The Meta-knowledge of Causality in Biomedical Scientific Discourse

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

Abstract

Causality lies at the heart of biomedical knowledge, being involved in diagnosis, pathology or systems biology. Thus, automatic causality recognition can greatly reduce the human workload by suggesting possible causal connections and aiding in the curation of pathway models. For this, we rely on corpora that are annotated with classified, structured representations of important facts and findings contained within text. However, it is impossible to correctly interpret these annotations without additional information, e.g., classification of an event as fact, hypothesis, experimental result or analysis of results, confidence of authors about the validity of their analyses etc. In this study, we analyse and automatically detect this type of information, collectively termed *meta-knowledge* (*MK*), in the context of existing discourse causality annotations.

Keywords: discourse analysis; biomedical causality; information extraction

1. Introduction

Statements regarding causal associations have been long studied in general language, mostly as part of more complex tasks, such as question answering (Girju, 2003; Blanco et al., 2008) and textual entailment (Ríos Gaona et al., 2010). To suport these efforts, several corpora containing causal annotations have also been created, the most prominent being the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008). Despite the large amount of work, a single, unified theory of causality has not yet emerged, be it in general or specialised language.

In contrast, until now, comparatively little work has been carried out on discourse relations in the biomedical domain. To our knowledge, there exist only two biomedical corpora that contain manually annotated causal associations. On the one hand, the BioCause corpus contains 850 annotations of causal associations over 19 full-text open-access journal articles from the domain of infectious diseases (Mihăilă et al., 2013). On the other hand, BioDRB (Prasad et al., 2011) contains annotations of 16 different discourse relations, one of which is causality, similar to the PDTB corpus. The total number of causal associations is 542, whilst another 23 are a mixture between causality and one of either background, temporal, conjunction or reinforcement associations.

In spite of the more focussed and powerful searching methods available today, typical discourse annotation efforts, such as BioCause and BioDRB, do not take into consideration the information regarding the context of discourse relations, although this is essential for their correct interpretation. For instance, negation is considered a universal property of all human languages (Greenberg et al., 1978), and plays an important role in contradiction detection. Vincze et al. (2008) report that around 13% of sentences found in biomedical research articles contain some form of negation, whilst Nawaz et al. (2013b) analyse three open access bio-event corpora to show that more than 6% of bio-events are negated. Additionally, determining the certainty level provides information about the confidence of authors in their statements. This could be because either there is uncertainty regarding the general truth value assigned to the relation or it is perceived that the relation does not hold all the time. Plus, it is necessary to automatically discover the novel parts of articles, as well as whether they are hypotheses, experiments, evaluations or results. The identification of such information is critical for several tasks in which biomedical researchers have to search and review the literature. One such example is the maintenance of models of biological processes, such as pathways (Oda et al., 2008). Take example (1) below, which contains a causal relation which is negated (*against*) and speculated (*argue*) analysis (*results*) attributed to previous work (*their*).

(1) SeMac functions like GAS M1 Mac in the inhibition of opsonophagocytosis of GAS by human PMNs [21].
Their results argue against the fact that the inhibition of the bactericidal activity of PMNs is not mediated by opsonophagocytosis or may be insignificant in whole blood.

The goal of capturing this type of interpretative information, explicitly or implicitly available in text, termed meta-knowledge (Thompson et al., 2011), is to extract as much useful information as possible about causal associations in their textual context. This will further support the development of information retrieval and extraction systems, the automatic discovery of new knowledge and the detection of contradictions.

In this work, we adapt an existing meta-knowledge annotation scheme (Thompson et al., 2011) from biomolecular events to biomedical discourse relations, apply it to the causal associations existing in the BioCause corpus and analyse the resulting annotations. Furthermore, we train classifiers to automatically recognise meta-knowledge information and evaluate their performance based on the human annotations. To our best knowledge, our method is the first that is able to automatically identify and classify

meta-knowledge information about causality in biomedical scientific discourse.

