing information to add another semantic layer to the model. Moreover, another advantage of having a gold standard annotation is the fact that it is now possible to separate the task of automatic causal trigger recognition from automatic named entity recognition and event extraction. The named entity and event annotation in the BioCause corpus is used to extract information about whether a token is part of a named entity or event trigger. Furthermore, the type of the named entity or event is included as a separate feature.

The second semantic knowledge source is WordNet.[6] Using this resource, the hypernym of every token in the text has been included as a feature. Only the first sense of every token has been considered, as no sense disambiguation technique has been employed.

Finally, tokens have been linked to the Unified Medical Language System (UMLS)[3] semantic types. Thus, we included a feature to say whether a token is part of a UMLS type and another for its semantic type if the previous is true.

3.3. *Experimental setup*

We explored the use of various machine learning algorithms and various settings for the task of identifying causal triggers.

On the one hand, we experimented with CRF,[15] a probabilistic modelling framework commonly used for sequence labelling tasks. In this work, we employed the

CRFSuite implementation.[a]

On the other hand, we modelled trigger detection as a classification task, using Support Vector Machines and Random Forests. More specifically, we employed the implementation in Weka[8,38] for RFs, and LibSVM[4] for SVMs.

Furthermore, we evaluated the best model on BioCause and BioDRB, cross-validated the models between BioCause and BioDRB and evaluated a model trained on both corpora.

4. Results and discussion

Several models have been developed and 10-fold cross-evaluated to examine the complexity of the task and the impact of various feature types (lexical, syntactic, semantic). Table 1 shows the performance evaluation of baseline systems and other classifiers. It should be noted that the dataset is highly skewed, with a ratio of positive examples to negative examples of approximately 1:52.

Table 1. Performance of various classifiers in identifying causal connectives

Classifier	P	R	F_1
Dict	0.08	1.00	0.15
Depend	0.08	0.77	0.14
Synt	0.15	0.20	0.17
Dict+Depend	0.14	0.75	0.24
Dict+Synt	0.22	0.20	0.21
CRF	0.89	0.74	0.79
SVM	0.88	0.61	0.70
RandFor	0.78	0.67	0.72

Several baselines have been devised. The first baseline is a dictionary-based heuristic, named *Dict*. A lexicon is populated with all annotated causal triggers and then this is used to tag all instances of its entries in the text as connectives. The precision of this heuristic is very low, 8.36%, which leads to an F-score of 15.43%, considering that the recall is 100%. This is mainly due to words and/or phrases which are rarely used as causal triggers, such as *and*, *by* and *that*.

Based on the previously mentioned observation about the lowest common ancestor for all tokens in a causal trigger, we built a baseline system that checks all constituent nodes in the lexical parse tree for the S, V, VP and ADV tags and marks them as causal triggers. The name of this system is *Depend*. Not only does *Depend* obtain a slightly lower precision than *Dict*, but it also performs worse in terms of

[a] *http://www.chokkan.org/software/crfsuite*

recall. The F-score is 13.68%, largely due to the high number of intermediate nodes in the lexical parse tree that have VP as their category.

The third baseline is a syntax-based approach, *Synt*. We extracted the syntactic patterns from all triggers, creating a set of 167 unique patterns. After experimenting with possible combinations of patterns to search for, the best performing pattern was found to be *V-C* (verb-complementiser), which occurs in 20.45% of triggers. It gives a precision of 14.61% and a recall of 20.45%, thus resulting in an F-score of 17.04%.

We then combined *Dict* and *Depend*: we considered only constituents that have the necessary category (S, V, VP or ADV) and include a trigger from the dictionary. Although the recall decreases slightly, the precision increases to almost twice that of both *Dict* and *Depend*. This produces a much better F-score of 24.03%. Similarly, the combination of *Dict* and *Synt* results in a precision of 21.88%, a recall of 20.45%, and thus in an F-score of 21.14%.

Table 2. Effect of feature types on the causal trigger recognition.

