A Distributional View of Discourse Encapsulation: Multifactorial Prediction of Coreference Density in RST

Amir Zeldes
Department of Linguistics, Georgetown University
amir.zeldes@georgetown.edu

Abstract

Early formulations of discourse coherence constraints postulated a connection between coreference likelihood and distance within a discourse parse, e.g. in the framework of Veins Theory (Cristea et al. 1998%CristeaIdeRomary1998), which proposes that coreference is expected to be encapsulated within tightly linked areas of discourse parses, called Domains of Referential Accessibility (DRAs). Using an RST dependency representation, this paper expands on previous work showing the relevance of DRAs to coreference likelihood. We develop a multifactorial model using both rhetorical and surface distance metrics, as well as confounds such as unit length and genre, and direct versus indirect rhetorical paths. We also explore coreferential accessibility as it applies to less studied types of coreference, including bridging and lexical coreference. The results show that rhetorical and surface distance, as well as direct linking, all influence coreference likelihood, and should not be treated as mutually exclusive or redundant metrics. Finally, we incorporate RST relation-specific tendencies that offer a more fine-grained model of coreference accessibility.

1 Introduction

Accessibility of discourse referents has been a major theme in discourse parsing frameworks since the beginning of the field. Polanyi (1988:616)%Polanyi1988 suggested that the stack of discourse units determined which discourse referents were available to be pronominalized; in Segmented Discourse Representation Theory (SDRT), the Right Frontier Constraint (Asher 1993%Asher1993) posited that newly attached discourse units could only link to the previous or immediately dominating segment, and later that anaphora was restricted to this domain (see Asher & Lascarides 2003%AsherLascarides2003); and in Rhetorical Structure Theory (RST, Mann & Thompson 1988%MannThompson1988), Veins Theory (Cristea et al. 1998%CristeaIdeRomary1998) was developed to identify Domains of Referential Accessibility (DRAs), said to constrain coreference relations. We can refer to the conjecture behind these approaches as the 'Discourse Encapsulation Hypothesis' (DEH), i.e. that discourse structure constrains domains of co-referentiality.

Empirical work examining different forms of the DEH has primarily focused on showing that some kind of discourse distance metric or domain definition is superior to surface distance as a predictor of coreferentiality, or to some other proposed metrics (e.g. Cristea et al. 1999%CristeaIdeMarcuEtAl1999, Tetreault & Allen 2003%TetreaultAllen2003, Chiarcos & Krasavina 2008%ChiarcosKrasavina2008). Surprisingly, there seems to be no work suggesting that rather than comparing DRA definitions to surface distance definitions, we could attempt to combine them, or even pool further predictors into a multifactorial model of coreferential accessibility – this will be the main goal of the present paper.

The idea that a multifactorial model may be more useful than categorical definitions of accessi-

Proceedings of the 6th Workshop Recent Advances in RST and Related Formalisms, pages 20–28,
Santiago de Compostela, Spain, September 4 2017. ©2017 Association for Computational Linguistics

ble domains gains credence from recent advances in the use machine learning for discourse annotation. While using cues from discourse parsing is still not standard in state of the art coreference resolution systems (Durret & Klein 2014%DurrettKlein2014, Clark & Manning 2015%ClarkManning2015, Wiseman et al. 2016%WisemanRushShieber2016), recent work in discourse parsing suggests that knowing about coreference can improve RST parsers (Surdeanu et al. 2015%SurdeanuEtAl2015, Braud et al. 2016%BraudPlankSoegaard2016), RST-based sentence compression (Durrett et al. 2016%DurrettBerg-KirkpatrickKlein2016), and discourse cohesion metrics (Iida & Tokunaga 2012%IidaTokunaga2012).

Different frameworks have applied some kind of DEH to different types of coreference: pronominal anaphora only (e.g. Tetreault & Allen 2003%TetreaultAllen2003, Chiarcos & Krasavina 2008%ChiarcosKrasavina2008), also lexical coreference (Cristea et al. 1999%CristeaIdeMarcuEtAl1999), or specific phenomena (e.g. discourse deictic and demonstrative *this/that*, Webber 1991%Webber1991). These approaches are in principle testable for any type of referentiality, and this paper will therefore compare coreference at large, pronominal anaphora, and bridging anaphora (Asher & Lascarides 1998%AsherLascarides1998).