2. Related Work

There exist several distinct efforts to capture various meta-knowledge dimensions in biomedical text, such as certainty (Kilicoglu and Bergler, 2008; Vincze et al., 2008), negation (Vincze et al., 2008; Nawaz et al., 2013b), manner (Nawaz et al., 2012) or source (Liakata et al., 2010; Sándor and de Waard, 2012; Nawaz et al., 2013a). However, most of them are focussed on biomedical events. Although identifying MK for bio-events is useful, the MK on the discourse relations connecting the spans of text containing bio-events is as necessary.

Regarding discourse, researchers have looked at articles as networks of hypotheses and evidence, and tried to identify the argumentation contained within a paper and the relationships between hypotheses, claims and evidence expressed in the article (de Waard et al., 2009). Others classified the discourse into discourse zones specific to scientific articles (e.g., background, methods, results) (Sándor, 2007). Another annotation scheme considers more than one aspect of meta-knowledge. For example, the ART corpus and its CoreSC annotation scheme (Liakata and Soldatova, 2009; Liakata et al., 2010) augment general information content categories with additional attributes, such as *New* and *Old*, to denote current or previous work. Despite these efforts, a study that takes into consideration multiple meta-knowledge dimensions, automatically identifies them and analyses their interaction has not been yet performed.

Considering the mentioned work, we decided to create a resource of biomedical discourse causality enriched with relevant meta-knowledge information. Furthermore, we train multiple classifiers to detect the MK information automatically.

3. Meta-knowledge annotation scheme

The original meta-knowledge annotation scheme is depicted in Figure 1. As can be noticed, it contains six dimensions which are centred on a biomedical event.

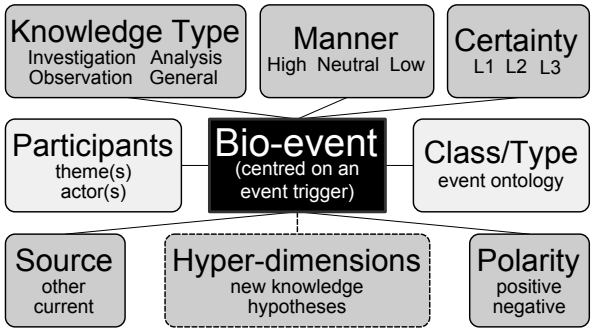

Figure 1: Meta-knowledge dimensions (from (Thompson et al., 2011)).

We adapted this meta-knowledge annotation scheme to the characteristics of discourse relations. All dimensions have been kept, with the exception of *Manner*, which is used to describe the change in intensity or speed of a biological process and does not have a correspondent in discourse. In what follows, we describe the adapted dimensions and categories.

3.1. Knowledge type

The *Knowledge Type* (*KT*) captures the general information about the content of the causal association, classifying it into five categories:

- *analysis*: inferences, interpretations, speculations or other types of cognitive analysis, always accompanied by lexical clues, typical examples of which include *suggest*, *indicate*, *therefore* and *conclude*.
- *fact*: events that describe general facts and well-established knowledge, and sometimes accompanied by lexical clues such as *known*.
- *investigation*: enquiries or investigations, which have either already been conducted or are planned for the future, typically accompanied by lexical clues like *examined*, *investigated* and *studied*.
- *observation*: direct observations, sometimes represented by lexical clues like *found*, *observed* and *report*, etc.
- *other*: the default category, assigned to associations that either do not fit into one of the above categories, do not express complete information, or whose *KT* is unclear or is unassignable from the context.

The original meta-knowledge *KT* dimension also includes a *Method* category, that is used to describe experimental methods, with clue words such as *stimulate* and *inactivate*. This category is not suitable for discourse, as intensity or speed does not apply to causality or other discourse relations.

3.2. Certainty

This dimension encodes the confidence or certainty level ascribed to the association in the given text. The epistemic scale is partitioned into three distinct levels:

- *L1*: explicit indication of either low confidence or considerable speculation towards the association or the association occurs infrequently or only some of the time.
- *L2*: explicit indication of either high (but not complete) confidence or slight speculation towards the association or the association occurs frequently, but not all of the time.
- *L3*: the default category. No explicit expression that either there is uncertainty or speculation towards the associations or that the association does not occur all of the time.

