

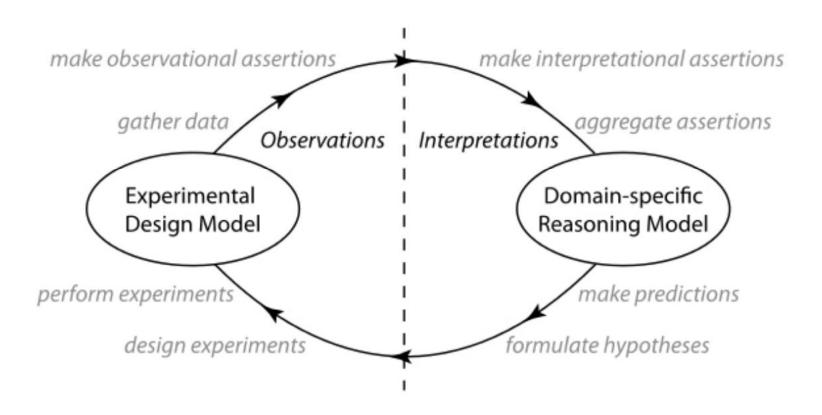


Jessica Cox, Corey A Harper, & Anita de Waard -- SAVE-SD April 24, 2018

The IMRaD Structure of A Paper

Cel Cell Introduction: "Create a Research Space" A Pleiotropica miR-31, Inhibits Establish a research territory amount the potential for extrusts From the shows a common of their in Cancer Metastas 201 human breast corear code (*201 Establish a niche MICCO-Figure 51AL Scott Valuatyse, ** Forenc Robbards, Nathon Be-Jame E. Brock, *Andrea L. Richardson, * and Robe If the not affect problemen in cite, but did on by 30 hidd and restlifts by 10-hidd (Figure 2A) Occupy the niche Whitehald Inethals for Stansactical Research, Contribute, M. D and S1Q. These effects were specifically attribut-Separatement of Biologics, Missourch's series treat also of Tarif to Prograd activities of milk S1, as indicated by the fact Available oversepression of a context militals, raffe t-diand the bellumence timesters are more than the property of a send darks need tails an estated 15 communicated militia for father phases. Ann. refl-21 expressing calls settlined 60% desir-Brighans and Woman's Househal, Braden, 58-301 W. Life. ad resistance to weekle-mediated cell death-(Figure 25). treet of Corner Basings, Suran Fastier Corner Institute, St Two extra court not be avorted to basisty weating have MF Ludwig Certer for Molecular Chomissos, Cardintiga, MA 8513 stopic nell-31 (Figure 510). The composences of milk-in "Correspondence with applications of the national property and the contract of appreciation warm not unique to 201 code; nell-21 mid-coal troson, mortifiy, and anothin walstance, yet did not affect profites in aggressive SLAR-150 human toward concer cells Methods and Results: Figure (22), Hamos, milk-St Impacks in vitro economies of mate-MicrofitNe are well outed to regulate turnor metatunis because of their capacity to coordinate) "Cycles of Scientific Investigation" repress numerous target games, thereby potential cause of its effects on it with trails associated with righ enabling their intervention at multiple steps of the ade malignancy, we smad whether actuals milk-bit could sited mode electe to offerwise appropries calls. Those, 201 calls ireasion-metastasis cascade. We identify a nioni prevates collect were injected this the orthologic also the (see below) FINA exemplifying these attributes, milital, who menary had pack-of mice. Unaquestedly, mill-31 enhanced many terms providing 1 is had an incompanied given a seed expression correlates inversely with metastasis human breast cancer patients. Overexpression Confidention (Figure 20) Places (OA), Control 26 Final polynomery milf-31 is otherwise appressive breast tumor or con displayed e-letterox of local invasion; however, milk-tri suppresses metastasis. We deploy a stable micr presently takens were well arrappedated and necknostics games 20 and 21). These changes were not accompanied by FINA sponge strategy to intstat miR-31 in white the allows otherwise-nonaggressive breast cancer cell and consequentiation (Figure 1996). to metertains. These phenotypes do not involve makele new ten contact of the nink species excelled size roll-31 was reduced by 3-bits in revisitatio DE-1 and Despite their strilly to garwents larger primary turnors, 201 pressing-wife I have a strikingly impaired in their capacity confounding influences on primary tumor developmost and are specifically attributable to mife? ad lung restantance, milk 21 expressing calls formed 85% mediated inhibition of several steps of metast mines than this controls \$20 steps after implentation F). This, milk it is appreciate to reduct unit from an orthoincluding local invasion, extraversation or Discussion: densitivy due, office of Impart, to its stiffly to impact survival at a distant site, and metastatic colonic Such pleiotropy is achieved via coordinate, ed the possibility that milk 21% impact on these sion of a cohort of metastasis-promotive artifications to closed vertation in our 201 cale including RhoA, Indeed, RhoA re-express \$1.05 in a single-cell-chartent population bereverses milli-31-imposed metastasis Statement of principal findings pertal 225 cade Figure Stat, Miles et al., These findings indicate that miR-31 on Kilothel otherbelooks with 25 weeren mechanisms to oppose metastasis. well-excapsulated primary tumors and For a video summary of this article date by 5-544 (Figures 549-640). Strengths and weaknesses of the study Flick file available with the orders for indrugs; mill-31 withercost petners apprentist functs were before Relation to other studies Figure 55. Tress characters Metarianes account for 80 Northy ers, as well as to seed fewer Unanswered questions and future research Managa k, 2000, yerborunden ce of the biological activities that govern reductate disc 1002 Col 107, 1002-108, June 12, 2009-02009 (Sector Inc. Carl 137, 1007-1049, June 12, 2009-02009 Element Inc., 1000

