Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 74–79
Marseille, 11–16 May 2020

Model-based annotation of coreference

Rahul Aralikatte and **Anders Søgaard**
University of Copenhagen
{rahul, soegaard}@di.ku.dk

Abstract

Humans do not make inferences over texts, but over *models* of what texts are about. When annotators are asked to annotate coreferent spans of text, it is therefore a somewhat unnatural task. This paper presents an alternative in which we preprocess documents, linking entities to a knowledge base, and turn the coreference annotation task – in our case limited to pronouns – into an annotation task where annotators are asked to assign pronouns to entities. Model-based annotation is shown to lead to faster annotation and higher inter-annotator agreement, and we argue that it also opens up for an alternative approach to coreference resolution. We present two new coreference benchmark datasets, for English Wikipedia and English teacher-student dialogues, and evaluate state-of-the-art coreference resolvers on them.

Keywords: Coreference resolution, Linguistic mental models

1. Introduction

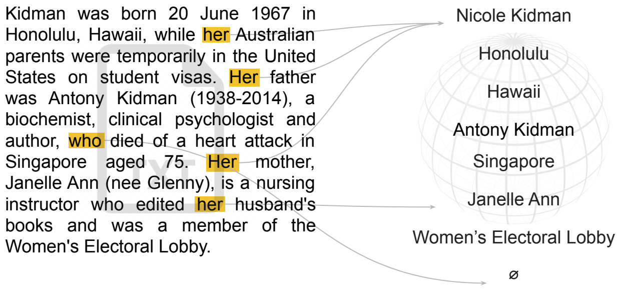
Language comprehension is often seen as the incremental update of a mental model of the situation described in the text (Bower and Morrow, 1990). The model is incrementally updated to represent the contents of the linguistic input processed so far, word-by-word or sentence-by-sentence. In this paper, we restrict ourselves to one central feature shared by most theories of mental models: they include a list of entities previously introduced in the text. This corresponds to the *constants* of first-order models or the referents associated with different *roles* in frame semantics. By models we thus simply mean a set of entities. Obviously, this is not sufficient to represent the meaning of texts, but focusing exclusively on annotating nominal coreference, we can ignore relations and predicates for this work. We will use the term *model-based annotation* to refer to linguistic annotation using model representations to bias or ease the work of the annotators.

Mental models have previously been discussed in linguistics literature on coreference (Runner et al., 2003). The motivation has often been that some pronouns refer to entities that are not explicitly mentioned in the previous text, but are supposedly available in the reader's mental model of the text, by inference. Consider, for example:

(1) I knocked on the door of room 624. *He* wasn't in.

The introduction of the referent of *he* in (1) is implied by the introduction of the entity *room 624*. In this paper, we present a new approach to annotating coreference that enables simple annotation of examples such as (1): Instead of asking an annotator to relate pronouns and previous spans of text, we ask the annotator to link pronouns and entities in document models. Moreover, we argue that model-based annotation reduces the cognitive load of annotators, which we experimentally test by comparing inter-annotator agreement and annotator efficiency across comparable annotation experiments. Fig. 1 showcases a concrete example from the collected dataset.

Contributions This paper makes a technical contribution, a conceptual contribution, and introduces a novel corpus annotated with coreference to the NLP community: (a) The technical contribution is a novel annotation methodology, where annotation is mediated through a model representation. We believe similar techniques can be developed for other NLP tasks; see §6 for discussion. (b) The conceptual contribution is a discussion of the importance of mental models in human language processing, and an argument for explicitly representing this level of representation in NLP models. (c) Our corpus consists of manually annotated sentences from English Wikipedia and QuAC (Choi et al., 2018). In addition to the model-based annotations, we also provide the coreference links obtained in our baseline experiments.



