ezCoref: A Scalable Approach for Collecting Crowdsourced Annotations for Coreference Resolution

Anonymous ACL submission

Abstract

Large-scale high-quality corpora are critical for advancing research in coreference resolution. Coreference annotation is typically time-consuming and expensive, since researchers generally hire expert annotators and train them with an extensive set of guidelines. Crowdsourcing is a promising alternative, but coreference includes complex semantic phenomena difficult to explain to untrained crowdsworkers, and the clustering structure is difficult to manipulate in a user interface. To address these challenges, we develop and release ezCoref, an easy-to-use coreference annotation tool and annotation methodology that facilitates crowdsourced data collection across multiple domains, currently in English. Instead of teaching crowdsworkers how to handle non-trivial cases (e.g., near-identity coreferences), ezCoref provides only a minimal set of guidelines sufficient for understanding the basics of the task. To validate this decision, we deploy ezCoref on Mechanical Turk to re-annotate 240 passages from seven existing English coreference datasets across seven domains, achieving an average rate of 2530 tokens per hour, for one annotator. This paper is the first to compare the quality of crowdsourced coreference annotations against those of experts, and to identify where their behavior differs to facilitate future annotation efforts. We show that it is possible to collect coreference annotations of a reasonable quality in a fraction of time it would traditionally require.

1 Introduction

Coreference resolution is the task of identifying all textual expressions that refer to the same discourse entity in a given document, and thus grouping such coreferent expressions (mentions) into clusters (entities). Systems trained to solve this task are often an integral component of the preprocessing pipeline for many downstream tasks, such as summarization (Azzam et al., 1999; Steinberger et al., 2007), question answering (Vicedo and Fernández, 2000; Dhingra et al., 2018), and machine translation (Hardmeier, 2012; Bawden et al., 2018).

Modern coreference systems are implemented via data-hungry neural network models (e.g., Lee et al., 2017; Moosavi and Strube, 2018; Joshi et al., 2019) trained on large-scale expert-annotated datasets such as OntoNotes (Weischedel et al., 2013).

Acquiring these datasets has traditionally been difficult, expensive and time-consuming, requiring linguists trained in fine-grained annotation schemas (e.g., Hovy et al., 2006; Poesio and Artstein, 2008; Uryupina et al., 2019). As such, coreference datasets exist only for a small set of languages (mostly English) and even then only for limited domains (mainly news and fiction). Furthermore, these datasets differ widely in their annotation guidelines, resulting in inconsistent annotations across languages and domains, with challenging cross-lingual and cross-domain issues (Poesio et al., 2021).

Is it possible to use crowdsourcing and non-expert annotators to generate high-quality coreference data? As in other types of linguistic annotations, crowdsourcing could reduce costs (Snow et al., 2008), allowing for larger scale datasets. Furthermore, by using a standard platform like Amazon Mechanical Turk, this approach is accessible to a wider range of researchers, who may wish to collect new data for different corpora or even languages.
To this end, we develop ezCoref, a crowd-sourced coreference annotation platform that is intuitive and easy to use for crowdworkers, and open-source for other researchers to utilize. Unlike existing crowdsourced coreference efforts (Chamberlain et al., 2016; Bornstein et al., 2020; Li et al., 2020), ezCoref simplifies the annotation task for workers by using automatically detected mention boundaries, is easily integrable with platforms like Amazon Mechanical Turk, and offers a short, effective, crowd-oriented interactive tutorial.

With this new interface, we turn to the question of how to define coreference to untrained crowdworkers. Expertly-collected datasets such as OntoNotes explain what should and should not be considered coreference via a lengthy set of guidelines (Weischedel et al., 2012) that covers many complex linguistic details (e.g., how to deal with head-sharing noun phrases, which premodifiers can and cannot corefer, or how to annotate generic mentions). Even if it was feasible to teach such guidelines to crowdworkers, existing coreference datasets differ widely in terms of what is considered as a mention and what types of links should be annotated (Poesio et al., 2021) and Table 1), so it is unclear what standards should be used. Instead, we explore whether we can collect high-quality coreference data with a minimal set of guidelines, only illustrating basic phenomena like pronoun resolution (Table 3). We use ezCoref to re-annotate a subset of documents from seven different English coreference datasets. Our crowdworkers obtain high agreement with most of the expert annotations, and that the quality of our annotations is higher than that of previous crowdsourced efforts. We conclude with a qualitative analysis of our annotators’ behavior and the types of annotation decisions they make, which we hope will inform future research into coreference platform development and guideline construction.

