Bridging Resolution: Making Sense of the State of the Art

Hideo Kobayashi and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{hideo,vince}@hlt.utdallas.edu

Abstract

While Yu and Poesio (2020) have recently demonstrated the superiority of their neural multi-task learning (MTL) model to rule-based approaches for bridging anaphora resolution, there is little understanding of (1) how it is better than the rule-based approaches (e.g., are the two approaches making similar or complementary mistakes?) and (2) what should be improved. To shed light on these issues, we (1) propose a hybrid rule-based and MTL approach that would enable a better understanding of their comparative strengths and weaknesses; and (2) perform a manual analysis of the errors made by the MTL model.

1 Introduction

Bridging resolution is an anaphora resolution task that involves identifying and resolving bridging/associative anaphors, which are anaphoric references to non-identical associated antecedents. To exemplify, consider the following sentences taken from the BASHI corpus (Rösiger, 2018a):

> Even if *baseball* triggers losses at CBS – and he doesn't think it will – "I'd rather see **the games** on our air than on NBC and ABC," he says .

In this example, a bridging link exists between the anaphor *the games* and its antecedent *baseball*, as the definite description cannot be interpreted correctly unless it is associated with *baseball*.

Bridging resolution is arguably more challenging than entity coreference resolution. The reason is that unlike in entity coreference, in bridging resolution there are typically no clear syntactic or surface clues for identifying the antecedent of a bridging anaphor. In many cases, resolution requires the use of context as well as commonsense inference.

Despite the difficulty of bridging resolution, the annotated corpora available for training bridging resolvers are much smaller than those for training entity coreference resolvers (e.g., OntoNotes (Hovy et al., 2006)). As a result, early work has focused on developing rule-based systems (e.g., Hou et al. (2014), Rösiger (2018b)). A key weakness of rule-based approaches is that the ruleset may have to be updated when it is applied to a new corpus (e.g., new rules may have to be added, and existing rules may have to be removed or modified), as different bridging corpora are annotated with slightly different guidelines (to cover different kinds of bridging links, for instance). In light of this weakness, Yu and Poesio (2020) have recently proposed a neural bridging resolver based on multi-task learning (MTL). Despite being trained on the relatively small amount of labeled data that are currently available, their resolver has achieved state-of-the-art results on three evaluation corpora.

In this paper, we seek to make sense of this state of the art by shedding light on two issues. First, how is the MTL model better than its rule-based counterparts? More specifically, while MTL is apparently making fewer mistakes than the rules, are the two approaches making similar or complementary mistakes? Second, given that the MTL model is the current state of the art, what needs to be improved in MTL?

To investigate the first issue, we propose a *hybrid* approach to bridging resolution: we first apply the hand-crafted rules to identify bridging links, and then employ the MTL-based model to resolve any (anaphoric) mentions that are not resolved by the rules. The design of this pipelined resolver is motivated in part by sieve-based approaches to entity coreference resolution (Raghunathan et al., 2010; Lee et al., 2013). Specifically, given our hypothesis that hand-crafted rules typically have higher precision and lower coverage than machine-learned patterns, we employ the rules as our first sieve and MTL as our second sieve. If our hybrid approach outperformed both the rule-based and learning-based approaches, that would provide suggestive evidence that these two approaches have

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1652–1659
June 6–11, 2021.

Corpora	Docs	Tokens	Mentions	Anaphors
ISNotes	50	40292	11272	663
BASHI	50	57709	18561	459
ARRAU RST	413	228901	72013	3777

Table 1: Statistics on different corpora.

different strengths and weaknesses and therefore should be viewed as *complementary* approaches to bridging resolution. Note that this would be an important ramification, as learning-based approaches and rule-based approaches to bridging resolution have thus far been viewed as *competing* approaches. For instance, when evaluating their MTL model, Yu and Poesio (2020) merely view the rule-based systems as baselines. To investigate the second issue, we perform a manual analysis of the major types of error made by MTL. Since interpretability remains a key weaknesses of neural models, we believe that our analysis could provide useful insights into what needs to be improved in MTL.

