Computing Reliability for Coreference Annotation

Rebecca J. Passonneau

Columbia University
Computer Science Department
New York, NY 10027
becky@cs.columbia.edu

Abstract

Coreference annotation is annotation of language corpora to indicate which expressions have been used to co-specify the same discourse entity. When annotations of the same data are collected from two or more coders, the reliability of the data may need to be quantified. Two obstacles have stood in the way of applying reliability metrics: incommensurate units across annotations, and lack of a convenient representation of the coding values. Given N coders and M coding units, reliability is computed from an N-by-M matrix that records the value assigned to unit M_j by coder N_k. The solution I present accommodates a wide range of coding choices for the annotator, while preserving the same units across codings. As a consequence, it permits a straightforward application of reliability measurement. In addition, in coreference annotation, disagreements can be complete or partial so I incorporate a distance metric to scale disagreements. This method has also been applied to a quite distinct coding task, namely semantic annotation of summaries.

1. Introduction

Coreference annotation is annotation of language corpora to indicate which expressions have been used to co-specify the same discourse entity. No matter how precise a language user might be, language interpretation is subjective. A given expression can be referentially ambiguous or vague. When annotations of the same data are collected from two or more coders, the reliability of the data should be quantified, particularly when creating a (set of) gold standard(s). Two obstacles have stood in the way of applying a reliability metric to coreference annotation.

1. Annotators can disagree about which expressions are referential
2. Different coders can relate a distinct set of expressions to a distinct number of entities

Given N coders and M coding units, reliability is computed from an N-by-M matrix that records the value assigned to unit M_j by coder N_k. Item 1 can be an obstacle to having the same units across annotations; in addition, sometimes annotators are allowed to choose which units to code. Item 2 can be an obstacle to comparing coding values across annotations. In fact, determining how to represent the *value* assigned to each unit in coreference annotation is a problem that has not been addressed before. The method I present permits a straightforward application of reliability tests. In addition, it is compatible with a reliability measure like Krippendorff's α (Krippendorff, 1980) that accommodates distance metrics, depending on the type of scale coding values fall within (e.g., nominal vs. ordinal, discrete vs. continuous). I propose a provisional distance metric.

2. Problem

Coreference annotation plays a role in the preparation of training corpora for applications from information extraction systems (Kibble and van Deemter, 2000) to dialogue (Poesio, 1993). Regardless of the language (e.g., German (Kunz and Hansen-Schirra, 2003)), the modality (e.g., spoken versus written), or the genre (e.g., newswire text versus narrative), different forms can refer to the same entity. (Whether to annotate zero references is a distinct question that should be addressed by the coreference annotation guidelines or tools.) For the sake of simplicity, the examples in this paper involve explicit noun phrases, whether proper names, descriptive NPS, or pronouns.

Figure 1 illustrates a newswire text annotated by three coders. Following the principles outlined in (Passonneau, 1994), I assume that all tokens to be annotated have been identified in advance by the investigator, either through a separate manual annotation process or automatically. I also assume the task of each annotator is to indicate for each token whether it refers to an existing discourse entity, in which case it should be coindexed with all the expressions that have already been indexed for this entity; or introduces a new discourse entity, in which case it receives a new index; or does not refer, in which case it receives no index.

All three cases are illustrated in Table 1, which shows how three annotators—RA.1, RA.2, RA.3—indexed the selected NPs for coreference. Two NPs receive the same index if they corefer, thus all three coders assign the same index to token A (*Gov. Price Daniel*) and token D (*Daniel*). Note that RA.1 and RA.3 assign a total of 4 indices whereas RA.2 assigns 5, and in addition assigns the NIL value to token J. Figure 2 represents the equivalence classes con-

Committee approval of [Gov. Price Daniel](A)'s [abandoned property act](B) seemed certain Thursday despite the protests of [Texas bankers](C). [Daniel](D) personally led the fight for [the measure](E). Under committee rules, [it](F) went automatically to a subcommittee for one week. But questions with which [committee members](G) taunted [bankers](H) appearing as [witnesses](I) left little doubt that [they](J) will recommend passage of [it](K).

- The NPs of interest have been bracketed
- Each bracketed NP token receives a unique index.