Kidman was born 20 June 1967 in Honolulu, Hawaii, while her Australian parents were temporarily in the United States on student visas. Her father was Antony Kidman (1938-2014), a biochemist, clinical psychologist and author, who died of a heart attack in Singapore aged 75. Her mother, Janelle Ann (nee Glenny), is a nursing instructor who edited her husband's books and was a member of the Women's Electoral Lobby.

Figure 1: Example of an annotation from the dataset.

2. Related Work

2.1. Annotation interfaces

The idea of easing the cognitive load of annotators by changing the way data is represented, is at the core of many papers on annotation interfaces. Early tools like MMAX2 (Müller and Strube, 2006) provide a clean user interface for annotators by highlighting mentions and connecting entity chains to visualize coreference along with helpful features like inter-annotator agreement checker, corpus querying, etc. Newer tools like WebAnno (Yimam et al., 2013; Day et al., 2004) ease the process of annotation by having support for flexible multi-layer annotations on a single document and also provide project management utilities. APLenty (Nghiem and Ananiadou, 2018) provides automatic annotations for easing annotator load and also has

an active learning component which makes the automatic annotations more accurate over time.
For relieving annotator load, these tools form clusters of coreference such that the annotator can choose to link a mention to one of these clusters. But this is possible only after the clusters are well-formed i.e. after some amount of annotation has taken place. One advantage of our approach is that we provide representatives for each cluster (the entities in the document) right from the start of the annotation process.

2.2. Mental models in NLP

Culotta et al. (2007) present a probabilistic first-order logic approach to coreference resolution that implicitly relies on mental models. Peng et al. (2015) focus on hard Winograd-style coreference problems and formulate coreference resolution as an Integer Linear Programming (ILP) to reason about likely models. Finkel and Manning (2008) also explore simple ILPs over simple first-order models for improving coreference resolution. They obtain improvements by focusing on enforcing transitivity of coreference links. In general, the use of first order models has a long history in NLP, rooted in formal semantics, going back to Fregean semantics. Blackburn and Bos (2005), for example, present a comprehensive framework for solving NLP problems by building up first order models of discourses.

2.3. Coreference datasets

The main resource for English coreference resolution, also used in the CoNLL 2012 Shared Task, is OntoNotes (Pradhan et al., 2012). OntoNotes consists of data from multiple domains, ranging from newswire to broadcast conversations, and also contains annotations for Arabic and Chinese. WikiCoref (Ghaddar and Langlais, 2016) is a smaller resource with annotated sentences sampled from English Wikipedia. Our dataset includes paragraphs from all pages annotated in WikiCoref, for comparability with this annotation project. See §5 for discussion. These are the datasets used below, but alternatives exist: GAP (Webster et al., 2018) is another evaluation benchmark, also sampled from Wikipedia and focuses on addressing gender bias in coreference systems. Phrase Detectives (Poesio et al., 2013) gamifies the creation of anaphoric resources for Wikipedia pages, fiction and art history texts. Cohen et al. (2017) annotate journal articles to create the CRAFT dataset which has structural, coreference and concept annotations. The annotation process of this dataset is similar in spirit to ours as their concept annotations link text mentions to curated ontologies of concepts and entities.

3. Data collection

We collect 200 documents[1] from two sources: (i) the summary paragraphs of 100 English Wikipedia documents (30 titles from WikiCoref and 70 chosen randomly), and (ii) the first 100 datapoints from the Question-Answering in Context (QuAC) dataset. Every QuAC document contains a Wikipedia paragraph and QA pairs created by two annotators posing as a student asking questions and a teacher answering the questions by providing short excerpts from the text. Thus the domain of all the documents is English Wikipedia.