### Conclusion

Table 1: Summary of seven datasets analyzed in this work, which differ in domain, size, annotator qualifications, mention detection procedures, types of mentions, and types of links considered as coreferences between these mentions. **Allows other types of mention only when this mention is an answer to a question.*** We interpret manual identification based on illustrations presented in the original publication (Chen et al., 2018b). ***See Footnote 8.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th>Annotators</th>
<th>Mention Detection</th>
<th>Mention Types</th>
<th>Coreference Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARRAU (Uryupina et al., 2019)</td>
<td>Multiple (552, 99K, 385K)</td>
<td>Single Expert</td>
<td>Manual</td>
<td>Yes</td>
<td>Special Link, No Link, Yes, Explicit</td>
</tr>
<tr>
<td>Phrase Detectives (PD) (Chamberlain et al., 2016)</td>
<td>Multiple (542, 100K, 400K)</td>
<td>Crowd (gamified) + 2 Experts</td>
<td>Semi Automatic</td>
<td>Yes</td>
<td>Special Link, Special Link, Yes, Implicit</td>
</tr>
<tr>
<td>GUM (Zelden, 2017)</td>
<td>Multiple (25, 6K, 20)</td>
<td>Experts (Linguistics Students)</td>
<td>Manual</td>
<td>Yes</td>
<td>Coref (Sub-Types), Coref (Sub-Types), Yes, None</td>
</tr>
<tr>
<td>PreCo (Chen et al., 2018a)</td>
<td>Multiple*** (30K, 3.58M, 12.5M)</td>
<td>Non-Expert, Non-Native</td>
<td>Manual***</td>
<td>Yes</td>
<td>Coref, Coref, Yes, None</td>
</tr>
<tr>
<td>OntoNotes (Hovy et al., 2006)</td>
<td>Multiple (1.6K, 94K, 950K)</td>
<td>Experts</td>
<td>Manual</td>
<td>Yes</td>
<td>Special Link, Special Link, Only with Pronouns, None</td>
</tr>
<tr>
<td>LitBank (Ramanan et al., 2020)</td>
<td>Single (100, 20K, 210K)</td>
<td>Experts</td>
<td>Manual</td>
<td>Yes</td>
<td>ACE (selected), Special Link, Only with Pronouns, None</td>
</tr>
<tr>
<td>QuizBowl (Guha et al., 2015)</td>
<td>Single (400-9.4K, 50K)</td>
<td>Domain Experts</td>
<td>Manual &amp; CRF***</td>
<td>Yes</td>
<td>Characters, Books, Authors*, Coref, Coref, If Applicable, None</td>
</tr>
</tbody>
</table>

ezCoref Pilot Dataset (this work) | Multiple | Crowd (paid) | Fully Automatic | Yes | None |

***See Footnote 8.

4Our platform’s code and collected data is available in supplementary materials, and will be released publicly after blind review.

5The syntax-based mention detector is our system’s only English-specific component.

6Our tutorial received overwhelmingly positive feedback. One annotator commented that it was “absolutely beautiful, intuitive, and helpful. Legitimately the best one I’ve ever seen in my 2 years on AMT! Awesome job.”

7Many others exist too; for example, see Jonathan Kummerfeld’s spreadsheet list (accessed Jan. 2022).
Table 2: A comparison of different coreference annotation tools. (* — ezCoref code will be open-sourced upon paper publication; Stenetorp et al. (2012) did not implement nested spans originally, but later added them with limited functionality. Yimam et al. (2013) have APIs for CrowdFlower integration, but suggest expert annotators.)

<table>
<thead>
<tr>
<th>System</th>
<th>Annotate all clusters</th>
<th>Pre-identified Mentions</th>
<th>Open Source</th>
<th>Webapp</th>
<th>Coref only</th>
<th>Keyboard and Mouse</th>
<th>MTurk Tested</th>
<th>Non-export Terminology</th>
<th>Nested Span Support</th>
<th>Interactive Tutorial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stenetorp et al. (2012)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Widlöcher and Mathet (2012)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Landragin et al. (2012)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Yimam et al. (2013)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Porro et al. (2013)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Girardi et al. (2014)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Kopec (2014)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Guha et al. (2015)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Oberle (2018)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Reiter (2018)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Borsenstein et al. (2020)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

ezCoref (this work)  ✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓

3 ezCoref: A Crowdsourced Coreference Annotation Platform

The ezCoref user experience consists of (1) a step-by-step interactive tutorial and (2) an annotation interface, which are part of a pipeline including automatic mention detection and Amazon Mechanical Turk integration.