Figure 1: A newsire text for coreference annotation

ID	TOKEN	RA.1	RA.2	RA.3
A	Gov. Price Daniel	1	1	1
B	abandoned property act	2	2	2
C	Texas bankers	3	3	3
D	Daniel	1	1	1
E	the measure	2	2	2
F	it	2	2	2
G	committee members	4	4	4
H	bankers	3	5	3
J	witnesses	4	NIL	3
K	they	4	4	3
L	it	2	2	2

Table 1: Coreference annotation for a newswire text

ID	RA.1	RA.2	RA.3
A	1	1	1
B	2	2	2
C	4	5	3
D	1	1	1
E	2	2	2
F	2	2	2
G	6	7	8
H	4	9	3
J	6	10	3
K	6	7	3
L	2	2	2

1:	{A, D}
2:	{B, E, F, L}
3:	{C, H, J, K}
4:	{C, H}
5:	{C}
6:	{G, J, K}
7:	{G, K}
8:	{G}
9:	{H}
10:	{J}

Table 2: Canonical form for coreference annotation

stituted by the tokens that corefer. Each class represents a single *entity* or index, and links the expressions that refer to that entity, as well as the predications the expressions occur in. Thus each class represents what the annotator takes to have been asserted about the presumed discourse entity. If annotators disagree on the member of an equivalence class, this may reflect a different interpretation of the discourse. From Table 1, it might seem that the referential indices could be used as variable values, assuming they can be *normalized* to a single set of symbols. RA.1 and RA.3 both use 4 distinct indices, making it possible to use these coding values for each coding. However, a closer look suggests that this would be a mistake.

A clear cut case of identical codings can be seen for tokens A and D: all three coders assign A and D to the same equivalence class. In the equivalence class representation we can see more clearly how RA.1 and RA.3 disagree, despite their use of the same number of referential indices. Only two of their equivalence classes are identical. In RA.1's coding, tokens C and H corefer, and do not corefer with any other tokens. In RA.3's coding, tokens C, H, J and K are assigned to the same equivalence class. If we let 3 be the index for C in RA.1, what would it mean for 3 to be the index for C in RA.2, given that coder RA.2 thinks there are two references to this entity compared with one for RA.1? The two coders may have very different conceptions of this entity. At the bottom of Figure 2 is the set of equivalence classes created by the union of all the codings. I propose to use these classes as the *values* assigned to each token in order to directly represent when two annotators have assigned the same referential value to a linguistic expression. This proposal thus results in ten *values*, instead of the four or six in the individual codings.

The proposed representation also makes it easier to compare disagreements. Where two codings disagree, the penalty assigned to the disagreement will depend on how different the equivalence classes are after removing the unit

RA.1 (N=4) {A, D} {B, E, F, L} {C, H} {G, J, K}
RA.2 (N=6) {A, D} {B, E, F, L} {C} {G, K} {H} {J}
RA.3 (N=4) {A, D} {B, E, F, L} {C, H, J, K} {G}

10 Coding Values: {A, D} {B, E, F, L} {C, H, J, K} {C, H} {C} {G, J, K} {G, K} {G} {H} {J}

Figure 2: Equivalence Classes from Three Annotations

being coded. Here, RA.1 and RA.2 assign different values to token C: {C, H} versus {C, H, J, K}: {H} is a subset of {H, J, K}. Intuitively, this difference in values should be penalized less than if the resulting difference sets were related by intersection rather than a subset relation, which in turn should be penalized less than if they were disjoint.

3. Proposed Solution

3.1. Representation of the Coded Data

Although I propose to use the equivalence classes that tokens are assigned to as the coding values, this representation can become unwieldy for large datasets. In this section, I introduce a level of indirection to make the representation more compact. I assign a unique index to each equivalence class, analogous to the primary key in a relational database. Table 2 shows the same data in this new representation that uses indices to point to the equivalence classes.

In Table 2, the row labels are the units (NP tokens) being coded; the column labels are the annotators; the cell contents indicate the value that a specific annotator assigned to a specific unit. In contrast to Table 1, this representation shows much more clearly the distribution of agreements and disagreements. The rows where each cell has the same value correspond to the tokens where all coders assigned the same values: the rows for tokens A, B, D, E, F, and L. Similarly, patterns of disagreement are directly discernible. Because no row that does not show perfect agreement has less than 3 symbols, we see easily that there are no cases where two coders agree and the third disagrees. Further, a symbol that occurs in only one column, e.g., *3*, indicates an equivalence class assigned by only one of the annotators (e.g., RA.3: {C, H, J, K}). What is missing from this representation is how to quantify the difference in values in a way that accords with the informal observation made above about the case where one coding subsumed the other. If we treat the cell values in Table 2 as nominal variables, meaning the difference between *2* and *7* is the same as the difference between *7* and *8*, then the inter-annotator reliability for the data in Table 2, using Krippendorff's α, is .45.[1] The distance metric for nominal data is binary: all values are either alike (delta=1) or not (delta=0). This is

[1] α is equivalent to Cohen's κ (Cohen, 1960), which is the ratio of: the observed agreements less the agreements expected by chance; to, 1 less the agreements expected by chance.

not the best way to compute reliability for coreference annotation because it fails to capture the intuition that some equivalence classes are more alike than others.