Rubber hits the road in Results: Cycles of Scientific Investigation



The Narrative Structure of Research Articles, Or, Why Science is Like a Fairy Tale



Anita de Waard, VP Research Data Collaborations Research Data Management Services, Elsevier

Similar to a fairy tale...

Story Grammar		The Story of Goldilocks and the Three Bears	Paper Grammar	The AXH Domain of Ataxin-1 Mediates Neurodegeneration through Its Interaction with Gfi-1/Senseless Proteins				
Setting	Time	Once upon a time	5 / /	The procedure was distinct COAA prother association and still prot falls.				
	Character	a little girl named Goldilocks	Background	The mechanisms mediating SCA1 pathogenesis are still not fully understood, but some general principles have emerged.				
	Location	She went for a walk in the forest. Pretty soon, she	Objects of study	the Drosophila Atx-1 homolog (dAtx-1) which lacks a polyQ tract,				
		Experiment	studied and compared in vivo effects and interactions to those of the					
Theme	Goal	She knocked and, when no one answered,	al setup	human protein				
			Research	Gain insight into how Atx-1's function contributes to SCA1				
	Attempt	she walked right in.	goal	pathogenesis. How these interactions might contribute to the disease process and how they might cause toxicity in only a subset of neurons in				
Episode N	Name	At the table in the kitchen,		SCA1 is not fully understood.				
		there were three bowls of porridge.	Hypothesis	Atx-1 may play a role in the regulation of gene expression				
	Subgoal	Goldilocks was hungry.	Name	dAtX-1 and hAtx-1 Induce Similar Phenotypes When Overexpressed in Files				
	Attempt	She tasted the porridge from the first bowl.	Subgoal	test the function of the AXH domain				
	Outcome	This porridge is too hot! she exclaimed.	Method	overexpressed dAtx-1 in flies using the GAL4/UAS system (Brand and Perrimon, 1993) and compared its effects to those of hAtx-1.				
	Attempt	So, she tasted the porridge from the second bowl.	Results	Overexpression of dAtx-1 by Rhodopsin1(Rh1)-GAL4, which drives expression in the differentiated R1-R6 photoreceptor cells (Mollereau et al., 2000 and O'Tousa et al., 1985), results in neurodegeneration in the				
Outcome This porridge is too she said		This porridge is too cold, she said		eye, as does overexpression of hAtx-1[82Q]. Although at 2 days after eclosion, overexpression of either Atx-1 does not show obvious				

Discourse Segment Type (DST) Classification

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.