Annotation structure: Two annotation approaches are prominent in the literature: (1) a local pairwise approach, annotators are shown a pair of mentions and asked whether they refer to the same entity (Hladká et al., 2009; Chamberlain et al., 2016; Li et al., 2020; Ravenscroft et al., 2021), which is time-consuming; or (2) a cluster-based approach (Reiter, 2018; Oberle, 2018; Borsenstein et al., 2020), in which annotators group all mentions of the same entity into a single cluster. In ezCoref we use the latter approach, which can be faster but requires the UI to support more complex actions for creating and editing cluster structures.

User interface: We spent two years iteratively designing, implementing, and user testing the interface in order to make it as simple and crowdsourcing-friendly as possible (Figure 1). Marked mentions are surrounded by color-coded frames with entity IDs. The currently selected mention (“the book”), is highlighted with a flashing yellow cursor-like box. The core annotation action is to select other mentions that corefer with the current mention, and then advance to a later unassigned mention; annotators can also re-assign a previously annotated mention to another cluster. Advanced users can exclusively use keyboard shortcuts, and undo and redo actions were added to allow error correction. Finally, ezCoref provides a side panel to show mentions of the entity currently being annotated, which helps to spot mentions assigned to the wrong cluster.

Coreference tutorial: To teach crowdworkers the basic definition of coreference and familiarize them with the interface, we develop a tutorial (aimed to take ~ 20 minutes) that familiarizes them with the mechanics of the annotation tool, then trains them in a minimal set of annotation guidelines (Table 3). The tutorial concludes with a quality control example to exclude poor quality financial compensation (Phrase Detectives; Chamberlain et al. (2016)).
Table 3: Phenomena explained in our tutorial.

<table>
<thead>
<tr>
<th>Phenomena Taught</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) personal pronouns</td>
<td>[John] doesn’t like [Fred], but [he] still invited [has/has] [the party].</td>
</tr>
<tr>
<td>(2) singulars</td>
<td>[This dog] likes to play [catch]. [It’s] better than other [dogs] at [this game].</td>
</tr>
<tr>
<td>(3) non-person entities (animals)</td>
<td>[I]s owner] is really proud.</td>
</tr>
<tr>
<td>(2) semantically similar expression which are not corefering</td>
<td>[This] dog has really similar [features] to [that] dog.</td>
</tr>
<tr>
<td>(3) non-person entities (time, item)</td>
<td>[This time] [he] often had to make [compromises].</td>
</tr>
<tr>
<td>(1) nested spans</td>
<td>[Director] [MacKenzie] spent [last two years] working on a [&quot;Young Adam&quot;]. During [this time] [he] often had to make [compromises].</td>
</tr>
<tr>
<td>(2) non-person entities (time, item)</td>
<td>[The office] wasn’t exactly small either.</td>
</tr>
<tr>
<td>(1) non-person entities (place)</td>
<td>[It]s owner] is really proud.</td>
</tr>
</tbody>
</table>

**Annotation workflow:** The annotators are presented with one passage (or “document”) at a time (Figure 1), and all mentions have to be annotated before proceeding to the next passage. There is no limitation to the length or language of the passage.

In this work, we divide an initial document into a sequence of shorter passages of complete sentences, on average 175 tokens, since shorter passages minimize the need to scroll, reducing annotator effort. While this obviously cannot capture longer distance coreference, a large portion of important coreference phenomena is local: within the OntoNotes written genres, for pronominal mentions, the closest antecedent is contained within the current or previous two sentences more than 95% of the time.

**Automatic mention detection:** To simplify the annotation task for crowdsworkers, we decide to automatically annotate mentions instead of forcing workers to mark mention boundaries in the text. This approach considerably reduces the annotator’s effort while speeding annotation; however, it relies heavily on the performance of the mention detection algorithm, which also can detect non-referring expressions (e.g., “hand” in “on the other hand”).