3.2. Krippendorff's α

I briefly present the formula for Krippendorff's α, mainly to illustrate where the distance metric figures in. Where p_{D_O} is the probability of observed disagreements, and p_{D_E} is the probability of expected disagreements:

$$\alpha = 1 - \frac{p_{D_O}}{p_{D_E}} = 1 - \frac{rm-1}{m-1}\frac{\sum_i \sum_b \sum_{c>b} n_{b_i} n_{c_i} \delta_{bc}}{\sum_b \sum_c n_b n_c \delta_{bc}}$$

Given a table of the form in Table 2 with m coders and r coding units, the agreement coefficient is given by summing disagreements within and across columns. For every pair of values b and c, δ_{bc} is the distance between the values; n_{b_i} is the number of times the value b occurred in row i. In nominal scales, $\delta = 0$ when $b = c$; otherwise $\delta = 1$. A full discussion is in Krippendorff (Krippendorff, 1980).

3.3. Distance Metrics for Sets

Now I will illustrate three cases of disagreement using examples where a pair of coders assigned distinct *referential values* to the same NP tokens. In the first example, the values are fairly similar: one is a subset of the other. Coder RA.1 assigned to unit C a value represented in Table 2 as *4* whereas RA.3's coding is represented as *3*. These correspond respectively to the sets {C, H, J, K} and {C, H}. In my coding scheme, every token is necessarily a member of the set that is its *referential value*, so to pose the question of how different two values are, I first remove the current token from the values assigned by the annotators; note that if this step is not taken, all values across annotators for the same unit will necessarily overlap. Here, in the case of the RA.1 and RA.3 codings of C, the set differences yield {H, J, K} and {H}. These two sets thus represent each annotator's decision about the co-specifying expressions for C: RA.3's coding of C subsumes RA.1's.

In the second example, we find a different set relation. Coder RA.1 assigned J a value encoded as 6, and RA.3 assigned *3*. Removing J from the equivalence classes that are the *actual* values yields {G, K} and {C, H, K}. Neither set subsumes the other, but the set intersection is non-null: {K}. In this case, the referential values of the two annotations overlap, so they are not in as much disagreement as in the case of disjoint difference sets. This third case is shown for token K: RA.2 and RA.3 assigned the values *7* and *3*, whose difference sets are {G} and {C, H, J}.

Intuitively, identity, subsumption, intersection and disjunction are ordered from most agreement to least agreement. To capture this intuition, I assign the δ values 0 for identity, .33 for subsumption, .67 for intersection, and 1 for disjunction. Applying this distance metric to the data in Table 2 yields a much higher α of .74.

Let us reconsider the distribution we see in Table 2 with respect to the two different values for α we get. If we look only at the columns with perfect agreement, we see quickly that 6 out of 11 columns exhibit this pattern, or roughly half the table. On a metric that treats all disagreements equally, than about half the data involves disagreement, which corresponds to the case where we treat the values as nominal data, and we get $\alpha = .45$, or close to half. However, if we treat the non-identical values within a column for a token as more or less different, depending on whether we find subsumption, intersection or disjunction relations, then we can capture the intuition that some disagreements should be weighted more heavily than others.

Narr	VRec	VPrec	α	EQ κ
1	.96	.96	.49	.85
2	.90	.93	.64	.65
4	.94	.98	.74	.89
5	.95	.99	.64	.89
6	.94	.97	.63	.83
8	.91	.96	.54	.84
9	.88	.96	.46	.75
11	.92	.95	.47	.79
12	.90	.92	.52	.74
15	.93	.93	.51	.80
16	.97	.98	.51	.93
17	.95	.96	.64	.86
18	.93	.96	.61	.84
19	.96	.93	.67	.85

Table 3: Comparing Metrics

4. Comparison with Other Coreference Scoring Schemes

Because of the obstacles to applying a reliability metric to coreference data, other investigators who have looked at how to compare coreference annotations have used recall and precision, notably (Vilain et al., 1995). As proposed here, they use equivalence classes to represent a coreference encoding, but their approach is otherwise entirely different. They compute recall and precision directly over equivalence classes. From a gold standard equivalence set $\mathcal{S}$ and a response set $\mathcal{R}$, they create a partition $\mathcal{P}$. Recall of a member of $\mathcal{S}$ (e.g., {C, H, J, K}) is its cardinality less the cardinality of its partition by the response set (e.g., {{C, H}, {J, K}), divided by one less than its cardinality (e.g., $\frac{4-2}{4-1}$). Recall of the entire equivalence set is a sum of the recall values for each member. Precision is computed by switching the gold standard and response.

To compare the two metrics, I used coreference data from a set of spoken monologues. As described in (Passonneau and Litman, 1997), we created a gold-standard coreference coding of a set of transcribed monologues known as the Pear stories (Chafe, 1980). In general, the method in (Vilain et al., 1995) cannot be used unless a gold standard already exists, because recall and precision are not symmetric. Unlike recall and precision, α takes into account the likelihood that two annotators will agree, given the rate at which values occur in the data; as pointed out in (Carletta, 1996) regarding percent agreement, it thus factors out chance agreement. I also report κ values computed over equivalence classes as in (Passonneau, 1997).

