Scalar Anaphora: Annotating Degrees of Coreference in Text

Bingyang Ye and **Jingxuan Tu** and **Kyeongmin Rim** and **James Pustejovsky**
Department of Computer Science
Brandeis University
Waltham, Massachusetts
{byye,jxtu,krim,jamesp}@brandeis.edu

Abstract

In this paper, we examine the concept of coreference in natural language text, and the challenge of identifying when two or more narrative entities should be resolved as anaphorically bound, and hence viewed as semantically identical or related. To help answer this question, we propose a coreference scale (*Scalar Anaphora*) for determining the degree of similarity between an anaphoric expression and its antecedent in narratives. We create a corpus of pairs of such anaphors and antecedents and annotate the relations between them based on the newly defined scale. Our data shows that the ratio of human annotators' agreement score aligns with the scale of coreference. We also present the baseline results of predicting the scales using recent T5 and GPT-4 models, which suggests that predicting such fine-grained scales is still a challenging task for large language models. We will make the code and the data publicly available.

1 Introduction

Anaphora resolution involves identifying the mentions that contain anaphoric or coreferential relations and predicting the correct relation for the extracted mentions. Conventional anaphora resolution corpora such as OntoNotes (Marcus et al., 2011) and ACE (Doddington et al., 2004) focus largely on coreference. However, there are many "anaphora-related" phenomena that are extremely important for facilitating deeper linguistic analysis and modeling by modern NLP systems.

Bridging (Clark, 1975; Asher and Lascarides, 1998; Hawkins, 2015) is one such phenomena. It refers to a set of non-identity anaphoric relations. Despite the recent growing attention on bridging, existing corpora and methods (Uryupina et al., 2020; Yu et al., 2022) still treat the bridging resolution and coreference resolution as two independent problems, overlooking the linguistic closeness between anaphoric phenomena and coreferential identity, which leads to discrepancies of annotations on the same mention across corpora (Recasens et al., 2010).

To alleviate the principal complexity resulting from using a binary distinction of identity and non-identity, Recasens et al. (2011) proposed the concept of "Near-Identity" which denotes partial identity relations between mentions. Many previous works acknowledged the importance of having a mid-ground as Near-Identity (Uryupina et al., 2020; Rösiger et al., 2018), but did not include it in their annotation schema or modeling implementation for fear of introducing too much uncertainty. Recasens et al. (2012) is the first attempt to create a public corpus of near-identity. However, the whole typology of near-identity was treated as a coarsed weak and strong classifications, still leaving some gaps between identity, near-identity, bridging and non-identity.

In this paper, we extend the themes from Recasens et al. (2010) and treat anaphora resolution as a continuum with a middle zone of near-identity relations. To supplement and enrich the notion of near-identity, we introduce *Scalar Anaphora*, a typology that categorizes near-identity with a simplified but more operationalized granularity, while unifying it with other anaphoric relations. Furthermore, we leverage the disagreements in the raw annotation of Phrase Detectives (PD) 3.0 (Yu et al., 2023) to create a dataset using Scalar Anaphora (SA) as the annotation schema. The presence of disagreement underscores the absence of a singular, unequivocal interpretation within a specific context for anaphora resolution, consistent with the concept of SA.

The major contributions outlined in this paper include:

- The introduction of *Scalar Anaphora*, a unified typology for anaphoric relations. Specifically, we define relations of *Coreference under Description* and *Coreference under Trans-*

formation to fill the gap between identity and non-identity while considering their semantic closeness on the scale of identity.

- We leverage the raw annotations in PD 3.0 release (Yu et al., 2023) to facilitate the detection of mentions with ambiguous anaphoric relation, and create a dataset of SA relations using our typology.
- We experiment with T5 and GPT-4 as baseline models for the evaluation of our anaphoric relations against human annotations. The results suggest that identifying ambiguous anaphoric relations in SA is still challenging.

2 Related Work

Anaphora resolution refers to the task of detecting the relation that holds between two textual entities in a text. Conventional linguistic anaphora designates coreference, where the two mentions refer to (denote) the same entity or concept. The Computational Linguistics literature has broadened this term to also allow for more general anaphoric relations, where the two mentions refer to different entities, but are linked via semantic, lexical, or encyclopedic relations (Hou et al., 2018). Most existing anaphora corpora only annotate coreference (Marcus et al., 2011; Yu et al., 2023). Within the wider definition of anaphora, however, the other major phenomenon of interest is bridging.

