A Hybrid Rule-Based and Neural Coreference Resolution System with an Evaluation on Dutch Literature

Andreas van Cranenburgh, Esther Ploeger, Frank van den Berg, Remi Thüss
University of Groningen
Department of Information Science
a.w.van.cranenburgh@rug.nl

Abstract

We introduce a modular, hybrid coreference resolution system that extends a rule-based baseline with three neural classifiers for the subtasks mention detection, mention attributes (gender, animacy, number), and pronoun resolution. The classifiers substantially increase coreference performance in our experiments with Dutch literature across all metrics on the development set: mention detection, LEA, CoNLL, and especially pronoun accuracy. However, on the test set, the best results are obtained with rule-based pronoun resolution. This inconsistent result highlights that the rule-based system is still a strong baseline, and more work is needed to improve pronoun resolution robustly for this dataset. While end-to-end neural systems require no feature engineering and achieve excellent performance in standard benchmarks with large training sets, our simple hybrid system scales well to long document coreference (>10k words) and attains superior results in our experiments on literature.

1 Introduction

This paper reports on a hybrid rule-based and neural coreference resolution system[1] evaluated on Dutch literary texts. We use neural classifiers for the following three subtasks:

1. Mention span identification;
2. Mention attributes: gender, animacy, number;
3. Pronoun resolution.

These subtasks have been selected based on the expected return on investment given the particular weaknesses of the rule-based model (Lee et al., 2017a) and specific challenges of literary coreference (Rösiger et al., 2018). To keep the approach as simple as possible, we implement these classifiers as independent modules operating in a pipeline.

[1]Code and models are available at https://github.com/andreasvc/dutchcoref

The classifiers can be trained on a laptop without GPU in ten minutes, and are therefore substantially less resource-intensive than state-of-the-art neural models; e.g., SpanBERT (Joshi et al., 2020) requires pre-training a BERT model with span representations on specialized hardware (TPU); moreover, end-to-end neural coreference resolution systems are generally memory and CPU intensive, especially when longer contexts are taken into account (Toshniwal et al., 2020).

The output of coreference resolution is a set of mention spans, partitioned into clusters (example based on Rudinger et al., 2018):

(1) [De chirurg]$_1$ kon [de patiënt]$_2$ niet behandelen. [Hij]$_2$ was [[haar]$_1$ zoon]$_2$!
[The surgeon]$_1$ couldn't treat [the patient]$_2$. [He]$_2$ was [[her]$_1$ son]$_2$!

2 Related Work

Rule-based coreference resolution provides a reasonable baseline (Lee et al., 2011, 2013), and its advantages are that it can exploit global features of entities based on the whole document. In contrast to end-to-end systems, information from parse trees and named-entity recognition can be used, as well as other components from the Natural Language Processing (NLP) pipeline. Feature-based models also use the NLP pipeline, but use machine learning classifiers that make local decisions (mention-pair and mention-ranking architectures), or attempt to take global context into account, but this runs into computational challenges with long documents. End-to-end neural systems do not need the NLP pipeline and are able to optimize all steps of coreference resolution jointly, which has enabled large advances in standard benchmarks (Lee et al., 2017b, 2018; Wu et al., 2020). However, there are several challenges with end-to-end neural models: long documents with long coreference chains (Joshi et al., 2019; Toshniwal et al., 2020), domain and

Proceedings of the 4th Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2021), pages 47–56
Punta Cana, Dominican Republic, November 10–11, 2021.

annotation differences across datasets (Zhu et al., 2021; Poot and van Cranenburgh, 2020), and needing a large number of training examples (Shalev-Shwartz and Shashua, 2016; Glasmachers, 2017). Moreover, gender bias is a general challenge in coreference resolution systems (Rudinger et al., 2018; Webster et al., 2018). Each of these areas is potentially easier to address with a well engineered rule-based or feature-based approach to coreference resolution, and we therefore choose to explore this direction.

Hybrid coreference resolution systems have been presented before; Lee et al. (2017a) present a system in which most steps of the rule-based system are implemented with random forest classifiers. They obtain improvements in accuracy and efficiency, but neural systems have since eclipsed these results. Their classifiers include mention detection and pronoun resolution, which we also pursue in this work.