3.1. Design Decisions

Some Wikipedia articles have short summaries with very few pronouns and some do not have summaries at all. Therefore, for each document chosen randomly, we first verify if it has a summary that contains at least five pronouns. If it does not, we choose another document and repeat this process till we get the required number of documents. We then extract all the entities from every document by parsing URL links present in the document which link to other Wikipedia pages or Wikidata entities. For QuAC documents, where all links are scrubbed, we parse their original Wikipedia pages to get the entities. Lastly we remove all markups, references and lists from the documents.
We collect a comprehensive list of English pronouns for linking. Some pronouns by their definition, almost never refer to entities. For example, (i) interrogative pronouns: 'what', 'which', etc., (ii) relative pronouns: 'as', 'who', etc., and (iii) indefinite pronouns: 'anyone', 'many', etc. For completeness, we do not remove these words from the list. We however allow the annotators to mark them specifically as *No Reference*.

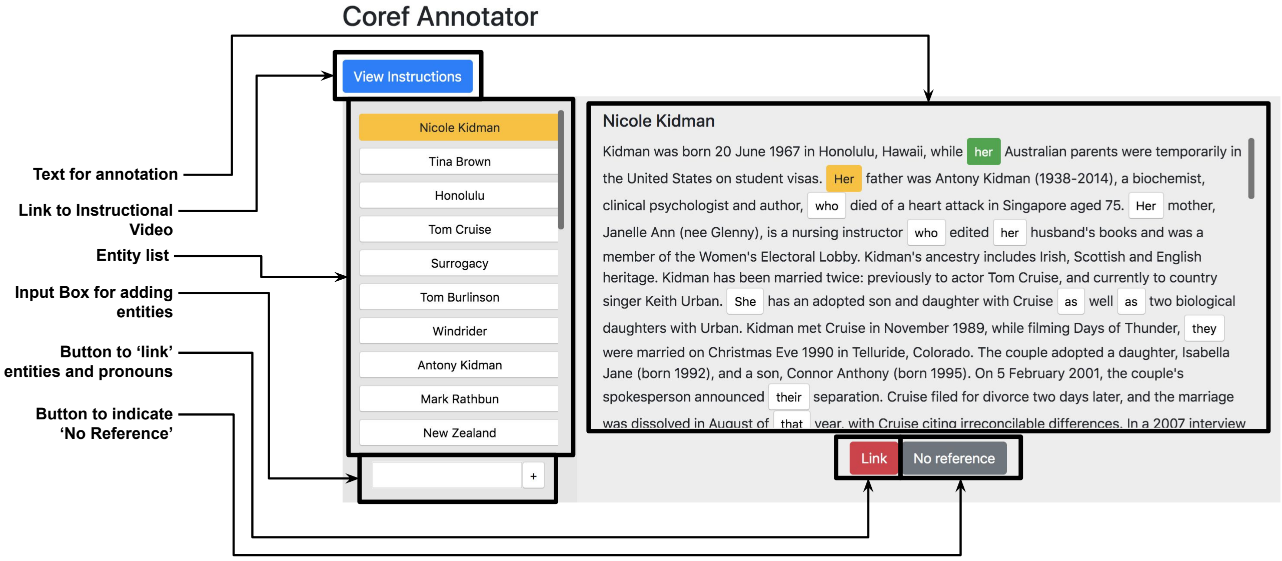
3.2. Annotation

To test our hypothesis that model-based coreference annotations are faster to create and more coherent, we pose two tasks on Amazon Mechanical Turk (AMT): (i) *Grounded task*: where all the parsed entities from a document are displayed to the annotator for linking with the pronouns, (ii) *Span annotation task*: where the entities are not shown and the annotator is free to choose any span as the antecedent. 30 documents from each source are doubly annotated to compute the inter-annotator agreement and the other 70 were singly annotated.
An annotation tool with two interfaces is built, one for each task, with slight differences between them as shown in Figures 2 and 3 respectively. The tool takes in a pre-defined list of mentions (pronouns in our case) which are markable. The annotators can link only these words with coreferent entities. This reduces the cognitive load on the annotators. The annotation process for the two tasks is briefly described below.