The Vieira / Poesio corpus (Poesio and Vieira, 1997) and GNOME (Poesio, 2004) are the two early attempts to annotate bridging. Since the release of the ARRAU corpus, more efforts have been dedicated to annotating bridging relations (Markert et al., 2012; Grishina, 2016; Rösiger, 2016; Zeldes, 2017; Rösiger et al., 2018). The Prague Dependency Treebank (Hajič et al., 2020) and the Polish Coreference Corpus (Ogrodniczuk et al., 2016) are other corpora annotating bridging in languages other than English. Due to the difficulty of detecting bridging (Poesio and Vieira, 1997; Vieira, 1998), most bridging corpora are still very small. Only ARRAU has a comparatively large annotations of bridging in English with 5,512 pairs of anaphor and antecedents. Moreover, these corpora all have rather diverse definitions and annotations of bridging which makes it even more difficult to do cross-corpus analysis and modeling (Rösiger et al., 2018).

Prior research on bridging resolution typically adopts two approaches: 1) incorporating bridging recognition within information status classification (Markert et al., 2012; Hou et al., 2013a); 2) focusing solely on antecedent selection, assuming prior completion of bridging recognition (Poesio et al., 2004; Hou et al., 2013b; Hou, 2018). Vieira and Poesio (2000) and Hou et al. (2014) also experimented with rule-based systems. More recently, there are a growing number of works using neural networks to tackle the problem (Yu and Poesio, 2020; Kantor and Globerson, 2019). Kantor and Globerson (2019) proposed the first neural model for full bridging resolution, leveraging a span-based neural model originally developed for entity coreference resolution. Hou (2020) proposed a neural approach to bridging resolution based on question answering.

Near-identity is also an anaphoric phenomenon that bears great linguistic values. The near-identity relations are akin to "bridging anaphora" as indirect connections requiring inference, yet distinct as they cannot be considered anything other than identity (Recasens et al., 2010). Since Recasens et al. (2011) introduced the concept of Near-Identity and proposed to redefine coreference as a scalar relation, a series of works on near-identity have been made. Recasens et al. (2010) proposed a typology of near-identity relations that comprised fifteen relations under five families. Preliminary annotation were also made to prove that the inter-annotator agreement is stable enough for a more extensive annotation of near-identity (Recasens et al., 2012) on NP4E corpus (Hasler et al., 2006). The granularity of typology in Recasens et al. (2010), however, was lost in the annotation as all the relations were labeled as either weak or strong near-identity. Ogrodniczuk et al. (2016) also contains near-identity relations in the corpus. These works, despite providing a strong theoretical base for research in near-identity, still lack empirical modeling and evaluation. Our paper presents a typology of anaphoric relations by merging and simplifying the typology of near-identity in Recasens et al. (2010). The Scalar Anaphora typology offers a means to establish a corpus with more nuanced subtypes of anaphoric relations, organized semantically in a hierarchical manner, as these SA relations correspond to varying degrees of identity on a scale. Additionally, our study delves into the modeling of SA relations and assesses their alignment with human annotations to explore the practicality of this schema.

Recasens et al. (2010) defined the anaphoric relation between one facet or attribute of an entity and the entity itself as a subtype of near-identity. It is also referred as metonymy in Markert and Nissim (2007) and Pustejovsky and Rumshisky (2009). In this work, we are only treating metonymy as a part of the near-identity. The type structure proposed in Generative Lexicon theory (Pustejovsky, 1995) could serve as the theoretical approach to further address the categorization of dot objects (systematic polysemies).

Recent work also studied the tracking of transformation or changes of entities within the frame of anaphora resolution. Fang et al. (2022) and Rim et al. (2023) annotated anaphoric relations including coreference and bridging for procedural texts. They treat the transformation of entities, e.g., *oil* mixed with *salt* being later referred to as a *mixture*, as a bridging relation. Rim et al. (2023) also defined the concept of Coreference under Transformation, which is the first attempt to introduce transformation of events into the scope of anaphora resolution. Oguz et al. (2022) presented a multimodal anaphora corpus on recipes where the transformation is annotated as a near-identity rather than bridging. Zeldes (2021) also argued the importance of tracking the change of entities over time for coreference resolution problem and proposed adding a new layer of annotation on "scope" for OntoNotes.

Learning from disagreements among coders has been a growing topic in the NLP field. An emerging trend in dataset creation involves moving beyond a solitary "gold" annotation to encompass the inclusion of the entirety of raw annotations provided by coders.Uma et al. (2021) and Leonardelli et al. (2023) posted shared tasks to model the disagreements among annotators in a variety of fields including coreference resolution, pos tagging, humour detection, etc. Recasens et al. (2012) also created the NIDENT corpus by automatically identifying near-identity relations using human coders' disagreements.

3 Defining Scalar Anaphora

For this paper, we propose a coreference scale called Scalar Anaphora, for determining the degree of similarity between an anaphoric expression and its antecedent in narratives. Figure 1 shows the typology of Scalar Anaphora as a decision tree, where neighboring nodes are semantically closer on the scale of identity. Following Recasens et al. (2010), we believe anaphoric binding relations in text are best viewed as expressing degrees of identity between the entities. In this paper, we go further and argue that these types can be partially ordered on a scale of referential similarity.