In addition, previous work shows that neural representations and surface features have complementary strengths (Moosavi and Strube, 2017). This is another sense in which our system is hybrid: we use both manually selected features as well as contextualized word embeddings. Parts of the neural architecture and features are inspired by Clark and Manning (2016), but we use BERT (Devlin et al., 2019) for embedding features instead of static word embeddings, since BERT representations have shown to bring about significant improvements in natural language tasks that rely on the context.

There has been some work on improving detection of mention attributes (animacy, gender, number) using external datasets and machine learning. Bergsma and Lin (2006) extract attributes from a large corpus with dependency parses using heuristic patterns. Orasan and Evans (2007) focus on animacy and use Wordnet and SemCor combined with machine learning. These methods aim to learn general patterns for detecting attributes of noun phrases. In contrast, we will annotate attributes of the entities in our coreference corpus in context and train the classifier on those annotations. We hope to handle more difficult, ambiguous cases which require context with this approach.

Although most coreference resolution systems are trained and evaluated on domains contained in benchmark datasets, such as news texts and phone conversations in the case of OntoNotes, we train and evaluate our hybrid system on Dutch literature. The reason we are interested in the literary domain is that, while literary texts are increasingly subject to computational analysis in the field of digital humanities, there is still a lot of work required to adapt NLP models to the literary domain, of which coreference resolution is a particularly challenging instance. Importantly, the literary domain contains unique characteristics, such as long coreference chains and dialogue, which do not appear in typical benchmark data for coreference resolution (Rösiger et al., 2018; Bamman et al., 2020).

3 Data

We use RiddleCoref (van Cranenburgh, 2019), with the same train/dev/test splits as used in Poot and van Cranenburgh (2020). The corpus consists of 162k tokens of contemporary (2007–2012) bestselling novels in Dutch (translated and original), with a total of 33 documents (fragments of novels), and an average of 4897.4 tokens per document. The entity coreference annotations follow the dutchcoref annotation guidelines. The 38,466 mentions in the corpus have been manually corrected and exclude non-referring expressions.

We did an additional round of corrections on the whole corpus, mostly to fix mention boundaries to exclude relative clauses and remove non-referring expressions (idioms, verbal expressions, negated mentions). We also made small improvements to the mention detection of the rule-based system: bare nouns in conjunctions are extracted as mentions, and subordinate clauses are removed from mention spans, in accordance with the annotation guidelines.

Three mention attributes, namely animacy, gender, and number, have been manually annotated for each of the 11,684 entities in the training and development sets. The gender attribute has four possible values: f (female), m (male), fm (unknown or mixed gender), and n (neuter, non-human). Any gender except neuter implies a human (person) entity; the animacy attribute is therefore implied. Note that Dutch has noun classes with grammatically gendered and neuter words; however, our annotations concern the gender with which individuals are identified. For example, the noun phrase *het meisje* (the girl) is grammatically neuter, but annotated as female, since it would be referred to by female pronouns.

The number attribute has two possible values: sg (singular) and pl (plural; an entity consisting of multiple individuals/objects). We annotate the

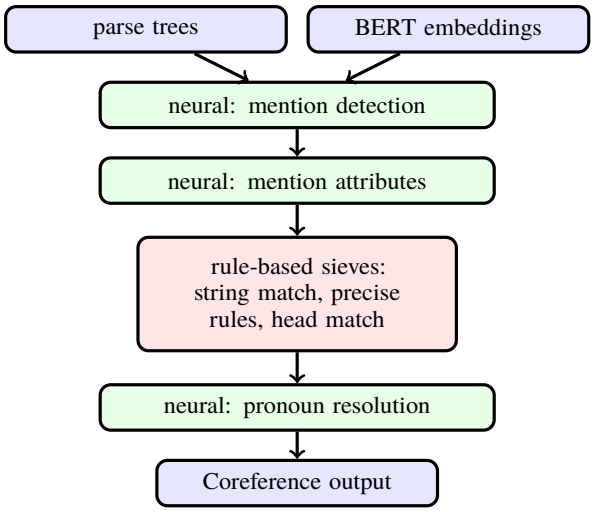

Figure 1: An overview of the hybrid system.