Table 1 shows that existing datasets use many methods and criteria to identify mentions, from completely manual to semi-automatic and fully automatic procedures. Instead of choosing an existing standard, we implement a simple automatic mention detection algorithm that yields a high average recall over all seven of these datasets. We consider all noun phrases (including proper nouns, common nouns, and pronouns) as markables, extracting them using the Stanza dependency parser.

### 4 Using ezCoref to Re-annotate Existing Coreference Datasets

To study annotator behavior in our setup, we deploy ezCoref on the Amazon Mechanical Turk (AMT) crowdsourcing platform to re-annotate 240 passages from seven existing datasets, covering seven unique domains. We compare our workers’ annotations to each other across domains, compare them to the previous gold standard annotations, and conduct our own qualitative analysis as well. In total, we collect annotations for 12,200 mentions and 42,108 tokens.

#### 4.1 Experimental Setup

**Datasets:** We collect coreference annotations for the seven existing datasets described in Table 1: OntoNotes (Hovy et al., 2006), LitBank (Bamman et al., 2020), PreCo (Chen et al., 2018a), ARRAU (Uryupina et al., 2019), GUM (Zeldes, 2017), Phrase Detectives (Chamberlain et al., 2016), and QuizBowl (Guha et al., 2015). The sample covers seven domains: news, opinionated magazines, weblogs, fiction, biographies, Wikipedia articles, and trivia questions from Quiz Bowl. For each dataset with multiple domains, we manually se-

---

5 Examples of the tutorial interface and the quality control example are provided in Appendix A.6.

6 We leave this for future work—for example, more sophisticated user interfaces to support longer documents, or merging coreference chains between short passages. As documents get progressively longer, such as book chapters or books, the task takes on aspects of cross-document coreference and entity linking (e.g. Bagga and Baldwin, 1998b; Fitzgerald et al., 2021; Logan IV et al., 2021).

7 The PreCo dataset is interestingly large, but seems difficult to access. In November 2018 and October 2021 we filled out the data request form at the URL provided by the paper, and attempted to contact the PreCo official email directly, but did not receive a response. To enable a precise research comparison, we scraped all documents from PreCo’s public demo in November 2018 (no longer available as of 2021); its statistics match their paper and our experiments use this version of the data. PreCo further suffers from data curation issues (Gebru et al., 2018; Jo and Gebru, 2020); it uses text from English reading comprehension tests collected from several websites, but the original document sources and copyright statuses are undocumented. When reading through PreCo documents, we found many domains including opinion, fiction, biographies, and news (Appendix A); we use our manual categories for domain analysis.
to gold mentions across seven datasets. Next, we measure the quality of our annotations (via inter-annotator agreement between our crowdworkers) and their agreement with other datasets. Finally, we discuss interesting qualitative results.

5.1 Mention Detector Evaluation

Datasets differ in the way they define their mentions’ boundaries. Hence, the boundaries for the same mention may differ. To fairly compare our mentions with the gold standards, we employ a headword-based comparison. In particular, we find the head of the given phrase by identifying, in the dependency tree, the most-shared ancestor of all tokens within the given mention. We consider two mentions as the same if their respective headwords match.

Table 4 compares our mention detector to the gold mentions in existing datasets. Our method obtains high recall across most datasets (>0.90). It has the lowest recall with ARRAU (0.84) and PreCo (0.88), which is to be expected as ARRAU marks all referring premodifiers (identified manually) and PreCo allows common noun modifiers, while we identify only the premodifiers which are proper nouns. Comparing precision, we observe a substantially lower score for OntoNotes, LitBank, and QuizBowl as these datasets restrict their mention types to limited entities (refer to Table 1). As a result, our algorithm identifies more mentions than in the original datasets, which also allows us to discover new entities. For the remaining datasets, the precision is >0.80, suggesting that the algorithm identifies most of the relevant mentions.

Finally, we compare mention density (number of mentions per token) between our detector and existing datasets, and find that while gold mention density varies considerably across the seven datasets due to their differing mention criteria (Table 1), it

8We allow only workers with a >= 99% approval rate and at least 10,000 approved tasks who are from the US, Canada, Australia, New Zealand, or the UK.

9We did not collect demographic data for the remaining eight individuals, from an earlier pilot experiment.