Formally, for two *narrtive entities*, e_1 and e_2, we identify five anaphoric relations on the scale based on the semantic closeness between them. The scale begins by distinguishing between the relation of (strict) *Identity* and (strict) *Non-identity*. If e_1 and e_2 are substitutible under both transparent and opaque contexts, then we say e_1 and e_2 are coreferential or Identical. For example, conventional coreference clusters, e.g., including *Clinton*, *Hillary Clinton*, and *she*, illustrate semantic substitutibility between all members of the cluster, including opaque contexts. Hence, in a belief context, such as *believes x is a good Senator*, any member of the cluster can be substituted without changing the truth value. Strict identity is the strongest relation of similarity.

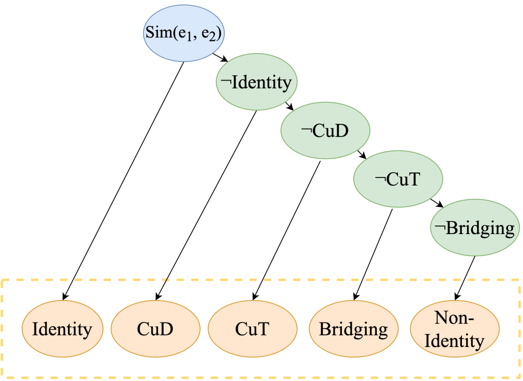


Figure 1: Typology of Scalar Anaphora.

The nearest relation to this comes about by identifying when a pair of entities is not substitutible under both opaque contexts and non-opaque. This arises with occupational and functional descriptions of entities, complicating substitutions.

(1) **Clinton**[ANTECEDENT], **the Senator**[ANAPHOR] from New York, voiced her concerns about the proposed bill during the congressional hearing.

For example, in 1, while the pair ("Clinton", "the Senator") are substitutible under non-opaque contexts (being female, American, medium height), the functional nominal *senator* can be embedded in an opaque context ("a very good senator"), while not allowing Clinton to necessarily be judged as good. Hence, we introduce a class of *Coreference under Description (CuD)* to describe the semantic relation between two entities, if they are substitutible only under non-opaque contexts. Hence, *CuD* is weaker than *Identity*.

The next class on the scale of similarity is defined by *Coreference under Transformation (CuT)*. This includes entities that are “substance identical", but are not formally identical. For entities that undergo changes by virtue of explicitly mentioned actions or processes (slicing, chopping, grilling), If two entities denote identical substances, regardless of individuation (form), we say they are substance identical. If the formal difference is the result of a transformative action, e.g., chopping or grilling, we say e_1 and e_2 are coreferential under transformation, e.g., *an onion* and *the chopped onion*. As a result, *CuT* is weaker than *CuD*.

The final distinction is defined by identifying whether the two entities are conceptually of the same type or different type. This of course includes Clark’s original examples of *bridging* relations, where we focus on tangible relations such as part-of, member-of and location. Clearly, *Bridging* is weaker than *CuT*. If none of these relations holds, we identify two entities as being in a strictly non-coreferential relation. The extremum of dissimilarity, *Non-Identity*, is therefore weaker than *Bridging*.

4 Corpus Annotation

Inspired by the method in Recasens et al. (2012) of automatically extracting mentions of near-identity relations by leveraging coders’ disagreement, we seek to use disagreement scores to speed up the process of identifying mention pairs with anaphoric relations rather than annotating exhaustively. Once the pairs are automatically extracted, we apply our schema of SA to them and annotate the scale of annaphora for each of the relation they hold.

4.1 Data Preparation

PD 3.0 (Yu et al., 2023) is a corpus collecting multiple human judgments about anaphoric reference crowdsourced in the form of Games-With-A-Purpose (Von Ahn, 2006) on fictions and Wikipedia texts. During annotation, players either aim at labeling antecedent for a given anaphor or they make a binary anaphoric judgment about other player’s annotation where the participants have to agree or disagree with the interpretation. We prepare the annotation by extracting mentions from the PD 3.0 corpus because of its rich annotations. Every mention in the texts is at least annotated by 8 players (20 in average). And for each different anaphoric judgment of two mentions, there are at least 4 players conducting the validation. The disagreement among the players for each pair is also reported.

We only use 35 Wikipedia texts from the PD 3.0 corpus gold data as our source data because comparing to fictions, Wikipedia tends to contain more proper nouns and less pronouns, which usually hold identity relation with their antecedents. The other reason is that Wikipedia requires more common knowledge than interpretation of the context, which results in less confusion among the annotators.