3.3. Source

The source of the knowledge expressed by the causal association is encoded as:

- *current*: the association makes an assertion that can be attributed to the current study. This is the default category, and is assigned in the absence of explicit lexical or contextual clues, although explicit clues such as *the present study* may be encountered.

- *other*: the association is attributed to a previous study. Explicit clues are usually present either as citations, or by using words such as *previously* and *recent studies*.

3.4. Polarity

This dimension identifies the truth value of the asserted causal association. A negated causal association is defined as one describing the non-existence or absence of a causal link between two spans of text. The recognition of such associations is vital, as it can lead to the correct interpretation of a causal association, completely opposite to that of a non-negated one.

- *positive*: no explicit negation of the causality. This is the default category, as most causal associations are expected to be positive.

- *negative*: the association has been negated according to the description above. The negation may be indicated through lexical clues such as *no*, *not* or *fail*.

4. Meta-knowledge of discourse causality

We have applied the adapted meta-knowledge annotation scheme to all 19 full papers in the BioCause corpus, previously annotated with discourse causality associations. Previous studies have shown that the annotator background does not affect the consistency of the resulting annotations of meta-knowledge (Thompson et al., 2011). Therefore, two annotators with background in computational linguistics and experience in meta-knowledge annotation have undertaken the annotation task. All causal associations have been annotated with meta-knowledge information. The two annotators have undergone a short training period, in which they have become accustomed to the annotation tool and guidelines and improved the agreement between them. High levels of inter-annotator agreement have been achieved, falling in the range of 0.88 - 0.95 Kappa, depending on the MK dimension. The Kappa scores for each MK dimension are given in Table 1. The lowest Kappa occurs in the case of *KT*, as it is the most complex dimension to annotate. The five possible values can be confusing with specific relations which lie at the border between labels. The highest score is obtained in the case of *Polarity*, as it is fairly easy to recognise whether a relation is negated or not. The few problems that arose were in cases where the negation is implicit to the trigger itself. All disagreements have been discussed after the annotation and a final option has been agreed for each such disagreement by both annotators.

The dataset is available under a Creative Commons BY-SA-NC licence from the site of the National Centre for Text Mining (NaCTeM).

Table 2 summarises the distribution of annotation for each category of each dimension in the MK scheme, together with their relative frequency. In Table 3, we give the most frequent clues for each category, together with their relative frequency for that category. The results included in both tables are discussed in what follows.

MK subdim.	Kappa
Knowledge type	0.88
Certainty	0.89
Polarity	0.95
Source	0.94

Table 1: Inter-annotator agreement per MK dimensions.

MK dimension	MK subdim.	Count (Freq.)
KT	Analysis	663 (82.88%)
	Fact	52 (6.50%)
	Investigation	4 (0.5%)
	Observation	62 (7.75%)
	Other	19 (2.38%)
Certainty	L1	78 (9.75%)
	L2	383 (47.88%)
	L3	339 (42.38%)
Polarity	positive	790 (98.75%)
	negative	10 (1.25%)
Source	current	722 (90.25%)
	other	78 (9.75%)

Table 2: Meta-knowledge category distribution in BioCause.

MK dim.	MK subdim.	Frequent clues
KT	Analysis	suggest (38.86%), indicate (21.68%)
	Fact	shown to (60%), known to (20%)
	Investigation	illuminate (100%)
	Observation	observe (45%), report (30%)
Certainty	L1	may (45%), might (30%), perhaps (8%)
	L2	suggest (51.8%), indicate (29.2%)
	L3	definitely (40%), firmly (30%)
Polarity	Negative	not (62.5%), no (15%), against (7%), rule out (4%)
Source	Current	in this study (67%), in this paper (17%)
	Other	citations (84%), previously (8%)

Table 3: Most frequent clues for each MK category with their respective relative frequency (computed over the number of explicit clues) for that category.

4.1. Manual analysis

Here we provide some key statistics regarding the causality annotation produced, together with a discussion of the characteristics of the corpus.