3.2.1. Grounded task

For this task, the interface (Fig. 2) is split into two parts. A larger part on the right contains the document text and the mention pronouns are highlighted in white. A sidebar on the left is populated with all the entities extracted from the document. In case of missing entities, the annotator has the option of adding one using the input box present at the bottom-left of the screen. The annotators are asked to link the mention pronouns in the document with one or more entities by: (i) clicking on a mention, (ii) clicking on one or more entities; , and (iii) clicking on the red *Link* button. If any mention does not have an antecedent, the annotators are asked to mark them with the grey *No reference* button. The color of the currently selected mention and entities are

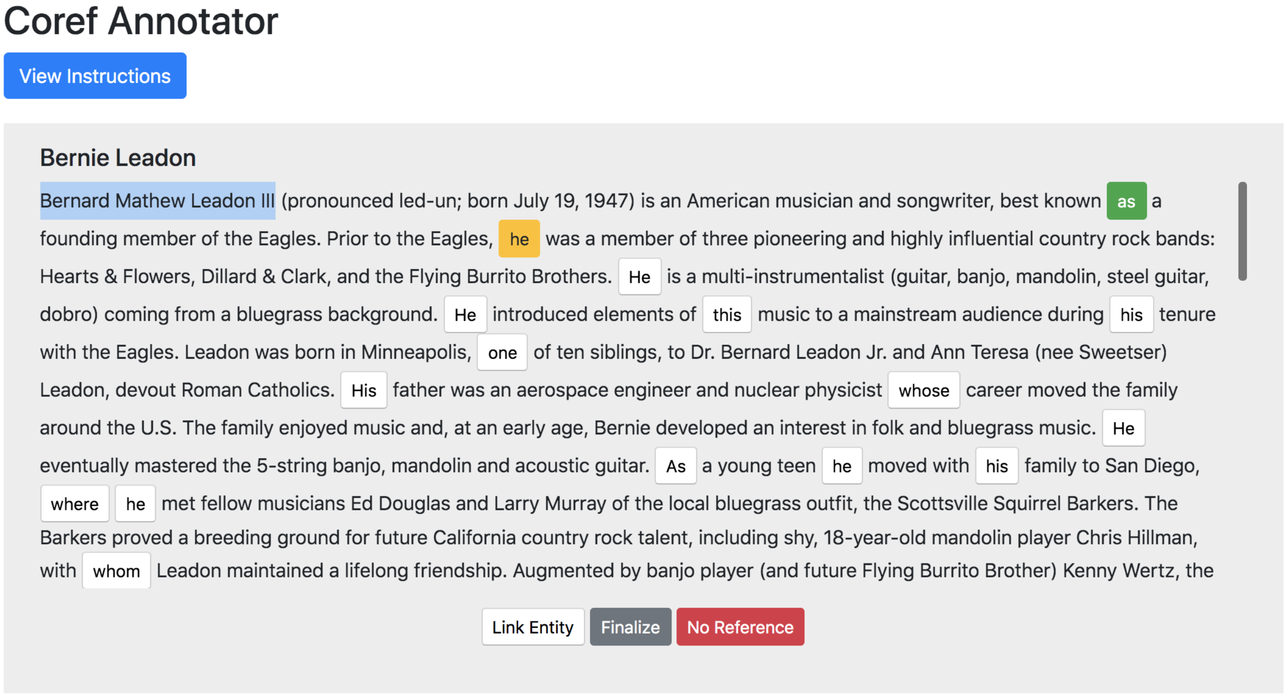
[1] We use the term *document* to denote a datapoint in our dataset.