10We made this decision as identifying automatically all premodifiers would result in many singletons and lead to more arduous annotation effort.
Table 4: Comparison of mentions identified by our mention detection algorithm with the gold mentions annotated in the respective datasets. We use head-word based comparison to compare mentions of different lengths.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Recall</th>
<th>Precision</th>
<th>Mentions / Tokens</th>
<th>Gold</th>
<th>This Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>OntoNotes</td>
<td>0.957</td>
<td>0.376</td>
<td>0.112</td>
<td>0.286</td>
<td></td>
</tr>
<tr>
<td>LitBank</td>
<td>0.962</td>
<td>0.415</td>
<td>0.121</td>
<td>0.280</td>
<td></td>
</tr>
<tr>
<td>QuizBowl</td>
<td>0.956</td>
<td>0.543</td>
<td>0.188</td>
<td>0.318</td>
<td></td>
</tr>
<tr>
<td>PD (Gold)</td>
<td>0.953</td>
<td>0.803</td>
<td>0.259</td>
<td>0.273</td>
<td></td>
</tr>
<tr>
<td>PD (Silver)</td>
<td>0.938</td>
<td>0.791</td>
<td>0.265</td>
<td>0.274</td>
<td></td>
</tr>
<tr>
<td>GUM</td>
<td>0.906</td>
<td>0.848</td>
<td>0.269</td>
<td>0.287</td>
<td></td>
</tr>
<tr>
<td>PreCo</td>
<td>0.881</td>
<td>0.883</td>
<td>0.287</td>
<td>0.287</td>
<td></td>
</tr>
<tr>
<td>ARRAU</td>
<td>0.840</td>
<td>0.870</td>
<td>0.289</td>
<td>0.279</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Comparison of mentions identified by our mention detection algorithm with the gold mentions annotated in the respective datasets. We use head-word based comparison to compare mentions of different lengths.

5.2 What domains are most suitable for crowdsourcing coreference?

Which domains yield the highest inter-annotator agreement (IAA) between our crowdworkers?

We use the B3 metric\(^{13}\) (Bagga and Baldwin, 1998a) to compute IAA for each domain, excluding singletons\(^{14}\) (see Table 6). We obtain the highest agreement on fiction (72.6\%) and biographies (72.4\%). This is because both domains contain a high frequency of pronouns (see examples \(a\) and \(b\) in Table 5), which our annotators found easier to annotate. We also observe that the fiction domain contains many well-known children stories (e.g., Little Red Riding Hood) that are likely familiar to our annotators, which may have made them easier to annotate. Annotators have the least agreement on Quiz Bowl coreference (59.73\%), as this dataset is rich in challenging cataphoras (example \(c\) in Table 5) and often require world knowledge about books, characters, and authors to identify coreferences (example \(e\) in Table 5).

5.3 Agreement with Existing Datasets

Having established relatively good agreement amongst our workers on most domains, we now turn to a different question: how well do crowdsourced annotations from ezCoref agree with gold annotations from existing datasets?

Aggregating annotations: To compare crowdsourced annotations with gold annotations, we first require an aggregation method that can combine annotations from multiple crowdworkers to infer coreference clusters. We use a simple aggregation method that determines whether a pair of mentions is coreferent by counting the number of annotators who marked the two mentions in the same cluster. Two mentions are considered as coreferent when the number of annotators linking them together is greater than a threshold (\(\tau\)). After inferring these pairs of mentions, we construct an undirected graph where nodes are mentions and edges represent coreference links. Finally, we find connected components in the graph to obtain coreference clusters.\(^{15}\) After aggregating our ezCoref annotations, we compare these annotations with gold annotations across the seven datasets using B3 scores (precision, recall, and F1), as illustrated in Figure 2.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Recall</th>
<th>Precision</th>
<th>Mentions / Tokens</th>
<th>Gold</th>
<th>This Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>QuizBowl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LitBank</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARRAU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

High agreement with OntoNotes, GUM, LitBank, ARRAU: Our annotators achieve the highest precision with OntoNotes, suggesting that most of the entities identified by crowdworkers are correct for this dataset. In terms of F1 scores, the datasets which are closest to crowd annotations are GUM, LitBank, and ARRAU, all of which are annotated by experts. This result shows that ezCoref facilitates high quality annotations from untrained crowdworkers.