Finally, previous approaches have explicitly disregarded the role of discourse function labels and utterance types in predicting coreferentiality domains, despite the relatively plausible proposition that certain relations or combinations of relations may influence coreference likelihood (e.g. we would expect coreference within an RST *Restatement*, but *Purpose* satellites may be less likely to co-refer to entities in their nuclei). In fact, many discourse connectives which signal specific relations have anaphoric components, e.g. causal connectives such as *therefore*, which imply event anaphora (see Stede & Grishina 2016%StedeGrishina2016).

In order to construct a multifactorial model of referent accessibility for coreference, anaphora and bridging, in Section 2 we present the data and scope of annotations used in this study. We then argue for the use of a dependency representation of RST, rather than traditional constituent trees for this task. Section 3 discusses the operationalization of discourse distance and the features used in our model, followed by the results in Section 4, and concluding with some discussion in Section 5.

2 Data

2.1 The GUM corpus

To model the DEH, we need data that is annotated for both RST and coreference, which narrows down the possible choices of corpus. The first natural choice for an RST corpus would normally have been the RST Discourse Treebank (Carlson et al. 2001%CarlsonEtAl2001), the largest available RST corpus, parts of which overlap with the coreference annotated portion of OntoNotes (Hovy et al. 2006%HovyMarcusPalmerEtAl2006). Although coreference annotations are available for 182 of the 380 Wall Street Journal documents in the RST Discourse Treebank (RSTDT), using OntoNotes coreference data to test the DEH is problematic, since OntoNotes rules out indefinite NPs as possible anaphors, as well as a variety of special situations, the most relevant of which are illustrated in (1)-(4) (all examples are from OntoNotes, but none are annotated as coreferent there).

(1) **Indefinite/generic:** [*Program trading*] *is "a racket,"*... [*program trading*] *creates ... swings*

(2) **Modifiers nouns:** *small investors seem to be adapting to greater* [*stock market*] *volatility ... Glenn Britta ... is "factoring"* [*the market's*] *volatility "into investment decisions."*

(3) **Metonymy:** *a strict interpretation ... requires* [*the U.S.*] *to notify foreign dictators of certain coup plots ...* [*Washington*] *rejected the bid ...*

(4) **Nesting:** *He has in tow* [*his prescient girlfriend, whose sassy retorts mark* [*her*] *...*]

Another phenomenon of interest that is not covered by OntoNotes data is bridging (see Asher & Lascarides 1998%AsherLascarides1998), shown in example (5), which will be evaluated separately in Section 4.

(5) *Mexico's President Salinas said* [*the country*]*'s recession had ended and* [*the economy*] *was growing again.*

In order to include these phenomena, we use the GUM corpus, containing 76 documents (64,000 tokens) in four genres of English from the Web (news, interviews, how-to guides and travel guides) annotated for RST, coreference, entities, syntax and a variety of other annotations (see Zeldes 2016%Zeldes2016).[1] The RST analyses in GUM use a fairly small, high-level inventory of 20 relations similar to the RSTDT's 16 top-level relation classes (see Section 4.3), while coreference relations cover 5 types: anaphora, cataphora (forward-pointing link), lexical coreference, apposition and bridging.

2.2 Rhetorical Structure Dependencies

In order to evaluate the DEH, we need to operationalize the notion of Rhetorical Distance (RD) in an RST graph. Here the argument will be presented that a 'flat' dependency-like structure offers a more intuitive way of calculating distances than fully hierarchical RST trees.

Because RST instantiates non-terminal nodes (spans and 'multinucs', i.e. multinuclear units), a direct comparison of surface distance and 'rhetorical distance' between elementary discourse units (EDUs) is non-trivial. An intuitive approach might be to count edges along the path between two EDUs, including transitions to non-terminal nodes (see Chiarcos & Krasavina 2008%ChiarcosKrasavina2008 for discussion). In this case, the RD between [1] and [3] in Figure 1 would be 3, which we write as *RD(u1,u3)=3*. However, there are both practical and theoretical problems with this way of counting.

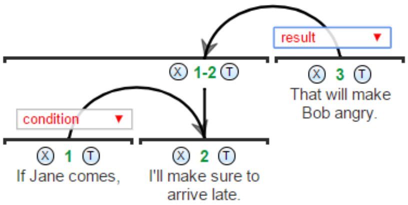

Figure 1. A simple RST example with a non-terminal span. *RD(u1,u3)=3* and *RD(u2,u3)=2*.

From a practical perspective, we note that *RD(u2,u3)=2*; this measurement is a direct result of the presence of the span [1-2], which is only needed due to the conditional in [1]. For the same two units with the same relation, *RD*=1 in Figure 2.

