ARRAU: Linguistically-Motivated Annotation of Anaphoric Descriptions

Olga Uryupina[1], Ron Artstein[2], Antonella Bristot, Federica Cavicchio[3], Kepa J. Rodriguez[4], Massimo Poesio[5,1]

[1]Department of Information Engineering and Computer Science, University of Trento,
[2]Institute for Creative Technologies, University of Southern California,
[3]Sign Language Lab, University of Haifa,
[4]Archives Division, Yad Vashem,
[5] School of Computer Science and Electronic Engineering, University of Essex

uryupina@gmail.com, artstein@ict.usc.edu, lucanto137@libero.it, federica.cavicchio@gmail.com, kepa.rodriguez@yadvashem.org.il, poesio@essex.ac.uk

Abstract

This paper presents a second release of the ARRAU dataset: a multi-domain corpus with thorough linguistically motivated annotation of anaphora and related phenomena. Building upon the first release almost a decade ago, a considerable effort had been invested in improving the data both quantitatively and qualitatively. Thus, we have doubled the corpus size, expanded the selection of covered phenomena to include referentiality and genericity and designed and implemented a methodology for enforcing the consistency of the manual annotation. We believe that the new release of ARRAU provides a valuable material for ongoing research in complex cases of coreference as well as for a variety of related tasks. The corpus is publicly available through LDC.

Keywords: Discourse, Anaphora, Coreference

1. Introduction

Coreference resolution is a crucial step in deep text understanding and as such is a vital prerequisite for a variety of high-level natural processing tasks, ranging from information extraction to summarization or machine translation. State-of-the-art statistical approaches to the task achieve robust performance on relatively easy cases of coreference, especially since the vast model optimization efforts have been undertaken by various research groups for the recent SemEval and CoNLL coreference resolution tracks (Recasens et al., 2010; Pradhan et al., 2011; Pradhan et al., 2012). More complex cases have been identified and investigated since the first years of research on coreference resolution, however, they have been out of the scope of the mainstream community till very recently. One of the main reasons is the lack of appropriate datasets.

Since its first release (Poesio and Artstein, 2008), the ARRAU corpus has been used, on one hand, for research on more complex coreference phenomena, and, on the other hand, as a reference point for annotating coreference corpora in other languages. The current paper presents the second release of ARRAU. For the second release we have not only focused on augmenting the number of covered documents, but also invested a considerable effort into improving the data quality. This involved annotating more attributes and designing a methodology for cleaning up the annotations. The former allows to use ARRAU for a variety of coreference-related problems, for example, for identifying and sub-classifying non-referential expressions. This is especially crucial when we want to train models covering complex cases of anaphora. The latter ensures the data quality and hence improves the performance of already existing models through retraining on cleaner data.

The rest of this paper is organized as follows. In the next section, we summarize the ARRAU annotation guidelines highlighting the differences between the two releases of ARRAU. In Section 3., we discuss the differences between ARRAU and other widely used coreferentially annotated corpora, ACE and OntoNotes. Section 4. presents our efforts at cleaning up the data and enforcing the annotation consistency. Section 5. summarizes our contributions and suggests directions for future work.

2. Two releases of ARRAU

The goal of the ARRAU project is to provide large-scale annotations of various linguistic phenomena related to anaphora. To this end, we have selected documents from four different domains: news (RST), task-oriented dialogues (TRAINS), fiction (PEAR) and medical leaflets (GNOME). We have opted for annotating already existing datasets with anaphoric information to enable studies involving several discourse-level phenomena. In particular, our news texts come from the Wall Street Journal dataset used in the RST Discourse Treebank (Carlson et al., 2002). Each document has been annotated for anaphoric information using the MMAX tool (Müller and Strube, 2006). The original annotation guidelines have been created and tested for the first release of ARRAU (Poesio and Artstein, 2008). In particular, the coders were requested to annotate each nominal markable, mark its referential status and basic morphological properties and then provide *coreference* and *bridging* links between (a) two markables or (b) a markable and a discourse segment. The guidelines have been designed in several stages to ensure the high inter-annotator agreement. We refer the reader to the original ARRAU-1 paper (Poesio and Artstein, 2008) for further details.