2 Evaluation Setup

Corpora. We use three English corpora that are arguably the most widely used corpora for bridging evaluation, namely ISNotes (composed of 50 WSJ articles in OntoNotes) (Markert et al., 2012) , BASHI (The Bridging Anaphors Hand-annotated Inventory, composed of another 50 WSJ articles in OntoNotes) (Rösiger, 2018a), and ARRAU (composed of articles from four domains, RST, GNOME, PEAR, and TRAINS) (Poesio and Artstein, 2008; Uryupina et al., 2020). Following previous work, we report results only on RST, the most comprehensively annotated segment of ARRAU. Table 1 shows the statistics on these corpora.

For ARRAU RST, we use the standard train-test split. For ISNotes and BASHI, we divide the available documents into 10 folds and report 10-fold cross validation results, following previous work (Hou, 2020; Yu and Poesio, 2020).

The hybrid approach. Recall that our hybrid approach is composed of a rule-based system and Yu and Poesio's (2020) (learning-based) MTL approach. Below we provide a brief overview of the MTL approach and the rules.

Yu and Poesio's (2020) MTL-based system is the first neural model for full bridging resolution.[1] They presented two extensions to Kantor and Globerson's (2019) span-based neural mention-ranking model (Denis and Baldridge, 2008) that was originally developed for entity coreference resolution. First, they provided gold mentions as input to the model, meaning that the model needs to learn the span representations but not the span boundaries. Second, they proposed to train the model to perform coreference and bridging in a MTL framework, where the span representation layer is shared by the two tasks so that information learned from one task can be utilized when learning the other task. Unlike feature-based approaches, where feature engineering plays a critical role in performance, this model employs only two features, the length of a mention and mention-pair distance.

Different rule-based systems have been developed for the three evaluation corpora. We used Hou's (2014) rules for ISNotes, and Rösiger's (2018) rulesets for BASHI and ARRAU.[2] Table 2 shows an example rule designed by Hou et al. (2014) for full bridging resolution in ISNotes.[3] As can be seen, a rule is composed of two conditions: one on the anaphor and the other on the antecedent. If two mentions satisfy these conditions, the rule will posit a bridging link between them. In the table, we express the rule in terms of its name, the condition on the anaphor, the condition on the antecedent, and the motivation behind its design.[4]

Setting. We report results for *full bridging resolution*. In this setting, a system is given as input not only a document but also the *gold* mentions in the document. The goal is to identify the subset of the gold mentions that are bridging anaphors and resolve them to their antecedents, which are also chosen from the gold mentions.

Postprocessing. Following previous work (Rösiger et al., 2018), we postprocess the output of a resolver by removing the gold coreferent anaphors from the predicted bridging anaphors.

Evaluation metrics. We report results for recognition and resolution in terms of precision, recall, and F-score. For recognition, recall is the fraction of gold anaphors that are correctly identified, whereas precision is the fraction of anaphors iden-

[1] In our experiments, we use their implementation publicly available from https://github.com/juntaoy/dali-bridging. All model parameter values are the same as those used in Yu and Poesio (2020).

[2] Rösiger et al. (2018) designed an additional rule for BASHI and another ruleset for ARRAU.

[3] The complete set of rules designed by Hou et al. (2014) and Rösiger et al. (2018) can be found in Appendix A.

[4] In our experiments, we use the implementation of these rule-based systems publicly available from https://github.com/InaRoesiger/BridgingSystem.

Rule	Description (anaphor)	Description (antecedent)	Motivation
Set: Percentage	Percentage NPs in subject position	Closest NP modifying another percentage NP via the preposition "of"	Percentage expressions can indicate set bridging

Table 2: Example rule for resolving bridging anaphors in ISNotes.