Low precision with Phrase Detectives and PreCo, low recall with Quiz Bowl: We observe that Phrase Detectives has a very low precision compared to all other datasets, implying that crowdworkers add more links compared to gold annotations. Our qualitative analysis reveals that PD annotators miss some valid links, splitting entities which are correctly linked together by our annotators (see Table 7). Another dataset with lower precision is PreCo, which also contains many missing links. In general, we observe more actual mistakes in PreCo and PD than in the other datasets, which is not surprising as they were not annotated by experts.\(^{16}\) This result is further validated by our agreement analysis of the fiction domain (Table 8), in which ezCoref annotations agree far more closely with

\(^{13}\)This method resolves to majority voting-based aggregation when the \(\tau\) is set so that more than half of annotators should agree. For \(\tau = N\), this method is very conservative, adding a link between two mentions only when all annotators agree unanimously. Conversely, for \(\tau = 1\), only a single vote is required to add a link between two mentions.

\(^{14}\)The agreement including singletons is substantially higher. The exact numbers are provided in Appendix A.3.
Table 5: Examples taken from respective datasets to illustrate their unique phenomena. Coreference mentions are marked with same color in each example.

Table 6: Inter Annotator Agreement (B3 %) across different domains. B3 scores are computed in accordance with the CoNLL script (Pradhan et al., 2014), excluding singletons. Bio (Biographies); Wiki (Wikipedia).

Table 7: Cases of split entities (missing links) in annotations provided with Phrase Detectives and PreCo datasets. Instead, our crowd annotators mark all mentions as referring to the same entity in each of these examples.

**Varying the aggregation threshold \( \tau \):** What is the effect of varying the aggregation threshold \( \tau \) on precision and recall with gold annotations? Figure 3 shows that the Quiz Bowl dataset has the highest drop in recall (36% absolute drop) when increasing \( \tau \) from 1 to 5.\(^{17}\) This indicates that the number of unanimous clusters \( \tau = 5 \) is considerably lower than the total number of clusters found individually by all annotators \( \tau = 1 \); as such, our annotators heavily disagree about gold clusters in the Quiz Bowl dataset. We observe a similar trend in other datasets.

\(^{17}\)We analyze variations in recall since it is more interpretable than precision, given that the denominator is fixed in recall with a variable number of annotators.
Table 9: Number of mistakes in our crowd annotations vs. gold datasets, obtained through a manual analysis.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Mistakes (our)</th>
<th>Mistakes (gold)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD (silver)</td>
<td>22</td>
<td>76</td>
</tr>
<tr>
<td>PreCo</td>
<td>12</td>
<td>33</td>
</tr>
<tr>
<td>GUM</td>
<td>48</td>
<td>25</td>
</tr>
<tr>
<td>OntoNotes</td>
<td>81</td>
<td>49</td>
</tr>
<tr>
<td>ARRAU</td>
<td>33</td>
<td>16</td>
</tr>
<tr>
<td>LitBank</td>
<td>21</td>
<td>13</td>
</tr>
<tr>
<td>QuizBowl</td>
<td>67</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 10: Examples of genuine ambiguity and generic “you” observed in our data.

```
[Fog] everywhere. [Fog] up [the river], where [it] flows among green
and meadows; [fog] down [the river], where [it] rolls defiled among
the tiers of shipping and the waterside pollutions of a great
(and dirty) city.
- Charles Dickens, Bleak House
```

```
Please, Ma’am, is this New Zealand or Australia? (and she tried to
curtsey as she spoke – fancy CURTSEYING as [you] ‘re falling
through the air! Do [you] think [you] could manage it?)
- Lewis Carroll, Alice in Wonderland
```

6 Conclusion

In this work, we present ezCoref, an intuitive and easy-to-use annotation tool to collect crowdsourced annotations for coreference resolution. Using ezCoref, we re-annotate a subset of documents from seven different English coreference datasets, each of which was created using a different set of complex linguistic guidelines. In contrast, our ezCoref re-annotation aims to collect high-quality coreference data with a minimal set of guidelines. Our results show that crowdworkers obtain high agreement with many expert annotations (e.g., GUM, ARRAU) and that our annotation quality is better than previous crowdsourced efforts (e.g., Phrase Detectives). We hope our ezCoref tool and observations will inform future research into coreference platform development and guideline construction.

7 Ethics Statement

The data collection protocol was approved by the coauthors’ institutional review board. All annotators were presented with a consent form prior to the annotation. They were also informed that only satisfactory performance on the screening example will allow them to take part in the annotation task. All data collected during the tutorial and annotations (including annotators’ feedback and demograph-
ics) will be released anonymized. We also ensure that the annotators receive at least $13.50 per hour. Since base compensation is per unit of work, not by time (the standard practice on Amazon Mechanical Turk), we add bonuses for workers whose speed caused them to fall below that hourly rate.