While ARRAU-1 provided valuable data for a variety of coreference and anaphora related applications, the corpus could be further improved. First, more data could be added, at least for some domains. The second release of ARRAU

domain	ARRAU 1			ARRAU 2		
	documents	tokens	markables	documents	tokens	markables
RST	204	146512	45590	413	228901	72013
PEAR	20	14059	3881	20	14059	4008
GNOME	5	21599	6215	5	21458	6562
TRAINS	35	25783	5198	114	83654	16999
total	264	184748	60884	552	348072	99582

Table 1: Corpus statistics for two releases of ARRAU

augments the corpus annotating all the documents available within the TRAINS and RST datasets. This has resulted in a significant increase in the data size. This quantitative improvement is extremely important for the TRAINS domain, since it provides a unique large collection of dialogues annotated with anaphoric information. More statistics for both releases of ARRAU are available in Table 1.

Second, we have focused on high-quality annotation of two phenomena closely related to anaphora: referential status and genericity of nominal markables. We have expanded our annotation of reference and genericity to all the domains, adopting a more principled approach. This has resulted in a more consistent annotation of reference: more than 10% of non-referring markables have been added to the documents already covered in ARRAU-1. For genericity, the first release only attempted a pilot annotation for the RST domain.

Moreover, we have further sub-classified non-referential and generic nominal phrases to allow for more refined linguistically motivated data driven analysis. As Table 3 shows, the second release of ARRAU provides a considerable amount of examples for non-referential and generic uses of nominal expressions and can thus be used as a valuable source for training statistical algorithms for these problems.

Finally, we have designed a methodology for enforcing the annotation consistency. Since the ARRAU scheme assumes simultaneous labeling of a variety of closely related phenomena, we have applied several constraints to ensure, for example, that a non-referential markable can not participate in a coreference chain. In a few cases, these constraints revealed intriguing cases of anaphoric expressions. Mostly, however, they have helped us identify and eliminate clear annotation errors. We will provide more details on our approach in Section 4. below.

3. ARRAU and other coreferentially annotated corpora

The ARRAU guidelines focus on more detailed representation of linguistic phenomena related to anaphora and coreference. In this section, we highlight the main differences between ARRAU and two other commonly used corpora annotated for coreference in English, ACE (Doddington et al., 2004) and OntoNotes (Pradhan et al., 2011; Pradhan et al., 2012). Table 2 provides a summary of the most distinctive features of ARRAU as opposed to ACE and OntoNotes.

The most prominent feature of ARRAU is its rich linguistically motivated annotation of markables. To start with, each nominal markables is shown with its minimal and maximal span. This solution is in line with the ACE annotation guidelines and has unfortunately been discarded for the OntoNotes dataset in order to decrease the annotation price and thus augment the corpus size. The maximal span corresponds to the full noun phrase, whereas the minimal span corresponds to the head noun or to the bare named entity for complex NE-nominals. With the latest development in the parsing technology, it might seem redundant to include minimal spans in the manual annotation directly: using dependencies or constituents with head-finding rules, one might expect to extract the minimal span for each NP rather reliably. It has been shown, however, that naive parsing-based heuristics do not lead to the best performance and a coreference resolver might benefit considerably from explicit or latent identification of minimal spans or heads (Zhekova and Kübler, 2013; Peng et al., 2015). Moreover, explicitly annotated minimal spans allow for better lenient matching that has been shown to improve the training procedure of coreference resolvers through better alignment of automatically extracted and gold markables (Kummerfeld et al., 2011). We believe therefore that the combination of minimal and maximal spans is the most reliable way of annotating markable boundaries for coreference. In the second release of ARRAU, we provide minimal and maximal spans for all the domains.

