Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns

Kellie Webster and **Marta Recasens** and **Vera Axelrod** and **Jason Baldridge**
Google AI Language
{websterk, recasens, vaxelrod, jasonbaldridge}@google.com

Abstract

Coreference resolution is an important task for natural language understanding, and the resolution of ambiguous pronouns a longstanding challenge. Nonetheless, existing corpora do not capture ambiguous pronouns in sufficient volume or diversity to accurately indicate the practical utility of models. Furthermore, we find gender bias in existing corpora and systems favoring masculine entities. To address this, we present and release GAP, a gender-balanced labeled corpus of 8,908 ambiguous pronoun–name pairs sampled to provide diverse coverage of challenges posed by real-world text. We explore a range of baselines that demonstrate the complexity of the challenge, the best achieving just 66.9% F1. We show that syntactic structure and continuous neural models provide promising, complementary cues for approaching the challenge.

1 Introduction

Coreference resolution involves linking referring expressions that evoke the same discourse entity, as defined in shared tasks such as CoNLL 2011/2012 (Pradhan et al., 2012) and MUC (Grishman and Sundheim, 1996). Unfortunately, high scores on these tasks do not necessarily translate into acceptable performance for downstream applications such as machine translation (Guillou, 2012) and fact extraction (Nakayama, 2008). In particular, high-scoring systems successfully identify coreference relationships between string-matching proper names, but fare worse on anaphoric mentions such as pronouns and common noun phrases (Stoyanov et al., 2009; Rahman and Ng, 2012; Durrett and Klein, 2013).

We consider the problem of resolving gendered ambiguous pronouns in English, such as **she**[1] in:

(1) In May, *Fujisawa* joined *Mari Motohashi*'s rink as the team's skip, moving back from Karuizawa to Kitami where **she** had spent her junior days.

With this scope, we make three key contributions:

- We design an extensible, language-independent mechanism for extracting challenging ambiguous pronouns from text.
- We build and release GAP, a human-labeled corpus of 8,908 ambiguous pronoun–name pairs derived from Wikipedia.[2] This data set targets the challenges of resolving naturally occurring ambiguous pronouns and rewards systems that are gender-fair.
- We run four state-of-the-art coreference resolvers and several competitive simple baselines on GAP to understand limitations in current modeling, including gender bias. We find that syntactic structure and transformer models (Vaswani et al., 2017) provide promising, complementary cues for approaching GAP.

Coreference resolution decisions can drastically alter how automatic systems process text. Biases in automatic systems have caused a wide range of underrepresented groups to be served in an inequitable way by downstream applications (Hardt, 2014). We take the construction of the new GAP corpus as an opportunity to reduce gender bias in coreference data sets; in this way, GAP can promote equitable modeling of reference phenomena complementary to the recent work of Zhao et al. (2018) and Rudinger et al. (2018).

[1] The examples throughout the paper highlight the ambiguous pronoun in **bold**, the two potential coreferent names in *italics*, and the correct one also underlined.

[2] http://goo.gl/language/gap-coreference.

Transactions of the Association for Computational Linguistics, vol. 6, pp. 605–618, 2018. Action Editor: Luke Zettlemoyer.
Submission batch: 7/2018; Published 12/2018.

Such approaches promise to improve equity of downstream models, such as triple extraction for knowledge-base populations.

2 Background

Existing datasets do not capture ambiguous pronouns in sufficient volume or diversity to benchmark systems for practical applications.

2.1 Data Sets with Ambiguous Pronouns

Winograd schemas (Levesque et al., 2012) are closely related to our work as they contain ambiguous pronouns. These are pairs of short texts with an ambiguous pronoun and a special word (in square brackets) that switches its referent:

(2) *The trophy* would not fit in *the brown suitcase* because **it** was too [big/small].

The Definite Pronoun Resolution Data Set (Rahman and Ng, 2012) comprises 943 Winograd schemas written by undergraduate students and later extended by Peng et al. (2015). The First Winograd Schema Challenge (Morgenstern et al., 2016) released 60 examples adapted from published literary works (Pronoun Disambiguation Problem)[3] and 285 manually constructed schemas (Winograd Schema Challenge).[4] More recently, Rudinger et al. (2018) and Zhao et al. (2018) have created two Winograd schema-style datasets containing 720 and 3,160 sentences, respectively, where each sentence contains a gendered pronoun and two occupation (or participant) antecedent candidates that break occupational gender stereotypes. Overall, ambiguous pronoun datasets have been limited in size and, most notably, consist only of manually constructed examples that do not necessarily reflect the challenges faced by systems in the wild.

In contrast, the largest and most widely used coreference corpus, OntoNotes (Pradhan et al., 2007), is general purpose. In OntoNotes, simpler high-frequency coreference examples (e.g., those captured by string matching) greatly outnumber examples of ambiguous pronouns, which obscures performance results on that key class (Stoyanov et al., 2009; Rahman and Ng, 2012). Ambiguous pronouns greatly impact main entity resolution in Wikipedia, the focus of Ghaddar and Langlais (2016a), who use WikiCoref, a corpus of 30 full articles annotated with coreferences (Ghaddar and Langlais, 2016b).