We parse the raw data from masxml files and extract candidate anaphoric relation pairs. Each pair has two mentions *anaphor* and *antecedent*, along with an array of human judgments agreeing or disagreeing with this interpretation. We calculate the DISAGREEMENT SCORE (DS) by dividing the number of disagreements by the total number of judgments. The higher the DS, the higher the ratio of disagreement among annotators.

Intuitively we hypothesize that the DS could indicate the scale of identity between the *anaphor* and *antecedent*, and the DS will be inversely proportional to the identity scale, i.e., with the DS increasing, the two mentions are less identical. In that sense, we are binning our set of pairs into three bins according to their DS assuming that different bins would show corresponding distributions of SA relations. We set the three bins as [0, 0.4], [0.4, 0.7] and [0.7, 1.0].

While there are 2,939 pairs extracted from the PD corpus, we do not have enough resources to annotate every one of them. To keep the topic diversity from the Wikipedia texts, for the document from each topic, we randomly sample three pairs from each DS bin. After careful examination, we exclude 7 cases where the sentence contexts are limited or missing for determining the anaphoric relation of the pair. Finally we have 308 pairs that are split into three batches.

4.2 Scalar Anaphora Annotation

Given a pair of *anaphor* and *antecedent*, and their sentence contexts, we ask annotators to annotate the pairwise anaphoric relation. After each round of annotation, annotators would adjudicate disagreements and create the harmonized annotation. All 308 pairs are dually annotated in three batches by two expert annotators from the linguistics and computer science departments of a US-based university. The annotation involves each annotator classifying the relation into the SA typology by judging the degree of identity between pairs of mentions. We design the annotation workflow based on the SA typology from Figure 1 and follow its decision-tree based methodology:

1. The annotator should first judge if the two mentions are strictly identical, which means they appear to denote the same individual. If yes, annotate IDENTITY.

2. If not, then check if one mention represents one facet or some attributes other than formal role of the other mention. For example, a company produces a product (i.e., a dot object with a metonymic interpretation (Pustejovsky, 1995)). If yes, annotate CUD.

3. Then check if one mention is substance identical to the other mention after some transformative actions where they are no longer strictly identical but still share some common characteristics. If yes, then annotate CUT.

4. Next, check if both mentions point to two entities that are conceptually of the same type or different type, while holding some relations such as part-of or location. If yes, then annotate BRIDGING.

5. Finally, if none of that above relations holds and the two mentions point to different entities, annotate NON-IDENTITY.

We use pairwise F1 and Cohen's Kappa as our metrics for Inter-Annotator Agreement (IAA). Table 1 shows the IAA from each round of the annotation. The complexity of annotating CuD and BRIDGING leads to most of the disagreement from the first round of the annotation. However, as annotators are getting more familiar with the SA typology, the IAA increases and reaches the highest in the last round.

The IAA for each relation in F1 is shown in Table 2. CUD, CUT and NON-IDENTITY constitute the relations with the highest disagreement in that they are of fewer instances and they tend to be more confusing because of their inherit ambiguity. For example, CUD are often mistaken as BRIDGING where the attribute of an entity is regarded as a relation:

	$F1$	Cohen's κ
Round 1	51.43	0.31
Round 2	66.67	0.51
Round 3	76.19	0.64

Table 1: IAA of each annotation round.

(2) **Laramie cigarettes** [ANTECEDENT], seeing an opportunity to sell **their products** [ANAPHOR] to children legally, offers to buy the rights to market tomacco for $150 million.

For CUT, the nature of narration in Wikipedia data exerts more subtlety onto the transformation unlike procedural texts. In 3, *Henry* undergoes a series of events including captivity and location change. However, one annotator overlooked the transformation and labeled CUT as IDENTITY.

(3) a. **Henry** [ANTECEDENT] was found off the coast of North Wales in a lobster pot, and is in captivity at the Blackpool Sea Life Centre in North West England;
b. **Henry** [ANAPHOR] is going to be in a new exhibit with an octopus at the Blackpool Sea Life Centre, entitled "Suckers".

The reason why the disagreement for NON-IDENTITY is low is that judgment is heavily context based. The anaphor and antecedent are usually similar strings but actually refer to two different entities after contextualization.

(4) An advantage of the knork is that it can be used easily by people who have **only one arm** [ANTECEDENT]; Roald Dahl reports in Boy how his father invented a knork precursor as a result of losing **his arm** [ANAPHOR].

We are pleased to report a high level of agreement in the annotation of both IDENTITY and BRIDGING instances. The robust concordance observed in IDENTITY annotations can be attributed to their relatively straightforward criteria, primarily involving exact string matching and explicit pronoun references. In the case of BRIDGING, the flexibility of its annotation criteria facilitates the discernment of tangible relations between mentions.