	CRF			RF			SVM		
Features	P	R	F_1	P	R	F_1	P	R	F_1
Lex	0.89	0.67	0.74	0.78	0.68	0.73	0.81	0.61	0.69
Syn	0.92	0.69	0.76	0.68	0.62	0.65	0.83	0.56	0.67
Sem	0.87	0.63	0.69	0.84	0.57	0.68	0.85	0.57	0.68
Lex+Syn	0.88	0.73	0.79	0.77	0.66	0.71	0.86	0.54	0.67
Lex+Sem	0.90	0.69	0.76	0.79	0.68	0.73	0.82	0.61	0.70
Syn+Sem	0.87	0.73	0.78	0.72	0.64	0.68	0.84	0.55	0.67
Lex+Syn+Sem	0.89	0.74	**0.79**	0.78	0.67	0.72	0.88	0.54	0.67

Table 2 shows the effect of different feature types on CRFs, RFs and SVMs. In the case of CRFs, as can be observed, the best performances, in terms of F-score, including the previously mentioned ones, are obtained when combining all three types of features, i.e., lexical, syntactic and semantic. The best precision and recall, however, are not necessarily achieved by using all three feature types. The best precision is obtained by using the syntactic features only, reaching over 92%, almost 3% higher than when all three feature types are used.

As can be noticed, the best performance of RFs is obtained when combining lexical and semantic features. Due to the fact that causal triggers do not have a semantic mapping to concepts in the named entity and UMLS annotations, the trees in the random forest classifier can easily produce rules that distinguish triggers from non-triggers. As such, the use of semantic features alone produce a very good precision of 84.34%. Also, in all cases where semantic features are combined with other feature types, the precision increases by 0.5% in the case of lexical features and 3.5% in the case of syntactic features. However, the recall of semantic features alone is the lowest. The best recall is obtained when using only lexical features.

For SVMs, we have experimented with two kernels, namely polynomial (second degree) and radial basis function (RBF) kernels. For each of these two kernels, we have evaluated various combinations of parameter values for cost and weight. Both these kernels achieved similar results, indicating that the feature space is not linearly separable and that the problem is highly complex.

The effect of feature types on the performance of SVMs is shown in Table 2. As can be observed, the best performance is obtained when combining the lexical and semantic feature types (69.85% F-score). The combination of all features produces the best precision, whilst the best recall is obtained by combining lexical and semantic features.

As we expected, the majority of errors arise from sequences of tokens which are only used infrequently as non-causal triggers. This applies to 107 trigger types, whose number of false positives (FP) is higher than the number of true positives (TP). In fact, 64 trigger types occur only once as a causal instance, whilst the average number of FPs for these types is 14.25. One such example is *and*, for which the number of non-causal instances (2305) is much greater than that of causal instances (1). Other examples of trigger types more commonly used as causal triggers, are *suggesting* (9 TP, 54 FP), *indicating* (8 TP, 41 FP) and *resulting in* (6 TP, 14 FP). For instance, example (2) contains two mentions of *indicating*, but neither of them implies causality.

(2) Buffer treated control cells showed intense green staining with syto9 (*indicating* viability) and a lack of PI staining (*indicating* no dead/dying cells or DNA release).

We have also evaluated our feature set using the BioDRB corpus. This corpus differs from the BioCause corpus in one important aspect: it does not contain any semantic annotation related to named entities or events. This means that, for the purpose of conducting experiments on the BioDRB in a similar manner, we need to include a pre-processing step that recognises named entities.

For this, we used a simple method that augments the annotation with the named entities present in the output of three named entity recognition tools (NERs), i.e. Metamap, NeMine[33] and OSCAR[10]. The types of entities in the output be each of the three tools, together with the NE types present in the UK PubMed Central (UKPMC)[17], are summarised in Table 3.

UK PubMed Central is an article database that extends the functionality of the original PubMed Central (PMC) repository.[b] Named entities in the UKPMC database were identified using NeMine, a dictionary-based statistical named entity recognition system. This system was later extended and used to recognise more types, such as phenomena, processes, organs and symptoms.[24] We used this most

[b]http://www.ncbi.nlm.nih.gov/pmc

Table 3. Named entity types and their source.

Type	UKPMC	NeMine	OSCAR
Gene	✓	✓	
Protein	✓	✓	
Gene—Protein	✓		
Disease	✓	✓	
Drug	✓	✓	
Metabolite	✓	✓	
Bacteria		✓	
Diagnostic process		✓	
General phenomenon		✓	
Indicator		✓	
Natural phenomenon		✓	
Organ		✓	
Pathologic function		✓	
Symptom		✓	
Therapeutic process		✓	
Chemical molecule			✓
Chemical adjective			✓
Enzyme			✓
Reaction			✓

recent version of the software as our second source of more diverse entity types.