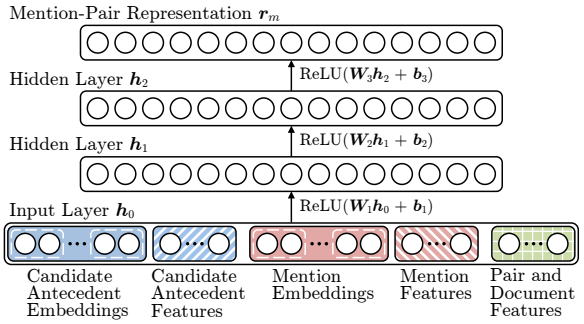

Figure 2: The mention-pair encoder for the pronoun resolution model; the other modules take a single mention as input; figure from Clark and Manning (2016)

semantic number (e.g., "the group" is plural since it is a collective noun that could be referred to by "they"), regardless of the syntactic number.

4 Method

The base system is a rule-based coreference resolution system (van Cranenburgh, 2019) which takes parse trees as input. We extend this system with neural classifiers for three subtasks; see Figure 1 for an overview of our hybrid system.

4.1 Rule-based system

The rule-based system starts by extracting mention candidates from parse trees based on rules. Mention attributes are heuristically assigned based on parse tree features and lexical resources. Mentions are then linked into entity clusters using several "sieves" for linking nominals, names, and finally pronouns. The pronoun resolution step is an implementation of the Hobbs (1978) system using heuristics of recency and syntactic prominence.

We use a feed-forward neural network classifier for the three subtasks (see Figure 2). The input consists of BERT token embeddings and several handpicked features. The network has two dense hidden layers with 500 and 150 neurons, respectively, both with ReLu activation and batch normalization. The output layer is a sigmoid function with the respective binary classification for the subtask and L_2 regularization of 0.05. We apply a dropout of 0.2 to the input layer and 0.5 to each hidden layer and fit the networks with a batch size of 32 and Adam with a learning rate of 0.0001. Each subtask is trained with early stopping until there are 5 successive epochs that do not show an improvement on the validation set.

BERT embeddings are produced using the monolingual, pre-trained BERTje model (de Vries et al., 2019). We use the BERT token embeddings from layer 9, since that layer was shown to perform best for the task of coreference resolution in Dutch (de Vries et al., 2020). For mentions consisting of multiple BERT tokens, we use the mean of the embeddings of all tokens as the mention representation. While neural systems (e.g., Lee et al., 2017b; Bamman et al., 2020) often use a recurrent layer (e.g., LSTM) to obtain contextualized representations of mentions, we follow Joshi et al. (2019) in using BERT embeddings directly. Unlike Joshi et al. (2019), we do not encode BERT embeddings for segments of multiple sentences, but encode each sentence independently.

4.2 Mention Span Classifier

To improve mention detection, we implement a mention span classifier that picks the best mention span from a list of candidates for a given head word (similar to Lee et al., 2017a), or classifies the spans as non-referring if none of them have a probability higher than a threshold (set at > 0.3 in our experiments based on experiments with development data).

Candidates are extracted based on the same syntactic rules as in van Cranenburgh (2019), but include alternative, shorter spans as candidates as well. Since mention spans that incorrectly include an adverb in the first position have been observed frequently in previous work (van Cranenburgh, 2019), we ensure that for each span (n, m), the span (n+1, m) is also considered. The system is trained on gold mention spans annotated in the corpus, as well as negative examples extracted from the parse

trees. During evaluation and prediction, only spans extracted by rules are used as candidates. While it would be possible to train a classifier that works with arbitrary spans as candidates, the rest of the rule-based system depends on parse tree features, and we therefore stick with candidates extracted by parse tree queries.

The features presented to the neural network are as follows:

1. The BERT token embeddings of the first and last token of the span;
2. Whether the rule-based mention detection would extract the span as mention;
3. The grammatical function of the constituent matching the span (subject, object, predicative, apposition);
4. Whether the span contains another NP;
5. Whether the head word of the span is a named entity (PER/LOC/ORG/MISC);
6. The POS tag of the head word (noun, name, pronoun, or verb), the first word (adverb, adjective, punctuation) and the last word (punctuation);
7. The number of words in the mention, histogrammed as in Clark and Manning (2016).