In our final corpus (Table 2), the ratio of near-identity (57.79%) in all annotations of anaphora is significantly higher than 12%-16% which is reported in (Recasens et al., 2012) as we have a more strict definition of Non-Identity and a broader definition of Bridging resulting in a shift from Non-Identity to Bridging.

	Count	Ratio (%)	IAA (F1)
Identity	114	37.0	75.98
CuD	31	10.1	40.68
CuT	18	5.8	37.04
Bridging	129	41.9	70.87
Non-Identity	16	5.2	36.36
Overall	308	100	65.31

Table 2: Statistics of annotation in terms of SA relation.

4.3 Correlation between Scales and Disagreement

To further understand whether the disagreement of the mention pairs from the PD 3.0 corpus can help identify high-quality candidates for our annotation, we investigate the possible correlations between the SA relation types and the DS of the mention pairs. We start by calculating the Spearman Rank Correlation coefficient (Spearman, 1961) between SA and DS. The score is 0.2248 which indicates there is a modest correlation between the two variables. Figure 2 details the distribution of SA relations when grouping them into the bins that are used in our annotation. Identity is the relation with the highest proportion, showing that it is less confusing. The proportions of CuD and CuT remain similar across bins of low and medium DS and decrease in the high DS bin. This indicates that the two relations tend to trigger low and medium disagreements among annotators and behave like Identity. Most Non-Identity cases are in the final bin, indicating that it is the most confusing among all relations. Bridging, being the most dominant relation in the medium and high DS bins, has a similar trend of appearing more in the bins of higher DS as Non-Identity.

Table 3 shows the average DS of each relation. The DS of Identity is lower while the other relations all demonstrate a comparatively high DS. Notably, the average DS increases as the relation becomes more towards non-identity on the anaphoric scale, which aligns with our hypothesis that DS could be inversely proportionate to identity. Overall, we believe that the DS is a useful resource for anaphoric relation annotation, and the correlation could be more statistically significant with more annotations.

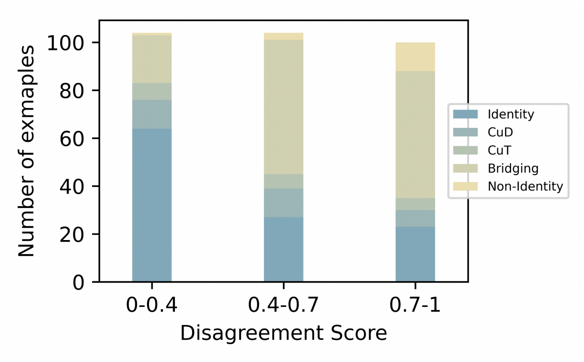


Figure 2: Distribution of SA relation in different DS bins (left inclusive).

	Average DS
Identity	0.376
CuD	0.465
CuT	0.477
Bridging	0.594
Non-Identity	0.727

Table 3: Average DS of each relation.

5 Scalar Anaphora Resolution

In this section, we present experiments of the task for anaphora resolution with fine-grained relations that we defined in the SA. We explore baselines from language models and provide further insights on our data. In our experiments, we formalize SA resolutions the task of identifying the SA relation between each mention pair given the sentence context of the entity.[1]

5.1 Data Processing

We use our annotated data for model training and evaluation. For all the pairs from each SA relations, we randomly sample 80% of the pairs for training and hold out the other 20% for testing. Table 4 shows the train test split for the experiments. Since some relation have much fewer pairs than the others, sampling by relation type is useful for ensuring the data balance between train and test.

Since PD only contains human-selected pairs where the two entities have associations, there is no "real" non-identity relation between the existing pairs. With that in mind, for each Wikipedia topic,

[1] Unlike conventional coreference resolution tasks, we provide gold mentions and only predict the relations between the mentions as our baselines are designed for providing insights on the new relations from SA.

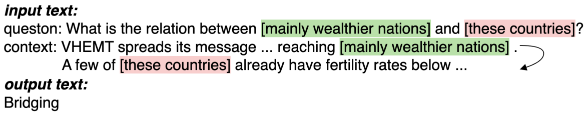
we generate two negative pairs with mentions randomly sampled from all the mentions from this topic. Those negative pairs will also be labeled as Non-identity in modeling. When reporting the results, we will label these pairs as Negative.

	Train	Test
Identity	91	23
CuD	25	6
CuT	14	4
Bridging	103	26
Non-Identity	13	3
Negative	56	14
Overall	302	76

Table 4: Train test split of the SA dataset.

5.2 Experiment 1: Scalar Anaphora with T5

Experiment Setup We use the recent sequence-to-sequence generation model T5 (Raffel et al., 2020) as the baseline. We set the input sequence as the question answering format with entities that are highlighted in the text. An example sequence is shown in Figure 3. The input includes the questions and the context where the mentions are wrapped by a pair of squared brackets ([...]).The output is the SA relation. We fine-tune the T5-base model on the training set, and evaluate the results on the testing set. Model performance was evaluated using precision, recall and F1-score.



input text:
queston: What is the relation between [mainly wealthier nations] and [these countries]?
context: VHEMT spreads its message ... reaching [mainly wealthier nations] .
A few of [these countries] already have fertility rates below ...
output text:
Bridging

Figure 3: Example of T5 model input and output for SA resolution task.

Results Table 5 shows the results of the pairwise SA relation classification on our test set. Identity and Bridging are the two relations that achieve relatively high F1 scores. The reasons are: 1. There are more training examples; 2. The two relations are relatively easy to categorize which aligns with human annotation. The result of randomly picked negative examples is also relatively high in that they are mostly just completely distinct entities thus also straightforward to distinguish. It is not surprising to see that the performance on CuD is low. T5 often times confuse it with Identity. The model also fail to predict any CuT or Non-Identity relation. Besides the fewer number of examples, the ambiguity of the two relations also contributes to the poor performance. Notably, T5 labels all examples of these two relations as Identity.

(5) It bought this land as **a standard-sized lot in 1903** [ANTECEDENT], but the City widened Pender Street in 1912 and expropriated 24 feet (7.3 m) of **the lot** [ANAPHOR].

(6) **The bulb** [ANTECEDENT] was officially listed in the Guinness Book of World Records as “the Most Durable Light”, in 1972, replacing **another bulb** [ANAPHOR] in Fort Worth, Texas.

For example, in 5, the relation of CuT is mistakenly predicted as Identity. This indicates that T5 model is unable to capture the transformative event the antecedent undergoes; while in 6 the model also failed to detect that *the bulb* and *another bulb* are distinct entities in a complicated context.

	P	R	F1
Identity	56.25	78.26	65.45
CuD	100	16.67	28.57
CuT	0	0	0
Bridging	60.00	57.69	58.82
Non-Identity	0	0	0
Negative	100	28.57	44.44
Overall	52.71	30.20	32.88

Table 5: Pairwise relation classification results on the test set with T5.

5.3 Experiment 2: Scalar Anaphora with GPT-4

Experiment Setup We experiment with GPT-4 (Brown et al., 2020; OpenAI, 2023) as another baseline for the SA resolution task. Comparing to the T5 baseline with fully supervised learning (§5.2), we use GPT-4 with few-shot prompt learning. In each prompt, we use a single set of 5 exemplars from the training set and a human-created instruction on how to perform the task. We conduct prompt tuning on a small subset of the training set, and evaluate the best prompt on the testing test. Similar to T5 baseline, Model performance was evaluated using precision, recall and F1-score.

Prompt Tuning We randomly sample 25 pairs from the training set as the “seeds” to evaluate the GPT-4 performance with different prompt formulations. Table 6 shows the prompt combinations

```
The following describes the task that predicts the relation between two phrases from the text.
The text spans of the phrase are wrapped within "[]". In this task, we define 5 types of relations:
    - Non-Identity: The two phrases point to different entities
    - Identity: The two phrases point to the same entity.  They have the same set of attributes,
                or one phrase represents the most important feature of the other phrase.
    - Role: One phrase represents one facet or some attributes of the other phrase. But this attribute should not be the most important one.
            For example, a company produces a product, is headquartered in a location, has a president, etc.
    - Transformation: Entity from one phrase undergo some transformation of which the outcome is no longer
    - Bridging: Both phrases point to two different entities, but these two entities usually related in a way that is not explicitly stated
```

```
The following describes the task that predicts the relation between two phrases from the text.
The text spans of the phrase are wrapped within "[]".
Please predicts the relation in the following order:
    1. Both phrases point to two different entities, but these two entities often holds some relations. E.g., one phrase refers to something
       that is part of the other phrase or one phrase could cause the other phrase to happen. Please predict Bridging.
    2. If both phrases point to two different entities and is not Bridging, see if one phrase represents one facet or
       some attributes of the other phrase. For example, a company produces a product, is headquartered in a location, etc. Please predict Role
    3. If both phrases point to two different entities and is not Bridging nor Role, please predict Non-Identity.
    4. If the two phrases point to the same entities, check if the entity from on phrase undergo any transformation or event.
       If so, please predict Transformation.
    5. If the two phrases point to the same entities, and it is not Transformation. Then please predict Identity.
```

Figure 4: Flat instruction (top) and hierarchical instruction (bottom) part of the prompt.

in the experiments. Each type of the prompt consists of an instruction and 5 exemplars (5-shot). *0-shot* only contains the instruction. *5-shot-random* contains random human-generated exemplars; *5-shot-domain* contains in-domain exemplars from the training set; *5-shot-CoT* adds additional chain of thought (Wei et al., 2022) to each in-domain exemplar. We generate two types of instructions for the prompts (Figure 4). The flat one instructs the model to predict the relations all as separate and individual classes, while the hierarchy one instructs the model to make decisions following several temporally ordered steps.

	Flat Instruct.	Hierarchical Instruct.
0-shot	✔	✔
5-shot-random	✔	✔
5-shot-domain	✔	✔
5-shot-CoT	✔	✔

Table 6: GPT-4 prompt combinations for the SA resolution baseline.

Table 7 shows the results on the 25 pairs using different prompts. We achieve the highest macro F1 score with few-shot tuning using CoT and Hierarchical instructions. For the following experiments with GPT-4, we will continue using this prompt setting.

Results Table 8 shows the results of GPT-4 using few-shot learning with CoT and hierarchical structure. The model achieves pretty good results on IDENTITY and CUT. However, the performances of the other relations are not very high.

Comparing the results of GPT-4 in table 8 with that of T5, we can notice that the overall performance slightly decreases as well as the performance for most individual relation. This is likely due to supervised learning outperforming few-shot learning since the task is non-trivial and it is naturally difficult to fully understand all the relations with just a few examples. The performance of relation IDENTITY is consistently high across the two models, while NON-IDENTITY still cannot be correctly predicted. This complies with our assumptions that IDENTITY relation is fairly easy to categorize and NON-IDENTITY is very confusing. We are glad to see that the F1 score of CUT increases significantly after explicitly asking the model to pay more attention to the transformative event. However, it is disappointing to see that the performances on CUD, BRIDGING and NEGATIVE all drop.

		P	R	F1
Flat	0-shot	32.41	**40.00**	33.63
	5-shot-random	36.94	33.33	26.79
	5-shot-domain	20.50	30.00	21.30
	5-shot-CoT	40.00	**40.00**	36.41
Hierarchy	0-shot	44.44	30.00	34.78
	5-shot-random	46.30	30.00	27.78
	5-shot-domain	41.32	30.00	34.76
	5-shot-CoT	**50.11**	36.67	**37.90**

Table 7: Pairwise relation classification results on 25 random examples with different prompt settings.

	P	R	F1
IDENTITY	46.81	95.65	62.86
CUD	10.00	16.67	12.50
CUT	40.00	50.00	44.44
BRIDGING	66.67	15.38	25.00
NON-IDENTITY	0	0	0
NEGATIVE	100	14.29	25.00
OVERALL	43.91	32.00	28.30

Table 8: Pairwise relation classification results on the test set with GPT-4.

6 Conclusion

We have proposed the Scalar Anaphora, a unified typology for anaphoric relations that can be identified and evaluated between coreference and non-coreference by considering their semantic closeness on the scale of identity. To that end, we have defined *Coreference under Description* and *Coreference under Transformation*, two additional granular relations that express difference semantic closeness on the scale of identity. We have constructed a new dataset that encodes manually annotated anaphoric relation between each mention pair, and our annotations have been able to show that the anaphoric relation correlates with human judgments on the closeness of each mention pair on the identity scale. We have also performed pairwise classification tasks on the anaphoric relations and presented baselines from recent T5 and GPT-4 models. The results have shown that the understanding of anaphoric relations remains challenging to current large language models. In future research, we intend to apply our method and annotation to more data and a broader range of text genres. We will also explore the validity and application of the anaphoric scale typology on the chain of cluster of entities and mentions by not limiting it to pairwise evaluation.

Ethics Statement

In conducting this research and preparing this paper, we want to affirm that our research has solely focused on scientific inquiry and there are no ethical concerns or issues that have arisen in the course of our study.

Limitations

Since the goal of PD annotation is to only annotate coreference, the raw data we process would be biased towards identity, and many cases of other SA relations could be omitted. And during data preparation stage, we exclude the human expert annotations and only focus on crowdsourced annotations in PD. In future work, we plan to take advantage of the expert annotations in determining the DS since it is of higher quality. Also due to the nature of narrative texts, the number of instances of CuT is low. A better understanding of how transformation affects the semantic closeness of an entity to its antecedent requires an extended annotation on procedural texts where tranformative events are more prevalent. To address the aforementioned issues, a new corpus of annotation of SA relations on different types of texts is needed.

References

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Herbert H. Clark. 1975. Bridging. In *Theoretical Issues in Natural Language Processing*.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Biaoyan Fang, Timothy Baldwin, and Karin Verspoor. 2022. What does it take to bake a cake? the RecipeRef corpus and anaphora resolution in procedural text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3481–3495, Dublin, Ireland. Association for Computational Linguistics.