In ARRAU, we focus on different types of noun phrases. In particular, we label markables that do not participate in coreference chains: singletons and non-referentials. The ACE guidelines restrict the annotation scope to referentials[1], whereas OntoNotes only marks co-referential (no singletons) markables. As Table 3 shows, non-referentials and singletons account for up to one third of all the markables. Again, restricting the annotation scope allows for reducing the manual effort per document and thus for increasing the corpus size. However, a dataset with all the nominal markables annotated provides material for training mention detection systems. Mention detection for OntoNotes is a non-trivial problem that is further aggravated by the fact that singletons are removed and thus the direct training becomes hardly possible. In addition, ARRAU provides a sub-classification of non-referentials, thus allowing for detailed linguistically-motivated data-driven studies of referentiality.

Each markable is annotated with its basic morphological

[1]Moreover, the ACE guidelines focus on specific semantic types of referential markables, motivated from the Information Extraction perspective: person, organization, location and so on.

	ACE-05	ARRAU	CONLL
corpus size (# tokens)	220K	350K	1.5M
different domains	-	+	+
min and max markable boundaries	+	+	-
markable type annotated	+	-	-
markable attributes annotated	±	+	-
singletons annotated	+	+	-
all (co)referential markables annotated	-	+	+
non-referentials	-	+	-
rich gold linguistic annotations of text	-	±	+

Table 2: Comparison across coreferentially annotated corpora

properties: number, gender and semantic class. This allows, again, for training markable-level classifiers to assign these features automatically. Similarly to minimal span, this task can be attempted via heuristics based on parse trees, however, one can expected a higher performance if such tasks are attempted in a data-driven way.

The text collections used in ARRAU have been annotated for a variety of relevant discourse-level properties by other projects. For example, our news documents are taken from the RST treebank and thus further annotations can be induced from RST to investigate possible interactions between coreference and rhetorical structure.[2] The OntoNotes dataset, on the contrary, provides valuable gold annotations of low-level phenomena (for example, gold part-of-speech tags or parse trees), but does not, to our knowledge, provide deep discourse-level annotations apart from coreference.

Finally, the ARRAU dataset combines identity coreference with bridging. Again, it is very important to have the same corpus annotated for both bridging and coreference since these two tasks are very interrelated.

To summarize, the ARRAU dataset provides a high-quality refined annotation of anaphora and related phenomena. It relies on much more detailed and specific annotation guidelines than other commonly used corpora. We believe therefore that while the OntoNotes corpus is of crucial importance for data-intensive modeling of linguistically easier cases of coreference, ARRAU can be valuable, on one hand, for deeper linguistically oriented analysis of complex cases and, on the other hand, for learning models for related phenomena (genericity, referentiality etc).

4. Enforcing annotation consistency

A significant effort has been devoted to improving not only the quantity, but also the quality of the material annotated within the ARRAU project. To this end, we have implemented the following measures for the second release of the dataset:

- Minimal and maximal spans, genericity and referentiality have been annotated for all the documents. This enforces consistency across domains and allows for more principled cross-domain studies of the relevant phenomena.
- All the unspecified attributes have been re-annotated.
- Morphological attributes have been checked across coreference chains. For example, a typical chain should not include two markables of different gender. All the violating cases have been assessed manually.
- Semantic type has been checked for consistency across coreference chains.
- All the non-referential markables have been checked to exclude their participation in coreference chains. While the annotation scheme does not allow non-referentials to be anaphors, no MMAX functionality prevents a non-referential markable from being selected as an antecedent.
- All the markables labeled as discourse-old have been assigned an antecedent.
- Basic bracketing constraints have been enforced: no nominal markables should intersect each other or sentence boundaries.

The result of this effort has been two-fold. On the one hand, we have identified and removed various typos and inconsistencies that inevitably arise as a result of manual annotation. On the other hand, we have identified a number of truly challenging cases of coreference. The linguistic analysis of such examples constitutes a part of our ongoing work. Note that producing a non-negligible amount of challenging example has only been made possible as a byproduct of our thorough linguistically motivated annotation, for example, through a clash between coreference and non-referentiality.