	ISNotes						BASHI						ARRAU RST					
	Recognition			Resolution			Recognition			Resolution			Recognition			Resolution		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Rules	68.6	17.5	27.9	47.9	12.2	19.5	47.8	24.1	32.1	24.8	12.5	16.6	37.0	17.8	24.0	25.6	12.3	16.6
MTL	58.3	35.1	43.8	33.5	20.2	25.2	35.3	34.9	35.1	18.2	18.0	18.1	37.6	35.9	36.7	24.6	23.5	24.0
Hybrid	57.3	43.6	**49.5**	34.7	26.4	**30.0**	35.2	47.6	**40.5**	17.5	23.7	**20.1**	32.9	43.2	**37.4**	22.4	29.4	**25.4**

Table 3: Full bridging recognition and resolution results in ISNotes, BASHI, and ARRAU RST.

tified by the system that are correct. For resolution, recall and precision are defined in a similar fashion.

3 Results

Bridging recognition and resolution results of the three approaches under comparison (i.e., Rules, MTL, and Hybrid) on the three evaluation corpora are shown in Table 3. The performance trends largely corroborate our hypothesis. On all three datasets, we see that the recall of Hybrid is substantially higher than those of Rules and MTL for both recognition and resolution, meaning that Rules and MTL are making different rather than similar mistakes and can therefore be used to complement each other's weaknesses. Moreover, Hybrid's F-scores on ISNotes and BASHI are better than those of Rules and MTL: on ISNotes, Hybrid outperforms MTL by 5.7% points and 4.8% points in F-score for recognition and resolution, respectively; and on BASHI, Hybrid outperforms MTL by 5.4% points and 2.0% points in F-score for recognition and resolution, respectively. On ARRAU RST, however, Hybrid's recognition and resolution F-scores are only slightly better than those of Rules and MTL. The failure of Hybrid to offer substantial gains on ARRAU RST w.r.t. F-score can be attributed to Rules's relatively low precision: unlike in ISNotes and BASHI, where Rules's precision is higher than MTL's, in ARRAU RST, Rules's precision are more or less at the same level as MTL's.

Next, we compare in Table 4 the performance of our three resolvers on different categories of anaphors defined by the rules used in the rule-based resolver.[5] Each rule category is identified using its rule ID (column 1).[6] Each fraction in column 2 is the ratio of the number of gold anaphors that satisfy the anaphor condition of a rule to the number of gold mentions that satisfy the same condition. Finally, the recognition and resolution results shown in the remaining columns are expressed in terms of precision (P), recall (R), and F-score (F). We believe that these results can reveal the comparative strengths and weaknesses of the resolvers.

A few points about the results in Table 4 deserve mention. On ISNotes (Table 4(a)), while Rules outperforms MTL on the majority of the rule categories in resolution F-score, MTL achieves the state of the art by resolving anaphors in the largest category, Rule 18 (Other), which consists of anaphors that cannot be handled by any of the rules. On BASHI (Table 4(b)), however, Rules outperforms MTL on only four rule categories. This is somewhat surprising because the rulesets used for ISNotes and BASHI are almost identical to each other.[7] A closer look at the numbers in the second column of Table 4 reveals an interesting observation: in a majority of the rules, the number of gold anaphors that satisfy a rule condition is smaller in BASHI than in ISNotes, whereas the number of gold mentions that satisfy an anaphor condition is larger in BASHI than in ISNotes. This is again somewhat surprising because both ISNotes and BASHI contain 50 WSJ news articles taken from OntoNotes that are annotated with very similar annotation schemes. Consequently, we computed the average length of a document in the two datasets and found that BASHI indeed has more tokens per document on average (1154 tokens/doc in BASHI compared to 805 tokens/doc in ISNotes). The fact that BASHI has longer documents could explain why more gold mentions satisfy the anaphor condi-

[5] Owing to space limitations, only the results on ISNotes and BASHI are shown in Table 4. The results on ARRAU RST can be found in Appendix B.

[6] The mapping between rule IDs and the rule categories can be found in Appendix A.

[7] As can be seen in Table 4, the ruleset for BASHI is simply the ruleset for ISNotes augmented with Rule 10, which handles comparative anaphors.