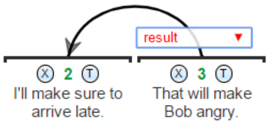

Figure 2. Without the conditional EDU, *RD(u2,u3)=1*.

This behavior is counter-intuitive, since for purposes of coreference likelihood, we would like to say that the rhetorical cohesion of the predicates in [2] and [3] is direct: Bob being angry in [3] is the result of arriving late in [2]. At least from a coreference-centric perspective, there is no reason to assume less tight juncture between referents in [2] and [3] due to having a further satellite to the left.

From a more theoretical standpoint, assuming equal distance regardless of the presence of peripheral modifiers is consistent with Marcu's (1996)%Marcu1996 compositionality criterion for discourse trees, which posits that 'spans can be joined in a larger span by a given rhetorical relation if and only if that relation holds also between the most salient units of those spans' (Marcu 1996:1070%Marcu1996; see also Zhang & Liu 2016%ZhangLiu2016 for an empirical study).

For these reasons, the present paper uses a conversion of the RST data from the GUM corpus into a dependency-style format, which contains no non-terminal nodes, linking only EDUs to each other such that relations emanating from spans are represented by edges linked to their nuclei. Several dependency representations have recently been suggested for RST, most notably by Hirao et al. (2013)%HiraoYoshidaNishinoEtAl2013 and Li et al. (2014)%LiWangCaoEtAl2014. A key difference between these representations is the handling of multinuclear relations (see Hayashi et al. 2016%HayashiHiraoNagata2016 for comparison and discussion). Figure 3, reproduced from Hayashi et al., illustrates the two approaches, which roughly correspond to propagating a multinuc's outgoing relation to its children, or using the multinuc relation name to connect its children. In this paper we follow Li et al.'s approach, which allows us to retain information about multinuclear relations (this will become important in Section 4).

[1] The corpus is available under a Creative Commons license at `https://corpling.uis.georgetown.edu/gum/`.

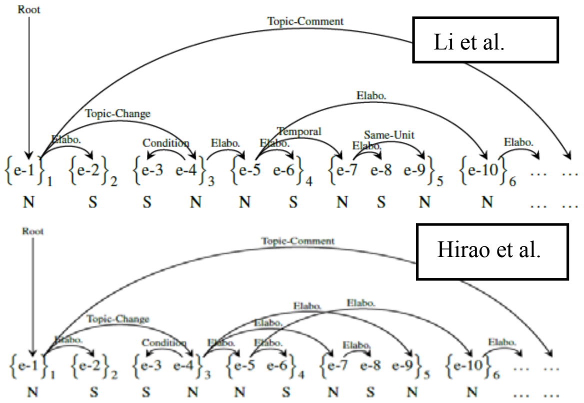

Figure 3. Rhetorical Structure Dependency representations (reproduced from Hayashi et al. 2016)

Using a dependency representation of the GUM data, calculating RD is simple, and hierarchy depth issues are avoided.[2] A limitation of this approach is that we no longer have access to the relative linking order of satellites: we could want to consider more closely nested satellites to be closer. For example, in the top representation in Figure 3, *RD(e-1,e-2)* = *RD(e-1,e-4)* = *1*. If the tree allows crossing edges, we no longer know whether e-2 or e-4 are more closely linked to e-1. Although Marcu's compositionality criterion suggests that this difference should be irrelevant, we reserve the possibility of RD metrics incorporating nesting depth in some way for future work; in any event, it seems reasonable that both *RD(e-1,e-2)* and *RD(e-1,e-4)* should be greater than *RD(e-1,e-3)*, and this assumption is respected by the suggested representation.

3 Setup

3.1 Operationalization

The dependent variable of interest in this study is the degree of coreferentiality between EDUs, but there are multiple ways of considering whether/to what extent coreference holds between any two units. One decision is whether coreferentiality constitutes binary (some coreference detected) or count data (how many coreferent entities, or entity mentions). Although more categorical formulations of the DEH may evoke interest in the binary option, a realistic corpus approach means expecting a range of different densities of coreferentiality at all distances, so that ignoring frequencies seems like an undesirable loss of information. We therefore choose to focus on count data modeling coreference density (but see Section 4.3 on binary prediction).

A second important distinction is whether we are interested in immediate antecedents or simply any members of a coreference chain. Clearly as distance grows, the immediate antecedent of an entity mention becomes unlikely across a pair of EDUs; however, distant EDUs may still discuss the same referents, which we will detect if we consider any distance in the coreference chain as an instance of the target phenomenon. As it is not clear which of these formulations is most interesting, we will tentatively examine both and compare the results in Section 4.

