An End-to-End Approach for Full Bridging Resolution

Joseph Renner, Priyansh Trivedi, Gaurav Maheshwari, Rémi Gilleron, Pascal Denis

An End-to-End Approach for Full Bridging Resolution

Joseph Renner, Priyansh Trivedi, Gaurav Maheshwari, Rémi Gilleron, Pascal Denis
Magnet, INRIA Lille, France
{joseph.renner, priyansh.trivedi, gaurav.maheshwari,
remi.gilleron, pascal.denis}@inria.fr

Abstract

In this article, we describe our submission to the CODI-CRAC 2021 Shared Task on Anaphora Resolution in Dialogues – Track BR (Gold)[1]. We demonstrate the performance of an end-to-end transformer-based higher-order coreference model finetuned for the task of *full bridging*. We find that while our approach is not effective at modeling the complexities of the task, it performs well on bridging resolution, suggesting a need for investigations into a robust anaphor identification model for future improvements.

1 Introduction

Anaphora is a discourse level phenomenon wherein a linguistic entity (referred to as **anaphor**) is associated with some other linguistic entity (referred to as an antecedent) within a document (Tognini-Bonelli, 2001). Broadly, the phenomenon is divided into coreference and bridging anaphora depending on whether the anaphoric references are linked to associated antecedents with an identical (*is-a*), or a non-identical (¬ *is-a*) relation, respectively. Following is an example of bridging anaphora – the focus of this article:

> Starbucks has a new take on the unicorn cappuccino. **One employee** accidentally leaked a picture of the secret new drink.

Here, the noun phrase "*One employee*" (bridging anaphor) is anaphorically linked to the antecedent – "*Starbucks*". In this instance, bridging anaphor can be thought of as an expression with an implicit argument, i.e., *one employee (of Starbucks)* (Rösiger et al., 2018).

The task of *full bridging* is that of identifying anaphors from a given set of linguistic entities in the document and linking them with their non-identical associated antecedent. This task is arguably more difficult and relatively understudied than that of entity coreference resolution – which involves identifying coreferent linguistic entities. This difficulty can be seen from the fact that Bridging anaphors are less likely to conform to syntactic or surface clues (Kobayashi and Ng, 2020), and the low inter-annotator agreement for bridging annotations (Markert et al., 2012). Hindrances to rapid progress on the task, however, also include lack of ample gold labeled data (Rösiger, 2018; Hou, 2020), and standardised evaluation schemes (Kobayashi and Ng, 2020). The CODI-CRAC 2021 shared task (Khosla et al., 2021) is thus a welcome addition to the existing set of datasets in the field as it provides a consistent benchmark across multiple gold-annotated datasets.

However, unlike the task variant that the majority of existing approaches tackle, the shared task also includes identification of the bridging anaphors as a part of the task. That is, most existing approaches (Lassalle and Denis, 2011; Hou et al., 2013; Hou, 2018a,b) assume the anaphor (here, "*one employee*") to be given, and are limited to selecting the correct antecedent for this phrase, amongst a predefined list of linguistic entities. The identification of anaphors compounds the difficulty of the task, owing to, amongst other things, the number of markables (linguistic entities) in a document (See Table 1), and low recall of anaphoric noun phrases in existing datasets (Rösiger et al., 2018).

That said, coreference resolution is a more complex task in terms of the decisions to be taken for correct predictions. Coreference chains are variadic, whereas relations representing bridging anaphors are binary. Further, contemporary approaches treat mention detection as a part of the task. The neural architectures thus proposed for

[1] Task home page - https://competitions.codalab.org/competitions/30312

Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue, pages 48–54
Punta Cana, Dominican Republic, November 10, 2021. ©2021 Association for Computational Linguistics

coreference resolution are often more expressive than the task of full bridging (with gold markables) requires. Based on this observation, we aim to find whether the aforementioned neural architectures can be adapted for solving a linguistically complex but simpler (in the above stated terms) task of full bridging. To that end, we experiment with the *independent* variant of the transformer based higher-order coreference model (Joshi et al., 2019).

We empirically find that the approach is inadequate in solving the task, and posit that the currently available amount of gold-labeled data is insufficient for this family of approaches. This suggests a need for more data, or the use of external information in solving the full bridging task when using this family of approaches. However, we observe that when tasked with only the resolution of bridging anaphors (i.e., the anaphoric markables are provided as a part of the task), this approach performs significantly better, suggesting that a two-step identification and resolution approach might be beneficial for the task.

