Optimized Machine Learning Methods Predict Discourse Segment Type in Biological Research Articles

Jessica Cox[1], Corey Harper[1], and Anita de Waard[1]

[1] Elsevier, Amsterdam, Netherlands
c.harper@elsevier.com

Abstract: To define salient rhetorical elements in scholarly text, we have earlier defined a set of Discourse Segment Types: semantically defined spans of discourse at the level of a clause with a single rhetorical purpose, such as Hypothesis, Method or Result. In this paper, we use machine learning methods to predict these Discourse Segment Types in a corpus of biomedical research papers. The initial experiment used features related to verb type and form, obtaining F-scores ranging from 0.41-0.65. To improve our results, we explored a variety of methods for balancing classes, before applying classification algorithms. We also performed an ablation study and stepwise approach for feature selection. Through these feature selection processes, we were able to reduce our 37 features to the 7 most informative ones, while maintaining F1 scores in the range of 0.63-0.65. Next, we performed an experiment with a reduced set of target classes. Using only verb tense features, logistic regression, a decision tree classifier and a random forest classifier, we predicted that a segment type was either a Result/Method or a Fact/Implication, with F1 scores above 0.8. Interestingly, findings from this machine learning approach are in line with a reader experiment, which found a correlation between verb tense and a biomedical reader's interpretation of discourse segment type. This suggests that experimental and concept-centric discourse in biology texts can be distinguished by humans or machines, using verb tense as a key feature.

Keywords: discourse segments, machine learning, sentence structure, linguistics

1. Introduction

To make sense of the overwhelming flood of scientific literature, a wealth of research has been done to support the development of online reasoning systems by analysing linear scholarly narratives and identifying salient components (see e.g. [1] for an overview of related work). As a first step in this analysis, the text needs to be parsed to identify the level of textual granularity that most closely defines what a 'salient component' is. Various different schemes for annotating discourse elements in scientific texts have been proposed (see e.g. [8] for an overview of other models of analysis).

To motivate our own choice of granularity, see sentences (1) – (3), taken from Voorhoeve et al. (2006):

(1) [An] escape from oncogene-induced senescence is a prerequisite for full transformation into tumor cells. (FACT)

(2) a. To identify miRNAs that can interfere with this process (GOAL)

b. and thus might contribute to the development of tumor cells, (HYPOTHESIS)

c. we transduced BJ/ET fibroblasts with miR-Lib (METHOD)

d. and subsequently transduced them with either RASV12 or a control vector (Figure 2B). (METHOD)

(3) After 2 or 3 weeks in culture, senescence-induced differences in abundance of all miR-Vecs were determined with the miR-Array. (RESULT)

Clearly, several distinct meanings are stated within these three single sentences: for example, in (2), the goal of the (sub)-experiment is first stated, followed by a hypothesis. After the comma, this is followed by a description of methods used. Given these definitions, it is clear that sentences are not the right level of granularity to qualify as Discourse Segments. Given this, we decided to identify Discourse Segments at approximately the level of a clause (i.e. a coherent sentence fragment containing a single verb). Next, we defined a small taxonomy of semantic (or pragmatic) segment types, with which to classify these Discourse Segments (see Table 1 for a definition of these types, taken from [3]). For further details on our segmentation and motivation for these Discourse Segment Types or DSTs, see [3].

Table 1: Discourse Segment Type Classification (DST)

Discourse Segment Type	Definition	Example
Goal	Research goal	*To examine the role of endogenous TGF-β signaling in restraining cell transformation,*
Fact	A known fact, a statement taken to be true by the author.	*Sustained proliferation of cells in the presence of oncogenic signals is a major leap toward tumorigenicity.*
Result	The outcome of an experiment	*Two largely overlapping constructs encoded both miRNA-371 and 372 (miR-Vec-371&2).*
Hypothesis	A claim proposed by the author	*These miRNAs could act on a factor upstream of p53 as a cellular suppressor to oncogenic RAS.*
Method	Experimental method	*We examined p53 mutations in exons five to eight in the primary tumors.*
Problem	An unresolved or contradictory issue	*The mechanism underlying this effect and its conservation to other tissues is not known.*
Implication	An interpretation of the results	*[This indicates that] miR-372/3 acts as a molecular switch.*

In earlier work, we identified three types of lexicogrammatical features that identify the various 'discourse realms' which these segments occupy, in particular: is the Discourse Segment related to *experimental* text (as in the case of Goals, Methods and Results), or *conceptual* text (as in the case of Hypotheses, Problems, Implications and Facts) (see [5] for a further definition of these ideas). The features we explored were

i. verb tense and form: Tense (i.e., Past/Present/Future tense), Verb Form (Perfective, Progressive, or unmarked) for each tense, and two nonfinite verb forms (To-infinitive or Gerund or '-ing' form) [5];
ii. a taxonomy of semantic verb classes [6], and
iii. a series of modality markers [11].