3.2 Features

Our dataset covers all possible pairs of EDUs within the same document in each of the documents in GUM. The corpus contains 4788 EDUs in 76 documents, which produce over 170K distinct EDU pairs. For each pair we collect:

- Name and genre of the document
- Surface distance in EDUs
- RD based on dependency representation
- Length in tokens
- Rough sentence type (10 types available in GUM, e.g. declarative, imperative, question, fragment..)
- Direct ancestry – a binary variable, whether one EDU is a direct ancestor of the other in the dependency tree
- Outgoing RST relation name
- Head POS and grammatical function
- Whether or not the EDU is a subordinate clause (values: attached left, right or none)
- Amount of coreferent mentions across the pair (excluding bridging; see below)
- Amount of direct antecedent relations across the pair (excluding bridging)
- The latter metric, but only for bridging

[2] Code generating the dependency representation from .rs3 files is available from `https://github.com/amir-zeldes/rst2dep`. The data itself is available from the GUM website.

Since bridging is not a transitive relation, we do not collect information about indirect chains containing a bridging link.

While collecting the count of direct antecedents is fairly straight-forward, computing indirect coreference is more complex. If, for example, an entity is mentioned twice in an EDU and once in a preceding EDU, we need to decide whether the coreference count is 1 or 2. Note that while each of the two mentions in the later EDU has an indirect antecedent in the earlier EDU, there is only one coreferent entity. However, collapsing the multiple mentions in an EDU means losing information – on some level, it makes intuitive sense that multiple subsequent mentions of the same entity should count as realizing an increase in cohesion. In the evaluation below, we therefore do not collapse multiple mentions and concentrate on *coreference density* as the metric for indirect coreferentiality. For direct antecedents and bridging, this issue does not arise: entity and mention density are the same.

4 Results

4.1 Coreference

Direct antecedent coreferentiality is a comparatively sparse phenomenon: in permuting all possible EDU pairs for the evaluation, very few mentions have their direct antecedent in any given pair, with the range in our data spanning only 0-6 coreferent mentions. At the same time, it is also highly correlated with EDU distance: direct antecedents are usually quite close to their present mention. Indirect coreference, by contrast, can be spread out throughout documents, and is much more frequent: while most EDUs share fewer than 4 mentions in common, outlier cases can have as many as 34 mentions in common (by repeating several identical mentions multiple times, usually only possible in longer EDUs). Figure 4 gives an overview of the relationship between EDU distance (bottom) or RD (top) and direct coreference (right) or indirect coreference (left).

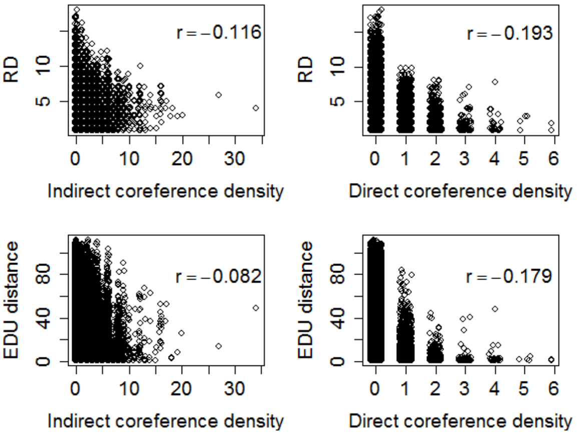

Figure 4. Direct and indirect coreference density as a function of EDU distance and RD.

As the correlation coefficients in the plots show, coreference is negatively correlated with distance in all cases; however for both direct and indirect density, RD is slightly more correlated than EDU surface distance. At the same time it should be noted that EDU distance and RD are significantly correlated (r = 0.243, p < 2.2e-16), and that high coreference density is in most cases connected to sentence length as well, since longer EDUs have a higher chance of matching multiple mentions. It is therefore difficult to evaluate the DEH without a multifactorial view of the data.