Nicole Kidman

Kidman was born 20 June 1967 in Honolulu, Hawaii, while her Australian parents were temporarily in the United States on student visas. Her father was Antony Kidman (1938-2014), a biochemist, clinical psychologist and author, who died of a heart attack in Singapore aged 75. Her mother, Janelle Ann (nee Glenny), is a nursing instructor who edited her husband's books and was a member of the Women's Electoral Lobby. Kidman's ancestry includes Irish, Scottish and English heritage. Kidman has been married twice: previously to actor Tom Cruise, and currently to country singer Keith Urban. She has an adopted son and daughter with Cruise as well as two biological daughters with Urban. Kidman met Cruise in November 1989, while filming Days of Thunder, they were married on Christmas Eve 1990 in Telluride, Colorado. The couple adopted a daughter, Isabella Jane (born 1992), and a son, Connor Anthony (born 1995). On 5 February 2001, the couple's spokesperson announced their separation. Cruise filed for divorce two days later, and the marriage was dissolved in August of that year, with Cruise citing irreconcilable differences. In a 2007 interview

Figure 2: Screen grab of the interface for the grounded-annotation task



Bernie Leadon

Bernard Mathew Leadon III (pronounced led-un; born July 19, 1947) is an American musician and songwriter, best known as a founding member of the Eagles. Prior to the Eagles, he was a member of three pioneering and highly influential country rock bands: Hearts & Flowers, Dillard & Clark, and the Flying Burrito Brothers. He is a multi-instrumentalist (guitar, banjo, mandolin, steel guitar, dobro) coming from a bluegrass background. He introduced elements of this music to a mainstream audience during his tenure with the Eagles. Leadon was born in Minneapolis, one of ten siblings, to Dr. Bernard Leadon Jr. and Ann Teresa (nee Sweetser) Leadon, devout Roman Catholics. His father was an aerospace engineer and nuclear physicist whose career moved the family around the U.S. The family enjoyed music and, at an early age, Bernie developed an interest in folk and bluegrass music. He eventually mastered the 5-string banjo, mandolin and acoustic guitar. As a young teen he moved with his family to San Diego, where he met fellow musicians Ed Douglas and Larry Murray of the local bluegrass outfit, the Scottsville Squirrel Barkers. The Barkers proved a breeding ground for future California country rock talent, including shy, 18-year-old mandolin player Chris Hillman, with whom Leadon maintained a lifelong friendship. Augmented by banjo player (and future Flying Burrito Brother) Kenny Wertz, the

Figure 3: Screen grab of the interface for the span-annotation task

changed to yellow for convenience. Mentions which are already annotated are marked in green.

3.2.2. Span annotation task

In this task, the interface does not have the sidebar (Fig. 3) and the annotators are free to mark one or more spans in the document as the antecedent(s) for a mention pronoun by selecting the span(s) with their pointers. In a scenario where one mention pronoun has to be linked with multiple antecedents, the annotators have to highlight the spans and click on the white *Link Entity* button multiple times. Therefore, an additional red *Finalize* button is provided to mark the end of one linking episode. Apart from the lack of the entity sidebar and inclusion of the previously mentioned *Finalize* button, all other features of the interface remain the same as those for the Grounded task.

3.2.3. AMT Details

The annotation tasks were open only to native English speakers whose approval rate was above 90%and they had ten minutes to annotate a document. Every fifth document annotated by an annotator was a secret test document for which annotations were known. The annotators were allowed to continue only if there was more than 90% match between the gold and their annotations. Each task was published 15 days apart to diversify the annotator pool.

	Exact Match	F_1 Score
Wiki grounded	0.70	0.74
Wiki free	0.50	0.65
QuAC grounded	0.65	0.67
QuAC free	0.52	0.64

Table 1: Inter-annotator agreement scores

4. Experiments

4.1. Inter-annotator agreement

As mentioned in Section 3.2., we doubly annotate 30 documents from each source to measure the inter-annotator agreement and the results are presented in Table 1. The numbers clearly indicate that the grounded tasks introduce less uncertainty about the antecedents and hence result in more agreements between the annotators. Ideally the exact match and F_1 scores for grounded tasks should be identical. However, the slight difference observed is because of mentions being linked to different similar looking entities. For example, in the sentence "Harry Potter is a global phenomenon. *It* has captured the imagination of ...", the mention *It* can be linked either to Harry Potter – the movies or Harry Potter – the books.