ARGUMENTATION IN THE RESULTS SECTION:

- 1. Importantly, our results so far indicate that the expression of miR-372&3 did not reduce the activity of RASV12, as these cells were still growing faster than normal cells and were tumorigenic, for which RAS activity is indispensable (Hahn et al, 1999 and Kolfschoten et al, 2005).
- 2. To shed more light on this aspect, we examined the effect of miR-372&3 expression on p53 activation in response to oncogenic stimulation.
- 3. We used for this experiment BJ/ET cells containing p14ARFkd because, following RASV12 treatment, in those cells p53 is still activated but more clearly stabilized than in parental BJ/ET cells (Voorhoeve and Agami, 2003), resulting in a sensitized system for slight alterations in p53 in response to RASV12.
- 4. Figure 4A shows that following RASV12 stimulation, p53 was stabilized and activated, and its target gene, p21cip1, was induced in all cases, indicating an intact p53 pathway in these cells.

Reg-clause	Fact	Goal	Method	Result	Implication
Hypothes	is (not sh	nown)	Problem (n	ot showr	1)

Potential Applications

- Text Summarization
- Hypothesis Formulation
- Citation Analysis
- Identifying Methods and Protocols
 - For extraction to Lab Notebooks
- Knowledge Extraction
 - Is extracted data a hypothesis, claim, or fact?
- Figure and Table Interpretation
 - Which figures represent your "results"

Networks of Claims and Evidence

Claim:

 sustained miR-31 activity is necessary to prevent the acquisition of aggressive traits by both tumor cells and untransformed breast epithelial

Evidence: Method:

 We transiently inhibited miR-31 in noninvasive MCF7-Ras cells with either antisense oligonucleotides or miRNA sponges.

Evidence: Result:

- Both approaches inhibited miR-31 function by >4.5-fold (Figure S7A).
- Suppression of miR-31 enhanced invasion by 20-fold and motility by 5-fold, but cell viability was unaffected by either inhibitor (Figure 3A; Figure S7B).
- The miR-31 sponge reduced miR-31 function by 2.5-fold, but did not affect the activity of other known antimetastatic miRNAs (Figures S8A and S8B).

```
Is it pertinent? -> Probably
Is it true? -> Sounds likely!
Is it new, but in agreement with what I know? -> Check/know
```

ELSEVIER | 10

Dataset of "Discourse Segment Type vs. Linguistic Features"

	Segments			Part_In_Doc					Verb_Form						GENERAL INFORMATION	N
	Seg_Text	Seg_Type	Seg_Type	Name	Name	Line	Section	Section	Verb	Verb Form	Verb Form	ModalPassive	ModalPassive		File extension xls File size 1 MB Uploaded 22-02-2017 License CC BY 4.0 URL https://data.mendeley.com/datasets/4bh33fdx4v/3 /files/246cc47e-18ef-4f3c-83a8-d15cd96f46b2/17 0223b_deWaard_DST_With_Text.xls?dl=1 REFERENCE THIS FILE	
	http://nar.oxfordjournals.org /content/34/6/1807.full	Header	10	В	1	2	н	1			0		0			
	PMCID: PMC1421503	Header	10	В	1	3	н	1			0		0			
	Btk expression is controlled by Oct and BOB.1/OBF.1	Header	10	В	1	4	н	1			0		0			
	Cornelia Brunner and Thomas Wirth	Header	10	В	1	5	н	1			0		0			
		Blank	0	В	1	6	В	0			0		0			
	Abstract	Header	10	В	1	7	Α	2			0		0	>	de Waard, Anita (2017),	
	BOB.1/OBF.1 is a lymphocyte-restricted transcriptional coactivator.	Fact	1	В	1	8	А	2	be a coactivator	Present	2		0		"Discourse Segmen Linguistic Features" Data, v3	
	It binds together with the Oct1 and Oct2 transcription factors to DNA	Fact	1	В	1	9	A	2	bind	Present	2		0		http://dx.doi.org/10.17632 /4bh33fdx4v.3#file-246cc47e-18ef-4f3c- 83a8-d15cd96f46b2	
	and enhances their transactivation potential.	Fact	1	В	1	10	Α	2	enhance	Present	2		0		Go to dataset	Download
	Mice deficient for the transcriptional coactivator BOB.1/OBF.1 show several defects in differentiation, function and signaling of B	Other- Result	17	В	1	11	A	2	show	Present	2		0		Visualise in Plotly	

https://data.mendeley.com/datasets/4bh33fdx4v/3

Features / Feature Groups

Frequently Used Verb										
Top 10 Verb					"Show" verb					
Verb Tense										
Future	C	Gerund	ı	Past		Past pa	articiple		Past perfect	
Past progressive	F	Present		Present perfect		Present progressive			To-infinitive	
Verb Class	Verb Class									
Cause and eff	ect	Change and growth	d Discourse verb		urse	Interpretation		Investigation		
None		Observation		Prediction I		Proce	Procedure Pro		operties	
Modality Mar	ker									
Modal		b class erpretation		Ruled by verb c interpretation		lass Referer internal			Reference external	
First person		lodal Po gnificant_ly		Possible_ility_ly		Potentia y	al_l	UN_Likely		
Sum_Adverb s_YesNO										