4.1.1. Knowledge type

The most frequent annotated value is *Analysis*, constituting almost two thirds of the total number of causal associations. This is not surprising, since most causal associations are the result of inference or interpretation of experimental results. Two other categories, *Observation* and *Fact*, are less frequently annotated, occurring in just over 2% of all annotations. *Investigation* appears even less, with only five instances in the entire corpus. The number of *Other* relations is 21 (2.63%).

There are several lexical clues that mark this MK category. The most common is *suggest*, which occurs in almost 39% of the *Analysis* cases. The second most common is *indicate*, which occurs in almost 22% of the *Analysis* cases. Other clues include *demonstrate*, *thus* and *therefore*.

4.1.2. Certainty

More than half of the causal associations in the corpus are expressed with some degree of uncertainty. That is, 57.62% of associations have been annotated with uncertainty clues, whilst 42.38% are certain or lack any uncertainty clue.

Under the speculated category, over 83% (47.88% per total) of associations are reported with slight speculation (*L2*), whilst just under 17% (10% per total) are annotated as having a high level of speculation (*L1*). This is again an expected result, since most authors express their analyses with a high level of confidence.

The most frequent clues that lead to uncertainty are verbs, such as *suggest* and *indicate*, and modals, e.g., *may* and *might*. Nevertheless, there are several other types of uncertainty clues, such as adverbs (*likely*, *maybe* and *perhaps*).

An interesting observation is that most of the uncertain associations (96.30%) belong to the *KT* type *Analysis*. There are very few instances of uncertain relation pertaining to other knowledge types. *Fact* has two relations (3.84%), whilst *Observation* has 12 relations (19.35%). Thus, almost 67% of all associations annotated as *Analysis* also have some degree of uncertainty.

Speculated relations are mostly part of the *Current* value of the *Source* dimension, and there are four negated speculated relations (44% of all negations).

4.1.3. Source

Very few associations belong to the *Other* category, when compared to *Current*. Just under 10% of all associations have their source in other articles, whilst 90% express knowledge created by the authors themselves.

Clues that are specific only to the *Other* category are citations to other articles. Other clues are phrases such as *previously reported* and *X proposes that*, where X substitutes the names of researchers.

Causal relations that have their source in other research are all positive from a *Polarity* point of view. However, they are not all completely certain: there are four instances which have *L1* as their *Certainty* level, whilst another 16 are *L2*. The rest of 57 are marked as *L3*.

The knowledge type of the causal relations is almost evenly split between *Analysis* (42 relations) and *Fact* (33 relations). There are one *Observation* and one *Other Knowledge type* relations from other sources. This fact is quite intuitive – most work already published tends to be treated as a fact or is analysed in connection with the research described in the current work.

4.1.4. Polarity

A small number of associations have been annotated with a *Negative* category in the *Polarity* dimension. Just over 1% of the annotations are marked as expressing a negated causality. This is to be expected, since, in scientific discourse, authors tend to present their positive results instead of negative ones. Nevertheless, it is vital to detect such information, since a simple negation completely changes the meaning of a causal relation.

Clues for negations are varied, some belonging to closed-class parts-of-speech, e.g. determiners (*no*), adverbs (*not*) or prepositions (*against*), whilst others belong to open-class parts-of-speech, such as verbs (*rule out*) and adjectives (*impossible*). Nevertheless, the adverbial *not* is the most frequent, accounting for almost two thirds of negated causal associations.

Negated causal relations always have the *Source* dimension set to *Current*. It is very unlikely that authors of one study directly contradict causal relations described in other research.

Furthermore, five out of the nine negated relations have the *Certainty* level set to *L3*. Two relation is set to *L1*, and another two to *L2*.

Looking at negated relations from a *Knowledge Type* perspective, seven relations are of type *Analysis*, whilst two are marked as *Observation*. The lack of occurrence of negative instances amongst the other types of *Knowledge Type* is to be expected, as it is usual that researchers investigate why events occur, and not why they do not.