GAP examples are not strictly Winograd schemas because they have no reference-flipping word. Nonetheless, they contain two person named entities of the same gender and an ambiguous pronoun that may refer to either (or neither). As such, they represent a similarly difficult challenge and require the same inferential capabilities. More importantly, GAP is larger than existing Winograd schema datasets, and the examples are from naturally occurring Wikipedia text. GAP complements OntoNotes by providing an extensive targeted dataset of naturally occurring ambiguous pronouns.

2.2 Modeling Ambiguous Pronouns

State-of-the-art coreference systems struggle to resolve ambiguous pronouns that require world knowledge and commonsense reasoning (Durrett and Klein, 2013). Past efforts have tried to mine semantic preferences and inferential knowledge via predicate–argument statistics mined from corpora (Dagan and Itai, 1990; Yang et al., 2005), semantic roles (Kehler et al., 2004; Ponzetto and Strube, 2006), contextual compatibility features (Liao and Grishman, 2010; Bansal and Klein, 2012), and event role sequences (Bean and Riloff, 2004; Chambers and Jurafsky, 2008). These usually bring small improvements in general coreference datasets and larger improvements in targeted Winograd datasets.

Rahman and Ng (2012) scored 73.05% precision on their Winograd dataset after incorporating targeted features such as narrative chains, Web-based counts, and selectional preferences. Peng et al. (2015)'s system improved the state of the art to 76.41% by acquiring ⟨subject, verb, object⟩ and ⟨subject/object, verb, verb⟩ knowledge triples.

In the First Winograd Schema Challenge (Morgenstern et al., 2016), participants used methods ranging from logical axioms and inference to neural network architectures enhanced with commonsense knowledge (Liu et al., 2017), but no system qualified for the second round. Recently, Trinh and Le (2018) have achieved the best results on the Pronoun Disambiguation Problem and Winograd Schema Challenge datasets, achieving 70% and 63.7%, respectively, which are 3 percentage

[3] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/PDPChallenge2016.xml.

[4] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection.xml.

points and 11 percentage points above Liu et al.'s (2017)'s previous state of the art. Their model is an ensemble of word-level and character-level recurrent language models, which, despite not being trained on coreference data, encode commonsense as part of the more general language modeling task. It is unclear how these systems perform on naturally occurring ambiguous pronouns. For example, Trinh and Le's (2018) system relies on choosing a candidate from a pre-specified list, and it would need to be extended to handle the case that the pronoun does not corefer with any given candidate. By releasing GAP, we aim to foster research in this direction, and set several competitive baselines without using targeted resources.

2.3 Bias in Machine Learning

Although existing corpora have promoted research into coreference resolution, they suffer from gender bias. Specifically, of the over 2,000 gendered pronouns in the OntoNotes test corpus, less than 25% are feminine (Zhao et al., 2018). The imbalance is more pronounced on the development and training sets, with less than 20% feminine pronouns each. WikiCoref contains only 12% feminine pronouns. In the Definite Pronoun Resolution Dataset training data, 27% of the gendered pronouns are feminine, and the Winograd Schema Challenge datasets contain 28% and 33% feminine examples. Two exceptions are the recent WinoBias (Zhao et al., 2018) and Winogender schemas (Rudinger et al., 2018) datasets, which reveal how occupation-specific gender bias pervades in the majority of publicly available coreference resolution systems by including a balanced number of feminine pronouns that corefer with anti-stereotypical occupations (see Example (3), from WinoBias). These datasets focus on pronominal coreference where the antecedent is a nominal mention, whereas GAP focuses on relations where the antecedent is a named entity.

(3) *The salesperson* sold some books to *the librarian* because **she** was trying the sell them.

The pervasive bias in existing datasets is concerning given that learned NLP systems often reflect and even amplify training biases (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2017). A growing body of work defines notions of fairness, bias, and equality in data and machine-learned systems (Pedreshi et al., 2008; Hardt et al., 2016; Skirpan and Gorelick, 2017; Zafar et al., 2017), and debiasing strategies include expanding and rebalancing data (Torralba and Efros, 2011; Buda, 2017; Ryu et al., 2017; Shankar et al., 2017), and balancing performance across subgroups (Dwork et al., 2012). In the context of coreference resolution, Zhao et al. (2018) have shown how debiasing techniques (e.g., swapping the gender of male pronouns and antecedents in OntoNotes, using debiased word embeddings, balancing Bergsma and Lin's [2006] gender list) succeed at reducing the gender bias of multiple off-the-shelf coreference systems.

We work towards fairness in coreference by releasing a diverse, gender-balanced corpus for ambiguous pronoun resolution and further investigating performance differences by gender, not specifically on pronouns with an occupation antecedent but more generally on gendered pronouns.

3 GAP Corpus

We create a corpus of 8,908 human-annotated ambiguous pronoun-name examples from Wikipedia. Examples are obtained from a large set of candidate contexts and are filtered through a multistage process designed to improve quality and diversity.

We choose Wikipedia as our base dataset given its wide use in natural language understanding tools, but are mindful of its well-known gender biases. Specifically, less than 15% of biographical Wikipedia pages are about women. Furthermore, women are written about differently than men: For example, women's biographies are more likely to mention marriage or divorce (Bamman and Smith, 2014), abstract terms are more positive in male biographies than female biographies (Wagner et al., 2016), and articles about women are less central to the article graph (Graells-Garrido et al., 2015).

