The Use of Granularity in Rhetorical Relation Prediction

Blake Stephen Howald and Martha Abramson
Ultralingua, Inc.
1313 SE Fifth Street, Suite 108
Minneapolis, MN 55414
{howald, abramson}@ultralingua.com

Abstract

We present the results of several machine learning tasks designed to predict rhetorical relations that hold between clauses in discourse. We demonstrate that organizing rhetorical relations into different granularity categories (based on relative degree of detail) increases average prediction accuracy from 58% to 70%. Accuracy further increases to 80% with the inclusion of clause types. These results, which are competitive with existing systems, hold across several modes of written discourse and suggest that features of information structure are an important consideration in the machine learnability of discourse.

1 Introduction

The rhetorical relations that hold between clauses in discourse index temporal and event information and contribute to a discourse's pragmatic coherence (Hobbs, 1985). For example, in (1) the NARRATION relation holds between (1a) and (1b) as (1b) temporally follows (1a) at event time.

(1) a. Pascale closed the toy chest.
b. She walked to the gate.
c. The gate was locked securely.
d. So she couldn't get into the kitchen.

The ELABORATION relation, describing the surrounding state of affairs, holds between (1b) and (1c). (1c) is temporally inclusive (subordinated) with (1b) and there is no temporal progression at event time. The RESULT relation holds between (1b-c) and (1d). (1d) follows (1b) and its subordinated ELABORATION relation (1c) at event time.

Additional pragmatic information is encoded in these relations in terms of granularity. Granularity refers to the relative increases or decreases in the level of described detail. For example, moving from (1b) to (1c), we learn more information about *the gate* via the ELABORATION relation. Also, moving from (1b-c) to (1d) there is a consolidation of information associated with the RESULT relation.

Through several supervised machine learning tasks, we investigate the degree to which granularity (as well as additional elements of discourse structure (e.g. tense, aspect, event)) serves as a viable organization and predictor of rhetorical relations in a range of written discourses. This paper is organized as follows. Section 2 reviews prior research on rhetorical relations, discourse structure, granularity and prediction. Section 3 discusses the analyzed data, the selection and annotation of features, and the construction of several machine learning tasks. Section 4 provides the results which are then discussed in Section 5.

2 Background

Rhetorical relation prediction has received considerable attention and has been shown to be useful for text summarization (Marcu, 1998). Prediction tasks rely on a number of features (discourse connectives, part of speech, etc.) (Marcu and Echihabi, 2002; Lapata and Lascarides, 2004). A wide range of accuracies are also reported - 33.96% (Marcu and Echihabi, 2002) to 70.70% (Lapata and Lascarides, 2004) for all rhetorical relations and, for individual relations, CONTRAST (43.64%) and CONTINUATION (83.35%) (Sporleder and Lascarides, 2005).

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 44–48,
Montréal, Canada, June 7-8, 2012.

We seek to predict the inventory of rhetorical relations defined in Segmented Discourse Representation Theory ("SDRT") (Asher and Lascarides, 2003). In addition to the relations illustrated in (1), we consider: BACKGROUND: *It was Christmas. Pascale got a new toy.*; EXPLANATION: *The aardvark was dirty. It fell into a puddle.*; CONSEQUENCE: *If the aardvark fell in the puddle, then it got dirty.*; ALTERNATION: *Pascale got an aardvark or a stuffed bunny.*; and CONTINUATION: *Pascale got an aardvark. Grimsby got a rawhide.*

Discourses were selected based on Smith (2003) who defines five primary discourse modes by: (1) the *situations* (events and states) they describe; (2) the overarching *temporality* (tense, aspect); and (3) the type of text *progression* (temporal - text and event time progression are similar; atemporal - text and event time progression are not similar). These contrastive elements inform the features selected for the machine learning tasks discussed in Section 3.2. The five modes, *narratives, reports* (news articles), *description* (recipes), *information* (scientific essays), and *argument* (editorials) were selected to ensure a balanced range of theoretically supported discourse types.

2.1 Granularity of Information

Granularity in discourse refers to the relative degree of detail. The higher the level of detail, the more informative the discourse is. We assume that there will be some pragmatic constraints on the informativeness of a discourse (e.g., consistent with Grice's (1975) Maxim of Quantity). For our purposes, we rely specifically on granularity as defined in Mulkar-Mehta et al. (2011) ("MM") who characterize granularity in terms of entities and events.