4.2. Automatic classification

We have experimented with several supervised machine learning algorithms in the task of automatically classifying causal discourse relations from the point of view of each MK dimension. The range of classifiers covers rules (JRip), trees (J48, Random Forests), support vector machines (SVM), Naïve Bayes and meta-classifiers (Vote). All classifiers are implemented in Weka (Witten and Frank, 2005; Hall et al., 2009). For Random Forests, we have evaluated different numbers of trees and random features, and the best results were obtained with 10 trees and 5 random features. The SVM classifier has been tested with different kernels, i.e. linear, polynomial and radial basis function. The best performance was obtained by using a second degree polynomial kernel, but very closely followed by the linear and RBF kernels. The Vote meta-classifier is configured to consider the decisions of the other five classifiers with a majority voting rule. In addition, we created a baseline rule for each dimension (named *Majority*), which marks all relations with the majority class label. This is due to the highly skewed data present in the corpus.

The learners have been trained on a large feature set, including the clues mentioned above. Lexical features are the most important, as they provide direct information to classifiers. Besides the surface expression of the tokens, we also include their lemmata, which is justified by the need of generalisation: some inflected lexemes may occur very

rarely (if at all) in the limited amount of training data, and, in a real-world deployment, a learner may be perplexed when encountering them. The tokenisation and lemmatisation steps are performed by employing the GENIA tagger (Tsuruoka et al., 2005) trained on MEDLINE. Having binary features that flag the presence of negation particles or modal verbs helps ML algorithms make better decisions.
Furthermore, syntax provides good support for the generalisation of triggers and their associated meta-knowledge. These are extracted from automatic parses created by the Enju system (Miyao and Tsujii, 2008) trained on GENIA. Syntactic features include the part-of-speech tag and syntactic category, as well as dependency, constituency and c-command information. C-command features, based on the definition of Barker and Pullum (1990), indicate whether a causal trigger c-commands, S-commands or VP-commands constituents containing relevant cues.
Besides lexical and syntactic features, the algorithms have learned using a semantic layer of annotations. These come from the gold standard named entities and events that are already present in the BioCause corpus, augmented with new information automatically obtained from UMLS (Bodenreider, 2004), OSCAR (Jessop et al., 2011), NeMine (Nobata et al., 2009), and EUPMC[1].
All features have been extracted from a context window spanning the full sentence in which the trigger is located.
We built separate models for each MK dimension, all of which have been 10-fold cross validated. The main results are given in Table 4. These represent the macro-average F-scores only. However, due to the skewed data, we also provide micro-average F-scores in the discussion of each individual dimension.

Algorithm	KT	Certainty	Polarity	Source
Majority	18.15%	21.53%	49.72%	47.51%
SVM	36.31%	87.40%	84.02%	68.25%
RandFor	34.76%	83.53%	79.97%	73.77%
JRip	29.28%	77.92%	84.02%	71.52%
J48	25.45%	83.75%	49.72%	47.51%
N. Bayes	32.96%	77.49%	61.17%	62.35%
Vote	41.69%	84.62%	79.97%	70.87%

Table 4: Macro-average F-scores achieved by various learners per each MK dimension.

4.2.1. Knowledge type

Table 5 lists the detailed performance of the employed classifiers in the task of detecting the *Knowledge Type* of causal relations. It includes the macro-average precision, recall and F-score, as well as the micro-average F-score. The large difference between the two scores comes from the fact that this is a five-way classification, corresponding to the five subdimensions of *KT*, and that the data is very skewed across these five subdimensions.
As can be noticed, all classifiers perform better than the baseline in a macro-average setting. However, in a micro-average context, Naïve Bayes is confused by the data imbalance and is outperformed by the Majority rule by almost 1%. The best performing classifier is the Vote meta-classifier, which reaches 83.80% micro-average F-score and 41.69% macro-average F-score. It also obtains the best precision and recall amongst all classifiers, in both macro- and micro-average settings.
Most errors arise because of the skewed distribution of the labels. For instance, for Vote, there are only eight false negatives for the *Analysis* label, but 82 false positives are generated. The two instances in the *Investigation* label are erroneously assigned to *Analysis*. This proves the tendency of the classifiers to assign most instances from minority classes to the majority class.