References


A Appendix

A.1 Details of our crowdsourced data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th>#Docs</th>
<th>#Passages</th>
<th>#Tokens</th>
<th>#Mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>OntoNotes</td>
<td>News</td>
<td>6</td>
<td>30</td>
<td>4925</td>
<td>1365</td>
</tr>
<tr>
<td></td>
<td>Opinion</td>
<td>12</td>
<td>20</td>
<td>3861</td>
<td>1157</td>
</tr>
<tr>
<td>LitBank</td>
<td>Fiction</td>
<td>4</td>
<td>30</td>
<td>5455</td>
<td>1494</td>
</tr>
<tr>
<td>QuizBowl</td>
<td>Quizzes</td>
<td>20</td>
<td>20</td>
<td>3304</td>
<td>1083</td>
</tr>
<tr>
<td>ARRAU</td>
<td>News</td>
<td>3</td>
<td>20</td>
<td>3336</td>
<td>885</td>
</tr>
<tr>
<td>GUM</td>
<td>Biographies</td>
<td>4</td>
<td>20</td>
<td>3422</td>
<td>1119</td>
</tr>
<tr>
<td></td>
<td>Fiction</td>
<td>4</td>
<td>20</td>
<td>3299</td>
<td>1008</td>
</tr>
<tr>
<td>Phrase</td>
<td>Wikipedia</td>
<td>7</td>
<td>20</td>
<td>3509</td>
<td>1003</td>
</tr>
<tr>
<td>Detectives</td>
<td>Fiction</td>
<td>4</td>
<td>20</td>
<td>4087</td>
<td>1063</td>
</tr>
<tr>
<td></td>
<td>Opinion</td>
<td>7</td>
<td>9</td>
<td>1692</td>
<td>495</td>
</tr>
<tr>
<td>ProCo</td>
<td>News</td>
<td>4</td>
<td>8</td>
<td>1318</td>
<td>369</td>
</tr>
<tr>
<td></td>
<td>Fiction</td>
<td>2</td>
<td>2</td>
<td>278</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>Biographies</td>
<td>1</td>
<td>1</td>
<td>152</td>
<td>53</td>
</tr>
<tr>
<td>Total</td>
<td>All</td>
<td>83</td>
<td>240</td>
<td>42108</td>
<td>12200</td>
</tr>
</tbody>
</table>

A.2 Detailed Mention Detection Algorithm

- We identify all noun phrases using the Stanza dependency parser (Qi et al., 2020). For each word with a noun-related part-of-speech tag, we recursively traverse all of its children in the dependency graph until a dependency relation is found in a whitelist. The maximal span considered as a candidate mention thus covers all words related by relations in the whitelist.

---

19 pronouns, nouns, proper nouns, and numbers.  
20 The whitelist includes all multi-word expression relations (i.e., compound, flat, and fixed) and modifier relations (i.e., determiners, adjectival modifiers, numeric modifiers, nominal modifiers, and possessive nominal modifiers).
Figure 4: Inter Annotator Agreement across different domains. B3 scores with Singletons included.

- Possessive nominal modifiers are also considered as candidate mentions. For instance, in the sentence “Mary’s book is on the table,” we consider both “Mary” and “Mary’s book” as mentions.

- Modifiers that are proper nouns in a multi-word expression are considered as mentions. For instance, in “U.S. foreign policy,” the modifier “U.S.” is also considered as a mention.

- All conjuncts, including the headword and other words depending on it via the conjunct relation, are considered mentions in a coordinated noun phrase. For instance, in the sentence, “John, Bob, and Mary went to the party:”, the detected mentions are “John,” “Bob,” “Mary,” and the coordinated noun phrase “John, Bob, and Mary.”

- Finally, we remove mentions if a larger mention with the same headword exists. We allow nested spans (e.g., [[my] hands]) but merge any intersecting spans into one large span (e.g, [western [Canadian] province] is merged into [western Canadian province]).

A.3 Inter-Annotator Agreement Among Our Annotators Across Domains

Figure 4 illustrates agreement among our annotators computed with B3 scores including singletons.

A.4 An illustrative example

An example of a single sentence annotated by two datasets, OntoNotes and ARRAU. These annotations differ widely from each other in kinds of mentions and links between mentions.