To illustrate, consider (2) where the rhetorical structure indicates that (2b) is an ELABORATION of (2a), the NARRATION relation holds between (2b) and (2c) and (2c) and (2d), and the RESULT relation between (2d) and (2e).

(2) a. The Pittsburgh Steelers needed to win.
b. Batch took the first snap.
c. Then he threw the ball into the endzone.
d. Ward caught the ball.
e. A touchdown was scored.

Entities and events can stand in *part-whole* and *causality* relationships with entities and events in subsequent clauses. A *positive* granularity shift indicates movement from whole to part (more detail) - e.g., Batch (2b) is a *part* of the *whole* Pittsburgh Steelers (2a). A *negative* granularity shift indicates movement from part to whole (less detail), or if one event *causes* a subsequent event (if an event is caused by a subsequent event, this is a *positive* shift) - e.g., Ward's catching of the ball (2d) *caused* the scoring of the touchdown (2e). *Maintained* granularities (not considered by MM) are illustrated in (2b-c) and (2c-d). Clauses (2b) through (2d) are temporally linked events, but there is no *part-whole* shift in, nor a *causal* relationship between, the entities or events; the granularity remains the same.

We maintain that there is a close relationship between rhetorical relations and granularity. Consequently, rhetorical relations can be organized as follows: *positive*: BACKGROUND, ELABORATION, EXPLANATION; *negative*: CONSEQUENCE, RESULT; and *maintained*: ALTERNATION, CONTINUATION, NARRATION. The machine learning tasks discussed in the remainder of the paper consider this information in the prediction of rhetorical relations.

3 Data and Methods

Five written discourses of similar sentence length were selected from each mode for 25 total discourses. The discourses were segmented by independent or dependent (subordinate) clauses, if the clauses contained discourse markers (*but, however*), and if the clauses were embedded in the sentence provided in the orginal written discourse (e.g., John, *who is the director of NASA*, gave a speech on Friday). The total number of clauses is 1090, averaging 43.6 clauses per discourse (σ=7.2).

3.1 Feature Annotation

For prediction, we use a feature set distilled from Smith's classification of discourses: TENSE and ASPECT; EVENT (from the TimeML annotation scheme (Pustejovksy, et al., 2005), *Aspectual, Occurence, States*, etc.); SEQUENCE information as the clause position normalized to the unit interval; and discourse MODE. We also include CLAUSE type - independent (*IC*) or dependent clauses (*DC*) with the inclusion of a discourse marker (*M*) or not,

Table 1: Distribution of Relations by Granularity Type.

Relation	Number (Avg.)
Positive	**515 (47%)**
BACKGROUND	315 (61%)
ELABORATION	161 (31%)
EXPLANATION	39 (7%)
Negative	**59 (5%)**
CONSEQUENCE	16 (26%)
RESULT	43 (71%)
Maintenance	**490 (44%)**
ALTERNATION	76 (14%)
CONTINUATION	30 (6%)
NARRATION	384 (78%)

embedded (*EM*) or not - and GRANULARITY shift categories which are an organization of the SDRT rhetorical relations (Asher and Lascarides, 2003), summarized in Table 1.

All 25 discourses were annotated by one of the authors using only a reference sheet. The other author independently coded 80% of the data (20 discourses, four from each mode). Average agreement and Cohen's Kappa (Cohen, 1960) statistics were computed and are within acceptable ranges: TENSE (99.65 / .9945), ASPECT (99.30 / .9937), SDRT (77.42 / .6850), and EVENT (75.88 / .6362).

These results are consistent with previously reported annotations for rhetorical relations (Sporleder and Lascarides, 2005; Howald and Katz, 2011), event verbs and durations, tense and aspect (Puscasu and Mititelu, 2008; Wiebe et al., 1997). *Positive, negative* and *maintained* granularities were not annotated, but MM report a Kappa between .8500 and 1. The distribution of these granularities, based on the organization of the annotated rhetorical relations is presented in Table 1.