Given the input of example (1), the candidates are (correct mention underlined): {De chirurg, chirurg}, {de patiënt, patiënt}, …, {mijn zoon, zoon}

We also experimented with adding an anaphoricity classifier (e.g., Clark and Manning, 2016; Moosavi and Strube, 2017), but initial experiments did not improve the results, so we leave this for future work. Moreover, mentions could also be classified as singleton or coreferent; however, following Lee et al. (2017a), we have not pursued this, since it is better to leave this decision to later sieves, at which point more global information is available.

4.3 Mention Attributes Classifier

The rule-based system (van Cranenburgh, 2019) detects mention attributes heuristically using parse tree features and several lexicons and lexical resources:

1. Named-entity category and grammatical features;
2. a list of the most common Dutch first names for men and women;
3. gender and animacy attributes for nouns from the Dutch Wordnet equivalent Cornetto;
4. and heuristic number and gender frequencies derived from English web text (Bergsma and Lin, 2006).

To improve the detected attributes, we train a supervised classifier that predicts these attributes for a given mention in a sentence. We manually annotated the mention attributes for each entity in the corpus based on the whole coreference chain. During training, we train and predict these entity attributes for each mention. This means that some data points will be difficult, e.g., predicting the gender of the mention "the person" is not possible without further context. Similarly, *ze* is both a third person singular female pronoun as well as a third person plural pronoun; when not in subject position, gender and number are ambiguous. In early experiments, attributes were only predicted for names and nominals; however, predicting attributes for all mention types (i.e., including pronouns) substantially improved performance. Furthermore, annotating and predicting number, despite being relatively reliably marked syntactically, also boosts performance.

The task is set up as a multi-label classification task such that a mention is assigned probabilities for all possible labels; this multi-task setup means that attributes are trained and predicted jointly. For each attribute, we assign all labels with a probability > 0.5. Experimenting with different thresholds did not improve results. Given this setup, it is possible for the classifier to predict no attributes for a mention, which is interpreted as the attributes being unknown by the system; or a combination of features such as female and neuter, which is not part of the annotations, this is again interpreted as an uncertain feature by the rule-based system.

The features from which the neural network predicts mention attributes are as follows:

1. The averaged BERT token embeddings for the mention;
2. The heuristically detected attributes for gender, animacy, and number;
3. Whether the mention is a subject or object;
4. Whether the mention contains another NP.

Given the input of (1), the expected output is: De chirurg: fm, sg; de patiënt: fm, sg; Hij: m, sg; haar f, sg; haar zoon: m, sg. However, based on the context, the predicted gender of the first two mentions could be more specific.

	recall	prec.	F1
Rule-based	88.3	84.9	86.6
Classifier	**88.8**	**87.2**	**88.0**

Table 1: Mention span classifier results on dev set (N=6434 mentions).

4.4 Pronoun Resolution Classifier

We train a binary classifier on predicting whether a pair of mentions is coreferent; i.e., a mention-pair architecture (Soon et al., 2001). Pairs consist of a pronoun and antecedent candidate. The pronoun is a third person personal, possessive, indefinite, or demonstrative pronoun. The antecedent candidate is a mention within the preceding 22 mentions words relative to the pronoun (this distance is applied during both training and prediction). Mentions with a grammatical function of appositive or determiner are not considered as candidates, since these often lead to incorrect links. We also filter out mention pairs based on binding constraints (i-within-i and co-argument restrictions). Mention pairs are assigned a probability. For each pronoun, the candidate with the highest probability is selected as its antecedent, unless the highest probability is less than a threshold, in which case no antecedent is selected. Based on experiments with the development data, we set the threshold at 0.2.

The features given to the neural network are as follows:

1. The averaged BERT token embeddings for the pronoun, and for the candidate;
2. Candidate mention type (pronoun, noun, name);
3. Whether the grammatical function (subj, obj, etc.) of the pair is the same;
4. Attribute compatibility (gender, animacy, number);
5. Person (1, 2, 3) of candidate, if it is a pronoun;
6. Whether pronoun or candidate occurs in quoted speech;
7. Distance in sentences and words between pronoun and candidate; number of words in candidate. The distances and lengths are histogrammed as in Clark and Manning (2016).