Rule	Anaphors / Mentions	Rules						MTL						Hybrid					
		Recognition			Resolution			Recognition			Resolution			Recognition			Resolution		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
1	8/35	80	100	89	60	75	67	60	38	46	40	25	31	73	100	84	55	75	63
2	7/98	62	71	67	50	57	53	75	43	55	75	43	55	60	86	71	50	71	59
3	19/67	82	95	88	64	74	68	77	53	62	46	32	37	76	100	86	60	79	68
4	35/241	84	60	70	64	46	53	69	71	70	61	63	62	67	86	75	53	69	60
5	8/76	100	62	77	100	62	77	100	50	67	75	38	50	100	62	77	100	62	77
6	11/14	91	91	91	73	73	73	100	9	17	100	9	17	91	91	91	73	73	73
7	56/393	70	41	52	42	25	31	77	48	59	29	18	22	68	68	68	36	36	36
8	2/7	100	100	100	100	100	100	100	100	100	50	50	50	100	100	100	100	100	100
9	102/772	47	25	32	21	11	14	64	48	55	24	18	20	54	61	57	19	22	20
18	415/9568	0	0	0	0	0	0	49	26	34	31	16	21	49	26	34	31	16	21

(a) ISNotes

Rule	Anaphors / Mentions	Rules						MTL						Hybrid					
		Recognition			Resolution			Recognition			Resolution			Recognition			Resolution		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
1	3/51	50	67	57	0	0	0	25	33	29	0	0	0	33	67	44	0	0	0
2	8/111	83	62	71	67	50	57	80	50	62	60	38	46	75	75	75	50	50	50
3	4/25	57	100	73	43	75	55	50	50	50	50	50	50	57	100	73	43	75	55
4	23/374	50	43	46	33	29	31	77	48	59	69	43	53	50	57	53	38	43	40
5	3/111	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	5/13	42	100	59	25	60	35	80	80	80	60	60	60	38	100	56	23	60	33
7	31/629	31	44	36	18	25	21	32	25	28	12	9	11	30	62	41	17	34	22
8	1/4	0	0	0	0	0	0	100	100	100	0	0	0	50	100	67	0	0	0
9	107/1018	35	26	30	13	9	11	42	41	41	15	15	15	35	54	42	13	21	16
10	55/116	74	69	72	40	37	38	78	40	53	50	26	34	73	79	76	39	42	40
18	206/16011	0	0	0	0	0	0	23	29	26	12	14	13	23	29	26	12	14	13

(b) BASHI

Table 4: Performance of the three resolvers on different rule categories on ISNotes and BASHI.

tions of the rules in BASHI than in ISNotes. However, we still could not explain why the number of gold anaphors that satisfy the anaphor conditions of the rules is smaller in BASHI than in ISNotes. To understand the reason, we took a closer look at the documents in BASHI and found that there are cases of bridging that are not being annotated. Examples of such missing bridging links are shown in Table 6, where the missing anaphors are boldfaced and their antecedents are italicized. We therefore speculate that the lower resolution precision achieved by Rules on BASHI has to do with the incomplete gold annotations on BASHI.

In Table 5, we quantify how different Rules and MTL are w.r.t. each rule category. Let GA_i be the set of gold anaphors that are covered by rule category i. We show for each i the percentage of GA_i that are (1) correctly recognized/resolved by both resolvers (B), (2) correctly recognized/resolved by Rules but not MTL (R), and (3) correctly recognized/resolved by MTL but not Rules (M). For both ISNotes and BASHI, the relatively large numbers under the "R" and "M" columns suggest that Rules and MTL are making different predictions; moreover, the fact that the numbers under "R" are larger than the corresponding numbers under "M" on a majority of categories implies that the number of gold anaphors that are solely recognized/resolved by Rules is larger than that by MTL.

4 Error Analysis

To better understand what areas of improvement are needed by the MTL model, we perform a manual analysis of its errors and discuss three major types of error in the following three subsections.

4.1 Recognition: Precision Errors

Precision errors in recognition refer to errors in misclassifying a mention as a bridging anaphor. Coreference anaphor errors are the most common type of precision errors, contributing to 14-30% of the overall precision errors in recognition. Coreference anaphor errors occur when a gold coreference anaphor is predicted as a bridging anaphor.