To address these confounds, we perform a linear mixed effects Poisson regression using the lme4 package in R, modelling the approximate shape of coreference density.[3] As fixed effects we initially consider EDU distance, RD, and EDU length of both units (z-score transformed). We also add two further predictors: the genre a document comes from and direct ancestry between the EDUs. Ancestry can be important, since RD does not capture an important distinction in measuring 'encapsulation': intuitively, a direct RST ancestor is more tightly connected to an RST child than units for which we must go 'up the tree and back down', even if the number of edges in both cases is identical. Genre is not strictly necessary, but it may be reasonable to assume that coreference likelihood and RD distance patterns vary systematically

[3] The Poisson distribution is a good fit for the left bounded distribution of coreference density bands: values under 0 are not possible, and the expected value is between 0 and 1 (I thank an anonymous reviewer for commenting on this).

across the genres represented in our data. Document identity is treated as a random effect introducing idiosyncratic noise into the data. Model coefficients are given below first for direct coreference.

```
Random effects:
 Groups Name        Variance Std.Dev.
 doc    (Intercept) 0.01434  0.1197
Number of obs: 172150, groups:  doc, 76

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.579493   0.056438  -10.27   <2e-16 ***
scale(len1)  0.221689   0.009478   23.39   <2e-16 ***
scale(len2)  0.193865   0.009436   20.55   <2e-16 ***
rsd_dist    -0.332126   0.012633  -26.29   <2e-16 ***
edu_dist    -0.139895   0.002778  -50.36   <2e-16 ***
genrenews   -0.056477   0.053785   -1.05   0.2937
genrevoyage -0.486155   0.056290   -8.64   <2e-16 ***
genrewhow   -0.096990   0.051096   -1.90   0.0577 .
directTrue   0.380319   0.035008   10.86   <2e-16 ***
---
p-val: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model shows that all of the relevant predictors are highly significant: even knowing both EDU lengths (which are clearly very important), as well as RD and EDU distance, and direct ancestry too, all predictors remain highly useful. Genre, by contrast, is less important, with only the travel guide genre (voyage) being associated with a lower coreferentiality baseline.

Looking at model coefficients, we see that length is likely to outweigh distance metrics in effect size as long as distance is moderate: an increase of one z-score in sentence length above the mean is associated with increases of about 0.2 coreferent mentions. Each EDU distance unit, by contrast, decreases coreferentiality by about 0.13 units compared to the intercept, which can however mount up. RD units have a stronger effect per unit (0.33), but a lower z value (-26 for RD, but -50 for EDU distance). This is understandable since for direct coreferentiality, distance can become overwhelming, and even units mentioning the same entities can score 0 due to the direct antecedent being elsewhere. Finally, being a direct ancestor (no going up and down the RST tree) offsets more than one unit of RD, suggesting that this relationship has a substantial effect. The overall model fit measured in r^2 for the correlation of fitted and actual values is 0.19, a respectable value considering we are predicting degree of coreferentiality without knowing anything about the contents of the EDUs; in other words, the model accounts for about a fifth of the variance in coreference density.

We can now compare the results above to what happens when we model indirect coreference, using the same predictors. In order for the model not to be skewed by comparatively rare outliers with over 15 coreferent mention pairs, the dependent variable in this case will be z-score scaled and fitted to a Gaussian distribution. Although the Gaussian model t-values cannot be translated into p-values directly due to inexact degrees of freedom (see Baayen 2008:269%Baayen2008), a conservative estimate treats values more extreme than ±2 as significant.

```
Random effects:
 Groups   Name        Variance Std.Dev.
 doc      (Intercept) 0.09789  0.3129
 Residual             0.82965  0.9109
Number of obs: 172150, groups:  doc, 76