[1] http://europepmc.org/

Algorithm	ma P	ma R	ma F_1	mi F_1
Majority	16.62%	20.00%	18.15%	75.40%
SVM	39.94%	33.38%	36.31%	82.60%
RandFor	39.12%	31.28%	34.76%	82.20%
JRip	40.96%	22.78%	29.28%	77.60%
J48	27.50%	23.68%	25.45%	77.90%
N. Bayes	30.52%	35.82%	32.96%	74.50%
Vote	54.64%	33.70%	41.69%	83.80%

Table 5: Performance of various classifiers in identifying the *Knowledge Type* of causal relations.

4.2.2. Certainty

A detailed account of the performance of the classifiers is given in Table 6. Unlike in the case of *Knowledge type*, the difference between macro- and micro-average is much smaller. This is due to the fact that there are only three possible labels that a classifier can assign.

Algorithm	ma P	ma R	ma F_1	mi F_1
Majority	15.90%	33.33%	21.53%	47.70%
SVM	89.20%	85.67%	87.40%	90.90%
RandFor	88.00%	79.50%	83.53%	87.70%
JRip	91.40%	67.90%	77.92%	87.70%
J48	87.27%	80.50%	83.75%	81.30%
N. Bayes	76.03%	79.00%	77.49%	83.70%
Vote	87.33%	82.07%	84.62%	88.60%

Table 6: Performance of various classifiers in identifying the *Certainty* of causal relations.

The best results are obtained by the SVM classifier, which reaches 90.90% micro F-score and 87.40% macro F-score. Class *L1* is recognised with the lowest precision and recall amongst the three classes, due to its low number of instances. The low scores of Naïve Bayes and J48 damages the performance of the Vote meta-classifier, which is the second best amongst all algorithms.
The most important features for this dimension are, as expected, the certainty clues previously described. The fact that triggers contain words such as *may*, *probably*, *suggest* or *can* is a good indicator for the correct certainty level.
Many of the error cases happen between the two uncertain classes, *L1* and *L2*. It is usually the case that *L1* relations are wrongly classified as *L2*. Furthermore, there are several instances of mostly *L2*, but also *L1*, classified as *L3* and vice-versa. For instance, in example (2), the causal relation

is speculated, but the model decided that it is certain and belongs to *L3*.

(2) [32] has shown that mutation of phosphotransferase system (PST) in extraintestinal pathogenic E. coli (ExPEC) *can cause* the loss of its colonization ability in extraintestinal organs, and bacteria are cleared rapidly from the bloodstream.

4.2.3. Polarity

The *Polarity* of causal relations is the most correctly recognised MK dimension amongst all four in terms of micro-average F-score, and the results for it are shown in Table 7. This is due to the fact that this dimension has the most skewed label distribution of all: 9 negative to 791 positive instances. As a consequence, the baseline is very high as well, reaching 98.30% micro F-score, but just under 50% macro F-score.

Algorithm	ma P	ma R	ma F_1	mi F_1
Majority	49.45%	50.00%	49.72%	98.30%
SVM	91.40%	77.75%	84.02%	99.30%
RandFor	89.70%	72.15%	79.97%	99.10%
JRip	91.40%	77.75%	84.02%	99.30%
J48	49.45%	50.00%	49.72%	98.30%
N. Bayes	58.90%	81.65%	61.17%	97.30%
Vote	89.70%	72.15%	79.97%	99.10%

Table 7: Performance of various classifiers in identifying the *Polarity* of causal relations.

The best overall results are obtained by SVM and JRip, in both macro- and micro-average settings. However, amongst all classifiers, Naïve Bayes manages to identify correctly most of the minority class instances, reaching a recall of 66.67%. In contrast, its recall for positive instances and precision for negative instances are the lowest, fact which affects the final micro-F-score, making it perform worse than the baseline rule in a micro setting. In addition, the low performance of Naïve Bayes, as well as that of J48, influence negatively the result of the Vote meta-classifier, which gets the second best result.