Our previous research established that there was a clear correlation between, in particular, verb tense and form and discourse realm. In particular, Methods and Results were correlated with Past Tense, Hypotheses with Modal Auxiliary Verbs, and Goals with To-Infinitives. We found both a correlation in our corpus study, as well as in a reader experiment: changing the verb tense changed the reader's interpretation of the discourse segment [3].

We saw that this work could be applied, for instance, to identify salient segment types, such as Implications, from large corpora of text: this means there might be applications from this work to support, for instance, automated summarization efforts. A first effort to scale up the identification of DSTs with text processing tools was done quite early on, with some promising results [4], but this was done at scale. In later work, we developed a classifier for these seven DSTs, and trained this on sentences in the Results sections of biology papers, achieving an overall F-score of 0.63 [1]. We wanted to improve on this score and enable large-scale text processing and identification of DSTs. In particular, we were interested to know if the verb features that were so significant for the reader studies could be used to identify discourse segments and help build tools that identify key conclusions or experimental segments from papers.

In this study, we present an exhaustive approach to automatically identifying discourse segment types using supervised machine learning methods. We work on the full text of a manually curated set of biomedical research papers, using a set of features derived from the corpus studies. The novelty of our approach rests in the fact that this dataset presents a typical challenge in employing classification algorithms, due to the severe class imbalance of our predicted classes. To improve model fitting and accuracy, we therefore first balanced the classes before subjecting them to classifiers, using several different under- and over-sampling methods [9]. This resulted in a set of 36 models that all used a different combination of class balancers and classification algorithms to predict segment discourse type. Next, we culled these models to select only those that include the most important features and produce the highest accuracy and F1 scores.

2. Methods and Results

In exploring this data, we first ran some pre-processing and filtering on our dataset and tackled some preliminary feature selection. We then ran 3 baseline classification algorithms on this preprocessed data, before proceeding with 4 experiments. Experiment 1 was a test of Class Balancing tools, experiments 2 and 3 were ablation studies to further limit our features, and experiment 4 was a separate and simplified classification problem prompted by our findings in the initial 3 experiments. In the methods and results section here, we walk through each of these experiments and the results. A summary of all 4 experiments is presented in Figure 1 at the end of the section.

2.1 Dataset curation

The dataset was based on a set of 10 papers in cell biology and electrophysiology, which was manually split into discourse segments and marked up with Discourse Segment Types described above, and in [3]. The full, manually curated dataset can be found on Mendeley Data at https://data.mendeley.com/datasets/4bh33fdx4v/3 [2].

The corpus started as a set of 3,239 Type-identified Discourse Segments, which were loaded into a Jupyter notebook. (See Suplementary Material, below.) Our predicted class is "Discourse Segment Type", as outlined in Table 1. Data points with the segment type "blank" (316), "header" (134), or "null" (1) were eliminated from the dataset, as were entries containing an empty verb type or verb form field (8). The missing data points were most likely due to an inability of the curators to identify the target class or feature. We determined *Header* to not be a useful discourse segment type for prediction in this case. We also eliminated those that were labeled "Intratextual" or "Intertextual" (taken from our earlier taxonomy, but not included in the current set of DST's, as their classes were particularly small relative to the size of the dataset, (71 and 14, respectively)). Our final dataset contained 2,695 points.

2.2 Feature Selection

We created a set of 32 features. These features were based on the three earlier linguistic explorations and can be grouped in three distinct classes: Verb form/tense, Verb Class, and Modality markers. Next to these, we created a new feature that reflects whether the verb used in the segment was in the top 10 most frequently used verbs: "show", "indicate", "demonstrate", "suggest", "use", "identify", "reduce", "suppress", "express", and "examine", as well as a separate feature, 'Show". All of our categorical features were converted to dummy variables so that they could be appropriately included within the model. These are described in Appendix 5.1.