Summary Statistics

ELSEVIER | 13

Class Distribution

```
In [6]: df.Seg Type name.value counts()
Out[6]: Result
                               625
         Implication
                               335
        Method
                               321
         Blank
                               316
        Reg-Implication
                               205
                               202
         Fact
                               197
         Hypothesis
         Other-Result
                               167
         Goal
                               138
         Header
                               134
        Other-Implication
                               120
         Problem
                                97
        Reg-Hypothesis
                                93
         Intratextual
                                72
                                61
         Reg-Result
         Other-Fact
                                53
         Other-Method
                                31
        Other-Hypothesis
                                26
         Intertextual
                                14
        Reg-Problem
                                10
         Reg-Fact
                                 6
         Other-Goal
                                 5
         Reg-Goal
                                 3
         Other-Problem
```

Collapsed and Filtered

- Subcategories normalized to parent
- Blank and Header rows removed
- Intertextual and Intratextual removed

```
df1['Seg Type name'].value co
In [24]:
Out[24]:
         Result
                           853
          Implication
                           660
          Method
                           352
                           316
          Hypothesis
                           262
          Fact
          Goal
                           149
          Problem
                           110
                            72
          Intratextual
          Intertextual
                            14
          Name: Seg Type name, dtype: i
```

Name: Seg_Type_name, dtype: int64

Verb Class & Modality Distributions

```
df. Verb Class. value counts()
Out[7]: Cause and Effect
                              716
                              536
        Interpretation
                              447
                              330
        Properties
        Procedure
                              315
                              217
        None
        Investigation
                              201
        Change and Growth
                              153
        Observation
                              148
        Discourse verb
                              105
        Prediction
                               70
        Discourse Verb
                                1
        Name: Verb Class, dtype: int64
        df.Modal.value counts()
In [8]:
Out[8]: 0
              3052
              186
        Name: Modal, dtype: int64
```

Machine Learning Models

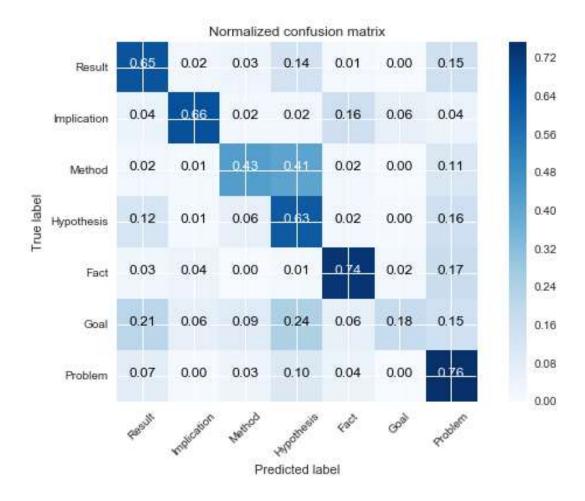
Methods Explored

- Baseline: Logistic Regression, Decision
 Tree, Random Forest
- Same classifiers using Class Balancers
- SciDT Neural Network for comparison
- Feature Reduction / Ablation
- Mini-experiment with binary targets and limited feature set

Baseline Random Forest Results

Accuracy score 0.6427688504326329

	precision	recall	f1-score	support
	0.50	0 67	0 50	0.0
Fact	0.53	0.67	0.59	88
Goal	0.79	0.60	0.68	50
Hypothesis	0.56	0.44	0.49	93
Implication	0.61	0.60	0.61	210
Method	0.78	0.75	0.77	106
Problem	0.60	0.18	0.28	33
Result	0.67	0.77	0.72	229
avg / total	0.64	0.64	0.64	809