The most salient features are the placement of negation particles in the vicinity of triggers. This leads to some error cases arising from those triggers which are negated not by the use of negating particles (e.g., *not*), but by using inherently negative triggers, such as in example (3). The verb *rule out* implicitly suggests a negative polarity. However, the sparse data in what regards relations negated by such means affects its correct recognition.

(3) Therefore, the DNA-induced resistance of biofilms requires both the cultivation and challenge under cation-limiting conditions.

These latter two observations *rule out* the possibility that negatively charged DNA simply interacts with cationic antimicrobial peptides and prevents their access to bacterial cells.

4.2.4. Source

The results of the classifiers in the case of the *Source* of causal relations are shown in Table 8. The best micro performance is achieved by the JRip classifier, at 90.20% F-score, whilst the best macro result is obtained by Random Forest, at 73.77%. The difference between these two classifiers is not that large, being less than 2% for macro and just 0.40% for the micro F-score. The main problem of these two classifiers is the low recall for the *Other* label, which is under-represented when compared to the *Current* label. The best recall for this class is achieved by Naïve Bayes, which captures 47.40% of its instances. However, thr precision drops significantly to only 24%, whilst JRip and Random Forest reach up to 80%.

Most errors occur when instances of *Other* are classified as *Current*.

Algorithm	ma P	ma R	ma F_1	mi F_1
Majority	45.25%	50.00%	47.51%	86.00%
SVM	77.30%	61.10%	68.25%	89.60%
RandFor	86.15%	64.50%	73.77%	89.80%
JRip	84.30%	62.10%	71.52%	90.20%
J48	45.25%	50.00%	47.51%	86.00%
N. Bayes	59.00%	66.11%	62.35%	83.40%
Vote	84.85%	60.85%	70.87%	89.90%

Table 8: Performance of various classifiers in identifying the *Source* of causal relations.

5. Conclusions and Future Work

This paper has described our approach to the enrichment of the BioCause corpus, which contains discourse causality associations, with meta-knowledge information. This type of contextual information regarding causal relations is crucial for their correct interpretation. Modifiers such as *not* and *might* completely alter the meaning and certainty of a relation, especially when placed in the context of a network of causal relations. Furthermore, it is important to recognise what type of knowledge the causal relations refer to and whether it is new or old knowledge. This helps the creation of new, testable hypotheses and the assignment of literature support to those relations which contain references.

We have adapted an existing meta-knowledge annotation scheme designed for biomedical events to the needs of discourse analysis. The annotation has been performed by two humans, and the inter-annotator agreement between them is high, ranging between 0.89 and 0.95 Kappa.

A manual analysis of how causality associations are expressed in the biomedical domain has been performed. This brought light into what phrases are used to convey negation, uncertainty, various knowledge types and source of statements.

Additionally, machine learners have been trained to automatically identify the value of each MK dimension for each causal relation. The algorithms base their decisions on a mixture of lexical, syntactic and semantic features, most of which are produced from automatic parses by off-the-shelf systems. Considering the skewness of the data, the classifiers perform reasonably well. SVM obtains the best

scores in the case of *Certainty* and *Polarity*, whilst Random Forest is the best at recognising the *Source* dimension. The best model for *Knowledge Type* considers all five algorithms combined by the Vote meta-classifier. Since the data is so sparse for some dimensions, more would be welcomed in order to be able to create more accurate models.

As future work, we plan to include in the annotation the 24 full-text articles from the BioDRB corpus. The larger amount of data will most likely provide a better understanding of the role played by meta-knowledge in biomedical scientific discourse analysis and how it can help improve automatic discourse analysis.

Acknowledgments

This work was partially funded by the EPSRC [grant number EP/P505631/1]; Medical Research Council; NESTA.

6. References

Barker, C. and Pullum, G. K. (1990). A theory of command relations. *Linguistics and Philosophy*, 13(1):1–34.