2.3 Model Construction

Given that many of the predicted classes are severely unbalanced, as shown in Table 3, we began by employing a variety of methods to account for these differences, by either under-sampling the majority class or over-sampling the minority classes. These methods were imported from the scikit-learn imblearn library [9]. We used 6 under-samplers, 4 over-samplers and 2 that used a combination of under- and over-sampling. The list and brief description of the 12 class balancers used are described in Appendix 5.2.

Table 3. Number of segments per DST

Segment Type	Number
Result	851
Implication	657
Method	351
Hypothesis	315
Fact	262
Goal	149
Problem	110

The data was split into a test and training set, using a test size equivalent to 30% of the total data. It was first fed through one of the 12 class balancers, and then to one of three classifiers (experiment 1) : logistic regression (LR), decision tree classifier (DTC) or random forest classifier (RFC). Logistic regression was performed using an LBFGS solver to handle multinomial loss. In addition to our experiment with class balancers, we also ran a baseline set of all 3 classifiers with no Class Balancer included. The accuracy, precision, recall and F1 scores were generated for each of these models and then compared, shown in Appendix 5.3.

The results in Appendix 3 show models ran with TomekLinks, SMOTE, SMOTEborderline, SMOTEborderline2 and SMOTETomek to have the highest performance. Highest accuracy was achieved by a decision tree and random forest classifier, using TomekLinks, with a score of 0.64. Highest precision was scattered across a few class balancers that used logistic regression with a score of 0.68. The highest recall score was 0.64, achieved by decision tree and random forest with TomekLinks. The highest F1 score was 0.65, also scattered across models that used a few class balancers and logistic regression.

Because these scores showed no improvement over our baseline model, we next sought to reduce model complexity and identify the most significant features. An ablation study (experiment 2) was performed using a random forest classifier. We looped through the model, and on each iteration removed the least informative feature in the dataset. During this process, our F1 Score stayed between .62-.64 until we'd removed more than half of our features. We ranked the features based on their significance to the model and then trimmed the features to the 9 most significant (Past', 'Present', 'To-infinitive', 'Interpretation', 'Investigation', 'Procedure', 'Modal', 'Verb_Class_Interpretaion', 'Ruled_by_VC_Interpretation'). These features were determined to be most significant by buidling a feature ranking list for each model, dropping the least significant , and stopping when we'd reched a significant drop in F1 score.

We then reran these features through a smaller set of class balancers, (TomekLinks, SMOTE, SMOTEborderline, SMOTEborderline2, SMOTETomek) followed by logistic regression, decision tree or random forest, reported in Table 5. These class balancers were chosen because they produced models with the highest overall metrics in the outcomes of experiment 1. We did not observe a vast improvement in performance in these models compared to those containing all of the features outlined in Appendix 5.3. However it is worth noting that we did not experience a drop-off in results either. This indicates that the majority of predictive power is coming a smaller number of features. In fact, if you look at the feature importance on this 9 feature model, nearly 45% of the descriptive confidence is coming from the first two features alone, past vs present tense verb.

Table 5. Ablation study model performance metrics

Classifier	Class balancer	Accuracy	Precision	Recall	F1
LR	TomekLinks	0.64	0.67	0.64	0.65
DTC	TomekLinks	0.65	0.66	0.65	0.64
RFC	TomeLinks	0.65	0.66	0.65	0.64
LR	SMOTE	0.64	0.67	0.64	0.65
DTC	SMOTE	0.63	0.66	0.63	0.64
RFC	SMOTE	0.63	0.66	0.63	0.63
LR	SMOTEborderline	0.60	0.66	0.60	0.62
DTC	SMOTEborderline	0.57	0.65	0.57	0.59
RFC	SMOTEborderline	0.57	0.65	0.57	0.59
LR	SMOTEborderline2	0.61	0.66	0.61	0.62
DTC	SMOTEborderline2	0.63	0.66	0.63	0.64
RFC	SMOTEborderline2	0.63	0.66	0.63	0.63
LR	SMOTETomek	0.64	0.67	0.64	0.65
DTC	SMOTETomek	0.63	0.66	0.63	0.64
RFC	SMOTETomek	0.63	0.65	0.63	0.64