4.2. Annotation times

We can estimate the cognitive load on the annotators by measuring the time taken for marking the documents. Figure 4 shows the mean annotation times and their standard deviations for annotating documents in different settings. In general, QuAC documents require more time and effort to annotate due to the presence of QA pairs which require the annotators to possibly re-read a portion of the context paragraph. Also, it is clear that grounding the document eases the load on annotators irrespective of the source of documents.

4.3. State-of-the-art

We run our data through three state-of-the-art coreference resolution systems and report the average precision, recall and F_1 scores of three standard metrics: MUC, B^3 and CEAFe (Cai and Strube, 2010), in Table 2.[2] While Clark and Manning (2016)[3] and Lee et al. (2018) train on OntoNotes 5 to perform both mention detection and entity linking, Aralikatte et al. (2019) use a multi-task architecture for resolving coreference and ellipsis posed as reading comprehension, which is also trained on OntoNotes 5, but uses gold bracketing of the mentions and performs only entity linking.[4] The results show that the dataset is hard even for the current state-of-the-art and thus a good resource to evaluate new research.

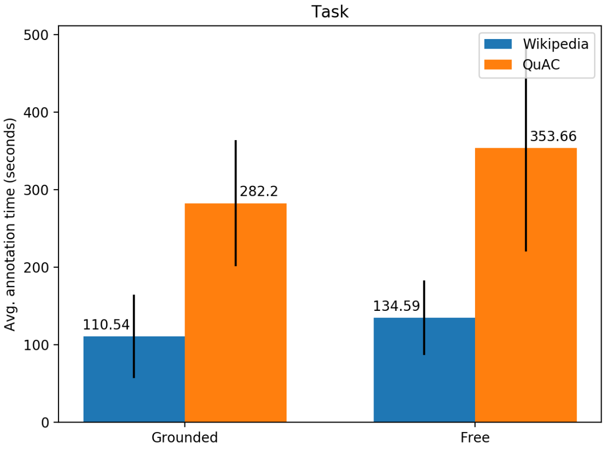


Figure 4: Average annotation times for the two tasks and settings

5. Discussion

The main purpose of this work is to study how humans annotate coreference with and without grounding. Therefore we give freedom to the annotators by asking them to abide by a minimal set of rules. We see interesting annotation patterns in our dataset: Generally, the indefinite pronoun 'all' is marked as having 'No Reference'. But for the sentence "...Harry Potter, and his friends Hermione Granger and Ron Weasley, *all* of whom ...", for example, the pronoun 'all' is linked as follows: (i) in the grounded task, the word is linked to three entities – Harry Potter, Hermione Granger and Ron Weasley, whereas (ii) in the span annotation task, the word is linked to the phrase "Harry Potter, and his friends Hermione Granger and Ron Weasley". We see that the annotation for the grounded task is cleaner than that for the span annotation task. This effect is observed throughout the dataset. Also, in span annotation tasks, while some annotators link mention pronouns to the first occurrence of an entity, some link them to the latest occurrence, sometimes resulting in multiple clusters instead of one. By design, this is not the case in the grounded tasks.

5.1. Comparison with WikiCoref

WikiCoref has 30 annotated pages from English Wikipedia. Our dataset contains 200 documents of which 30 titles are the same as those of WikiCoref. WikiCoref uses the full Wikipedia page for annotation, whereas we extract only the summary paragraphs from each page. WikiCoref doubly annotates only 3 documents for reporting inter-annotator agreement, whereas we do it for 30 documents. The inter-annotator agreements themselves are not comparable because they only report the Kappa coefficient for mention identification which does not occur in our tasks.