Blanco, E., Castell, N., and Moldovan, D. (2008). Causal relation extraction. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., and Tapias, D., editors, *In Proceedings of the 6th International Conference on language Resources and Evaluation (LREC)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270.

de Waard, A., Shum, S. B., Carusi, A., Park, J., Samwald, M., and Sándor, Á. (2009). Hypotheses, evidence and relationships: The hyper approach for representing scientific knowledge claims. In *Proceedings 8th International Semantic Web Conference, Workshop on Semantic Web Applications in Scientific Discourse. Lecture Notes in Computer Science*, Berlin, October. Springer Verlag.

Girju, R. (2003). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering - Volume 12*, MultiSumQA '03, pages 76–83, Stroudsburg, PA, USA. Association for Computational Linguistics.

Greenberg, J., Ferguson, C., and Moravcsik, E. (1978). *Universals of Human Language: Method & theory*. Language science and national development series. Stanford University Press.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November.

Jessop, D., Adams, S., Willighagen, E., Hawizy, L., and Murray-Rust, P. (2011). OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, 3(1):41.

Kilicoglu, H. and Bergler, S. (2008). Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9(Suppl 11):S10.

Liakata, M. and Soldatova, L. (2009). ART corpus. http://hdl.handle.net/2160/1979.

Liakata, M., Teufel, S., Siddharthan, A., and Batchelor, C. (2010). Corpora for the conceptualisation and zoning of scientific papers. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Mihăilă, C., Ohta, T., Pyysalo, S., and Ananiadou, S. (2013). BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14(1):2, January. Highly Accessed.

Miyao, Y. and Tsujii, J. (2008). Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34(1):3580, March.

Nawaz, R., Thompson, P., and Ananiadou, S. (2012). Identification of manner in bio-events. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3505–3510.

Nawaz, R., Thompson, P., and Ananiadou, S. (2013a). Something old, something new: identifying knowledge source in bio-events. *International Journal of Computational Linguistics and Applications*, 4(1):129–144.

Nawaz, R., Thompson, P., and Ananiadou, S. (2013b). Negated bioevents: Analysis and identification. *BMC Bioinformatics*, 14(1):14, January. Highly Accessed.

Nobata, C., Sasaki, Y., Okazaki, N., Rupp, C., Tsujii, J., and Ananiadou, S. (2009). Semantic search on digital document repositories based on text mining results. In *International Conferences on Digital Libraries and the Semantic Web 2009 (ICSD2009)*, pages 34–48.

Oda, K., Kim, J.-D., Ohta, T., Okanohara, D., Matsuzaki, T., Tateisi, Y., and Tsujii, J. (2008). New challenges for text mining: mapping between text and manually curated pathways. *BMC Bioinformatics*, 9(S-3).

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., and Tapias, D., editors, *In Proceedings of the 6th International Conference on language Resources and Evaluation (LREC)*, pages 2961–2968.

Prasad, R., McRoy, S., Frid, N., Joshi, A., and Yu, H. (2011). The biomedical discourse relation bank. *BMC Bioinformatics*, 12(1):188.

Ríos Gaona, M. A., Gelbukh, A., and Bandyopadhyay, S. (2010). Recognizing textual entailment using a machine learning approach. In Sidorov, G., Hernández Aguirre, A., and Reyes García, C., editors, *Advances in Soft Computing*, volume 6438 of *Lecture Notes in Computer Science*, pages 177–185. Springer Berlin / Heidelberg.

Sándor, A. and de Waard, A. (2012). Identifying claimed knowledge updates in biomedical research articles. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, ACL '12, pages 10–17, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sándor, Á. (2007). Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts. *Revue Française de Linguistique Applique*, 200(2):97–109.

Thompson, P., Nawaz, R., McNaught, J., and Ananiadou, S. (2011). Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12:393.

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., and Tsujii, J. (2005). Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, volume 3746 of *LNCS*, pages 382–392. Springer-Verlag, Volos, Greece, November.

Vincze, V., Szarvas, G., Farkas, R., Mora, G., and Csirik, J. (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Witten, I. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann.