AMBI COREF: Evaluating Human and Model Sensitivity to Ambiguous Coreference

Yuewei Yuan and Chaitanya Malaviya and Mark Yatskar
University of Pennsylvania
{yuewei, cmalaviy, myatskar}@seas.upenn.edu

Abstract

Given a sentence “Abby told Brittney that she upset Courtney”, one would struggle to understand who “she” refers to, and ask for clarification. However, if the word “upset” were replaced with “hugged”, “she” unambiguously refers to Abby. We study if modern co-reference resolution models are sensitive to such pronominal ambiguity. To this end, we construct AMBI COREF, a diagnostic corpus of minimal sentence pairs with ambiguous and unambiguous referents. Our examples generalize psycholinguistic studies of human perception of ambiguity around particular arrangements of verbs and their arguments. Analysis shows that (1) humans are less sure of referents in ambiguous AMBI COREF examples than unambiguous ones, and (2) most coreference models show little difference in output between ambiguous and unambiguous pairs. We release AMBI COREF as a diagnostic corpus for testing whether models treat ambiguity similarly to humans.

1 Introduction

Ambiguity is a fundamental feature of language (Wasow et al., 2003) that some linguists believe arises because of a pressure for efficient communication (Haywood et al., 2005; Piantadosi et al., 2012). Recently, several works have highlighted the existence of ambiguity in tasks such as question answering (Min et al., 2020; Guo et al., 2021), frame disambiguation (Dumitrache et al., 2019), anaphora resolution (Poesio and Artstein, 2005) and language modeling (Aina and Linzen, 2021). Yet systematic evaluation of how models react to ambiguity across many types of language processing problems is missing. We contribute one such study about coreference resolution.

Coreference resolution is crucial to natural language understanding, especially in long contexts, such as dialog. Ambiguity may arise naturally in dialog, but existing models do not have well-defined target behavior for such coreferences. In contrast, when people encounter coreferential ambiguity, they recognize it, and can ask for clarification. Existing resources, such as OntoNotes (Weischedel et al., 2013), do not provide fine-grained annotations of such instances to evaluate model behavior. This may result in models not being calibrated to handle the uncertainty in interpretations of ambiguous statements. In this work, we ask how sensitive to ambiguity are models trained on these resources?

To understand how existing coreference models react to ambiguity, we construct a diagnostic corpus, AMBI COREF. AMBI COREF is composed of minimal pairs with ambiguous and unambiguous referents, created from four types of templates. Ambiguity is achieved by reducing context sizes to one sentence, and creating sentences where participating verbs under-constrain the interpretation of their arguments. For example, in Table 1, line 2, our first template leverages ambiguity around verbs expressing subjective experiences.

The templates are designed by drawing on psycholinguistic studies (Springston, 1976; Caramazza et al., 1977; Rohde and Kehler, 2014) and a core contribution of our work is to generalize their observations to create thousands of instances. We achieve this by identifying VerbNet (Schuler, 2005) classes that are likely to contain appropriate verbs, and manually assigning them to templates. Combined with variability we introduce using noun lists, AMBI COREF contains over 96 thousand sentences.

We verify that humans perceive instances in AMBI COREF in intended ways by crowdsourcing judgements (§3). Annotators are asked to find the coreferent for a pronoun in a sentence, and rate their confidence, to account for the gradience in ambiguity judgements (Schutze, 1995). We find
that, for unambiguous instances, humans strongly associate the pronoun with the intended noun but for ambiguous ones, they show reduced confidence across all templates, where the majority of participants are either not confident or mark them as ambiguous. This suggests that humans process ambiguous and unambiguous sentences in AMBI COREF in qualitatively different ways.

AMBI COREF can be used to evaluate model behavior in the presence of ambiguity. We analyze five representative English models: three in CoreNLP (Manning et al., 2014), SpanBERT (Joshi et al., 2020), and NeuralCoref 4.0 (Wolf et al., 2020) (§4). Our main evaluation involves comparing coreference cluster assignments of the pronoun, between ambiguous and unambiguous samples. Out of the 5 models we analyze show almost no behavioral change. Unlike humans, coreference models largely do not alter their decisions in the presence of ambiguity. Our analysis implies models likely need to explicitly account for ambiguity to achieve human-like behavior in the face of ambiguous input.

2 Dataset Construction

To understand model sensitivity towards corefential ambiguity, we build AMBI COREF using four types of templates, shown in Table 1. The templates are created in minimal pairs, and the only difference between the ambiguous and unambiguous counterparts lies in the choice of verb phrase. Note that while ambiguity is a graded phenomenon, we use the term “ambiguous” for instances that are more likely to elicit ambiguous human judgments and vice-versa. Verb phrases are extracted from suitable verb classes in VerbNet (Schuler, 2005), identified by manual annotation of VerbNet clusters. Each template is instantiated with verbs, names, noun-phrases, and gender-appropriate pronouns, greatly expanding the variation in cases identified in previous studies.

2.1 Template Types

Experiencer Constraint for Objects (ECO)
Springston (1976) propose the Experiencer Constraint for complement constructions which we operationalize in our templates. Verbs that mark their object as the experiencer of an emotion restrict the assignment of an object position pronoun to the subject of a declarative communication verb. Conversely, the assignment is unconstrained when the subject is an experiencer verb. For example, in row 2 of Table 1, a pronoun in the subject position of “bored” is ambiguous (but would not be so in the object position). If the main verb does not impose an experiencer constraint, row 1, then a pronoun in the subject position is ambiguous. We instantiate two variants with names (rows 1,2) and general entities (rows 3,4).

Experiencer Constraint for Subjects (ECS)
The Experiencer Constraint also suggests that verbs that mark their subjects as the experiencer of the emotion restrict the assignment of a subject position pronoun. The assignment of the pronoun is unconstrained when it is in the object position. For

Table 1: Summary of the six template pairs that make up AMBI COREF. Template slot are indicated in square bracket, and clusters are marked with subscripts and color. All templates pair an unambiguous sentence with an ambiguous sentence, where they differ only in the choice of verb phrase.

<table>
<thead>
<tr>
<th>Type</th>
<th>Ambig.</th>
<th>Template</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Experiencer Obj (ECO-1)</td>
<td>✗</td>
<td>[Emily]_A told [Jessica]_B that [she]_A [saw] [Brian].</td>
<td>11336</td>
</tr>
<tr>
<td>2 Experiencer Obj (ECO-1)</td>
<td>✓</td>
<td>[Emily]_A told [Jessica]_B that [she]_A [bored] [Brian].</td>
<td>11336</td>
</tr>
<tr>
<td>3 Experiencer Obj (ECO-2)</td>
<td>✗</td>
<td>[The mother]_A told [the sister]_B that [she]_A [saw] the client.</td>
<td>11336</td>
</tr>
<tr>
<td>4 Experiencer Obj (ECO-2)</td>
<td>✓</td>
<td>[The mother]_A told [the sister]_B that [she]_B [bored] the client.</td>
<td>11336</td>
</tr>
<tr>
<td>5 Experiencer Sub (ECS-1)</td>
<td>✗</td>
<td>[The aunt]_A told [Sarah]_B that [the daughter] [met with] [her]_A.</td>
<td>4472</td>
</tr>
<tr>
<td>6 Experiencer Sub (ECS-1)</td>
<td>✓</td>
<td>[The aunt]_A told [Sarah]_B that [the daughter] [liked] [her]_A.</td>
<td>4472</td>
</tr>
<tr>
<td>7 Experiencer Sub (ECS-2)</td>
<td>✗</td>
<td>[The father]_A told [the son]_B that the client [met with] [him]_A.</td>
<td>4472</td>
</tr>
<tr>
<td>8 Experiencer Sub (ECS-2)</td>
<td>✓</td>
<td>[The father]_A told [the son]_B that the client [liked] [him]_A.</td>
<td>4472</td>
</tr>
<tr>
<td>9 Implicit Causality (IC)</td>
<td>✗</td>
<td>[Abby]_A [called] [Jane]_B because [she]_A [wanted to apologize].</td>
<td>8424</td>
</tr>
<tr>
<td>10 Implicit Causality (IC)</td>
<td>✓</td>
<td>[Abby]_A [called] [Jane]_B because [she]_B [is leaving soon].</td>
<td>8424</td>
</tr>
<tr>
<td>11 Transfer (TOP)</td>
<td>✗</td>
<td>[Daniel]_A [baked] [the boy]_B [a cake] [after] [he]_B [asked for one].</td>
<td>8424</td>
</tr>
<tr>
<td>12 Transfer (TOP)</td>
<td>✓</td>
<td>[Daniel]_A [baked] [the boy]_B [a cake] [before] [he]_B [had lunch].</td>
<td>8424</td>
</tr>
</tbody>
</table>
example, in Table 1, row 6, “liked” is ambiguous when a pronoun is placed in the object position (but not in the subject position). We instantiate variants with names (rows 5,6) and entities (rows 7,8).

**Implicit Causality (IC)** Caramazza et al. (1977) hypothesize that implicit causality of a verb can determine the direction of pronoun assignment. For example, in Table 1 row 9, the phrase “wanted to apologize” establishes a cause for why “Emily called,” so the pronoun is constrained to the subject of “call”. Conversely, in row 10, the phrase “is leaving soon” fails to create such a relationship, leaving the pronoun ambiguous. For these templates (rows 9,10), we vary the names of the entities involved, and pair verbs (i.e. called) with constructed phrases that imply causality (i.e. apologizing), manually.

**Transfer of Possession (TOP)** Rohde and Kehler (2014) suggests that in transfer-of-possession contexts such as, “John passed the comic to Bill. He...”, the pronoun is equally likely to refer back to subject and non-subject. We draw upon this observation, and create a template around verbs that involve source-goal possession transfers. We distill the example to one sentence and pair the transfer event with a reason. For example, in Table 1 row 11, the phrase “asked for one” constrains the pronoun to be the receiver of “bake”. Conversely, before having lunch provides no such constraint, because either the receiver or giver could have “had lunch” before the event. Templates vary the names, verbs, objects, reasons, and preposition (rows 11,12).

### 2.2 Filling Template Slots

For each template, we construct a list of appropriate verb phrases, reasons (for IC and TOP templates), and shared list of gendered names and noun-phrases. Verb phrases were constructed by manually inspecting VerbNet classes. To control for name bias, we randomly sample names from popular name lists from the last 50 years, and reuse gendered noun-phrase lists from Winobias (Zhao et al., 2018). Excluding name and noun-phrase variations, templates have 114, 45, 81, 82 instances for ECO, ECS, IC, and TOP, respectively.

### 3 Human Judgements

The templates used to create AMBIREF generalize several psycholinguistic studies using lexical resources. Next, we verify that humans perceive ambiguity in these examples in the intended ways. We extract a subset of data for each template and ask Amazon Mechanical Turk workers which person a pronoun refers to (marked as A or B in Table 1) and assign confidence (definitely, or likely). Annotators were also allowed to mark the referent as entirely ambiguous. One sentence was sampled for each template and verb slot, uniformly at random. We collected 3 annotations per instance.  See Appendix A for details on the collection of human judgements.

Figure 1 summarizes our results. Human judgements for unambiguous templates favor the intended coreference decision. For unambiguous ECO, ECS, IC, TOP instances, the intended reading is selected as likely or definitely, 83.2%, 91.9%, and 85.8%, 68.3% of the time, respectively. For ambiguous instances, annotations display a substantial shift toward ambiguity. As shown in previous work, humans display substantial disagreement on ambiguous instances (Poesio et al., 2019). This is reflected in many templates, such as TOP, where humans produce almost uniform responses.

---

4 https://www.ssa.gov/oact/babynames/decades/  
5 In ambiguous cases, annotators do not reliably annotate a particular category, but often guess with low confidence. As such, we do not only report a majority opinion per instance, but instead simply report multiple annotations per sentence to see overall trends.
Figure 2: Percentage of ambiguous (■) and unambiguous (▲) instances that fall into each of our five cases for the SpanBERT-based model across all templates. All other models show negligible shifts (red and grey distributions are almost identical). The ground truth for unambiguous instances, from left to right, corresponds to A, A, A, A, A, B.

4 Model Evaluation

We now examine if we can detect sensitivity to ambiguity in existing coreference resolution models by evaluating on AMBiCOREF. We experiment with five representative models: NeuralCoref 4.0 model from Hugging Face, SpanBERT (Joshi et al., 2020) representation within the independent framework for end-to-end coreference (Joshi et al., 2019), and the three models in Stanford CoreNLP (Manning et al., 2014): deterministic (Lee et al., 2013), statistical (Clark and Manning, 2015) and neural mention ranking (Clark and Manning, 2016). All models were trained on the CoNLL 2012 dataset (Pradhan et al., 2012).

Here, we evaluate the model’s final predictions, not their distribution over possible choices. The reason is two-fold: (1) not all models produce a distribution and (2) initial analysis revealed that the models are miscalibrated, as in other settings (Desai and Durrett, 2020; Jiang et al., 2021), making it unreliable to interpret their output scores directly.

4.1 Setup

In this section, we ask, are there differences between how models process similar unambiguous and ambiguous examples? As our examples are synthetically generated, we use the unambiguous examples as a form of control. If a model is unable to link the pronoun with the correct noun on unambiguous examples for at least 40% of examples, we omit that template during evaluation.

We analyze model behavior by breaking it into cases that cover all possible cluster assignments for the pronoun in a single sentence. We compute the percentage of time a model outputs a cluster with:

- case **A**: the pronoun and noun A
- case **B**: the pronoun and noun B
- case **S**: the pronoun as a singleton
- case **M**: the pronoun, noun A, and noun B
- case **O**: the pronoun and any other span

For example, Figure 2 contains SpanBERT’s output distribution over these cases for each template. For each such distribution where the model’s performance is above threshold, we compare ambiguous (red bar) and unambiguous (grey bar) distributions using Earth Mover’s Distance (EMD) (Pele and Werman, 2009). Table 2 reports the number of templates above threshold, and their mean EMD.

Table 2: Mean Earth Mover’s Distance between matched ambiguous and unambiguous case distributions and the number of templates where models get at least 40% of unambiguous cases correct.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean EMD %</th>
<th>Templates</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpanBERT</td>
<td>11.7</td>
<td>5</td>
</tr>
<tr>
<td>CoreNLP Neural</td>
<td>3.5</td>
<td>5</td>
</tr>
<tr>
<td>NeuralCoref 4.0</td>
<td>4.0</td>
<td>5</td>
</tr>
<tr>
<td>CoreNLP Statistical</td>
<td>1.2</td>
<td>3</td>
</tr>
<tr>
<td>CoreNLP Deterministic</td>
<td>0.6</td>
<td>5</td>
</tr>
</tbody>
</table>

4.2 Results

Overall, most models we evaluated show essentially no change in output distribution over cases between ambiguous and unambiguous templates, as evidenced by near zero EMD. Most models are evaluated on five of six templates, but TOP is often excluded, representing a hard unambiguous case for most systems in its own right.

Of the models we evaluated, only SpanBERT shows significant deviation in behavior with ambiguous inputs. Figure 2 breaks down SpanBERT’s performance on each template. While average EMD is higher than for other models, it still largely doesn’t change predictions. When deci-

---

6 Roughly one week of continuous Colab GPU compute.
7 https://github.com/huggingface/neuralcoref

8 Earth Mover’s distances represent the amount of probability mass required to match two probability distributions. Hence, they help us compare distributions for ambiguous and unambiguous instances in a more interpretable way, than other possible measures like KL divergence.
isions change, often the pronoun is linked with the other noun. For example, in ambiguous cases of ECO-1, SpanBERT reduces merged outputs, and instead links the pronoun with noun B more frequently. In ambiguous cases, other models largely link the first noun-phrase (A) to the pronoun.

5 Discussion and Conclusion

Overall, our results suggest that model behavior significantly deviates from how human treat ambiguous coreference. We lend more evidence that models miss aspects of how people understand language, especially in discourse (Upadhye et al., 2020). The reason is likely in part that models are trained on resources which do not account for distributions in judgments. As a result, models do not have well-defined behavior when ambiguity arises and are poorly calibrated.