Consider the first example in Table 7. In this example, the gold coreference anaphor *the stake* is predicted as a bridging anaphor and resolved to *the ground*, but it has a coreference link with *a big iron stake*. By definition, a bridging anaphor (especially referential bridging) should not be a coreference anaphor. We speculate that MTL makes these mistakes because it is trained on coreference and

Rule	Recognition			Resolution		
	B	R	M	B	R	M
1	38	62	0	25	50	0
2	29	43	14	29	29	14
3	47	47	5	16	58	16
4	46	14	26	40	6	23
5	50	12	0	38	25	0
6	9	82	0	9	64	0
7	21	20	27	4	21	14
8	100	0	0	50	50	0
9	12	13	36	3	8	15
18	0	0	26	0	0	16

(a) ISNotes

Rule	Recognition			Resolution		
	B	R	M	B	R	M
1	33	33	0	0	0	0
2	38	25	12	25	25	12
3	50	50	0	50	25	0
4	33	10	14	24	5	19
5	0	0	0	0	0	0
6	80	20	0	40	20	20
7	6	38	19	0	25	9
8	0	0	100	0	0	0
9	13	13	28	1	8	14
10	31	39	10	15	23	11
18	0	0	29	0	0	14

(b) BASHI

Table 5: Percentages of gold anaphors in each rule that are correctly recognized/resolved by both Rules and MTL (B), by Rules only (R), and by MTL only (M) in ISNotes and BASHI.

bridging in the multi-task setting.

4.2 Recognition: Recall Errors

Recall errors in recognition refer to the model's failure to identify bridging anaphors. Indefinite expression errors are the most common type of recall errors, contributing to 48-71% of the overall recall errors in recognition on the three datasets. Indefinite expression errors occur when a system misclassifies an indefinite bridging anaphor as a mention having the NEW information status.[8]

Consider the second example in Table 7. In this example, the indefinite bridging anaphor *production* is not detected by the MTL model. The reason is that the syntactic forms of many NEW instances and indefinite bridging anaphors are the same. Thus, it is not easy for model to distinguish between them. This observation has also been made by Hou et al. (2018).

4.3 Resolution: Precision Errors

Precision errors in resolution refer to errors in identifying the antecedent for a bridging anaphor. Unmodified expression errors are the most common type of precision errors, contributing to 23-63% of the overall precision errors in resolution. Unmodified expression errors occur when a predicted anaphor is a short mention without modifiers. Such a mention is semantically less rich than those that are modified and is therefore harder to resolve.

Consider the third example in Table 7. In this example, the anaphor *any divisions* is resolved to a wrong antecedent *their imprisonment* rather than the correct antecedent *the ANC*.

[8]Bridging is a subcategory of the MEDIATED.

When Michael S. Perry took the podium at *a recent cosmetics industry event*, more than 500 executives packing **the room** snapped to attention .
Folk doctors also prescribe it for kidney , bladder and urethra problems , duodenal ulcers and hemorrhoids . **Some** apply it to gouty joints .

Table 6: Examples of unannotated bridging links in BASHI.

After three Sagos were stolen from his home in Garden Grove , "I put *a big iron stake* in *the ground* and tied the tree to **the stake** with a chain , " he says proudly.
Currently, *Boeing* has a backlog of about $80 billion, but **production** has been slowed by a strike of 55,000 machinists , which entered its 22nd day today .
In addition, the government is figuring that the releases could create a split between the internal and external wings of *the ANC* and between the newly freed leaders and those activists who have emerged as leaders inside the country during *their imprisonment*. In order to head off **any divisions** , Mr. Mandela , in a meeting with his colleagues before they were released, instructed them to report to the ANC headquarters in Lusaka as soon as possible .

Table 7: Examples illustrating the three major types of recognition and resolution errors made by MTL.

5 Conclusion

In this paper, we sought to make sense of the state of the art in bridging resolution. We combined the hand-crafted rules and the MTL model in a pipelined fashion, showing that (1) the rules and MTL were making complementary mistakes and (2) the resulting hybrid approach achieved state-of-the-art results on three standard evaluation datasets. In addition, we performed a manual error analysis to determine what needed to be improved in MTL. Finally, our findings suggested that BASHI's annotation quality may need to be reassessed.