We then tried a third experiment, using another manual approach to feature reduction, in which we used forward feature selection, incorporating each feature in turn into a random forest model, and comparing metrics (experiment 3). This resulted in 13 features to be included ('Past', 'Procedure', 'To-infinitive', 'Modal', 'Properties', 'Investigation', 'Future', 'show_verb', 'Observation', 'Verb_Class_Interpretaion', 'Interpretation', 'Ruled_by_VC_Interpretation', and 'Past Progressive'). Of note, this list has a fairly

consistent overlap with the features appeared in our initial ablation study. Again, we reran these features through the same smaller set of class balancers and then through logistic regression, decision tree and random forest classifier. Metrics are reported in Table 6.

Table 6. Forward selection study model performance metrics

Classifier	Class balancer	Accuracy	Precision	Recall	F1
LR	TomekLinks	0.65	0.68	0.65	0.65
DTC	TomekLinks	0.66	0.68	0.66	0.66
RFC	TomekLinks	0.66	0.68	0.66	0.66
LR	SMOTE	0.65	0.68	0.65	0.65
DTC	SMOTE	0.62	0.69	0.62	0.64
RFC	SMOTE	0.62	0.69	0.62	0.64
LR	SMOTEborderline	0.57	0.65	0.57	0.59
DTC	SMOTEborderline	0.57	0.64	0.57	0.59
RFC	SMOTEborderline	0.57	0.64	0.57	0.59
LR	SMOTEborderline2	0.56	0.65	0.56	0.56
DTC	SMOTEborderline2	0.65	0.67	0.65	0.65
RFC	SMOTEborderline2	0.65	0.67	0.65	0.65
LR	SMOTETomek	0.65	0.68	0.65	0.65
DTC	SMOTETomek	0.62	0.69	0.62	0.64
RFC	SMOTETomek	0.62	0.69	0.62	0.64

Again, we did not observe a vast improvement in performance in this subset of features. However, these scores do mirror and slightly improve those presented in Tables 4 and 5, reinforcing the significance of these specific features. A confusion matrix was generated for each model presented in these tables. This provides a more nuanced view of the breakdown of class prediction. These matrices are available in our supplemental dataset in Mendeley: http://dx.doi.org/10.17632/tds3k5kyvg.1.

2.4 Verb tense experiment

Due to concerns regarding dataset size and predicted class distribution, we designed a fourth experiment that diverged from the others in two ways. First, we subsampled from the larger dataset, specifically data labeled with a segment type of “Result”, “Method”, “Fact” or “Implication”. The second difference is that we limited our feature set to verb tense features

("Future", "Gerund", "Past", "Past participle", "Past perfect", "Past progressive", "Present", "Present perfect", "Present progressive", "To-infinitive"). In earlier manual corpus work [5] we established that Present was the predominant tense for Fact and Implication statements, and Method and Result were predominantly described with a Past tense. We interpreted these results from a cognitive linguistics standpoint as tense markings being specific for a specific 'discourse realm', where experimental segments (such as Method and Result) are predominantly described in a (narrative) past, which lies within the author's personal experience, whereas factual or conceptual statements are presented in the 'gnomic', eternal Present, also used to describe faculties of statements of fact in other forms of discourse, such as mythological text.

For this experiment, we ended up with a "Result/Method" class with 1205 data points, and "Fact/Implication" class with 922 data points. Because these classes are more balanced, we did not need to first subject them to a class balancer. We again ran the data through three classifiers: logistic regression, decision tree and random forest, shown in Table 7. Tables 8-10 present the confusion matrices generated for each of the three models represented in Table 7. The results presented in Tables 6-9 show that these models perform remarkably similarly and all achieved an F1 score of 0.80-0.81.

Table 7: Performance metrics of 3 models to evaluate segment type based on verb tense.