2 Task Description

In this section, we introduce domain specific terms used throughout the article, and formalise the different variants of tasks corresponding to identification and association of bridging anaphors. We use the nomenclature used in (Rösiger et al., 2018).

Markables: A set of linguistic entities in a document of which anaphors and antecedents are both subsets.

Bridging Anaphor: A markable whose interpretation depends upon an antecedent, or more generally, which is implicitly linked to an antecedent with a non-identity relation. We refer to them as simply *anaphors* in the rest of this article.

Antecedent: The markable which is related to the anaphor with an implicit non-identity relation.

2.1 Task Variants

There are two primary variants of the task, defined as follows:

- **Bridging Resolution**: Given a document, and bridging anaphor markables, the task is that of finding the associated antecedent corresponding to each given bridging anaphors, from one of the markables preceding it.
- **Full Bridging**: Given a document, identify the bridging anaphors and find their associated antecedent from one of the markables preceding it.

If the markables are provided alongwith the corpus annotations, we call them **gold** markables. Otherwise, the approaches solving the task are expected to predict these markables. We refer to the latter as **predicted** markables.

The *CODI-CRAC 2021 Shared-Task: Anaphora Resolution in Dialogues* - Track BR (Gold) that we target is thus that of **Full Bridging with Gold Markables**[2].

3 Approach

3.1 Full Bridging with Higher-Order coreference model

Our full bridging system is based on the independent version (Joshi et al., 2019) of the higher-order coreference resolution model described in (Lee et al., 2018). As their model is designed for the coreference resolution with **predicted** markables, we adapt it for the full bridging task with gold markables. In this sub-section, we provide a brief overview of our augmented model and the associated problem formulation.

Following the footsteps of (Lee et al., 2018), we formulate the problem as selecting an antecedent y_i, from the set $\gamma(i)$, for each markable m_i in the document. The set includes a dummy antecedent ϵ and all the markables in the document before m_i, that is $\gamma(i) = \{\epsilon, m_1, ..., m_{i-1}\}$. A non-dummy assignment represents an anaphor-antecedent link between m_i and y_i, while a dummy assignment means that the markable has no antecedent in the document i.e., the markable is not a bridging anaphor.

For each markable m_i, the model learns a distribution $P(y_i)$ over all the previous markable set $\gamma(i)$:

$$P(y_i) = \frac{e^{s(m_i,y_i)}}{\sum_{y' \in \gamma(i)} e^{s(m_i,y')}} \qquad (1)$$

The s(x,y) is the scoring function consisting of three parts defined as:

$$s(x,y) = s_m(x) + s_m(y) + s_p(x,y) \qquad (2)$$

$$s_m(x) = \text{FFNN}_m(\mathbf{x}) \qquad (3)$$

$$s_p(x,y) = \text{FFNN}_p(\mathbf{x}, \mathbf{y}, \phi(x,y)) \qquad (4)$$

[2] For the purposes of further analysis, we also consider **Bridging Resolution with Gold Markables** over provided train sets.

Here $\mathbf{x}$ and $\mathbf{y}$ are the encoded representation of the two markables. These encodings are obtained by concatenating the transformer's output at the start and end of the span along with the attention vector computed over the output representation of the tokens in the span. $\phi(x, y)$ refers to the hand crafted features (the genre indicating the dataset to which this document belongs , speaker-ids, length of the two markables, and the number of tokens between them), while FFNN$_m$ and FFNN$_p$ represents feed forward network.

Recall that the approach in (Lee et al., 2018) is factored into a two-staged beam search. The first stage is responsible for predicting markables: a beam of up to M potential markables is computed based on the spans with the highest markable scores $s_m(x)$ out of all possible text spans, up to a certain width, in a document. In the second stage, the pairwise scores $s_p(x, y)$ are then only computed between the top markables, in a coarse-to-fine manner. Since we have access to the **gold** markables, we repurpose the markable scorer $s_m(x)$ to score markables as possible antecedents or anaphors. The number of markables can be large in documents, thus the beam search is necessary to keep memory costs down, otherwise the pairwise scoring would not be feasible.