3.1 Extraction and Filtering

Extraction targets three patterns, given in Table 1, that characterize locally ambiguous pronoun contexts. We limit to singular mentions, gendered non-reflexive pronouns, and names whose head tokens are different from one another. Additionally, we do not allow intruders: There can be no other compatible mention (by gender, number, and entity type) between the pronoun and the two names.

To limit the success of naïve resolution heuristics, we apply a small set of constraints to focus on those pronouns that are truly hard to resolve.

Type	Pattern	Example
FINALPRO	(Name, Name, Pronoun)	*Preckwinkle* criticizes *Berrios'* nepotism: [...] County's ethics rules don't apply to **him**.
MEDIALPRO	(Name, Pronoun, Name)	*McFerran*'s horse farm was named Glen View. After **his** death in 1885, *John E. Green* acquired the farm.
INITIALPRO	(Pronoun, Name, Name)	Judging that **he** is suitable to join the team, *Butcher* injects *Hughie* with a specially formulated mix.

Table 1: Extraction patterns and example contexts for each.

Dimension	Values	Ratio
Page coverage		1 per page per pronoun form
Gender	masc. : fem.	1 : 1
Extraction Pattern	final : medial : initial	6.2 : 1 : 1
Page Entity	true : false	1.3 : 1
Coreferent Name	nameA : nameB	1 : 1

Table 2: Diversity statistics in final corpus.

- **FINALPRO**. Both names must be in the same sentence, and the pronoun may appear in the same or directly following sentence.
- **MEDIALPRO**. The first name must be in the sentence directly preceding the pronoun and the second name, both of which are in the same sentence. To decrease the bias for the pronoun to be coreferential with the first name, the pronoun must be in an initial subordinate clause or be a possessive in an initial prepositional phrase.
- **INITIALPRO**. All three mentions must be in the same sentence and the pronoun must be in an initial subordinate clause or a possessive in an initial prepositional phrase.

From the extracted contexts, we sub-sample those to send for annotation. We do this to improve diversity in five dimensions:

- **Page Coverage**. We retain at most three examples per page–gender pair to ensure a broad coverage of domains.
- **Gender**. The raw pipeline extracts contexts with a m:f ratio of 9:1. We oversampled feminine pronouns to achieve a 1:1 ratio.[5]
- **Extraction Pattern**. The raw pipeline output contains seven times more FINALPRO contexts than MEDIALPRO and INITIALPROcombined, so we oversampled the latter two to lower the ratio to 6:1:1.
- **Page Entity**. Pronouns in a Wikipedia page often refer to the entity the page is about. We include such examples in our dataset but balance them 1:1 against examples that do not include mentions of the page entity.
- **Coreferent Name**. To ensure that mention order is not a cue for systems, our final dataset is balanced for label — namely, whether Name A or Name B is the pronoun's referent.

We applied these constraints to the raw extractions to select 8,604 contexts (17,208 examples) for annotation that were globally balanced in all dimensions (e.g., 1:1 gender ratio in MEDIALPRO extractions). Table 2 summarizes the diversity ratios obtained in the final dataset, whose compilation is described next.

3.2 Annotation

We used a pool of in-house raters for human annotation of our examples. Each example was presented to three workers, who selected one of five labels (Table 3). Full sentences of at least 50 tokens preceding each example were presented as context (prior context beyond a section break is not included). Rating instructions accompany the dataset release.

Despite workers not being expert linguists, we find good agreement both within workers and between workers and an expert. Inter-annotator agreement was $\kappa = 0.74$ on the Fleiss et al. (2003) kappa statistic; in 73% of cases there was full agreement between workers, in 25% of cases two

[5] In doing this, we observed that many feminine pronouns in Wikipedia refer to characters in film and television.

Label	Raw	Final
Name A	2,913	1,979
Name B	3,047	1,985
Neither Name A nor Name B	1,614	490
Both Name A and Name B	1,016	0
Not Sure	14	0
Total	8,604	4,454

Table 3: Consensus label counts for the extracted examples (Raw) and after further filtering (Final).

of three workers agreed, and only in 2% of cases was there no consensus. We discard the 194 cases with no consensus. On 30 examples rated by an expert linguist, there was agreement on 28 and one was deemed to be truly ambiguous with the given context.

To produce our final dataset, we applied additional high-precision filtering to remove some error cases identified by workers,[6] and discarded the "Both" (no ambiguity) and "Not Sure" contexts. Given that many of the feminine examples received the "Both" label from referents having stage and married names (Example (4)), this unbalanced the number of masculine and feminine examples.

(4) *Ruby Buckton* is a fictional character from the Australian Channel Seven soap opera Home and Away, played by *Rebecca Breeds*. **She** debuted ...

To correct this, we discarded masculine examples to re-achieve 1:1 gender balance. Additionally, we imposed the constraint that there be one example per Wikipedia article per pronoun form (e.g., **his**), to reduce similarity between examples. The final counts for each label are given in the second column of Table 3. Given that the 4,454 contexts each contain two annotated names, this constitutes 8,908 pronoun–name pair labels.

4 Experiments

We set up the GAP challenge and analyze the applicability of a range of off-the-shelf tools. We find that existing resolvers do not perform well and are biased to favor better resolution of masculine pronouns. We empirically validate the observation that Transformer models (Vaswani et al., 2017) encode coreference relationships, adding to the results by Voita et al. (2018) on machine translation, and Trinh and Le (2018) on language modeling. Furthermore, we show they complement traditional linguistic cues such as syntactic distance and parallelism.