Table 3 shows the results for the new annotator on fourteen Pear narratives; it should be noted that our gold

standard enforces strict semantic distinctions regarding set membership, and that there are varying sets of **boys** and **pears** which introduce referential ambiguity for expressions such as *the boys* and *the pears*. Given the sensitivity of α to the distance metric used, there is no absolute α value for high agreement, but Krippendorff (Krippendorff, 1980) cites a variety of work indicating that values below .67 are inconclusive, as in this data. Impressionistically, the differences in α correlate with differences in coherence across narratives; thus narrative 9, with the lowest α in the table, and the second lowest κ, has many incomplete utterances as shown in 19.2 (each line is an intonational phrase with a "." for a final fall, "?" for a rise, else ","); in 19.3 the phrase *another girl* introduces the only female character:

19.1 And he's driving along the road,
19.2 {clears throat} and he comes up [pause] uh [pause] oh.
19.3 There's another girl,
... ...

It also has awkward constructions, as in the following three intonational phrases meaning *the hat belonging to the boy who had the pears*; it is *the hat* that is in the road:

27.1 The boy who had the pears',
23.2 hat,
27.3 [pause] is way back on the road,

While precision for narrative 9 (.88) was lower than for other narratives, recall was rather high (.96); it is not clear how to interpret this. The overall distribution for recall and precision exhibits little differentiation, with values ranging from .88 to .97 for recall and .92 to .99 for precision.

The distance metric I use here is the result of a combination of a principled difference between partial versus complete overlap in values, and experimentation. For example, I applied several distance metrics proposed by readers of earlier drafts of this paper that took into account the relative sizes of two equivalence classes being compared, but the resulting κ values were surprisingly low. My hypothesis is that what matters is not the relative frequency an entity is referred to, hence the relative size of its equivalence class, but what has been asserted about it. To arrive at a distance metric that captures what is subjectively distinct about the equivalence classes, it might be necessary to compare the semantic properties associated with each equivalence class, that is, the descriptive nouns, adjectives and predications that specify the respective discourse entities.

5. Conclusion

In sum, disadvantages to recall and precision metrics as applied directly to the type of equivalence class representation illustrated in Figure 2 are several: they cannot be applied unless a gold standard already exists; they cannot apply to multiple coders; the rate at which different referential values occur is not represented; and finally, they do not allow for distance metrics. The annotation method proposed here addresses all these problems.

The proposal presented here applies to any coding where annotators create sets from the units being coded, and are free to create any number of sets. Although I know of few coding tasks that fit this criterion, and I originally began this work for coreference annotation, I was motivated to take it a step further for a recent semantic annotation task. In (Passonneau and Nenkova, 2003), we faced an identical problem in a semantic annotation method for annotating content units in summaries. To evaluate the interannotator reliability of our Summary Content Units (SCUs), I applied a slight variant of the approach presented here.

Acknowledgements. This work was funded by NSF grant IRI-9528998 and DARPA grant N66001-00-1-8919. I thank Vera Horvath, Ani Nenkova, Owen Rambow, Andrew Rosenberg and Advaith Siddharthan for comments.

6. References

Carletta, Jean, 1996. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*:249–54.

Chafe, Wallace L., 1980. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex Publishing Corporation.

Cohen, Jacob, 1960. A coeffiecient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Kibble, Rodger and Kees van Deemter, 2000. Coreference annotation: Whither? In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Krippendorff, Klaus, 1980. *Content Analysis: An Introduction to Its Methodology*. Beverly Hills, CA: Sage Publications.

Kunz, Kerstin and Silva Hansen-Schirra, 2003. Coreference annotation of the TIGER treebank. In *Poster Session Presentation at Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*.

Passonneau, Rebecca and Ani Nenkova, 2003. Evaluating content selection in human- or machine-generated summaries: The pyramid method. Technical Report CUCS-025-03, Columbia University.

Passonneau, Rebecca J., 1994. Protocol for coding discourse referential noun phrases and their antecedents. Technical report, Columbia University.

Passonneau, Rebecca J., 1997. Applying reliability metrics to co-reference annotation. Technical Report CUCS-017-97, Columbia University.

Passonneau, Rebecca J. and Diane Litman, 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23.1:103–139. Special Issue on Empirical Studies in Discourse Interpretation and Generation.

Poesio, M., 1993. A situation-theoretic formalization of definite description interpretation in plan elaboration dialogues. In P. Aczel, D. Israel, Y. Katagiri, and S. Peters (eds.), *Situations Theory and its Applications, vol.3*, chapter 12. Stanford: CSLI, pages 339–374.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman, 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference*. San Francisco: Morgan Kaufmann.