Fixed effects:
              Estimate Std. Error t value
(Intercept)  0.2695836  0.0723038    3.73
scale(len1)  0.2043943  0.0023432   87.23
scale(len2)  0.1833124  0.0023811   76.99
rsd_dist    -0.0511588  0.0014351  -35.65
edu_dist    -0.0015377  0.0001168  -13.17
genrenews   -0.0348780  0.0997936   -0.35
genrevoyage -0.2161897  0.1047555   -2.06
genrewhow    0.0969725  0.1016942    0.95
directTrue   0.2280120  0.0091334   24.96
```

Again, genre is not a strong predictor, with 'voyage' somewhat below the intercept. Sentence lengths are now even more significant (largest t-values), and effect sizes per z-score unit are much larger than for the distance metrics. However the most interesting part of the result is the disparity between the very weak (but significant) effect of EDU distance, compared to a 50 times more influential contribution of RD. An RD shift of four units is as strong as a standard deviation in sentence length, but EDU shifts needs to be more than 10 times as large for the same effect. This suggests that a large part of the effect found for the direct model simply reflects the proximity of immediate antecedents. Finally, direct ancestry still plays a role, comparable to just over one standard deviation in EDU length. The total model r^2 is 0.17, a slightly worse fit, but unsurprising considering the reduced informativity of surface distance.

4.2 Bridging and pronominal anaphora

Following the results for coreference at large, we can also ask whether bridging and pronominal anaphora pattern in the same way. From a discourse cohesion point of view, bridging is a very similar phenomenon to coreference, since resolving bridging reference requires recourse to antecedents. Due to the non-transitive nature of the relation, the distribution is very sparse: Only 601 out of over 170,000 possible EDU pairs exhibit some bridging (one or more cases). This highly skewed distribution makes a regression on the complete dataset problematic: even if we cast the problem as binomial (bridging present or absent), the regression will inevitably learn to guess 'no bridging', a majority baseline which achieves over 99% accuracy. For bridging we therefore opt to concentrate on the distribution of those pairs that do exhibit some bridging. Figure 5 shows a log-log scatter plot of RD and EDU distance for bridging cases, distinguishing direct and indirect rhetorical dominance paths. Each circle represents an EDU pair, with circle size corresponding to the number of bridging instances for that pair.

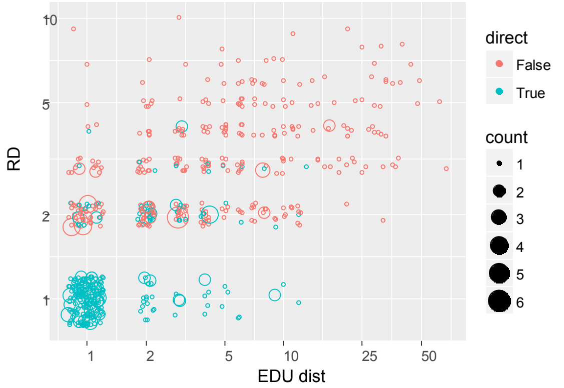

Figure 5. RD vs. EDU distance for pairs with bridging, also showing direct rhetorical ancestry (log-log scale).

The figure shows that most of the data has immediate proximity (*RD*=*ED*=1, 30.2% of pairs, covering 32.1% of bridging cases). However much like for coreference, bridging covers a wide range of EDU distances, and remains somewhat well attested at range: the mean EDU distance is 5.27 (comparable to direct coreference: 5.23), whereas RD, which only reaches 10, is strongly concentrated in the region below 4 or 5, with a mean of *RD*=2.45 (a small, but significant difference to direct coreference: 2.62).

Long-distance direct ancestry is unsurprisingly rare, especially for high RD, and cases are concentrated at the bottom of the plot. However the preponderance of direct ancestry in bridging cases is particularly high: 45.7% of EDU pairs exhibiting bridging are in a rhetorical ancestry relation, covering 48% of bridging instances. By contrast, 43.2% of direct coreference EDU pairs (and 45.7% of coreference instances) have direct ancestry, and a much lower 14.3% and 15.6% respectively for indirect coreference. In sum, it seems that while bridging is too rare to build a complete multifactorial model, it has similar distance and direct ancestry effects to regular coreference.

For pronominal anaphora, data is less sparse, but a negative baseline (always say 0) for testing whether any pair of EDUs has a direct anaphoric link still scores over 98% accuracy. We therefore again focus on the distribution of cases exhibiting some anaphoric links in Figure 6.

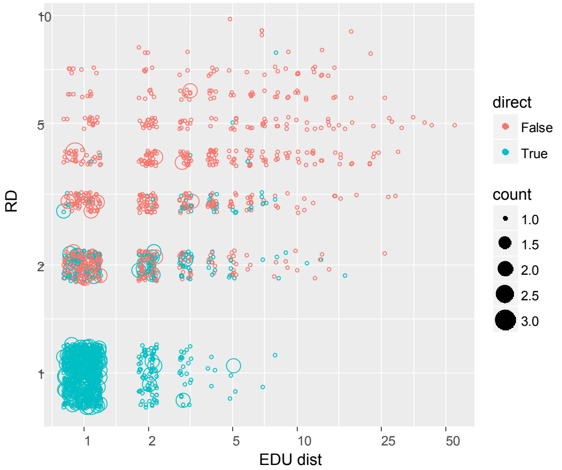

Figure 6. RD versus EDU distance for anaphora.