The Open-Source Chemistry Analysis Routines (OSCAR) software is a toolkit for the recognition of named entities and data in chemistry publications. Currently in its fourth version, it uses three types of chemical entity recognisers, namely regular expressions, patterns and Maximum Entropy Markov models.

After augmenting the existing NEs by running the two NER tools on the corpus, the outputs were combined to give a single "silver" annotation list. This operation was performed by computing the mathematical union of the three individual annotation sets, as shown in Eq. 1.

$$\mathbb{A}_{\mathsf{Silver}} = \mathbb{A}_{\mathsf{Metamap}} \cup \mathbb{A}_{\mathsf{Oscar}} \cup \mathbb{A}_{\mathsf{NeMine}} \cup \mathbb{A}_{\mathsf{UKPMC}} \tag{1}$$

For reasons of fairness, the gold standard semantic annotation in the BioCause corpus has been removed and replaced with automatic NER results.

For the evaluation, we used the best performing algorithm and its parameter settings, i.e. CRF with all three types of features. We created different models and evaluated them in various ways, and the results of these tests are given in Table 4. The first two columns of the table show the training corpus and the test corpus, respectively, for that respective test. In the case of 10-fold cross validation, *10X* is used.

As can be observed, the model trained on the BioDRB corpus obtains a higher precision than the one trained the BioCause. This is mainly due to the smaller set of unique connectives present in the BioDRB. The recall is, however, lower, and, overall, the F-score for the BioDRB model is 1% lower than the F-score for the BioCause model.

Table 4. Results of the evaluationwith BioDRB.

Train	Test	P	R	F_1
BioCause	*10X*	0.84	0.69	0.74
BioDRB	*10X*	0.88	0.67	0.73
BioCause	BioDRB	0.68	0.61	0.63
BioDRB	BioCause	0.77	0.57	0.61
BioCause+BioDRB	*10X*	0.81	0.66	0.71

The second type of evaluation is a cross validation between the two corpora: training is carried out on one and testing on the other. In the first case, we trained a model on BioCause and tested it on BioDRB. The second case is the opposite, training on the BioDRB and testing on BioCause. There are significant differences in precision and recall between the two tests, but the resulting F-scores are approximatively equal. The precision is lower in the first case by 5%, whilst the recall is 4% lower because of the wider variety of causal triggers that are present in BioCause and do not occur in BioDRB.

Finally, we trained CRF on the combination of the BioCause and BioDRB corpora. The results of the 10-fold cross-validation are slightly worse than those achieved for each of the individual corpora, but much better than for the cross evaluation between the two corpora. It can be noticed that both precision and recall are moderately lower than those obtained for each of the two corpora.

We also hypothesised that an increase in the size of the training data increases the performance. Therefore, we extracted random subsets of the combined corpus at various percentages. For each of the six corpus sizes, varying from 50% to 100% in intervals of 10%, we created five random subsets. These subsets have been 10-fold cross-validated using the best performing algorithm and its parameter settings, CRF with all three types of features.

Fig. 3 shows the F-score achieved for each of the 30 evaluated subsets with circles. Also depicted is a thick black line that shows the second-degree polynomial increase of the F-score trend. The co-efficient of determination, R^2, which shows how closely the trendline fits with the data points, has the value of 0.9761, indicating that the trend line is very reliable.

Furthermore, we tested the statistical significance of this increase by using the Anova Single Factor test. At an α of 0.05, we obtained an $F_{statistic} = 15.12$, much larger than the corresponding $F_{crit} = 2.62$, a fact which rejects the null hypothesis that all the F-scores are equal in favour of the alternate hypothesis that at least two of the means are different. The resulting p-value is 9.53E-7, which again allows us to reject the null hypothesis. Taken together, these results strengthen our hypothesis that the more data there is, the better the system performs.

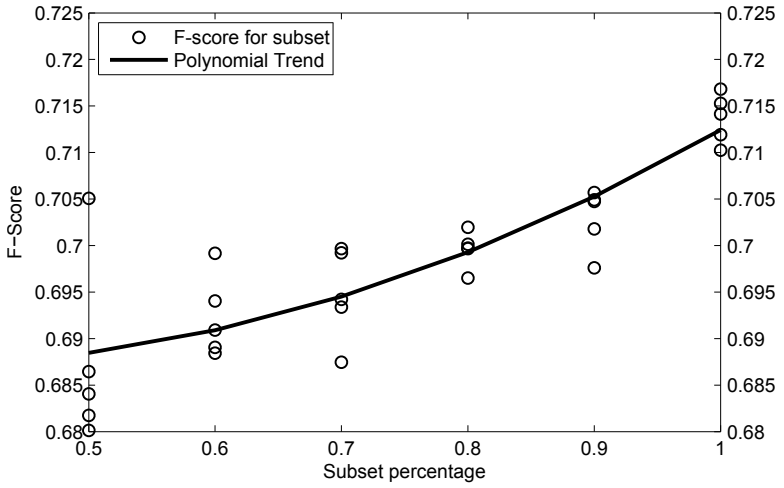

Fig. 3. F-score distribution and trend for the various subsets of the combination of BioCause and BioDRB.

5. Concluding remarks

We have described our approach to automatically recognising triggers of causal discourse relations in biomedical scientific text. The complexity of this task has proven to be very high, posing many challenges. Shallow approaches, such as dictionary matching and lexical parse tree matching, perform extremely poorly, with F-scores of approximately 15% each and 24% when combined, due to the high ambiguity of causal triggers. We have explored various algorithms that automatically learn to distinguish tokens into triggers or non-triggers and we have evaluated the impact of multiple lexical, syntactic and semantic features. The performance obtained by SVMs shows that the task of identifying causal triggers is indeed complex. The best performing classifier is CRF-based and combines lexical, syntactical and semantic features in order to obtain a final F-score of 79.35%.

Furthermore, we tested our feature set on another corpus, BioDRB, in order to check whether a data insufficiency problem exists and, if so, estimate the optimal amount of necessary data. We discovered a polynomial increase in F-score when the two datasets are combined. Thus, it might be necessary to produce more gold standard data by employing experts or to develop a method for automatically bootstrapping more data as accurately as possible.

As future work, more evaluations against the general domain need to be performed, in order to establish the differences in expressing causality in the biomedical domain. One possible source for this is the PDTB corpus. This will allow researchers to easily take off-the-shelf end-to-end discourse parsers produced for the general domain and adapt them to biomedicine.

A more difficult task that needs attention is that of identifying implicit triggers.

These occur much more rarely than explicit triggers, but their role is as important. Finally, our system needs to be extended in order to identify the two arguments of causal relations, the cause and effect, thus allowing the creation of a complete discourse causality parser.

Acknowledgments

This work was partially funded by the EPSRC [grant number EP/P505631/1]; Medical Research Council; Europe PubMed Central Funders (led by Wellcome Trust).

References

1. Ananiadou S, McNaught J (eds.), *Text Mining for Biology And Biomedicine*, Artech House, Inc., 2006. ISBN 158053984X.
2. Batista-Navarro RTB, Ananiadou S, Building a coreference-annotated corpus from the domain of biochemistry, *Proceedings of BioNLP 2011*, pp. 83–91, 2011.
3. Bodenreider O, The unified medical language system (UMLS): integrating biomedical terminology, *Nucleic Acids Research* **32**(suppl 1):D267–D270, 2004. doi: 10.1093/nar/gkh061.
4. Chang CC, Lin CJ, LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology* **2**:27:1–27:27, 2011.
5. Cohen KB, Hunter L, Getting started in text mining, *PLoS Computational Biology* **4**(1):e20, 2008. doi:10.1371/journal.pcbi.0040020.
6. Fellbaum C (ed.), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA, 1998. ISBN 978-0-262-06197-1.
7. Fukuda Ki, Tsunoda T, Tamura A, Takagi T, Toward information extraction: Identifying protein names from biological papers, *Proceedings of the Pacific Symposium on Biocomputing*, pp. 707–718, 1998.
8. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH, The WEKA data mining software: an update, *SIGKDD Explor Newsl* **11**:10–18, 2009. doi: http://doi.acm.org/10.1145/1656274.1656278.
9. Ibn Faiz S, Mercer RE, Identifying explicit discourse connectives in text, in *Advances in Artificial Intelligence*, eds., Zaïane OR, Zilles S, Springer Berlin Heidelberg, pp. 64–76, 2013.
10. Jessop D, Adams S, Willighagen E, Hawizy L, Murray-Rust P, Oscar4: a flexible architecture for chemical text-mining, *Journal of Cheminformatics* **3**(1):41, 2011. doi: 10.1186/1758-2946-3-41.
11. Kano Y, Baumgartner W, McCrohon L, Ananiadou S, Cohen KB, Hunter L, Tsujii J, U-compare: share and compare text mining tools with UIMA, *Bioinformatics* **25**(15):1997–1998, 2009.
12. Kemper B, Matsuzaki T, Matsuoka Y, Tsuruoka Y, Kitano H, Ananiadou S, Tsujii J, Pathtext: a text mining integrator for biological pathway visualizations, *Bioinformatics* **26**(12):i374–i381, 2010.
13. Kleinberg S, Hripcsak G, A review of causal inference for biomedical informatics, *Journal of Biomedical Informatics* **44**(6):1102 – 1112, 2011. doi:10.1016/j.jbi.2011.07.001.
14. Kontonatsios G, Korkontzelos I, Ananiadou S, Developing multilingual text mining workflows in UIMA and U-Compare, *Proceedings of the 17th International conference on Applications of Natural Language Processing to Information Systems*, 2012.
15. Lafferty JD, McCallum A, Pereira FCN, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proceedings of the Eighteenth Inter-*