Classifier	**Accuracy**	**Precision**	**Recall**	**F1 score**
Logistic Regression	0.80	0.81	0.80	0.80
Decision Tree Classifier	0.81	0.82	0.81	0.81
Random Forest	0.81	0.82	0.81	0.81

Table 8. Logistic regression model confusion matrix

True label	*Result/Method*	238	37
	Fact/Implication	91	271
		Result/Method	*Fact/Implication*
		Predicted label	

Table 9. Decision tree classifier model confusion matrix

True label	*Result/Method*	238	37
	Fact/Implication	86	276
		Result/Method	*Fact/Implication*
		Predicted label	

Table 10. Random forest classifier model confusion matrix

True label	*Result/Method*	236	39
	Fact/impliImplication cation	84	278
		Result/Method	*Fact/Implication*
		Predicted label	

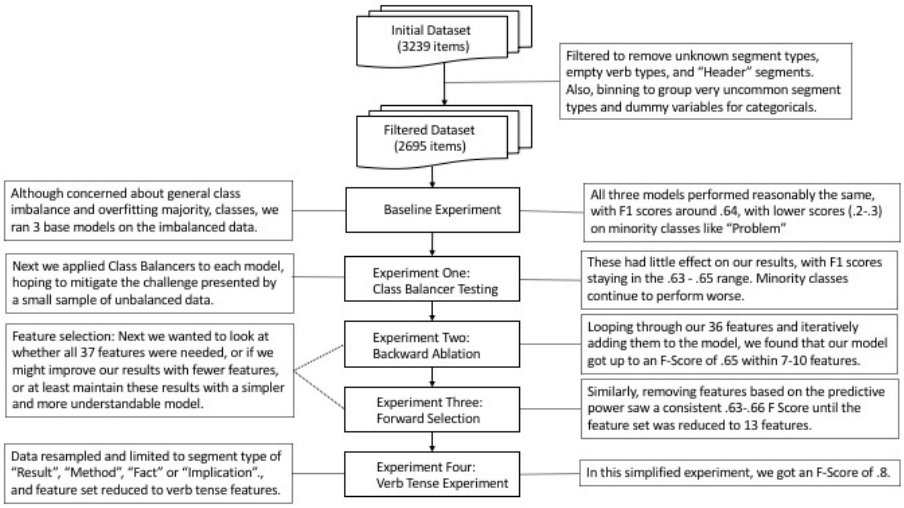

Fig. 1. A figure describing the 4 separate experiments undertaken for this research

3. Discussion

The current work presents a methodical machine learning approach to classifying segment discourse type, using verb tense features of the segment. We approached this problem using a variety of machine learning methods. The first challenge we encountered was the severe class imbalance amongst the predicted segment types. The Result and Implication classes both had greater than 650 data points each, while the five remaining classes had less than 351 each. The smallest class, Problem, only had 110. While this is reflective of the breakdown of segment discourse within a biomedical research paper, it makes for an obstacle when using machine learning algorithms that rely on relatively balanced classes.

We turned to the imblearn package in scikit-learn [9] to aid in balancing the classes before running our classifiers. This package provides a suite of sampling methods that can be used to either over-sample the minority class by synthetically generating these data points, or under-sample the majority class by throwing some points out. If the classes remained unbalanced, we would observe bias towards the majority classes. To avoid this, we decided to try a combination of all of these methods in conjunction with our classifiers and to evaluate the differences between them. This approach is similar to what [7] used in their paper, where they compared performance metrics of different combinations of sampling methods and classifiers on two different datasets.

In our first experiment, we observed that models that performed more poorly than our baseline all used the same balance method, and there didn't appear to be large differences in what classifier was used. While the remaining balancers didn't seem to have a huge impact on results. Specifically, these poorly performing balancers appeared to boost precision over recall. For example, we observed use of ClusterCentroids in conjunction with the three classifiers produced accuracy and F1 scores of 0.35-0.55, while the precision scores were all above 0.55. ClusterCentroids is an undersampling method that replaces a cluster of samples with its cluster centroid, as calculated by the KMeans algorithm. Because the majority class is being replaced with its cluster centroid, it is unsurprising that we observe higher precision than recall. This illustrated the need for refining our feature selection, which we approached with two different experiments. Additionally, we used the metrics from the first experiment to limit or sampling methods to SMOTE, TomekLinks, SMOTEborderline and SMOTETomek.

SMOTE (synthetic minority over-sampling technique) generates synthetic data points for the minority classes. Given how small the dataset is in general (~2600) it is unsurprising that this technique yielded the best results by increasing the numbers within our dataset. SMOTEborderline specifically oversamples those on the border of the majority and minority classes. TomekLinks is an under-sampling method that removes data points that are near the border of the majority and minority classes. This allows for cleaning of the data in such a way that noise is reduced and the training set is improved. SMOTETomek is a combination of performing SMOTE on the minority class and TomkeLinks on the majority class.