3.2 Machine Learning

Three supervised machine learning tasks were constructed to predict SDRT relations. The first task (**Uncollapsed**) created a 8-way classifier to predict the SDRT relations based on the feature set, omitting the GRANULARITY feature. The second task (**Collapsed**) created a 3-way classifier to predict the GRANULARITY categories (the SDRT feature was omitted). The third task (**Combined**) included the GRANULARITY feature back into the **Uncollapsed** 8-way classifier. We utilized the WEKA toolkit (Witten and Frank, 2005) and treated each clause as a vector of information (SDRT, EVENT, TENSE, ASPECT, SEQUENCE, CLAUSE, MODE, GRANULARITY), illustrated in (3)[1]:

(3) a. The Pittsburgh Steelers needed to win.
START, *State, Pa., N, .200, IC, NA, start*
b. Batch took the first snap.
ELAB., *Occ., Pa., N, .400, IC, NA, pos.*
c. Then he threw the ball into the endzone.
NAR., *Asp., Pa., N, .600, IC-M, NA, main.*
d. Ward caught the ball.
NAR., *Occ., Pa., N, .800, IC, NA, main.*
e. A touchdown was scored.
RESULT, *Occ., Pa., Perf., 1.00, IC, NA, neg.*

Table 2: Relation Prediction - Combined Modes.

Feature	J48	K*	NB	MCB
Uncollapsed	**58.99**	55.41	56.69	35
Collapsed	69.90	**70.18**	69.81	41
Combined	78.62	71.92	**80.00**	35 (70)

We report results from the Naïve Bayes (NB), J48 (C4.5 decision tree (Quinlan, 1993)) and K* (Cleary and Trigg, 1995) classifiers, run at 10-fold cross-validation.

4 Results

Table 2 indicates that the best average accuracy for the **Uncollapsed** task is 58.99 (J48). The accuracy increases to 70.18 (K*) for the **Collapsed** task. The accuracy increases further to 80.00 (NB) for the **Combined** task. All accuracies are statistically significant over majority class baselines ("MCB"): **Uncollapsed** (MCB = 35) - χ^2 = 15.11, d.f. = 0, $p \leq$ .001; **Collapsed** (MCB = 41) - χ^2 = 20.51, d.f. = 0, $p \leq .001$; and **Combined** (treating the best **Collapsed** accuracy as the new baseline (MCB = 70)) - χ^2 = 1.43, d.f. = 0, $p \leq .001$.

As shown in Table 3, based on the NB 8-way **Combined** classifier, the prediction accuracies of

[1]Note that what is being predicted is the rhetorical relation, or associated granularity, with the second clause in a clause pair. Tasks were performed where clause information was paired, but this did not translate into improved accuracies.

Table 3: Individual Relation Prediction Accuracies (%).

Relation	A	I	D	N	R	T
NAR.	73	55	100	100	94	**96**
RES.	75	88	85	100	100	**93**
BACK.	93	92	96	87	94	**92**
ELAB.	57	41	69	21	48	**69**
CONSEQ.	20	0	0	0	0	**37**
ALTER.	50	42	0	0	43	**27**
CONTIN.	8	0	0	0	0	**23**
EXPLAN.	0	20	0	9	0	**2**
Total	**68**	**72**	**92**	**74**	**74**	**80**

the individual modes are no more than 12 percentage points off of the average (80.00). Accuracies range from 68% **A(rgument)** (σ=-12) to 92% **D(escription)** (σ=+12) with **N(arrative), R(eport)**, and **I(nformation)** being closest to average (σ=-6-8). For individual relation predictions, NARRATION, RESULT and BACKGROUND have the highest total accuracies followed by ELABORATION and CONTRAST. Performing less well is CONSEQUENCE, ALTERNATION and CONTINUATION with EXPLANATION performing the worst. All accuracies are statistically significant above baseline (χ^2 = 341.89, d.f. = 7, $p \leq .001$).

5 Discussion and Conclusion

Using the **Collapsed** performance as a baseline for the **Combined** classifier, we discuss the features contributing to the 10 percentage point increase as well as the optimal (minimal) set of features for prediction. The best accuracies for the **Combined** experiment only require CLAUSE and GRANULARITY information; achieving 79.08% (NB - 44 above MCB, *f*-score=.750). Both CLAUSE and GRANULARITY are necessary. Relying only on CLAUSE achieves a 48.25% accuracy (J48) and relying only on GRANULARITY achieves 70.36% for all classifiers, but this higher accuracy is an artifact of the organization as evidenced by the *f*-score (.585).

The relationship between CLAUSE and the rhetorical relations is straightforward. For example, the CONSEQUENCE relation is often an "intersentential" relation (*if the aardvark fell in the puddle, then it got dirty*), each of the 16 CONSEQUENCE relations are embedded. Similarly, 93% of all ELABORATION relations, which are temporally subordinating, are embedded. Clause types appear to be a viable source of co-varying information in rhetorical relation prediction in the tasks under discussion.