ELSEVIER | 19

RandomUnderSampler	Undersamples the majority classes by randomly picking samples
Tomeklinks	Undersamples the majority classes by removing Tomek's links
ClusterCentroids	Under samples the majority classes by replacing a cluster of the majority samples by the cluster centroid of a KMeans algorithm
CondensedNearestNeighbor	Under samples the majority classes using the condensed nearest neighbor method
OneSidedSelection	Uses one-sided selection method on majority classes
InstanceHardnessThreshold	Samples with lower probabilities are removed from the majority class
RandomOverSampler	Randomly generates new samples from the minority classes
SMOTE	Synthetic Minority Oversampling Technique; generates new samples of minority class by interpolation
SMOTEborderline	Generates new samples of minority class specific to the borders between two classes.
SMOTEborderline2	Generates new samples of minority class specific to the borders between two classes.
SMOTETomek	Combines use of SMOTE on minority class and Tomek Links on majority class
SMOTEENN	Combines use of SMOTE on minority class and Edited

Class balancers had minimal to no effect

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

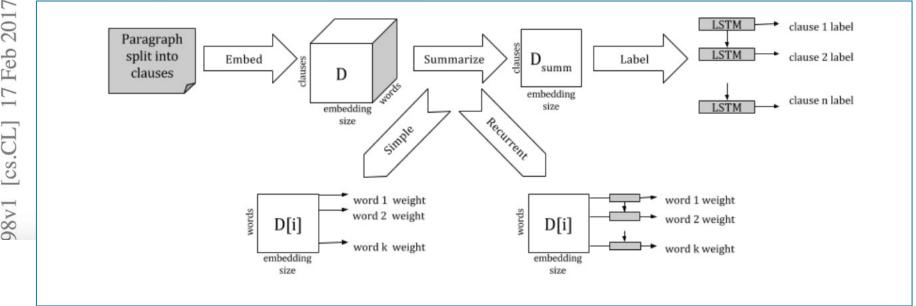
98v1 [cs.CL] 17 Feb 2017

Scientific Discourse Tagger

Experiment Segmentation in Scientific Discourse as Clause-level Structured Prediction using Recurrent Neural Networks

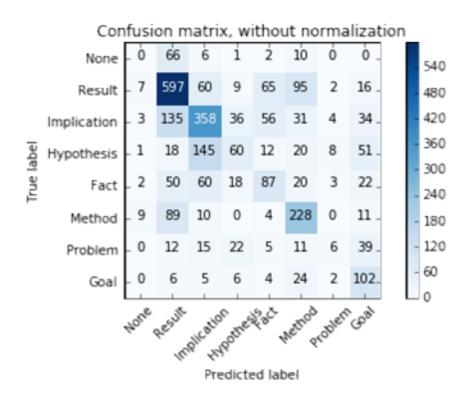
Pradeep Dasigi¹, Gully A.P.C. Burns², Eduard Hovy¹, and Anita de Waard³

¹Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA ²Information Sciences Institute, Viterbi School of Engineering, University of Southern California, Marina del Rey, CA 90292, USA ³Elsevier Research Data Services, Jericho, VT 05465, USA



https://github.com/edvisees/sciDT https://arxiv.org/pdf/1702.05398.pdf

	precision	recall	f1-score	support
0	0.00	0.00	0.00	85
1	0.61	0.70	0.65	851
2	0.54	0.54	0.54	657
3	0.39	0.19	0.26	315
4	0.37	0.33	0.35	262
5	0.52	0.65	0.58	351
6	0.24	0.05	0.09	110
7	0.37	0.68	0.48	149
avg / total	0.49	0.52	0.49	2780



ELSEVIER | 23

Feature Reduction and Forward Ablation

• Reduce to 13 Features: F1 increases .65 before dropping

Accuracy score 0.6477132262051916

	precision	recall	f1-score	support
Fact	0.43	0.81	0.56	88
Goal	0.69	0.72	0.71	50
Hypothesis	0.64	0.47	0.54	93
Implication	0.65	0.65	0.65	210
Method	0.86	0.76	0.81	106
Problem	0.27	0.24	0.25	33
Result	0.78	0.65	0.71	229
avg / total	0.68	0.65	0.65	809

Feature Reduction and Forward Ablation

• Forward Ablation to 9 Features before F1 levels off.