Given the input of (1), the candidates are (correct antecedent underlined): Hij: {de patiënt, De chirurg}; haar: {Hij, de patiënt, De chirurg}.

Unfortunately, several simple features reported to be useful in previous work (e.g., string match, pronoun type, position in sentence) actually decreased development scores in our experiments, which is why we end up with this relatively small list of features. We also considered frequency features: how frequent is the candidate entity in the preceding context or whole document. We have not pursued this since it complicates the implementation as it makes predictions dependent on previous predictions. Another feature which is left for future work is incorporating external knowledge on selectional preferences, as used successfully by Zhang et al. (2019).

	RB	NEU	support
nonhuman	83.3	**94.5**	3073
human	85.0	**95.0**	3361
female	59.1	**73.4**	1347
male	85.8	**89.9**	3002
neuter	80.1	**94.5**	3073
singular	96.7	**98.6**	5224
plural	90.8	**93.6**	1210
macro avg	83.0	**91.3**	20290

Table 2: F1-scores for mention attributes on development set with rule-based baseline using word lists (RB) and the neural classifier (NEU).

5 Results

We first report the results for each module on the development set, and then report the results for the system in various configurations on the development and test sets.

The results for the mention span classifier are shown in Table 1. We obtain a decent improvement over the rule-based method: a difference of 1.4% F1 points for mentions, mainly due to higher precision. The mention recall is limited by the rules for mention candidate extraction and parse tree errors.

The results for the mention attributes classifier are shown in Table 2. We obtain a solid improvement over the baseline, with a macro averaged F1 improvement of 8.3% points and consistent improvement for each label. Female mentions show the largest improvement, but also remain the most difficult to detect. There is a striking contrast with male and neuter mentions, which show higher scores. Animacy detection is also improved substantially, and number to a lesser extent, since the baseline is already high for this attribute.

For coreference evaluation, we use the averaged CoNLL score (Pradhan et al., 2011) and the LEA coreference metric (Moosavi and Strube, 2016; Moosavi et al., 2019). In addition we report mention scores and pronoun accuracy. Pronoun accuracy includes demonstrative and indefinite pronouns in addition to third person personal and possessive pronouns.

See Table 3 for the main coreference results, presented incrementally. The original rule-based model is listed as "dutchcoref", with the modules proposed in this paper listed as span (mention span classifier), attr (mention attributes classifier), and pron (pronoun resolution), respectively. The line "dutchcoref+span" means that the mention span classifier is used, but the rest of the system remains rule-based. For transparency, we report results both on the development and test sets; the parameters and models were tuned only on the development set. Since the annotations and the rule-based system were improved, we report results from Poot and van Cranenburgh (2020) for comparison. Each neural module improves performance scores on the development set, across all metrics. Unfortunately, on the test set the results are less consistent. On several metrics, the rule-based "dutchcoref" performs best, while the pronoun resolution classifier does not improve the pronoun accuracy with respect to the previous line "dutchcoref+span,attr" with results for the rule-based model with neural mention detection and mention attributes, but rule-based pronoun resolution; however, the mention span and attributes modules perform well.

In order to isolate the effect of mention detection (which is known to introduce pipeline errors), we also perform an evaluation on the test set with gold mentions, see Table 4. Here we find that the mention attribute classifier improves the performance across the board. Again the pronoun resolution module does not improve the results compared to the result for 'dutchcoref+attr'. (it does improve the results compared with the purely rule-based system).

All our evaluations include singletons; evaluating without singletons does not change the ranking of the systems on each metric. We conclude that the mention attribute classifier robustly improves the performance, but that the pronoun resolution classifier yields inconsistent results. Finally, while the result is puzzling, a similar result was reported by Poot and van Cranenburgh (2020), where the rule-based system performed better on the test set than on the development set, while the end-to-end neural system showed the opposite effect (better on development set than on test set).