The aspects of syntactic-semantic form and pragmatic function in the relationship between granularity and rhetorical relations is of central interest in this investigation. Asher and Lascarides represent discourses hierarchically through coordination and subordination of information which corresponds to changes in granularity. However, while the notion of granularity enters into the motivation and formulation of the SDRT inventory, it is not developed further. These results potentailly allow us to say something deeper about the structural organization of discourse as it relates to granularity.

In particualr, while there is some probabilistic leverage in collapsing categories, it is not the case that arbitrary categorizations will perform similarly. This observation holds true even for theoretically informed categorizations. For example, organizing the SDRT inventory into *coordinated* and *subordinated* relations yields lower performance on relation prediction. *Coordinated* and *subordinated* can be predicted with 80% accuracy, but the prediction of the individual relations given the category performs only at 70%. Since the granularity-based organization presented here performs better, we suggest that the pragmatic *function* of the relation is more systematic than the syntactic-semantic *form* of the relation.

Future research will focus on more data, different machine learning techniques (e.g. unsupervised learning) and automatization. Where clause, tense, aspect and event are readily automatable, rhetorical relations and granularity are less so. Automatically extracting such information from an annotated corpus such as the Penn Discourse Tree Bank is certainly feasible. However, the distribution of genres in this corpus is somewhat limited (i.e., predominately news text (Webber, 2009)) and calls into question the generalizeability of results to other modes of discourse. Overall, we have demonstrated that the inclusion of a granularity-based organization in the machine learning prediction of rhetorical relations increases performance by 37%, which is roughly 14% above previous reported results for a broader range of discourses and relations.

Acknowledgments

Thank you to Jeff Ondich and Ultralingua for facilitating this research and to four anonymous *SEM reviewers for insightful and constructive comments.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge, UK.

John G. Cleary and Leonard E. Trigg 1995. K*: An Instance-based Learner Using an Entropic Distance Measure. In *Proceedings of the 12 International Conference on Machine Learning*, 108–113.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

H. Paul Grice. 1975. *Logic and Conversation*. In *Syntax and Semantics, Vol. 3, Speech Acts*, 43–85. Academic Press, New York.

Jerry R. Hobbs. 1985. On The Coherence and Structure of Discourse. *CSLI Technical Report*, CSLI-85-37.

Blake Stephen Howald and Graham Katz. 2011. The Exploitation of Spatial Information in Narrative Discourse. In *Proceedings of the Ninth International Workshop on Computational Semantics*, 175–184.

Mirella Lapata and Alex Lascarides. 2004. Inferring Sentence Internal Temporal Relations. In *Proceedings of the North American Association of Computational Linguistics (NAACL-04) 2004*, 153–160.

Daniel Marcu. 1998. Improving Summarization Through Rhetorical Parsing Tuning. In *Proceedings of The 6th Workshop on Very Large Corpora*, 206–215.

Daniel Marcu and Abdessamad Echihabi. 2002. An Unsupervised Approach to Recognizing Discourse Relations. In *Proceedings of the Association of Computational Linguistics (ACL-02) 2002*, 368–375.

Rutu Mulkar-Mehta, Jerry R. Hobbs and Eduard Hovy. 2011. Granulairty in Natural Language Discourse. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011) 2011*, 195–204.

Georgiana Puscasu and Verginica Mititelu. 2008. Annotation of WordNet Verbs with TimeML Event Classes. *Proceedings of the Sixth International Language Resources and Evaluation (LREC08)*

James Pustejovsky, José Castaño, Robert Ingria, Roser Saur, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2005. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Proceedings of the Fith International Conference on Computational Semantics (IWCS 2005)*

Ross Quinlan. 1993 *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco, CA.

Carlota Smith. 2003. *Modes of Discourse: The Local Structure of Texts*. Cambridge University Press, Cambridge, UK.

Caroline Sporleder and Alex Lascarides. 2005. Exploiting Linguistic Cues to Classify Rhetorical Relations. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-05)*, 532–539.

Caroline Sporleder and Alex Lascarides. 2008. Using Automatically Labelled Examples to Classify Rhetorical Relations: An Assessment. *Natural Language Engineering*, 14:369–416.

Janyce Wiebe, Thomas O'Hara, Thorsten Öhrström-Sandgren and Kenneth McKeever. 1997. An Empirical Approach to Temporal Reference Resolution. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-97)*, 174–186.

Ian Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Techniques with Java Implementation (2nd Ed.)* Morgan Kaufmann, San Francisco, CA.

Bonnie Webber 2009. Genre Distictions for Discourse in the Penn TreeBank. In *Proceedings of the 47th ACL Conference*, 674–682.