After identifying the top performing models in Table 4, we designed two additional experiments to refine our feature selection and reduce model complexity. Our ablation study identified 9 significant features: 'Past', 'Present', 'To-infinitive', 'Interpretation', 'Investigation', 'Procedure', 'Modality Markers', 'Verb_Class_Interpretation', 'Ruled_by_VC_Interpretation'. Our forward selection experiment identified 13 significant features: 'Past', 'Procedure', 'To-infinitive', 'Modal', 'Properties', 'Investigation', 'Future',

'show_verb', 'Observation', 'Verb_Class_Interpretaion', 'Interpretation', 'Ruled_by_VC_Interpretation', and 'Past Progressive'.

We did not observe a remarkable improvement in model performance, largely due to the limitation of our sample size, though our overall F1 Scores seemed to improve slightly from an average of 63-64 to an average of 65-66. Some of the class sizes in our test set classes had less than 50 data points.

Interestingly, these experimental results suggest that indeed, verb class and verb form are key markers for Discourse Segment Type: Past, Present, Modal and To-infinitive are important markers for identifying the realm of the Discourse, as found in the reader study described above. Of the Verb Classes, again, these were predominant markers, and therefore match reader experiments [3]. This implies in any event that verb tense should not be discarded or ignored in text mining experiments, as is often done.

Given the small sample size and average model results, we took on a second approach in which we subsetted the data. We grouped together the Result and Method classes and the Fact and Implication classes and dropped all other points to explore correlations between the two most frequent experimental Discourse Segment types and the two predominant conceptual Discourse Segment Types, and limited our features to those that were related to verb tense. Table 7 lists the performance metrics of the three models, in which the scores were very similar to one another, and significantly higher than what we observed in the first few experiments. This was in line with our expectations, given the tighter classes and the selected features. The confusion matrices also illustrate how the classes are predicted in our test set, and all three models tend to classify them the same.

In earlier discourse work, we investigated whether these tense correlations were perceived to be defining of discourse realm for a reader. In [3] we conducted a reader experiment, where 21 subjects with a biology background were asked to identify Discourse Segment Type for a set of segments which presented either in unmodified form, of with a modified tense. We found that significantly, verb tense was strongly correlated with segment type, especially for Implications and Results. This bears a striking similarity with the machine learning results found in this study.

We are exploring a number of future directions for this research. One line of research is to experiment with other corpora, e.g. in other domains, or other document types. Our data set is hand-coded and it remains to be seen how these results apply to unknown data. The challenge here is getting labels of discourse type assigned to test data. The features themselves likely don't need to be hand-curated and can be generated with standard natural language processing (NLP) techniques. However, labeling segment types correctly requires more work. Initial explorations look fruitful, and we are exploring the use of "Snorkel" [10] to produce noise aware generative models to help bootstrap additional training data in

other domains. Additional future work involves applying the segment types on the sentences of citations to other papers. In combination with graph and network analysis and other term frequency analysis, this would support the classification of reason and type of citation.

In summary, our main objective was to predict segment discourse types based on lexicogrammatical features, and in doing so, we have found a good correlation with corpus studies. In the process of doing this, we contribute to the development of methods used to examine an unbalanced dataset in linguistic discourse analysis. Future work includes applying our models on additional datasets and combining with other research, such as citing sentence and citation graph analyses.

4. Supplemental material

Full manually curated dataset can be found here: de Waard, Anita (2017), "Discourse Segment Type vs. Linguistic Features", Mendeley Data, v3 http://dx.doi.org/10.17632/4bh33fdx4v.3

Jupyter notebooks containing steps to reproduce, analyze and view output are available here: Cox, Jessica (2017), "Optimised Machine Learning Methods Predict Discourse Segment Type in Biological Research Articles", Mendeley Data http://dx.doi.org/10.17632/tds3k5kyvg.1

5. Appendices

Appendix 5.1. Starting feature list and descriptions

Feature Class	Feature	Included in Experiment #
Frequently Used Verb	Top 10 Verb	1
Frequently Used Verb	'Show' Verb	1, 3
Verb Tense	Future	1, 3, 4
Verb Tense	Gerund	1, 4
Verb Tense	Past	1, 2, 3, 4
Verb Tense	Past participle	1, 4
Verb Tense	Past perfect	1, 4
Verb Tense	Past progressive	1, 3, 4
Verb Tense	Present	1, 2, 4
Verb Tense	Present perfect	1, 4
Verb Tense	Present progressive	1, 4
Verb Tense	To-infinitive	1, 2, 3, 4
Verb Class	Cause and effect	1
Verb Class	Change and growth	1
Verb Class	Discourse verb	1