All experiments use the Google Cloud NL API[7] for pre-processing, unless otherwise noted.

4.1 GAP Challenge

GAP is an evaluation corpus and we segment the final dataset into a development and test set of 4,000 examples each;[8] we reserve the remaining 908 examples as a small validation set for parameter tuning. All examples are presented with the URL of the source Wikipedia page, allowing us to define two task settings: *snippet-context* in which the URL may not be used, and *page-context* in which it may. Although name spans are given in the data, we urge the community not to treat this as a *gold mention* or Winograd-style task. That is, systems should detect mentions for inference automatically, and access labeled spans only to output predictions.

To reward unbiased modeling, we define two evaluation metrics: F1 score and Bias. Concretely, we calculate F1 score **O**verall as well as by the gender of the pronoun (**M**asculine and **F**eminine). **B**ias is calculated by taking the ratio of feminine to masculine F1 scores, typically less than 1.[9]

4.2 Off-the-Shelf Resolvers

The first set of baselines we explore are four representative off-the-shelf coreference systems: the rule-based system of Lee et al. (2013) and three neural resolvers—Clark and Manning (2015),[10] Wiseman et al. (2016),[11] and Lee et al. (2017).[12] All were trained on OntoNotes and run in as close to their out-of-the-box configuration as possible.[13] System clusters were scored against GAP examples according to whether the cluster

[6] For example, missing sentence breaks, list environments, and non-referential personal roles/nationalities.

[7] https://cloud.google.com/natural-language/.

[8] All examples extracted from the same URL are partitioned into the same set.

[9] http://goo.gl/language/gap-coreference.

[10] https://stanfordnlp.github.io/CoreNLP/download.html.

[11] https://github.com/swiseman/nn_coref.

[12] https://github.com/kentonl/e2e-coref.

[13] We run Lee et al. (2017) in the `final` (single-model) configuration, with NLTK preprocessing (Bird and Loper, 2004); for Wiseman et al. (2016) we use Berkeley preprocessing (Durrett and Klein, 2014); and the Stanford systems are run within Stanford CoreNLP (Manning et al., 2014).

	M	F	B	O
Lee et al. (2013)	55.4	45.5	0.82	50.5
Clark and Manning	58.5	51.3	0.88	55.0
Wiseman et al.	**68.4**	59.9	0.88	64.2
Lee et al. (2017)	67.2	**62.2**	**0.92**	**64.7**

Table 4: Performance of off-the-shelf resolvers on the GAP development set, split by **M**asculine and **F**eminine (**B**ias shows F/M), and **O**verall. **Bold** indicates best performance.

	M	F	B	O
Lee et al. (2013)	47.7	53.2	1.12	49.2
Clark and Manning	64.3	**63.9**	**0.99**	**64.2**
Wiseman et al.	61.9	58.0	0.94	60.6
Lee et al. (2017)	**68.9**	51.9	0.75	63.4

Table 5: Pronoun-name F1 score, by gender, of off-the-shelf systems on the OntoNotes test set. Scores based on 2,091 masculine pronoun–named entity pairs (in 403 clusters) and 1,095 feminine pairs (in 104 clusters). **Bold** indicates best performance.

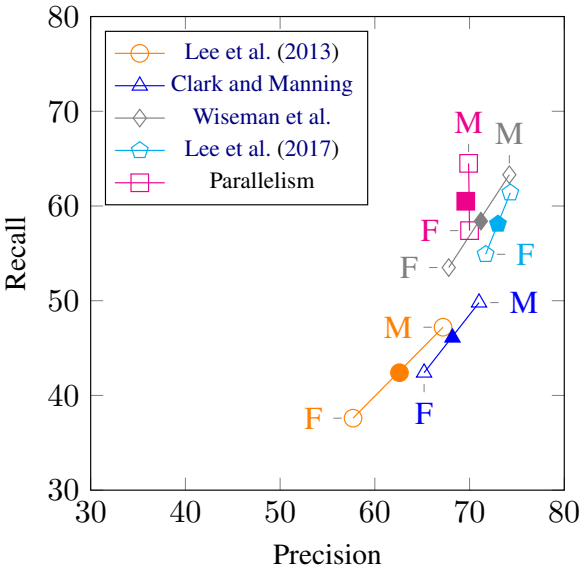

Figure 1: Precision-Recall on the GAP development data set—Overall (solid markers), **M**asculine, **F**eminine—for off-the-shelf resolvers and Parallelism.

containing the target pronoun also contained the correct name (**TP**) or the incorrect name (**FP**), using mention heads for alignment. We report here their performance on GAP as informative baselines, but expect retraining on Wikipedia-like texts to yield an overall improvement in performance. (This remains as future work.)

Table 4 shows that all systems struggle on GAP. That is, despite modeling improvements in recent years, ambiguous pronoun resolution remains a challenge. We note particularly the large difference in performance between genders, which traditionally has not been tracked but has fairness implications for downstream tasks using these publicly available models.

Table 5 provides evidence that this low performance is not solely due to domain and task differences between GAP and OntoNotes. Specifically, with the exception of Clark and Manning (2015), the table shows that system performance on pronoun–name coreference relations in the OntoNotes test set[14] is not vastly better than GAP. One possible reason that in-domain OntoNotes performance and out-of-domain GAP performance are not very different could be that state-of-the-art systems are highly tuned for resolving names rather than ambiguous pronouns.