*national Conference on Machine Learning*ICML '01, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 282–289, 2001. ISBN 1-55860-778-1.
16. Lin Z, Ng HT, Kan MY, A pdtb-styled end-to-end discourse parser, *Natural Language Engineering* **FirstView**:1–34, 2012. doi:10.1017/S1351324912000307.
17. McEntyre JR, Ananiadou S, Andrews S, Black WJ, Boulderstone R, Buttery P, Chaplin D, Chevuru S, Cobley N, Coleman LA, Davey P, Gupta B, Haji-Gholam L, Hawkins C, Horne A, Hubbard SJ, Kim JH, Lewin I, Lyte V, MacIntyre R, Mansoor S, Mason L, McNaught J, Newbold E, Nobata C, Ong E, Pillai S, Rebholz-Schuhmann D, Rosie H, Rowbotham R, Rupp CJ, Stoehr P, Vaughan P, UKPMC: a full text article resource for the life sciences, *Nucleic Acids Research* **39**(suppl 1):D58–D65, 2011.
18. Mihăilă C, Ohta T, Pyysalo S, Ananiadou S, BioCause: Annotating and analysing causality in the biomedical domain, *BMC Bioinformatics* **14**(1):2, 2013.
19. Miwa M, Sætre R, Miyao Y, Tsujii J, Protein-protein interaction extraction by leveraging multiple kernels and parsers, *International Journal of Medical Informatics* **78**(12):e39–e46, 2009. doi:10.1016/j.ijmedinf.2009.04.010.
20. Miwa M, Thompson P, Ananiadou S, Boosting automatic event extraction from the literature using domain adaptation and coreference resolution, *Bioinformatics* **28**(13):1759–1765, 2012. doi:10.1093/bioinformatics/bts237.
21. Miwa M, Thompson P, McNaught J, Kell DB, Ananiadou S, Extracting semantically enriched events from biomedical literature, *BMC Bioinformatics* **13**:108, 2012. doi:10.1186/1471-2105-13-108.
22. Miyao Y, Ohta T, Masuda K, Tsuruoka Y, Yoshida K, Ninomiya T, Tsujii J, Semantic retrieval for the accurate identification of relational concepts in massive textbases, *ACL*, 2006.
23. Miyao Y, Tsujii J, Feature forest models for probabilistic HPSG parsing, *Computational Linguistics* **34**(1):3580, 2008.
24. Nobata C, Sasaki Y, Okazaki N, Rupp CJ, Tsujii J, Ananiadou S, Semantic search on digital document repositories based on text mining results, *International Conferences on Digital Libraries and the Semantic Web 2009 (ICSD2009)*, pp. 34–48, 2009.
25. Pitler E, Nenkova A, Using syntax to disambiguate explicit discourse connectives in text, *ACL/AFNLP (Short Papers)*, pp. 13–16, 2009.
26. Prasad R, Dinesh N, Lee A, Miltsakaki E, Robaldo L, Joshi A, Webber B, The Penn Discourse TreeBank 2.0, Calzolari N, Choukri K, Maegaard B, Mariani J, Odjik J, Piperidis S, Tapias D (eds.), *In Proceedings of the 6th International Conference on language Resources and Evaluation (LREC)*, pp. 2961–2968, 2008.
27. Prasad R, McRoy S, Frid N, Joshi A, Yu H, The biomedical discourse relation bank, *BMC Bioinformatics* **12**(1):188, 2011.
28. Pyysalo S, Ohta T, Kim JD, Tsujii J, Static relations: a piece in the biomedical information extraction puzzle, *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*BioNLP '09, BioNLP '09, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1–9, 2009. ISBN 978-1-932432-30-5.
29. Pyysalo S, Ohta T, Rak R, Sullivan D, Mao C, Wang C, Sobral B, Tsujii J, Ananiadou S, Overview of the infectious diseases (ID) task of BioNLP shared task 2011, *Proceedings of the BioNLP Shared Task 2011 Workshop*, Association for Computational Linguistics, Portland, Oregon, USA, pp. 26–35, 2011.
30. Rak R, Rowley A, Black W, Ananiadou S, Argo: an integrative, interactive, text mining-based workbench supporting curation, *Database: The Journal of Biological Databases and Curation* **2012**, 2012.
31. Ramesh PB, Prasad R, Miller T, Harrington B, Yu H, Automatic discourse connective