Verb Class	Interpretation	1, 2, 3
Verb Class	Investigation	1, 2, 3
Verb Class	None	1
Verb Class	Observation	1, 3
Verb Class	Prediction	1
Verb Class	Procedure	1, 2, 3
Verb Class	Properties	1, 3
Modality Marker	Modal	1, 2, 3
Modality Marker	Verb class interpretation	1, 2, 3
Modality Marker	Ruled by verb class interpretation	1, 2, 3
Modality Marker	Reference internal	1
Modality Marker	Reference external	1
Modality Marker	First person	1
Modality Marker	Modal significant_ly	1
Modality Marker	Possible_ility_ly	1
Modality Marker	Potential_ly	1
Modality Marker	UN_Likely	1
Modality Marker	Sum_Adverbs_YesNO	1

Appendix 5.2. Description of sampling methods used.

Sampling Method	Description	Method
RandomUnderSampler	Undersamples the majority classes by randomly picking samples	Undersampler
Tomeklinks	Undersamples the majority classes by removing Tomek's links	Undersampler
ClusterCentroids	Under samples the majority classes by replacing a cluster of the majority samples by the cluster centroid of a KMeans algorithm	Undersampler
CondensedNearestNeighbor	Under samples the majority classes using the condensed nearest neighbor method	Undersampler
OneSidedSelection	Uses one-sided selection method on majority classes	Undersampler
InstanceHardnessThreshold	Samples with lower probabilities are removed from the majority class	Undersampler
RandomOverSampler	Randomly generates new samples from the minority classes	Oversampler
SMOTE	Synthetic Minority Oversampling Technique; generates new samples of minority class by interpolation	Oversampler
SMOTEborderline	Generates new samples of minority class specific to the borders between two classes.	Oversampler

SMOTEborderline2	Generates new samples of minority class specific to the borders between two classes.	Oversampler
SMOTETomek	Combines use of SMOTE on minority class and Tomek Links on majority class	Over and undersampler
SMOTEENN	Combines use of SMOTE on minority class and Edited Nearest Neighbors on majority class	Over and undersampler

Appendix 5.3. Accuracy, precision, recall and F1 scores of all 36 models tested.

Classifier	Class Balancer	Accuracy	Precision	Recall	F1
LR	No Class Balancer	0.62	0.68	0.63	0.64
DTC	No Class Balancer	0.64	0.64	0.64	0.64
RFC	No Class Balancer	0.64	0.65	0.65	0.64
LR	RandomUnderSampler	0.58	0.64	0.58	0.59
DTC	RandomUnderSampler	0.55	0.64	0.55	0.56
RFC	RandomUnderSampler	0.57	0.63	0.56	0.57
LR	Tomeklinks	0.63	**0.68**	0.63	0.64
DTC	Tomeklinks	**0.64**	0.64	**0.64**	0.64
RFC	Tomeklinks	**0.64**	0.64	**0.64**	0.64
LR	ClusterCentroids	0.55	0.64	0.55	0.55
DTC	ClusterCentroids	0.35	0.48	0.35	0.32
RFC	ClusterCentroids	0.38	0.47	0.38	0.35
LR	CondensedNearestNeighbor	0.62	0.67	0.62	0.62
DTC	CondensedNearestNeighbor	0.53	0.59	0.53	0.53
RFC	CondensedNearestNeighbor	0.55	0.60	0.55	0.55
LR	OneSidedSelection	0.60	0.65	0.6	0.61
DTC	OneSidedSelection	0.47	0.47	0.47	0.46
RFC	OneSidedSelection	0.48	0.43	0.48	0.45
LR	InstanceHarnessThreshold	0.46	0.58	0.46	0.5
DTC	InstanceHarnessThreshold	0.37	0.61	0.37	0.41
RFC	InstanceHarnessThreshold	0.40	0.61	0.4	0.44
LR	RandomOverSampler	0.63	**0.68**	0.63	0.64
DTC	RandomOverSampler	0.60	0.64	0.6	0.61
RFC	RandomOverSampler	0.61	0.64	0.61	0.61
LR	SMOTE	0.63	**0.68**	0.63	0.64