Acknowledgments

We thank the three anonymous reviewers for their detailed and insightful comments on an earlier draft of the paper. This work was supported in part by NSF Grants IIS-1528037 and CCF-1848608.

References

Pascal Denis and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 660–669.

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

Yufang Hou, Katja Markert, and Michael Strube. 2014. A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 2082–2093.

Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Volume 1 (Long Papers)*, pages 795–804.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1170–1174.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501.

Ina Rösiger. 2018a. BASHI: A corpus of wall street journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Ina Rösiger. 2018b. Rule- and learning-based methods for bridging resolution in the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 23–33.

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*, pages 3516–3528.

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

Juntao Yu and Massimo Poesio. 2020. Multi-task learning based neural bridging reference resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3534–3546.

A Rules for Bridging Resolution

Table 8 enumerates the list of heuristic rules manually designed for bridging resolution on ISNotes, BASHI, and ARRAU RST. As mentioned in Section 2, each rule is composed of a rule ID, an anaphor condition, and an antecedent condition. To enable the reader to better understand these rules, we describe in the last column of the table the motivation behind the design of each rule.

B Results on ARRAU RST

Results on ARRAU RST are shown in Tables 9 and 10. Specifically, Table 9 shows the performance of the three resolvers (Rules, MTL, and Hybrid) on different rule categories. This table is formatted in the same way as Table 4 and therefore can be interpreted in the same manner as Table 4. Comparing Table 4 and Table 9, we can see that Rules 11−17 are specifically designed for bridging resolution on ARRAU RST. Nevertheless, three of these seven rules are not fired on the ARRAU RST test set. Among these three rules, Rules 13 and 16 may have captured infrequent bridging phenomena, as they fail to cover any gold anaphors in the test set, whereas Rule 17 may have overfitted the training set, as it fails to recognize any gold anaphors in the test set. Overall, the Rules's results are somewhat disappointing: Rules outperforms MTL on only two of the rule categories, specifically the two defined by Rule 12 and Rule 14. In fact, our hypothesis that hand-crafted rules tend to have higher precision than machine-learned patterns fails on

ID	Rule	Condition on anaphor	Condition on antecedent	Motivation
1	Building part	Common NPs whose head is a building part without nominal pre-modifications	NP with the strongest semantic connectivity to the anaphor	Building part is often involved in meronymy
2	Relative person	Non-generic NPs whose head is a relative without nominal/adjective pre-modifications	Closest non-relative person NP	Handles relative nouns, which tend to be bridging
3	GPE job title	Job titles with country pre-modifications (e.g., Italian mayor)	Most salient GPE (e.g., Italy)	Some job title NPs implicitly refer to the globally salient GPE
4	Professional role	Professional role NPs (e.g., professor)	Most salient organization name	A more general rule than "Relative person" and "GPE job title"
5	Set: Percentage	Percentage NPs in subject position	Closest NP modifying another percentage NP via "of" (e.g., 22% of the firms)	Percentage expressions can indicate set bridging
6	Set: Number or indefinite pronoun	Number expressions (e.g., two dogs) or indefinite pronouns (e.g., some ...)	Closest plural NP in subject position. If not found, closest plural NP in object position	Numbers or indefinite pronouns can indicate set bridging
7	Argument-taking NPs 1	NPs with high argument ratio and without nominal/adjective pre-modifications or indefinite determiners	1. take all nominal modifiers of NPs whose head is same as anaphor's head. 2. closest NP that is a realization of these modification	Different instances of the same noun predicate likely maintain the same argument fillers indicated by nominal modifiers
8	Argument-taking NPs 2	NPs in subject position with high argument ratio and without nominal/adjective pre-modifications	NP with the strongest semantic connectivity to the anaphor	A NP in subject position that is likely to take arguments tends to be bridging anaphor
9	Meronymy relation	Unmodified definite NPs	NP classified as meronym with the anaphor by a relation classifier trained using WordNet	Handles meronym bridging
10	Comparative anaphora	NPs with comparative markers	Closest NP with same head and semantic category	Comparative anaphors are typically indicated by certain markers
11	Subset or element-of relation	NPs modified by noun, adjective, or relative clause	Closest NP with same head and semantic category	Anaphor is typically more specific than antecedent in subset or element-of bridging
12	Time subset	Expressions whose semantic category is TIME (e.g., 1920s)	Closest NP with TIME category and same decade number	Handles time expressions
13	One anaphora	Common noun starting with "one" (e.g., one committee member)	Closest plural NP with same semantic category and same common noun part	Handles one-anaphors
14	Locations	NPs with semantic category GPE or ORG	Closest NP with same category and has WordNet PartHolonym relation with anaphor	Handles links between cities/areas and their state/country
15	Same head	Singular and short NPs	Closest plural NP with same head and semantic category	Complements the "subset or element-of" rule
16	The rest	NPs whose string is "the rest"	Closest number expression	"the rest" is often annotated as bridging
17	Person	Person expressions with appositions (e.g., David Baker, vice president)	Closest plural person NP whose head is the same as head of anaphor's apposition	Handles person expressions with appositions
18	Other	NPs that cannot be handled by any of the rules		