OntoNotes: Lloyd’s, once a pillar of the world insurance market is being shaken to its very foundation.

ARRAU: Lloyd’s, once a pillar of the world insurance market is being shaken to its very foundation.

A.5 Consent

Before participating in our study, we requested every annotator to provide their consent. The annotators were informed about the purpose of this research study, any risks associated with it, and the qualifications necessary to participate. The consent form also elaborated on task details describing what they will be asked to do and how long it will take. The participants were informed that they could choose as many documents as they would like to annotate (by accepting new Human Intelligence Tasks at AMT) subject to availability, and they may drop out at any time. Annotators were informed that they would be compensated in the standard manner through the Amazon Mechanical Turk crowdsourcing platform, with the amount specified in the Amazon Mechanical Turk interface. As part of this study, we also collected demographic information, including their age, gender, native language, education level, and proficiency in the English language. We ensured our annotators that the collected personal information would remain confidential in the consent form.

A.6 Details of Tutorial

12
Coreference Annotation Tutorial

Welcome!

This is a paid tutorial for the "Large-Scale Coreference Annotation Task."

In this tutorial you will learn how to annotate coreferences, that is, words and phrases that refer to the same person or things.

Upon completing the tutorial, you will get a completion code. You MUST enter this code in the textbox below and submit the HIT in order to receive the payment. Depending on your performance, you might be invited to participate in our "Large-Scale Coreference Annotation Task."

Before proceeding to the tutorial, please fill in the following survey:

What is your gender?

________________________

What is your age?

________________________

What is your native language?

________________________

How is your English level?

○ Native speaker
○ Advanced (near native)
○ Intermediate
○ Beginner
○ Absolute Beginner

What is your education level?

○ Primary
○ Secondary
○ College (No Degree)
○ Bachelor's
○ Masters
○ Ph.D. or higher

Click this link to begin.

[OPTIONAL] We would love to hear your feedback about this tutorial. All participants who provide meaningful suggestions will receive a bonus of $0.5.

Submit your code below:

Submit

Figure 5: Screenshot of tutorial task invitation on AMT with detailed instructions.
Coreference Tutorial Mode

Welcome to the coreference tutorial mode. Here you will learn how to use the interface efficiently to label text for coreferences.

What are coreferences?
A coreference is when two words or spans (sequence of words) refer to the same thing.

In the examples below, the following words are coreferences (they refer to the same “thing”):
(1) "John" and "He"
(2) "Robert" and "He"
(3) "Alice" and "Her"

John is cool. He is nice.
Robert loves Alice. He talks to her everyday.

Let’s get started.

Figure 6: Tutorial Interface (Introductory prompt)

Select Spans (Task 1 of 10)

Step 1 of 2

Observe how the border around “Mary” is flashing. This means the span “Mary” is the current target. Click on all the spans that refer to the target “Mary.”

Mary is fun. She jokes a lot. That’s why Mark likes her.

Figure 7: Tutorial interface: A sample prompt teaching tool functionality.
Annotation Examples Task

Remember:
- If the current target does NOT have any coreferences go to the next target.
- You should annotate all the spans that refer to the current target before moving onto the next target.
- Once you have finished annotating the current passage, click on the Continue button to move on to the next passage.

![Tutorial interface: A sample prompt teaching basic coreferences.](image)

Figure 8: Tutorial interface: A sample prompt teaching basic coreferences.

Remember:
- If the current target does NOT have any coreferences go to the next target.
- You should annotate all the spans that refer to the current target before moving onto the next target.
- Once you have finished annotating the current passage, click on the Continue button to move on to the next passage.

![Tutorial interface: quality control example.](image)

Figure 9: Tutorial interface: quality control example.
Latest Coreference Annotation Task

Welcome to the coreference annotation task. In this task you will be asked to annotate a short paragraph for coreferences. If you need to review the tutorial, please follow this link.

What are coreferences?

A coreference is when two words or spans (sequence of words) refer to the same thing.

In the examples below, the following words are coreferences (they refer to the same "thing"): 

1. "John" and "He"
2. "Robert" and "He"
3. "Alice" and "Her"

John is cool. He is nice.
Robert loves Alice. He talks to her everyday.

Click this link to begin annotation

[OPTIONAL] We would love to hear your feedback. Let us know if anything was unclear or particularly challenging.

Submit your code below:

Submit