5.2. Generalization to other NLP tasks

Our first annotation experiments have been limited to coreference for pronouns, but obviously the same technique can be used to annotate other linguistic phenomena involving relations between noun phrases, e.g., other forms of coreference, nominal ellipsis, implicit arguments, or roles of semantic frames. Our models only include individuals or con-

[2] Converting our grounded data to the OntoNotes format is in some cases lossy, since entity aliases may not perfectly match previous mentions.

[3] We use an improved implementation available at https://github.com/huggingface/neuralcoref.

[4] This explains the comparatively higher numbers. See discussion in their paper for more details.

System	Wiki			QuAC		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**
Clark and Manning (2016)	24.72	32.87	27.95	20.15	27.98	23.39
Lee et al. (2018)	21.38	37.90	26.67	17.42	39.07	23.79
Aralikatte et al. (2019)*	43.88	48.58	45.96	46.18	46.23	46.14

Table 2: The macro-averages of MUC, B^3, and CEAF$_{\phi_4}$. (*assumes gold brackets for mentions.)

stants, but if we extend our models to also include propositions holding for individuals or between individuals, we could potentially also do grounded annotation of complex verbal phenomena such as VP ellipsis, gapping, sluicing, etc.

6. Conclusion

We propose a new way of annotating coreference by grounding the input text to reduce the cognitive load of the annotator. We do this by making the annotators choose the antecedent for mentions from a pre-populated entity list rather than having to select a span manually. We empirically show that annotations performed in this manner are faster and more coherent with higher inter-annotator agreements. We benchmark the collected data on state-of-the-art models and release it in the open domain at https://github.com/rahular/model-based-coref.

Acknowledgement

We would like to thank all the members of CoAStaL NLP, Matt Lamm and Vidhyashree Murthy for their valuable feedback on the annotation interface. This work was supported by a Google Focused Research Award.

Bibliographical References

Aralikatte, R., Lamm, M., Hardt, D., and Søgaard, A. (2019). Ellipsis and coreference resolution as question answering. *CoRR*, abs/1908.11141.

Blackburn, P. and Bos, J. (2005). *Representation and Inference for Natural Language*. CSLI, Stanford, CA.

Bower, G. and Morrow, D. (1990). Mental models in narrative comprehension. *Science*, 247(4938):44–48.

Cai, J. and Strube, M. (2010). Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 28–36. Association for Computational Linguistics, Sep.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October-November. Association for Computational Linguistics.

Clark, K. and Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas, November. Association for Computational Linguistics.

Cohen, K. B., Verspoor, K., Fort, K., Funk, C., Bada, M., Palmer, M., and Hunter, L. E., (2017). *The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation in the Biomedical Domain*, pages 1379–1394. Springer Netherlands, Dordrecht.

Culotta, A., Wick, M., and McCallum, A. (2007). First-order probabilistic models for coreference resolution. In *Proceedings of NAACL*.

Day, D., Mchenry, C., Kozierok, R., and Riek, L. (2004). Callisto: A configurable annotation workbench. 01.

Finkel, J. and Manning, C. (2008). Enforcing transitivity in coreference resolution. In *Proceedings of ACL*.

Ghaddar, A. and Langlais, P. (2016). Wikicoref: An english coreference-annotated corpus of wikipedia articles. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana, June. Association for Computational Linguistics.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, et al., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Nghiem, M.-Q. and Ananiadou, S. (2018). Aplenty: annotation tool for creating high-quality datasets using active and proactive learning. In *Proceedings of EMNLP*.

Peng, H., Khashabi, D., and Roth, D. (2015). Solving hard coreference problems. In *Proceedings of NAACL*.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1):3:1–3:44, April.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40. Association for Computational Linguistics, Jul.

Runner, J. T., Sussman, R. S., and Tanenhaus, M. K. (2003). Assignment of reference to reflexives and pronouns in picture noun phrases: evidence from eye movements. *Cognition*, 89(1):B1 – B13.

Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. (2018). Mind the gap: A balanced corpus of gendered ambiguous pronouns. In *Transactions of the ACL*, page to appear.

Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of ACL*.