detection in biomedical text, *Journal of the American Medical Informatics Association* **19**(5):800–808, 2012. doi:10.1136/amiajnl-2011-000775.
32. Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboué AP, Weng W, Wilbur W, Hatzivassiloglou V, Friedman C, Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data, *Journal of Biomedical Informatics* **37**(1):43 – 53, 2004. doi:10.1016/j.jbi.2003.10.001.
33. Sasaki Y, Tsuruoka Y, McNaught J, Ananiadou S, How to make the most of NE dictionaries in statistical NER, *BMC Bioinformatics* **9**(Suppl 11):S5, 2008.
34. Savova GK, Chapman WW, Zheng J, Crowley RS, Anaphoric relations in the clinical narrative: corpus creation, *Journal of the American Medical Informatics Association* **18**(4):459–465, 2011. doi:10.1136/amiajnl-2011-000108.
35. Settles B, ABNER: An open source tool for automatically tagging genes, proteins, and other entity names in text, *Bioinformatics* **21**(14):3191–3192, 2005.
36. Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, McNaught J, Ananiadou S, Tsujii J, Developing a robust part-of-speech tagger for biomedical text, in *Advances in Informatics - 10th Panhellenic Conference on Informatics*, eds., Bozanis P, Houstis EN, Springer-Verlag, Volos, Greece, pp. 382–392, 2005.
37. Wellner B, *Sequence Models and Ranking Methods for Discourse Parsing*. PhD Thesis, Brandeis University, 2009.
38. Witten I, Frank E, *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*, Morgan Kaufmann, 2005.
39. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB, Frontiers of biomedical text mining: current progress, *Briefings in Bioinformatics* **8**(5):358–375, 2007.

Claudiu Mihăilă received his B.Sc. degree in Computer Science and M.Sc. degree in Computational Linguistics from the "Al.I. Cuza" University of Iaşi, Romania, in 2008 and 2010, respectively. After obtaining his M.Sc. degree, he started his PhD research at The National Centre for Text Mining in the School of Computer Science, The University of Manchester. In this research, he is looking at automatically recognising causal relations in biomedical scientific discourse.

Sophia Ananiadou is Professor of Computer Science in the School of Computer Science, the University of Manchester and director of the National Centre for Text Mining (NaCTeM). She has established a strong track record over the past decade, building NaCTeM with +16 staff, substantial infrastructure in text mining and active collaborations with industry and academia internationally. She holds a PhD in Natural Language Processing from the University of Manchester. Her research includes the development of large-scale resources for biology, data integration using text mining, event extraction for drug discovery, pathway reconstruction and association mining, text mining tools for education, chemistry, social sciences, institutional repositories and systematic reviews. She is leading the text mining work of EuropePubMedCentral, funded by Wellcome Trust. She has authored over 250 publications.