6 Analysis

6.1 Analysis of Differences

It could be that the development and test set differ in difficulty. We consider several basic statistics to compare the two sets. We first consider differences in the out-of-vocabulary (OOV) rate and word frequencies with the Jensen-Shannon distance. We find that the development set actually has a lower OOV rate than the test set, with respect to the training set (16.3% and 13.3%, respectively). The Jensen-Shannon distance shows the same pattern (0.307 vs 0.290, respectively). Moreover, the average sentence length is similar between the development and test sets (18.39 and 18.26, respectively), but higher than the train set (15.51). The development set does have a lower number of mentions (6548 vs 6869) and entities (2643 vs 3008). Finally, the development set has a higher percentage of names (14.9 vs 9.1).

Genre is another potential explanation for the difference. There are four different genres in the RiddleCoref dataset: (Literary) Fiction, Suspense, Romance, and Other. The development set contains 4 Fiction and 1 Other novel, while the test set contains 3 Fiction, 1 Romance, and 1 Suspense novel. We now take a closer look at the difficulty of these genres using an out-of-domain training set. We evaluate the pronoun resolution module on each genre in RiddleCoref; the results are in Table 5. We evaluate on two novels for each genre. The two documents of the Other genre are three chapters from Harry Potter and The Hunger Games, and are therefore considerably longer than the other documents. Since these documents from varying genres all originate from the RiddleCoref training set, we train the pronoun resolution model on a different corpus: SoNar-1 (Schuurman et al., 2010). This model achieved a CoNLL score of 70.76 on the RiddleCoref test set, which is comparable to that of the model trained on RiddleCoref. This is in line with our expectations: on the one hand, SoNaR-1 is much larger with 1 million tokens, providing more training data, but on the other hand, there is a difference in domain. Moreover, as noted by van Cranenburgh (2019) and Poot and van Cranenburgh (2020), there are differences in the annotation

System	set	Mentions			LEA			CoNLL	Pron
		R	P	F1	R	P	F1		Acc
Poot and van Cranenburgh (2020):									
e2e-Dutch	dev	83.12	87.65	85.33	48.37	50.99	49.65	64.81	
dutchcoref	dev	86.85	85.84	86.34	49.18	58.03	53.24	65.91	
This work:									
dutchcoref	dev	88.34	84.87	86.57	50.60	58.06	54.07	66.55	52.16
dutchcoref+span	dev	**88.76**	**87.23**	**87.99**	50.54	59.33	54.58	67.32	53.37
dutchcoref+span,attr	dev	**88.76**	**87.23**	**87.99**	52.27	60.98	56.29	68.80	61.30
dutchcoref+span,attr,pron	dev	**88.76**	**87.23**	**87.99**	**53.68**	**61.27**	**57.23**	**69.37**	**65.38**
Poot and van Cranenburgh (2020):									
e2e-Dutch	test	81.95	89.00	85.33	44.82	50.48	47.48	63.55	
dutchcoref	test	87.65	90.80	89.20	50.83	64.78	56.97	69.86	
This work:									
dutchcoref	test	89.41	90.35	89.88	52.90	**64.93**	**58.30**	70.90	62.82
dutchcoref+span	test	**89.71**	**90.47**	**90.09**	52.02	63.10	57.03	70.07	64.96
dutchcoref+span,attr	test	**89.71**	**90.47**	**90.09**	53.46	63.39	58.00	**71.00**	**68.73**
dutchcoref+span,attr,pron	test	**89.71**	**90.47**	**90.09**	**54.29**	61.34	57.60	70.90	67.81

Table 3: Coreference results on the RiddleCoref dataset (predicted mentions, including singletons).

System	set	Mentions			LEA			CoNLL	Pron
		R	P	F1	R	P	F1		Acc
dutchcoref	test	100	100	100	59.55	69.70	64.22	77.75	69.59
dutchcoref+attr	test	100	100	100	61.16	**70.69**	**65.58**	**78.88**	**74.07**
dutchcoref+attr,pron	test	100	100	100	**61.87**	67.08	64.37	78.44	71.44

Table 4: Coreference results on the RiddleCoref dataset (gold mentions, including singletons).