We refer interested readers to (Lee et al., 2018) for a more detailed explanation of the model, including the coarse-to-fine pairwise scorer. The model is trained by the marginal log-likelihood of the possible correct antecedents. We further add a binary cross entropy based supervision over the outputs of the markable scorer $s_m(i)$, labeling markables as 0 if its neither an antecedent or an anaphor, and 1 if it is either of the two.

3.2 Bridging Resolution with Higher-Order coreference model

We adapt the model explained above to also solve the Bridging Resolution task (See Sec. 2.1). That is, since the set of anaphor markables is given, we do not intend for the model to identity anaphors, as well as resolve their antecedents, but only do the latter. This allows simplifications to the model and training setting: first, we can pass each anaphor, together with the document (up to the end of the anaphor sentence), into the model at a time, predicting one antecedent for the input anaphor given the relevant part of the document, instead of passing the entire document into the model and predicting all the anaphors and their antecedents at the same time.

This change in setting considerably alleviates memory constraints, since the pairwise scorer $s_p(x, y)$ is only computed between the one given anaphor and the possible antecedent markables, as opposed to an *n by n* pairwise scorer in the full bridging setting. This eliminates the need for both the mention scorer $s_m(x)$ and the coarse part of the coarse to fine pairwise scorer, leaving just the higher order, "fine" pairwise scorer described in (Lee et al., 2018). Also, this allows the use of a cross entropy loss over all possible antecedents (excluding the dummy class, as we know each labeled anaphor has an antecedent) for each anaphor. Finally, we remove the auxiliary supervision over the markable scorer outputs, as the scorer is no longer used.

4 Experiments

In this first experiment, we perform **full bridging** (with gold mentions) over the provided datasets (Sec. 4.1).

4.1 Datasets

The shared task is comprised of conversational documents from four domains, annotated with bridging anaphors using the Universal Anaphora format (Poesio et al., 1999). The five domains are:
Switchboard: A subset of the Switchboard Dialog Act Corpus (Godfrey et al., 1992), this dataset consists of transcribed phone conversations between two participants about varied topics including child care, recycling and news media. We filter out transcribed speech disfluencies (such as "emm", "ahh", "uh", etc) based on a hand-crafted list of bi-grams as a pre-processing step.
Light: Light is a collection of "character driven, human-human crowdworker interaction involving action, emotes and dialogue" (Urbanek et al., 2019) in the context of a fantasy text adventure game.
Persuasion: A collection of crowdsourced online conversations where a persuader tries to convince the persuadee to donate to a charity were introduced in (Wang et al., 2019). An annotated subset of these conversations are a part of this shared task.
AMI: Some of the transcripts of multi-speaker office meetings (Carletta, 2006) were annotated with bridging anaphors. Generally, these conversations are the longest of the four.

In all four cases, the test set is held out but

Dataset	Anaphors	Documents	Avg. Words	Avg. Markables	Avg. Words Between
Switchboard	603	11	1362 ± 339	366 ± 100	131 ± 228
Light	381	20	575 ± 79	194 ± 28	71 ± 100
Persuasion	245	21	474 ± 98	131 ± 30	41 ± 28
AMI	851	7	4820 ± 2258	1276 ± 595	197 ± 560
Trains-91	67	15	956 ± 691	190 ± 132	99 ± 165
Trains-93	610	94	726 ± 408	148 ± 83	66 ± 100

Table 1: Some statistics about the train sets of the datasets used in the shared task. From left to right, we report (i) the number of *anaphors* across all documents, (ii) number of documents, (iii) avg. length of each document, (iv) avg. number of markables in a document, and, (v) avg. number of tokens between an anaphor and its antecedent.

documents from the train set are annotated with markables, bridging anaphors and their antecedents. Apart from these, annotated instances from Trains-1993 (Allen, James and Heeman, Peter A., 1995), and Trains-1991 (Gross et al., 1993) Spoken Dialog Corpus, a subset of the ARRAU corpus (Uryupina et al., 2020) were used for training the models as well. Table 1 contains further statistics on the train set of these datasets.

4.2 Experimental Setup

We use Entity-F1 (Pradhan et al., 2012) as our metric for this experiment. We initialize our model, as outlined in Sec. 3.1 with a transformer based encoder with `bert-base-uncased` weights (Devlin et al., 2019) provided on the HuggingFace Model Hub (Wolf et al., 2020), and freeze it before subsequent fine-tuning. We use a two layer network with its hidden dimension and dropouts specified below, and a ReLU activation in the feed forward layers indicated in Eqn. 4.

During training, we set the batch size as 1, vary the hidden dimension of feed forward subnetworks between $\{256, 512\}$, and their dropout between $\{0.0, 0.3, 0.5\}$. We also experiment with multiple class weights for auxiliary supervision over the mention scorer's outputs, to compensate for the imbalance between anaphoric and non-anaphoric markables. We vary the inclusion of hand-crafted features (see $\phi(x, y)$ in Eq. 4 and the description of hand crafted features in Sec. 3.1) as a part of the grid search.

This experiment also represents our submission to the shared task. Corresponding to each of the aforementioned datasets, we submit a separate model. The hyperparameters for these models are found by running a 5-fold cross validation based grid search where, in each fold, the models are trained on 80% of instances of the corresponding dataset, and 100% of train instances of the remaining datasets[3], and is evaluated on the held out instances from the corresponding dataset. Once the hyperparameters are fixed corresponding to a dataset, we retrain the model from scratch on all training instances of all datasets for up to 20 epochs. The performance of the approach can be found in Table 2.

4.3 Results

We find that our end-to-end approach performs sub-optimally on all four datasets. Upon closer inspection, this performance is indicative of the challenge arisen by the amount of markables in a document. For instance, a document in Persuasion (F1: 16.28) contains only 134 markables on average, whereas a document in AMI (F1: 6.00) contains 1381 markables. This alludes to an inverse correlation between the average number of markables in document of a dataset and the entity F1 score on it. Also, the "Avg. words between" column in Table 1 indicates that anaphors and antecedents lie closer to each other in Persuasion when compared to other datasets. We hypothesise that this correlation, modeled as a part of hand-crafted features is actively exploited by our approach to increase a higher relative performance on the task.

Moreover, full bridging can be thought of as a combination of anaphor identification and bridging resolution. In order to ascertain whether our approach falters disproportionately on either of the two tasks, we perform a subsequent bridging resolution experiment over the train sets of these datasets (as they contain identified anaphors).

4.4 Bridging Resolution

We make the changes to the model and training procedure as outlined in Sec. 3.2. We keep the

[3] Including Trains-91 and Trains-93 datasets

Test Set	Ent F1	Track	Setting	Baselines	Learning Framework	Markable Identification Model	Training Set	Development Set
Switchboard	7.79	Bridging	Gold	-	-	-	All$_{train}$	5CV$_{dev}$
Light	9.35	Bridging	Gold	-	-	-	All$_{train}$	5CV$_{dev}$
Persuasion	16.28	Bridging	Gold	-	-	-	All$_{train}$	5CV$_{dev}$
AMI	6.00	Bridging	Gold	-	-	-	All$_{train}$	5CV$_{dev}$

Table 2: Results, and settings of our submission to the shared task as detailed in Sec. 4.2. As mentioned, we use the Entity-F1 metric to report the performance. All$_{train}$ refers to the collection of all instances from the training set of the datasets mentioned in Sec. 4.1, and 5CV$_{dev}$ refer to the *development* subsets, as they occur in each fold of the aforementioned 5-fold cross-validation based grid search in Sec. 4.2.

Approach	Switchboard Acc	Light Acc	Persuasion Acc	AMI Acc
Random	3.96	6.8	6.94	7.29
Skip-Gram	21.44	25.21	36.33	18.10
Unnamed	34.00	33.08	55.51	31.02

Table 3: Results of the experiment outlined in Sec. 4.4.

same hyperparameters (and grid search based hyperparameter search) as above, and train one model to make the predictions. However, unlike before, this experiment is performed in a 5-fold cross validation setup. Specifically, in each fold, we treat 20% of instances from each dataset as the test set. Another 10% are reserved for hyperparameter optimisation, and the remaining 70% of the instances are used for training the model. The performance reported in Table 3 is averaged over the five folds.

4.4.1 Baselines

We also report the performance of two baselines, outlined below.

Random: We select one of the markables at random which appears either in the first sentence of the document, or up to two sentences behind the anaphor. This sentence based strategy seemed to perform better than the one used in (Rösiger et al., 2018; Poesio et al., 2004).