Training models with finer-grained coreference judgments could allow models to better align with human behavior. Techniques to improve model calibration could also be effective, allowing models to abstain or seek clarification when ambiguity arises. We hope that AMBI COREF can serve as a diagnostic set for future modeling approaches in evaluating their sensitivity to instances of ambiguity in language.

6 Limitations

Our study focuses entirely on coreference in the English language with models trained in high-resource settings. Furthermore, the cases of ambiguity we identify are English-specific and the names we insert into templates are popular American names. It is an open question as to how our results generalize to low-resource non-American-English settings.

The language we use to evaluate models is tautological. While we make an effort to account for unnatural data, by only evaluating templates models do well at, models struggle to completely solve all our unambiguous examples. This presents a challenge for future model builders. On the other hand, our templates may not reflect a particular real world distribution that models will be tested on.

Acknowledgements

We thank Chris Callison-Burch and the PennNLP group for their helpful comments on this work.

References


A Human Judgement Tests

In all our human judgement tests, we required annotators to be based primarily in English-speaking countries: the US, UK, Canada or Australia. Further, annotators needed to have at least 1000 approved HITs and a HIT acceptance rate of at least 98%. Each HIT contained 10 examples, and we estimated the completion time for each HIT to be $\sim 5$ minutes, so we paid $1.25$ per HIT, for a pay rate of $15$ per hour.

For our human judgement tests, we first ran a qualification round to ensure high-quality annotations. In this round, we asked annotators to complete a single HIT with 10 examples (5 unambiguous, 5 ambiguous randomly ordered). For each annotator who completed this round, we compute their accuracy by measuring how often they responded with the correct referent (or the ambiguous label), while ignoring their confidence. The top 100 annotators were qualified to work on the main task.

For our main task, we had 625 sentences labeled in total, with 3 assignments per sentence. Each annotator was asked to work on not more than 5 HITs, so that we get a diverse set of judgements. Similar to the qualification round, we asked each annotator to label the referent (or the ambiguous label) and their confidence. We group the annotations into 5 options: (Noun A, definitely), (Noun A, likely), Ambiguous, (Noun B, likely), and (Noun B, definitely). The human judgement labels for each template type were aggregated by computing the fraction of annotations in each of the five options. Our annotation interface for the main task is shown in Figure 3.
You have been qualified for the final task on coreference resolution.

Please work on at most 5 HITs for this task (ideally 3 to 4 HITs). We will not be able to compensate you for your work on more than 5 HITs.

You will be presented 10 sentences.
- Each sentence contains at least two nouns, and a nominative third-person singular pronoun (he/she).
- If three nouns appear in the sentence, please only consider the first two nouns as possible candidates.
- Following each sentence, there will be a question asking you which noun the specified pronoun refers to, or if you find the pronoun ambiguous.

For example, **Chloe told Emma that she was sad. Who was sad?**
- **Definitely Chloe** - I feel confident that “she” refers to “Chloe”;
- **Probably Chloe** - I think “she” refers to “Chloe” as opposed to “Emma” but I feel unsure;
- **Mostly ambiguous** - I find it completely ambiguous whether “she” refers to “Chloe” or “Emma”;
- **Probably Emma** - I think “she” refers to “Emma” as opposed to “Chloe” but I feel unsure;
- **Definitely Emma** - I feel confident that “she” refers to “Emma”;

Different people may have different judgments and tolerance for ambiguity, so please feel free to use your intuitive judgments.

---

Are you a native speaker of English?
- Yes
- No

1. **Matthew bought Joshua a pizza after he asked for more food.**
   - **Who asked for more food?**
     - Definitely Matthew
     - Probably Matthew
     - Mostly ambiguous
     - Probably Joshua
     - Definitely Joshua

2. **Kimberly told the aunt that she chatted with the granddaughter.**
   - **Who chatted with the granddaughter?**
     - Definitely Kimberly
     - Probably Kimberly
     - Mostly ambiguous
     - Probably the aunt
     - Definitely the aunt

---

Figure 3: Annotation interface for the human judgement tests, presented in section 2.