The picture is similar to bridging, but more dense, with 40.5% of pairs/42.5% of cases having *ED*=*RD*=1. Somewhat higher RD values are seen even at close EDU proximity, suggesting surface proximity is more influential for anaphora, and close RD is more critical to bridging.

4.3 Predicting coreference density

So far we have only considered unlabeled RST distance, without looking at specific RST relations or properties of the underlying units other than length. Although the DEH does not presuppose any expectations for these factors directly, it is interesting to consider which RST relations and what kinds of EDUs play into the DEH, and which are less in line with the hypothesis. To test this, we first examine which RST relations are more likely to exhibit coreference between head and dependent.

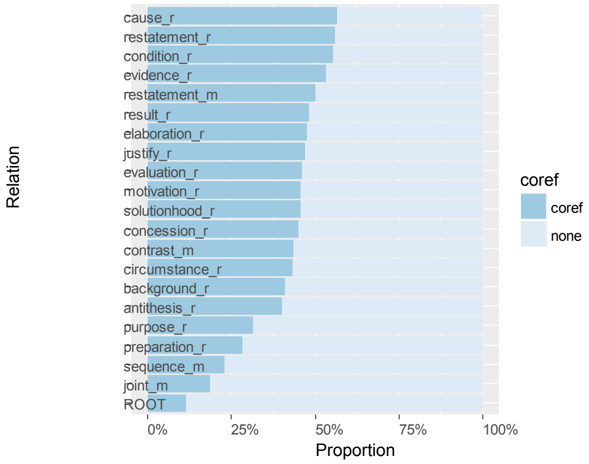

Figure 7. Proportion of EDUs showing coreference with their dependency heads by relation.

Figure 7 shows a rather broad variation in proportion of coreferentiality by relation, especially in the bottom 5 relation types (from *Purpose* down). *Cause* and *Restatement* are unsurprisingly at the top, while typically coordinating multinuclear relations such as *Sequence* and *Joint* are at the bottom. These results suggest that relation type may be a relevant predictor modulating domain or path effects on coreference likelihood.[4]

Given everything we've seen above, it seems likely that we can create a multifactorial model to predict how likely an EDU is to contain the antecedent of a given mention, which could outperform a binary 'accessible/inaccessible' DRA definition. To test this, we generate a randomized test set of 10% of EDU pairs (~17K) in the data, stratified by coreference prevalence (same proportions of single coreferent mention, 2, 3, 4… as in the rest of the data). Using the Python implementation in sklearn, we train an Extra Trees Random Forest regressor (Geurts et al. 2006%GeurtsErnstWehenkel2006) on the features outlined in Section 3.2 to predict exact coreference degree (number of coreferent mentions) and a classifier for the presence of coreference (yes/no). We also train baseline classifiers (clf) and regressors (reg) on RD and EDU distance only. Table 1 shows the results.

features	RMSE (reg)	accuracy (clf)
EDU	95.01	78.36
RD	94.53	78.79
all	**71.07**	**86.83**

Table 1. Classification accuracy for binary coreferentiality and root mean square error for regression on exact mention pair count for unseen EDU pairs.

The regressor with all features achieves a root-mean-square error of 71%, meaning it is usually about 0.71 mentions away from the true coreferent mention count. Using only EDU distance or RD is worse, at about 0.95 RMSE (Root Mean Square Error). For classification of binary coreferentiality, using all features gives a gain of ~8% accuracy, close to 87% vs. close to 79% for RD and closer to 78% for EDU distance. Classifier feature importances based on Gini indices are shown in Figure 8.

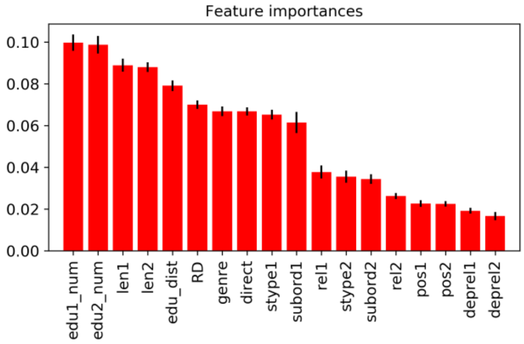

Figure 8. Variable importances for the binary classifier. Error bars give standard deviations for each feature.

The most relevant predictors, before examining any distance metrics, are the positions of the two EDUs (EDU1 is the earlier, antecedent EDU) and their lengths. This is not surprising, since late EDUs in a text have a chance to refer to more mentions, and long EDUs have more mentions. These predictors are not relevant to the DEH framework, but they are important confounds that have gone largely ignored to date. Immediately following, we see the two distance measures, with EDU distance slightly ahead of RD, and genre (another critical confound) and direct ancestry next. The remaining variables give more information about the function of the specific EDUs, including RST relations (cf. Figure 7), utterance types, clause subordination information and grammatical functions. All of these have some influence on coreference likelihood (see e.g. Trnavac & Taboada

[4] For *Purpose*, a partial reason may be that the frequent infinitive '… (in order) to do X' suppresses the unexpressed infinitive subject (i.e. the 'doer' is not expressed and cannot be pronominalized). I thank Paul Portner for this suggestion.