Skip-Gram: We take the mean of pretrained Skip-Gram (Mikolov et al., 2013) embeddings of every token in a markable to create its vector representation. Then, using the cosine distance as a measure of anaphora, we select the antecedent which lies closest to the anaphor.

4.4.2 Results

We find that while the task is far from solved, our approach is significantly better at bridging resolution, compared to full bridging. That is, our model is unable to perform anaphor identification with a reasonable accuracy leading to a much worse performance on the full bridging task. Interestingly, the performance gap between the random and skip-gram baseline suggests that anaphor and antecedent markables often lie close in the vector spaces in a non-trivial amount of cases.

Across different datasets, we observe a similar trend as in the previous experiment. Our approach (as well as the skip-gram baseline) achieve its highest score on Persuasion, followed by Switchboard, Light and AMI. The performance gap between AMI and the next best approach is not as stark here (-9.3% here; -23% in Exp. 1). This can be explained by observing the random baseline. Its performance suggests that the sentences around the anaphor in AMI conversations have the least amount of markables to choose from, thereby making the task slightly easier.

5 Conclusion

In this work, we experiment with a higher-order coreference model based end-to-end approach for the full bridging task over conversational documents. We find that it is unable to model the task's complexities, however, its performance on bridging resolution is significantly better. This suggests a different approach to anaphor detection is needed, whether it be a stand alone anaphor detection model or a more guided adaptation of the higher-order coreference model (Joshi et al., 2019) that better suits the bridging task. We leave investigations along this line for the future. We also aim to experiment with approaches that can prime a model towards conversational documents, such as including the SpanBERT (Joshi et al., 2020) pretraining objective as a pre-finetuning step.

References

Allen, James and Heeman, Peter A. 1995. Trains spoken dialog corpus.

Jean Carletta. 2006. Announcing the ami meeting corpus. *The ELRA Newsletter*, 11(1):3–5.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

John J. Godfrey, Edward Holliman, and Jane McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '92, San Francisco, California, USA, March 23-26, 1992*, pages 517–520. IEEE Computer Society.

Derek Gross, James Allen, and David Traum. 1993. The trains 91 dialogues.

Yufang Hou. 2018a. A deterministic algorithm for bridging anaphora resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1938–1948. Association for Computational Linguistics.

Yufang Hou. 2018b. Enhanced word representations for bridging anaphora resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 1–7. Association for Computational Linguistics.

Yufang Hou. 2020. Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1428–1438. Association for Computational Linguistics.

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

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel S. Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5802–5807. Association for Computational Linguistics.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. The codi-crac 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue, Association for Computational Linguistics*.

Hideo Kobayashi and Vincent Ng. 2020. Bridging resolution: A survey of the state of the art. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3708–3721. International Committee on Computational Linguistics.

Emmanuel Lassalle and Pascal Denis. 2011. Leveraging different meronym discovery methods for bridging resolution in french. In *Anaphora Processing and Applications - 8th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2011, Faro, Portugal, October 6-7, 2011. Revised Selected Papers*, volume 7099 of *Lecture Notes in Computer Science*, pages 35–46. Springer.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 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. Association for Computational Linguistics.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*, pages 795–804. The Association for Computer Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Massimo Poesio, Florence Bruneseaux, and Laurent Romary. 1999. The mate meta-scheme for coreference in dialogues in multiple languages. In *ACL'99 Workshop Towards Standards and Tools for Discourse Tagging*, pages 65–74.

Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* *21-26 July, 2004, Barcelona, Spain*, pages 143–150. ACL.

Sameer Pradhan, Alessandro Moschitti, and Nianwen Xue, editors. 2012. *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Proceedings of the Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, EMNLP-CoNLL 2012, July 13, 2012, Jeju Island, Korea*. ACL.

Ina Rösiger. 2018. BASHI: A corpus of wall street journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Ina Rösiger, Arndt Riester, and Jonas Kuhn. 2018. Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3516–3528. Association for Computational Linguistics.

Elena Tognini-Bonelli. 2001. *Corpus linguistics at work*, volume 6. John Benjamins Publishing.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 673–683. Association for Computational Linguistics.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa Joseba Rodríguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Nat. Lang. Eng.*, 26(1):95–128.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5635–5649. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.