Table 8: Complete set of hand-crafted rules for bridging resolution on ISNotes, BASHI, and ARRAU RST.

Rules 5, 6, 14, and 15. Consequently, the improvement of Hybrid over MTL on ARRAU RST is the smallest of the three evaluation datasets.

In Table 10, we attempt to quantify how different Rules and MTL are w.r.t. each rule category on ARRAU RST by showing the percentages of gold anaphors covered by each rule category that are correctly recognized/resolved correctly by both Rules and MTL (B), by Rules only (R), and by MTL only (M). This table is formatted in the same way as Table 5 and therefore can be interpreted in the same way as Table 5. As we can see, the largest values in the "R" column for both recognition and resolution are associated with Rules 12 and 14, meaning that these are the rule categories in which Rules has unique strength. This observation is consistent with the results of Rules 12 and 14 in Table 9. Other than these two rule categories, Rules

Rule	Anaphors / Mentions	Rules						MTL						Hybrid					
		Recognition			Resolution			Recognition			Resolution			Recognition			Resolution		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
1	1/17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	8/92	56	62	59	11	12	12	100	75	86	50	38	43	64	88	74	18	25	21
6	1/2	50	100	67	0	0	0	100	100	100	0	0	0	50	100	67	0	0	0
10	56/110	53	42	47	38	30	34	53	67	59	38	49	43	48	74	59	32	49	39
11	215/2653	37	17	23	28	13	18	36	42	39	23	27	25	33	47	39	22	32	26
12	3/119	33	100	50	33	100	50	0	0	0	0	0	0	30	100	46	30	100	46
13	0/9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	26/54	67	96	79	61	88	72	100	36	53	78	28	41	67	96	79	61	88	72
15	255/5163	16	8	11	11	6	8	32	27	29	20	17	18	25	31	28	17	20	18
16	0/0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	14/41	0	0	0	0	0	0	56	64	60	44	50	47	56	64	60	44	50	38
18	79/1987	0	0	0	0	0	0	30	21	24	20	14	17	30	21	24	20	14	17

Table 9: Performance of the three resolvers on different rule categories in ARRAU RST.

Rule	Recognition			Resolution		
	B	R	M	B	R	M
1	0	0	0	0	0	0
5	50	12	25	12	0	25
6	100	0	0	0	0	0
10	35	7	33	21	9	28
11	12	6	30	7	6	20
12	0	100	0	0	100	0
13	0	0	0	0	0	0
14	36	60	0	24	64	4
15	4	4	23	2	4	14
16	0	0	0	0	0	0
17	0	0	64	0	0	50
18	0	0	21	0	0	14

Table 10: Percentages of gold anaphors in each rule that are correctly recognized/resolved by both Rules and MTL (B), by Rules only (R), and by MTL only (M) in ARRAU RST.

manages to uniquely recognize/resolve just a few anaphors covered by rule categories 5, 10, 11, and 15. In contrast, the number of gold anaphors that are uniquely recognized/resolved by MTL is larger than that by Rules. Overall, we can infer from the results in Table 10 that the use of Rules does not add a lot of value to MTL on ARRAU RST.