Further, the relative performance of the four systems is different on GAP than on OntoNotes. Particularly interesting is that the current strongest system overall for OntoNotes, namely, Lee et al. (2017), scores best on GAP pronouns but has the largest gender bias on OntoNotes. This perhaps is not surprising given the dominance of masculine examples in that corpus. It is outside the scope of this paper to provide an in-depth analysis of the data and modeling decisions that cause this bias; instead, we release GAP to address the measurement problem behind the bias.

Figure 1 compares the recall/precision trade-off for each system split by Masculine and Feminine examples, as well as combined (Overall). Also shown is a simple syntactic Parallelism heuristic in which subject and direct object pronoun are resolved to names with the same grammatical role (see §4.3). In this visualization, we see a further factor contributing to the low performance of off-the-shelf systems, namely, their low recall. That is, whereas personal pronouns are overwhelmingly anaphoric in both OntoNotes and Wikipedia texts, OntoNotes-trained models are conservative. This observation is consistent with the results for Lee et al. (2013) on the Definite Pronoun Resolution Dataset (Rahman and Ng, 2012),

[14]For each gendered pronoun in a gold OntoNotes cluster, we compare the system cluster with that pronoun. We count a **TP** if the system entity contains at least one gold coreferent NE mention; **FP** if the system entity contains at least one non-gold NE mention, and **FN** if the system entity does not contain any gold NE mention.

	M	F	B	O
Random	43.6	39.3	0.90	41.5
Token Distance	50.1	42.4	0.85	46.4
Topical Entity	51.5	43.7	0.85	47.7
Syntactic Distance	63.0	56.2	0.89	59.7
Parallelism	**67.1**	**63.1**	**0.94**	**65.2**
Parallelism+URL	**71.1**	**66.9**	**0.94**	**69.0**
Transformer-Single	58.6	51.2	0.87	55.0
Transformer-Multi	59.3	52.9	0.89	56.2

Table 6: Performance of our baselines on the development set. Parallelism+URL tests the page-context setting; all other test the snippet-context setting. **Bold** indicates best performance in each setting.

	M	F	B	O
Random	47.5	50.5	1.06	49.0
Token Distance	50.6	47.5	0.94	49.1
Topical Entity	50.2	47.3	0.94	48.8
Syntactic Distance	66.7	66.7	**1.00**	66.7
Parallelism	**69.3**	**69.2**	**1.00**	**69.2**
Parallelism+URL	**74.2**	**71.6**	**0.96**	**72.9**
Transformer-Single	59.6	56.6	0.95	58.1
Transformer-Multi	62.9	61.7	0.98	62.3

Table 7: Performance of our baselines on the development set in the *gold-two-mention* task (access to the two candidate name spans). Parallelism+URL tests the page-context setting; all others test the snippet-context setting. **Bold** indicates best performance in each setting.

on which the system scored 47.2% F1,[15] failing to beat a random baseline due to conservativeness.

4.3 Coreference-Cue Baselines

To understand the shortcomings of state-of-the-art coreference systems on GAP, the upper sections of Table 6 consider several simple baselines based on traditional cues for coreference.

To calculate these baselines, we first detect candidate antecedents by finding all mentions of PERSON entity type, NAME mention type (headed by a proper noun), and, for structural cues, that are not in a syntactic position which precludes coreference with the pronoun. We do not require gender match because gender annotations are not provided by the Google Cloud NL API and, even if they were, gender predictions on last names (without the first name) are not reliable in the *snippet-context* setting. Second, we select among the candidates using one of the heuristics described next.

For scoring purposes, we do not require exact string match for mention alignment—that is, if the selected candidate is a substring of a given name (or vice versa), we infer a coreference relation between that name and the target pronoun.[16]

Surface Cues Baseline cues that require only access to the input text are:

- **RANDOM**. Select a candidate uniformly at random.
- **TOKEN DISTANCE**. Select the closest candidate to the pronoun, with distance measured as the number of tokens between spans.
- **TOPICAL ENTITY**. Select the closest candidate that contains the most frequent token string among extracted candidates.

The performance of RANDOM (41.5 Overall) is lower than an otherwise possible guess rate of ∼50%. This is because the baseline considers all possible candidates, not just the two annotated names. Moreover, the difference between masculine and feminine examples suggests that there are more distractor mentions in the context of feminine pronouns in GAP. To measure the impact of pronoun context, we include performance on the artificial *gold-two-mention* setting, where only the two name spans are candidates for inference (Table 7). RANDOM is indeed closer here to the expected 50% and other baselines are closer to gender-parity.

TOKEN DISTANCE and TOPICAL ENTITY are only weak improvements above RANDOM, validating that our dataset creation methodology controlled for these factors.

Structural Cues Baseline cues that may additionally access syntactic structure are:

- **SYNTACTIC DISTANCE**. Select the syntactically closest candidate to the pronoun. Back off to TOKEN DISTANCE.
- **PARALLELISM**. If the pronoun is a subject or direct object, select the closest candidate with the same grammatical argument. Back off to SYNTACTIC DISTANCE.

Both cues yield strong baselines comparable to the strongest OntoNotes-trained systems (cf. Table 4). In fact, Lee et al. (2017) and PARALLELISM produce remarkably similar output: of the 2,000

[15] Calculated based on the reported performance of 40.07% Correct, 29.79% Incorrect, and 30.14% No decision.