5. Conclusion

In this paper, we have presented the second release of ARRAU—a corpus of anaphoric information. The dataset spans over four different domains and contains manually labeled data for such tasks as coreference resolution, bridging, mention detection, referentiality and genericity. Building upon the first release of ARRAU almost a decade ago, we have invested a considerable effort into improving the corpus, both qualitatively and quantitatively. To this end,

[2] We do not include RST annotations in the ARRAU distributions. The relevant information can be extracted through straightforward corpora alignment.

	RST	TRAINS	GNOME	PEAR
all	72013	16999	6562	4008
singletons	39431	5308	2887	1320
generic	2772	3071	819	122
generic	1434	728	12	74
episodic-no	-	4	-	-
no-generic	385	1285	470	-
operator-conditional	89	231	184	2
operator-instruction	15	163	205	-
operator-iquant	7	6	-	-
operator-modal	437	1078	142	16
operator-question	54	429	39	10
operator-tquant	16	4	-	-
underspecified-decease	-	-	75	-
underspecified-generic	1	3	-	-
underspecified-replicable	35	1	2	21
underspecified-substance	684	431	160	-
non-referential	9477	2351	1033	607
coordination	2408	231	326	37
expletive	444	851	75	122
idiom	638	148	29	42
incomplete	-	149	1	36
predicate	4252	145	352	79
quantifier	1724	817	250	132
unknown	9	6	-	159

Table 3: Distribution of markables in the second releases of ARRAU

we have doubled the amount of textual material, added several important attributes and designed the methodology to enforce the annotation consistency.

As a part of our ongoing work, we are analyzing the outcome of our efforts, in particular, of the measures undertaken to improve the data quality. On the one hand, we focus on providing a linguistic account for complex cases that have been identified through this methodology. On the other hand, we are running extensive evaluation experiments to assess the impact of data cleaning on the performance level of coreference resolvers.

For the third release of ARRAU, we plan to focus on bridging. We will expand the bridging annotation to all the domains and design consistency constraints for bridging aiming at reducing errors. We will also investigate the interaction between bridging and coreference.

The already released dataset opens up several research directions that we are planning to follow in the nearest future. First, the alignment between ARRAU and other corpora, in particular, RST, allow for joint modeling of coreference and discourse structure. Second, we plan to investigate a model that combines a data-intensive OntoNotes-style processing with deeper analysis based on ARRAU .

6. Acknowledgments

This work has been partially supported by the EC projects #671625 H2020 CogNet and #288024 FP7 LiMoSINe; EPSRC grant GR/S76434/01: Anaphora Resolution and Underspecification (ARRAU), as well as by the Autonomous Province of Trento project LiveMemories.

7. References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. Rst discourse treebank ldc2002t07.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassell, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proceedings of the Language Resources and Evaluation Conference*.

Jonathan K Kummerfeld, Mohit Bansal, David Burkett, and Dan Klein. 2011. Mention detection: Heuristics for the ontonotes annotations. In *Conference on Natural Language Learning*, pages 102–106. Association for Computational Linguistics.

Cristoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In *Corpus Technology and Language Pedagogy*, pages 197–214. S. Braun and K. Kohn and J. Mukherjee.

Haoruo Peng, Kai-Wei Chang, and Dan Roth. 2015. A joint framework for coreference resolution and mention head detection. *CoNLL 2015*, 51:12.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the arrau corpus. In *Proceedings of the Language Resources and Evaluation Conference*.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*, Portland, Oregon, June.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Marta Recasens, Lluís Màrquez, Emili Sapena, M.Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, Uppsala, Sweden.

Desislava Zhekova and Sandra Kübler. 2013. Machine learning for mention head detection in multilingual coreference resolution. In *RANLP*, pages 747–754.