	Fiction	Romance	Suspense	Other
tokens	9664	4046	4533	34,354
avg sent len	15.5	14.9	17.5	16.1
ment. / ent.	3.0	2.6	2.2	3.5
ent. / tok.	0.08	0.10	0.11	0.07
% pronoun	35.1	42.6	37.9	42.0
% nominals	54.8	45.1	50.2	41.2
% names	10.2	12.2	11.9	16.8
CoNLL score	67.76	71.30	72.45	70.22

Table 5: Evaluation of the pronoun resolution module trained on SoNaR-1, evaluated on different genres with two documents per genre.

scheme of RiddleCoref and SoNaR-1; however, for pronoun resolution, these differences do not prevent the model from achieving a decent score. Comparing the results for the different genres in Table 5 reveals that the genre Fiction resulted in the lowest scores and Suspense resulted in the best scores, the difference being 4.69 percentage points in the CoNLL score. It is quite noticeable that the genres with more tokens in this experiment performed worse. This is in line with the performance of the end-to-end neural model from Poot and van Cranenburgh (2020) where a similar effect was noticed. Furthermore there does not seem to be a clear correlation between the percentage of pronouns and the CoNLL score in this experiment. As some of the most common link errors involve pronouns, genres with more pronouns were expected to result in lower scores. This, however, does not seem to be the case in this sample, as the second best performing genre contained the highest percentage of pronouns and the worst performing genre contained the lowest percentage of pronouns. Lastly, the length of a sentence does not seem to have a substantial effect on the scores. A longer sentence could possibly be more complex with more mentions and therefore create more room for mistakes for the model, how-

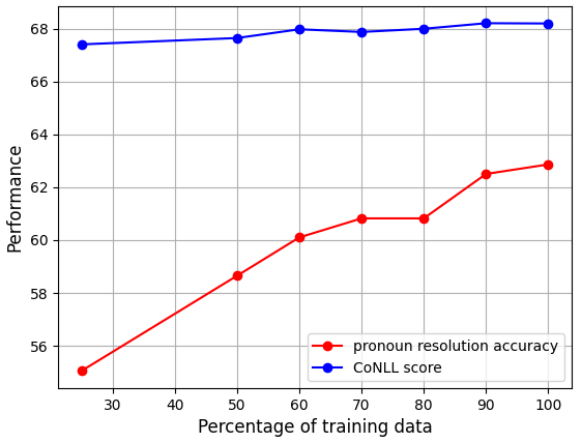

Figure 3: A training curve showing the effect of varying the amount of training data.

ever this does not seem to be the case here. The genre with the highest number of tokens per sentence actually performed the best.

Figure 3 shows training curves for the pronoun resolution model (i.e., these results do not include the mention span and attribute modules), with predicted mentions. From the curve we see the benefit of more training data for pronoun resolution accuracy (and to a lesser extent for the overall CoNLL score). Since the training curve keeps rising without reaching a plateau, we expect that adding more training data will improve pronoun resolution more. Still, since the curve gradually becomes less steep with supplying more data, we expect that there will be diminishing returns as more annotated training data is added.

6.2 Probing for gender bias

The results for the mention attribute classifier show that recognizing female mentions is most difficult. We suspect that the difficulty may lie in mentions that can be both male or female, in which the system may assume male as the most likely label. As a simple probe for gender bias, we experiment with the running example (1) which is a *Winogender* sentence (Rudinger et al., 2018).

Interestingly, the rule-based system correctly identifies *De chirurg* (the surgeon) as male or female (based on the Cornetto lexical resource), while the neural mention attribute classifier predicts it as male. How did this gender bias get introduced? The training data contains only one instance of *chirurg*, which is correctly annotated as male or female, since the context does not identify the surgeon's gender. Another potential source of gender bias is the BERT embeddings. If we present BERT with the following sentence:

(2) De chirurg kon [MASK] patiënt niet behandelen.
The surgeon couldn't treat [MASK] patient.

We find that BERT considers *de, zijn* (the, his) as overwhelmingly most probable, with *een, deze, die* (a, this, that) as distant runner ups, but no female possessive pronoun in the top 5. We therefore conclude that the pre-trained BERTje model has introduced a source of gender bias, which is in line with previous results for Dutch (Chávez Mulsa and Spanakis, 2020). Unless an effective bias mitigation technique is applied, this presents a dilemma: the goal is either to maximize overall accuracy, in which case for example the gender most commonly associated with an occupation is assumed, or gender bias is removed using constraints that lower overall performance.