[16] Note that requiring exact string match drops recall and causes only a small difference in F1 performance.

Head \ Layer	L0	L1	L2	L3	L4	L5
H0	46.9	47.4	45.8	46.2	45.8	45.7
H1	45.3	46.5	46.4	46.2	49.4	46.3
H2	45.8	46.7	46.3	46.5	45.7	45.9
H3	46.0	46.3	46.8	46.0	46.6	48.0
H4	45.7	46.3	46.5	47.8	45.1	47.0
H5	47.0	46.5	46.5	45.6	46.2	52.9
H6	46.7	45.4	46.4	45.3	46.9	47.0
H7	43.8	46.6	46.4	55.0	46.4	46.2

Table 8: Coreference signal of a Transformer model on the validation dataset, by encoder attention layer and head.

example pairs in the development set, the two have completely opposing predictions (i.e., Name A vs. Name B) on only 325 examples. Further, the cues are markedly gender-neutral, improving the Bias metric by 9 percentage points in the standard task formulation and to parity in the *gold-two-mention* case. In contrast to surface cues, having the full candidate set is helpful: mention alignment via a non-indicated candidate successfully scores 69% of PARALLELISM predictions.

Wikipedia Cues To explore the *page-context* setting, we consider a Wikipedia-specific cue:

- **URL**. Select the syntactically closest candidate that has a token overlap with the page title. Back off to PARALLELISM.

The heuristic gives a performance gain of 2% overall compared to PARALLELISM. That the feature is not more helpful again validates our methodology for extracting diverse examples. We expect future work to greatly improve on this baseline by using the wealth of cues in Wikipedia articles, including page text.

4.4 Transformer Models for Coreference

The recent Transformer model (Vaswani et al., 2017) demonstrated tantalizing representations for coreference: When trained for machine translation, some self-attention layers appear to show stronger attention weights between coreferential elements.[17] Voita et al. (2018) found evidence for this claim for the English pronouns *it*, *you*, and *I* in a movie subtitles dataset (Lison et al., 2018). GAP allows us to explore this claim on Wikipedia for ambiguous personal pronouns. To do so, we investigate the heuristic:

- **TRANSFORMER**. Select the candidate that attends most to the pronoun.

The Transformer model underlying our experiments is trained for 350k steps on the 2014 English-German NMT task,[18] using the same settings as Vaswani et al. (2017). The model processes texts as a series of **subtokens** (text fragments the size of a token or smaller) and learns three multi-head attention matrices over these, two self-attention matrices (one over the subtokens of the source sentences and one over those of the target sentences), and a cross-attention matrix between the source and target. Each attention matrix is decomposed into a series of feed-forward **layers**, each composed of discrete **heads** designed to specialize for different dimensions in the training signal. We input GAP snippets as English source text and extract attention values from the source self-attention matrix; the target side (German translations) is not used.

We calculate the attention between a name and pronoun to be the mean over all subtokens in these spans; the attention between two subtokens is the sum of the raw attention values between all occurrences of those subtoken strings in the input snippet. These two factors control for variation between Transformer models and the spreading of attention between different mentions of the same entity.

TRANSFORMER-SINGLE Table 8 gives the performance of the TRANSFORMER heuristic over each self-attention head on the development dataset. Consistent with the observations by Vaswani et al. (2017), we observe that the coreference signal is localized on specific heads and that these heads are in the deep layers of the network (e.g., L3H7). During development, we saw that the specific heads which specialize for coreference are different between different models.

The TRANSFORMER-SINGLE baseline in Table 6 is the one set by L3H7 in Table 8. Despite not having access to syntactic structure, TRANSFORMER-SINGLE far outperforms all surface cues above.

[17] See Figure 4 at https://arxiv.org/abs/1706.03762.

[18] http://www.statmt.org/wmt14/translation-task.html.

		PARALLELISM	
		Correct	Incorrect
TRANSF.	Correct	48.7%	13.4%
	Incorrect	21.6%	16.3%

Table 9: Comparison of the predictions of the PARALLELISM and TRANSFORMER-SINGLE heuristics over the GAP development dataset.

	M	F	B	O
Lee et al. (2017)	67.7	60.0	0.89	64.0
Parallelism	69.4	64.4	0.93	66.9
Parallelism+URL	72.3	68.8	0.95	70.6

Table 10: Baselines on the GAP challenge test set.

That is, we find evidence for the claim that Transformer models implicitly learn language understanding relevant to coreference resolution. Even more promising, we find that the instances of coreference that TRANSFORMER-SINGLE can handle is substantially different from those of PARALLELISM; see Table 9.

TRANSFORMER-MULTI We learn to compose the signals from different self-attention heads using extra tree classifiers (Geurts et al., 2006).[19] We choose this classifier because we have little available training data and a small feature set. Specifically, for each candidate antecedent, we:

- Extract one feature for each of the 48 Transformer heads. The feature value is True if there is a substring overlap between the candidate and the prediction of TRANSFORMER-SINGLE.
- Use the χ^2 statistic to reduce dimensionality. We found $k = 3$ worked well.
- Learn an extra trees classifier over these three features with the validation dataset.