2012%TrnavacTaboada2012 on the importance of subordination for cataphora).

5 Discussion

The results of the models in the previous section, as well as individual correlations with predictors shown in Figures 4-7 demonstrate that a binary model of accessibility in DRAs is unnecessarily impoverished. We can get much better prediction accuracy for coreference domains using a multifactorial model, which is also intuitively plausible: sentence length and position are expected to have an influence, and not all RST relations and sentence types are equal with respect to coreference likelihood. The results also support the conclusion that RD and EDU distance metrics are both useful, and can be used in conjunction.

It is important to note that the features examined in this paper are EDU based, and RST graph-based, since our focus has been on properties that make a pair of EDUs likely to form a domain of coreference. It goes without saying that actual prediction of coreferentiality should take into account the inventory and properties of referring expressions within those EDUs. Thus although the classifier above is far from being able to predict exact coreference density using our features, its prediction accuracy may be considered surprisingly good considering the fact that it knows nothing about the entity types, agreement class compatibility, or even count of nominal expressions in each EDU. Although this remains outside of the scope of this paper, it seems likely that this type of information can be integrated in approaches using RST based features for prior coreference likelihood, together with established coreference resolution features.

References

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer.

Nicholas Asher and Alex Lascarides. 1998. Bridging. *Journal of Semantics* 15(1):83–113.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. (Studies in Natural Language Processing.) Cambridge: Cambridge University Press.

R. Harald Baayen. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Chloe Braud, Barbara Plank and Anders Søgaard. 2016. Multi-View and Multi-Task Training of RST Discourse Parsers. In *Proceedings of COLING 2016*. Osaka, 1903–1913.

Lynn Carlson, Daniel Marcu and Mary Ellen Okurowski. 2001. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*. Aalborg, 1–10.

Christian Chiarcos and Olga Krasavina. 2008. Rhetorical Distance Revisited: A Parametrized Approach. In Anton Benz and Peter Kühnlein (eds.), *Constraints in Discourse*. Amsterdam and Philadelphia: John Benjamins, 97–115.

Kevin Clark and Christopher D. Manning. 2015. Entity-Centric Coreference Resolution with Model Stacking. In *Proceedings of ACL-IJCNLP 2015*. Beijing, 1405–1415.

Dan Cristea, Nancy Ide, Daniel Marcu and Valentin Tablan. 1999. Discourse Structure and Co-Reference: An Empirical Study. In *Proceedings of the Workshop on the Relationship Between Discourse/Dialogue Structure and Reference*. College Park, MD, 46–53.

Dan Cristea, Nancy Ide and Laurent Romary. 1998. Veins Theory: A Model of Global Discourse Cohesion and Coherence. In *Proceedings of ACL/COLING*. Montreal, Canada, 281–285.

Greg Durrett, Taylor Berg-Kirkpatrick and Dan Klein. 2016. Learning-Based Single-Document Summarization with Compression and Anaphoricity Constraints. In *Proceedings of ACL 2016*. Berlin, 1998–2008.

Greg Durrett and Dan Klein. 2014. A Joint Model for Entity Analysis: Coreference, Typing, and Linking. *Transactions of the ACL* 2:477–490.

Pierre Geurts, Damien Ernst and Louis Wehenkel. 2006. Extremely Randomized Trees. *Machine Learning* 63(1):3–42.

Katsuhiko Hayashi, Tsutomu Hirao and Masaaki Nagata. 2016. Empirical Comparison of Dependency Conversions for RST Discourse Trees. In *SIGDIAL 2016*. Los Angeles, 128–136.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda and Masaaki Nagata. 2013. Single-Document Summarization as a Tree Knapsack Problem. In *EMNLP 2013*. Seattle, 1515–1520.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. New York, 57–60.

Ryu Iida and Takenobu Tokunaga. 2012. A Metric for Evaluating Discourse Coherence based on Coreference Resolution. In *Proceedings of COLING 2012*. Mumbai, 483–494.