Moreover, while the mention attribute classifier mistakenly classifies *De chirurg* (the surgeon) as male, the neural pronoun resolution module ignores this misclassification, and correctly links *haar*. This demonstrates the advantage of the neural classifiers which exploit mention attributes as features, but do not treat them as hard constraints, as the rule-based pronoun resolution sieve does.

7 Discussion and Conclusion

We have presented a hybrid coreference resolution system that extends a rule-based baseline with three simple neural classifiers. The classifiers substantially increase the coreference performance in our experiments on Dutch literature, except for pronoun resolution on the test set. The strongest improvements is on pronoun accuracy, which is especially important in longform narrative text.

There are several areas in which the system can be improved. In our approach we erred on the side of simplicity, but in the case of pronoun resolution the approach was too simple, leading to an improvement on the development set but not on the test set. The simplicity can be relaxed in several ways. The modules are trained with gold standard input, but using predictions of previous modules may give better results. If possible, the modules should be trained jointly. Adding more and more varied training data, such as from SoNaR-1 can be expected to yield better results. BERT performs better when finetuned and when encoding segments of 128 to-

kens, as reported by Joshi et al. (2019). Finally, other modules could be added. Anaphoricity classifiers are used in most state-of-the-art systems. In literature, dialogue is particularly important; annotating and predicting speakers of direct speech will help in resolving first and second person pronouns.

Future work should investigate in more detail the trade-offs between rule-based systems using an NLP pipeline and modern end-to-end neural models, especially in the challenging case of long-document coreference in narrative text.

Acknowledgments

We are grateful to Tommaso Caselli and three anonymous reviewers for helpful comments.

References

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of LREC*, pages 44–54.

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

Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating bias in Dutch word embeddings. In *Proceedings of GeBNLP*, pages 56–71.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of ACL*, pages 643–653.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. arXiv:1912.09582.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about BERT's layers? A closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of EMNLP*, pages 4339–4350.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Tobias Glasmachers. 2017. Limits of end-to-end learning. In *Asian Conference on Machine Learning*, pages 17–32. PMLR.

Jerry R Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of EMNLP-IJCNLP*, pages 5807–5812.

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.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of CoNLL*, pages 28–34.

Heeyoung Lee, Mihai Surdeanu, and Dan Jurafsky. 2017a. A scaffolding approach to coreference resolution integrating statistical and rule-based models. *Natural Language Engineering*, 23(5):733–762.

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

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of NAACL*, pages 687–692.

Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. 2019. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In *Proceedings of ACL*, pages 4168–4178.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. In *Proceedings of ACL*, pages 632–642.

Nafise Sadat Moosavi and Michael Strube. 2017. Use generalized representations, but do not forget surface features. In *Proceedings of CORBON*, pages 1–7.

Constantin Orasan and Richard J Evans. 2007. NP animacy identification for anaphora resolution. *Journal of Artificial Intelligence Research*, 29:79–103.

Corbèn Poot and Andreas van Cranenburgh. 2020. A benchmark of rule-based and neural coreference resolution in Dutch novels and news. In *Proceedings of CRAC*, pages 79–90.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of CoNLL*, pages 1–27.

Ina Rösiger, Sarah Schulz, and Nils Reiter. 2018. Towards coreference for literary text: Analyzing domain-specific phenomena. In *Proceedings of LaTeCH-CLfL*, pages 129–138.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of ACL*, pages 8–14.

Ineke Schuurman, Véronique Hoste, and Paola Monachesi. 2010. Interacting semantic layers of annotation in SoNaR, a reference corpus of contemporary written Dutch. In *Proceedings of LREC*, pages 2471–2477.

Shai Shalev-Shwartz and Amnon Shashua. 2016. On the sample complexity of end-to-end training vs. semantic abstraction training. arXiv preprint arXiv:1604.06915.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of EMNLP*, pages 8519–8526.

Andreas van Cranenburgh. 2019. A Dutch coreference resolution system with an evaluation on literary fiction. *Computational Linguistics in the Netherlands Journal*, 9:27–54.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of ACL*, pages 6953–6963.

Hongming Zhang, Yan Song, and Yangqiu Song. 2019. Incorporating context and external knowledge for pronoun coreference resolution. In *Proceedings of NAACL*, pages 872–881.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres. In *Proceedings of ACL*, pages 461–467.