That TRANSFORMER-MULTI is stronger than TRANSFORMER-SINGLE in Table 6 suggests that different self-attention heads encode different dimensions of the coreference problem. Though the gain is modest when all mentions are under consideration, Table 7 shows a 4.2 percentage point overall improvement over TRANSFORMER-SINGLE for the *gold-two-mention* task. Future work could explore filtering the candidate list presented to Transformer models to reduce the impact of distractor mentions in a pronoun's context—for example, by gender in the *page-context* setting. It is also worth stressing that these models are trained on very little data (the GAP validation set). These preliminary results suggest that learned models incorporating such features from the Transformer and using more data are worth exploring further.

[19] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html.

Difficulty	Agreement with Gold	#	%
Green	4	631	28.7
Yellow	3	469	21.3
	2	420	19.1
	1	353	16.0
Red	0	328	14.9

Table 11: Analysis of the GAP development examples by the number of systems (out of 4) agreeing with gold.

4.5 GAP Benchmarks

Table 10 sets the baselines for the GAP challenge. We include the off-the-shelf system that performed best Overall on the development set (Lee et al., 2017), as well as our strongest baseline for the two task settings, PARALLELISM[20] and URL.

We note that strict comparisons cannot be made between our *snippet-context* baselines given that Lee et al. (2017) has access to OntoNotes annotations that we do not, and we have access to pronoun ambiguity annotations that Lee et al. (2017) do not.

5 Error Analysis

We have shown that GAP is challenging for both off-the-shelf systems and our baselines. To assess the variance between these systems and gain a more qualitative understanding of what aspects of GAP are challenging, we use the number of off-the-shelf systems that agree with the rater-provided labels (Agreement with Gold) as a proxy for difficulty. Table 11 breaks down the name-pronoun examples in the development set by

[20] We also trained an Extra Tree classifier over all explored coreference-cue baselines (including Transformer-based heuristics), but its performance was similar to PARALLELISM and the predictions matched in the vast majority of instances.

Category	Description	Example (abridged)	#
NARRATIVE ROLES	Inference involving the roles people take in described events	As *Nancy* tried to pull *Hind* down by the arm in the final meters as what was clearly an attempt to drop **her** [...]	28
COMPLEX SYNTAX	Syntactic cues are present but in complex constructions	*Sheena* thought back to the 1980s [...] and thought of idol *Hiroko Mita*, who had appeared on many posters for medical products, acting as if **her** stomach or head hurt	20
TOPICALITY	Inference involving the entity topicality, inc. parentheticals	The disease is named after *Eduard Heinrich Henoch* (1820–1910), a German pediatrician (nephew of *Moritz Heinrich Romberg*) and **his** teacher	15
DOMAIN KNOWLEDGE	Inference involving knowledge specific to a domain, e.g. sport	The half finished 4–0, after Hampton converted a penalty awarded against *Arthur Knight* for handball when *Fleming*'s powerful shot struck **his** arm.	6
ERROR	Annotation error, inc. truly ambiguous cases	When **she** gets into an altercation with *Queenie*, *Fiona* makes her act as Queenie's slave [...]	6

Table 12: Fine-grained categorization of 75 *Red* examples from the GAP development set (no system agreed with the worker-selected name). Underlining indicates the rater-selected name.

Agreement with Gold (the smaller the agreement the harder the example).[21]

Agreement with Gold is low (average 2.1) and spread. Less than 30% of the examples are successfully solved by all systems (labeled *Green*), and just under 15% are so challenging that none of the systems gets them right (*Red*). The majority are in between (*Yellow*). Many *Green* cases have syntactic cues for coreference, but we find no systematic trends within *Yellow*.

Table 12 provides a fine-grained analysis of 75 *Red* cases. When labeling these cases, two important considerations emerged: (1) labels often overlap, with one example possibly fitting into multiple categories; and (2) GAP requires global reasoning—cues from different entity mentions work together to build a snippet's interpretation. The *Red* examples in particular exemplify the challenge of GAP, and point toward the need for multiple modeling strategies to achieve significantly higher scores on the data set.

6 Conclusions

We have presented a data set and a set of strong baselines for a new coreference task, GAP. We designed GAP to represent the challenges posed by real-world text, in which ambiguous pronouns are important and difficult to resolve. We highlighted gaps in the existing state of the art, and proposed the application of Transformer models to address these. Specifically, we show how traditional linguistic features and modern sentence encoder technology are complementary.

Our work contributes to the emerging body of work on the impact of bias in machine learning. We saw systematic differences between genders in analysis; this is consistent with many studies that have called out differences in how men and women are discussed publicly. By rebalancing our data set for gender, we hope to reward systems that are able to capture these complexities fairly.

It has been outside the scope of this paper to explore bias in other dimensions, to analyze coreference in other languages, and to study the impact on downstream systems of improved coreference resolution. We look forward to future work in these directions.

Acknowledgments

We would like to thank our anonymous reviewers and the Google AI Language team, especially Emily Pitler, for the insightful comments that contributed to this paper. Many thanks also to the Data Compute team, especially Ashwin Kakarla, Henry Jicha, and Daphne Luong, for their help with the annotations, and thanks to Llion Jones for his help with the Transformer experiments.

References

David Bamman and Noah A. Smith. 2014. Unsupervised discovery of biographical structure from text. *Transactions of the ACL*, 2:363–376.