DTC	SMOTE	0.62	0.64	0.63	0.63
RFC	SMOTE	0.63	0.64	0.63	0.63
LR	SMOTEborderline	0.63	**0.68**	0.63	**0.65**
DTC	SMOTEborderline	0.63	0.64	0.63	0.63
RFC	SMOTEborderline	0.62	0.63	0.62	0.62
LR	SMOTEborderline2	0.63	**0.68**	0.63	0.64
DTC	SMOTEborderline3	0.63	0.64	0.63	0.63
RFC	SMOTEborderline4	0.62	0.64	0.62	0.62
LR	SMOTETomek	0.63	**0.68**	0.63	**0.65**
DTC	SMOTETomek	0.63	0.64	0.63	0.63
RFC	SMOTETomek	0.63	0.65	0.63	0.63
LR	SMOTEENN	0.50	0.63	0.50	0.52
DTC	SMOTEENN	0.42	0.65	0.42	0.45
RFC	SMOTEENN	0.44	0.63	0.44	0.46

References

1. Burns, Gully A.P.C., Dasigi, Pradeep, de Waard, Anita, Hovy, Eduard H. (2016) Automated detection of discourse segment and experimental types from the text of cancer pathway results sections, *Database*, Volume 2016, 1 January 2016, baw122, https://doi.org/10.1093/database/baw122
2. de Waard, Anita (2017) Manually curated dataset of papers into segments and DSTs: "Discourse Segment Type vs. Linguistic Features", Mendeley Data, v3 http://dx.doi.org/10.17632/4bh33fdx4v.3
3. de Waard, Anita, Pander Maat, Henk (2012) Verb form indicates discourse segment type in biological research papers: Experimental evidence, Journal of English for Academic Purposes, Volume 11, Issue 4, 2012.
4. de Waard, Anita, Buitelaar, Paul and Eigner, Thomas (2009) Identifying the epistemic value of discourse segments in biology texts. In Proceedings of the Eighth International Conference on Computational Semantics (IWCS-8 '09), Harry Bunt, Volha Petukhova, and Sander Wubben (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 351-354.
5. de Waard, Anita (2010) Realm traversal in biological discourse: from model to experiment and back again, Multidisciplinary Perspectives on Signalling Text Organisation, MAD 2010, Moissac, 17-20 March 2010, p 136, https://hal.archives-ouvertes.fr/hal-01391515/document#page=139
6. de Waard, Anita, Pander Maat, Henk (2010). A Classification of Research Verbs to Facilitate Discourse Segment Identification in Biological Text, in: Proceedings from The Interdisciplinary Workshop on Verbs. The identification and representation of verb features. Pisa, Italy, 2010. http://linguistica.sns.it/Workshop_verb/papers/de%20Waard_verb2010_submission_69.pdf

7. Elhassan T, Aljurf M, Al-Mohanna F and Shoukri M. (2016). Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. Journal of Informatics and Data Mining. 1(2):11, 1-12. 2016. http://datamining.imedpub.com/classification-of-imbalance-data-using-tomek-linktlink-combined-with-random-undersampling-rus-as-a-data-reduction-method.pdf
8. Liakata, Maria, Paul Thomson, Anita de Waard, et al. (2012), A Three-Way Perspective on Scientific Discourse Annotation for Knowledge Extraction, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 37–46, Jeju, Republic of Korea, 12 July 2012, http://www.aclweb.org/anthology/W12-4305
9. Lemaitre, Guillaume, Fernando Nogueira, Christos K. Aridas (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research. 18(17): 1-5, 2017. http://jmlr.org/papers/v18/16-365
10. Ratner, Alexander, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, Christopher Ré (2017) Snorkel: Rapid Training Data Creation with Weak Supervision. Proceedings of the VLDB Endowment, 11(3): 269-282, 2017. https://arxiv.org/abs/1711.10160
11. de Waard, Anita, Pander Maat, Henk (2012) Epistemic modality and knowledge attribution in scientific discourse: a taxonomy of types and overview of features. In Proceedings of the Workshop on Detecting Structure in Scholarly Discourse (ACL '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 47-55. https://dl.acm.org/citation.cfm?id=2391180