Accuracy score 0.6489493201483313

F Score 0.6445074039326106

	precision	recall	f1-score	support
Fact	0.45	0.75	0.56	88
Goal	0.84	0.64	0.73	50
Hypothesis	0.67	0.38	0.48	93
Implication	0.61	0.69	0.64	210
Method	0.79	0.82	0.81	106
Problem	0.40	0.18	0.25	33
Result	0.74	0.68	0.71	229
avg / total	0.67	0.65	0.64	809

ELSEVIER | 25

Minimal Change to Confusion Matrix

0.72

0.64

0.56

0.48

0.40

0.32

0.24

0.16

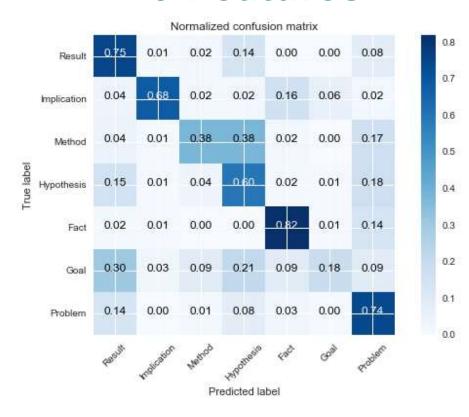
0.08

0.00

All Features

Normalized confusion matrix 0.02 0.03 0.00 0.15 0.14 0.01 Result 0.04 0.02 0.02 0.16 0.06 0.04 Implication 0.02 0.01 0.02 0.00 0.11 Method 0.01 0.06 0.63 0.02 0.00 0.12 0.16 Hypothesis 0.74 0.03 0.04 0.00 0.01 0.02 0.17 Fact 0.06 0.09 0.24 0.06 0.18 0.21 0.00 0.03 0.04 0.00 0.76 0.07 0.10 Problem Predicted label

9 Features



Significant Features

Feature Class	Feature				
Verb Tense	Past				
Verb Tense	Present				
Verb Tense	To-infinitive				
Verb Class	Interpretation				
Verb Class	Investigation				
Verb Class	Procedure				
Modality Marker	Modal				
Modality Marker	Verb class				
	interpretation				
Modality Marker	Ruled by verb class				
	interpretation				

Verb Tense Mini-Experiment

Binary Verb Tense Features

Future

Gerund

Past

Past participle

Past perfect

Past progressive

Present

Present perfect

Present progressive

To-infinitive

Binary Class Labels

| 27

Result/Method

Fact/Implication

ELSEVIER | 28

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 10. Random forest classifier model confusion matrix

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

Conclusions And Future Work

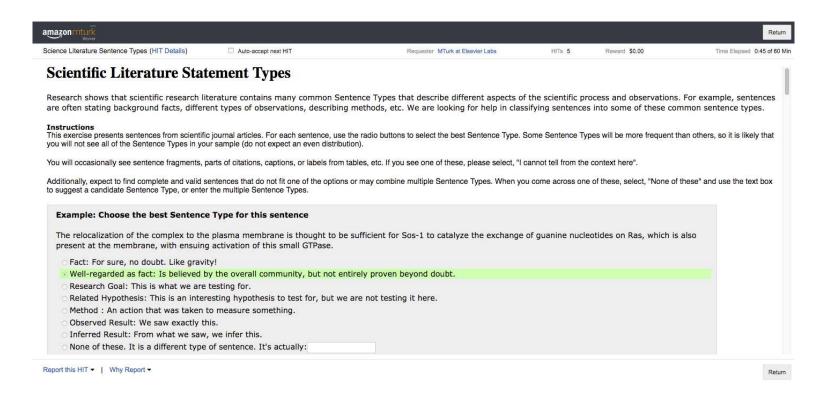
Conclusions

 Verb tense, class, and modality are reasonably good predictors of discourse segment type

- Class prediction success doesn't line up with majority / minority classes
- Verb tense alone is very good at distinguishing Result/Method from Fact/Implication
- There are some limits to our methodology

Future Work

- Build larger training sets of segment types
 - Mechanical Turk experiment underway



Future Work

- Evaluate automated extraction of features
- Looking at Snorkel as a a tool for merging multiple noisy classifiers
 - Majority voting: F1 of .64
 - Generative model: F1 of .58
 - Based on Sujit Pal's post:
 http://sujitpal.blogspot.co.uk/2018/01/cleaning-up-noisy-labels-using-snorkels.html

Corey A Harper

@chrpr

c.harper@elsevier.com



We're hiring!