[21] Given that system predictions are not independent for the two candidate names for a given snippet, we only focus on the positive coreferential name-pronoun pair when the gold label is either "Name A" or "Name B"; we use both name–pronoun pairs when the gold label is "Neither".

Mohit Bansal and Dan Klein. 2012. Coreference semantics from web features. In *Proceedings of ACL*, pages 389–398, Jeju Island, Korea.

David Bean and Ellen Riloff. 2004. Unsupervised learning of contextual role knowledge for coreference resolution. In *Proceedings of HLT-NAACL*, pages 297–304, Boston, Massachusetts.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of ACL*, pages 33–40, Sydney, Australia.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, Barcelona, Spain.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of NIPS*, pages 4349–4357, Barcelona, Spain.

Mateusz Buda. 2017. A systematic study of the class imbalance problem in convolutional neural networks. Master's thesis, KTH Royal Institute of Technology.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL: HLT*, pages 789–797, Columbus, Ohio.

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of ACL*, pages 1405–1415, Beijing, China.

Ido Dagan and Alon Itai. 1990. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of COLING*, pages 330–332, Helsinki, Finland.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of EMNLP*, pages 1971–1982, Seattle, Washington.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the ACL*, 2:477–490.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Proceedings of ITCS*, pages 214–226.

Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. 2003. *The Measurement of Interrater Agreement*, 3 edition. John Wiley and Sons Inc.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Abbas Ghaddar and Philippe Langlais. 2016a. Coreference in Wikipedia: Main Concept Resolution. In *Proceedings of CoNLL*, pages 229–238, Berlin, Germany.

Abbas Ghaddar and Philippe Langlais. 2016b. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of LREC*, Portorož, Slovenia.

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: Gender bias in Wikipedia. In *Proceedings of the 26th ACM Conference on Social Media*, pages 165–174, Guzelyurt, Northern Cyprus.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference 6: A brief history. In *Proceedings of COLING*, pages 466–471, Copenhagen, Denmark.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the EACL*, pages 1–10, Avignon, France.

Moritz Hardt. 2014. How big data is unfair: Understanding unintended sources of unfairness in data driven decision making. https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of NIPS*, pages 3323–3331, Barcelona, Spain.

Andrew Kehler, Douglas Appelt, Lara Taylor, and Aleksandr Simma. 2004. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of HLT-NAACL*, pages 289–296, Boston, Massachusetts.

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.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of EMNLP*, pages 188–197, Copenhagen, Denmark.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of KR*, pages 552–561, Rome, Italy.

Shasha Liao and Ralph Grishman. 2010. Large corpus-based semantic feature extraction for pronoun coreference. In *Proceedings of the 2nd Workshop on NLP Challenges in the Information Explosion Era (NLPIX)*, pages 60–68, Beijing, August.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of LREC*, pages 1742–1748, Miyazaki, Japan.

Quan Liu, Hui Jiang, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2017. Combing context and commonsense knowledge through neural networks for solving Winograd schema problems. In *Proceedings of AAAI*, pages 315–321, San Francisco, California.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of ACL System Demonstrations*, pages 55–60, Baltimore, Maryland.

Leora Morgenstern, Ernest Davis, and Charles L. Ortiz. 2016. Planning, executing, and evaluating the Winograd Schema Challenge. *AI Magazine*, 37(1):50–54.

Kotaro Nakayama. 2008. Wikipedia mining for triple extraction enhanced by co-reference resolution. In *Proceedings of the ISWC Workshop on Social Data on the Web*, Berlin, Germany.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of KDD*, pages 560–568, Las Vegas, Nevada.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. In *Proceedings of NAACL*, pages 809–819, Denver, Colorado.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of HLT-NAACL*, pages 192–199, New York, New York.

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 CoNLL: Shared Task*, pages 1–40, Jeju, Republic of Korea.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of ICSC*, pages 446–453, Irvine, California.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd Schema Challenge. In *Proceedings of EMNLP-CoNLL*, pages 777–789, Jeju, Republic of Korea.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of NAACL*, New Orleans, Louisiana.

Hee Jung Ryu, Margaret Mitchell, and Hartwig Adam. 2017. Improving smiling detection with race and gender diversity. ArXiv e-prints v2 https://arxiv.org/abs/1712.00193.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017.

No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *Proceedings of the NIPS Workshop on Machine Learning for the Developing World*, Long Beach, California.

Michael Skirpan and Micha Gorelick. 2017. The authority of “fair" in machine learning. In *Proceedings of the KDD Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, Halifax, Nova Scotia.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of ACL-IJCNLP*, pages 656–664, Singapore.

Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *Proceedings of CVPR 2011*, pages 1521–1528, Colorado Springs, Colorado.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. ArXiv e-prints v1 https:/arxiv.org/abs/1806.02847.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, pages 6000–6010, Long Beach, California.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of ACL*, Melbourne, Australia.

Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: Gender assymetrics in Wikipedia. *EPJ Data Science*, 5.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning global features for coreference resolution. In *Proceedings of NAACL-HLT*, pages 994–1004, San Diego, California.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2005. Improving pronoun resolution using statistics-based semantic compatibility information. In *Proceedings of ACL*, pages 165–172, Ann Arbor, Michigan.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of WWW*, pages 1171–1180, Perth, Australia.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of EMNLP*, pages 2941–2951, Copenhagen, Denmark.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL*, New Orleans, Louisiana.