Yulia Grishina. 2016. Experiments on bridging across languages and genres. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 7–15.

Jan Hajič, Eduard Bejček, Jaroslava Hlavacova, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague dependency treebank - consolidated 1.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.

Laura Hasler, Constantin Orasan, and Karin Naumann. 2006. NPs for events: Experiments in coreference annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

John Hawkins. 2015. *Definiteness and indefiniteness: A study in reference and grammaticality prediction*. Routledge.

Yufang Hou. 2018. A deterministic algorithm for bridging anaphora resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1938–1948, Brussels, Belgium. Association for Computational Linguistics.

Yufang Hou. 2020. Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438, Online. Association for Computational Linguistics.

Yufang Hou, Katja Markert, and Michael Strube. 2013a. Cascading collective classification for bridging anaphora recognition using a rich linguistic feature set. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 814–820.

Yufang Hou, Katja Markert, and Michael Strube. 2013b. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 907–917, Atlanta, Georgia. Association for Computational Linguistics.

Yufang Hou, Katja Markert, and Michael Strube. 2014. A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2082–2093, Doha, Qatar. Association for Computational Linguistics.

Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy. Association for Computational Linguistics.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

RWEHM Marcus, Martha Palmer, RBSPL Ramshaw, and Nianwen Xue. 2011. Ontonotes: A large training corpus for enhanced processing. *Joseph Olive, Caitlin Christianson, andJohn McCary, editors, Handbook of Natural LanguageProcessing and Machine Translation: DARPA GlobalAutonomous Language Exploitation*.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804.

Katja Markert and Malvina Nissim. 2007. SemEval-2007 task 08: Metonymy resolution at SemEval-2007. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 36–41, Prague, Czech Republic. Association for Computational Linguistics.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2016. Polish coreference corpus. In *Human Language Technology. Challenges for Computer Science and Linguistics: 6th Language and Technology Conference, LTC 2013, Poznań, Poland, December 7-9, 2013. Revised Selected Papers 6*, pages 215–226. Springer.

Cennet Oguz, Ivana Kruijff-Korbayova, Emmanuel Vincent, Pascal Denis, and Josef van Genabith. 2022. Chop and change: Anaphora resolution in instructional cooking videos. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 364–374, Online only. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

M Poesio. 2004. The mate/gnome scheme for anaphoric annotation. In *Proceedings of SIGDIAL*, pages 168–175.

Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 143–150.

Massimo Poesio and Renata Vieira. 1997. A corpus-based investigation of definite description use. *arXiv preprint cmp-lg/9710007*.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

James Pustejovsky and Anna Rumshisky. 2009. SemEval-2010 task 7: Argument selection and coercion. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 88–93, Boulder, Colorado. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2010. A typology of near-identity relations for coreference (NIDENT). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Marta Recasens, Eduard H. Hovy, and Maria Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121:1138–1152.

Marta Recasens, M. Antònia Martí, and Constantin Orasan. 2012. Annotating near-identity from coreference disagreements. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 165–172, Istanbul, Turkey. European Language Resources Association (ELRA).

Kyeongmin Rim, Jingxuan Tu, Bingyang Ye, Marc Verhagen, Eben Holderness, and James Pustejovsky. 2023. The coreference under transformation labeling dataset: Entity tracking in procedural texts using event models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12448–12460, Toronto, Canada. Association for Computational Linguistics.

Ina Rösiger. 2016. Scicorp: A corpus of english scientific articles annotated for information status analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1743–1749.

Ina Rösiger, Maximilian Köper, Kim Anh Nguyen, and Sabine Schulte im Walde. 2018. Integrating predictions from neural-network relation classifiers into coreference and bridging resolution. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 44–49.

Charles Spearman. 1961. The proof and measurement of association between two things.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Natural Language Engineering*, 26(1):95–128.

Renata Vieira. 1998. Definite description resolution in unrestricted texts. *Unpublished doctoral dissertation. University of Edinburgh, Centre for Cognitive Science*.

Renata Vieira and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Luis Von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Juntao Yu, Sopan Khosla, Ramesh Manuvinakurike, Lori Levin, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2022. The CODI-CRAC 2022 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–14, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Juntao Yu, Silviu Paun, Maris Camilleri, Paloma Garcia, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2023. Aggregating crowdsourced and automatic judgments to scale up a corpus of anaphoric reference for fiction and Wikipedia texts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 767–781, Dubrovnik, Croatia. Association for Computational Linguistics.

Juntao Yu and Massimo Poesio. 2020. Multitask learning-based neural bridging reference resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3534–3546, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes. 2021. Can we fix the scope for coreference? problems and solutions for benchmarks beyond ontonotes. *arXiv preprint arXiv:2112.09742*.