

Coreference Resolution

Performance of state-of-the-art coreference models on different referential expression types

Authors: Hugo Robalino and Liubov Karpova

Abstract

This paper explores the question whether a state-of-the-art coreference model performs differently depending on the type of referential expressions. In order to assess that performance, the model *neuralcoref Neuralcoref 4.0* n.d. was chosen, which uses *spacy spacy models* n.d. language models for English; in this case `en_core_web_sm` and `en_core_web_lg`. The GUM corpus Zeldes 2017 was chosen to test how well *neuralcoref* performs in general. For assessing its general performance with coreference evaluation metrics, the official implementation of the revised coreference scorer used for CoNLL-2011/2012 shared tasks on coreference resolution Pradhan et al. 2014 was used. Finally, a simple mention error analysis for each type of referential expression was done.

1 Introduction

Coreference resolution is the task of finding all expressions that refer to the same entity in a text. Keselj 2009

One can speak of at least four types of referring expressions which can be addressed as entities (definite and indefinite Noun Phrases, pronouns, and proper names). The entities can have a different information status: they can be discourse-new or discourse-old depending on whether the entity was mentioned in the text (text file or corpus file) before. Keselj 2009

There are also more challenging cases of coreference that cannot be so easily detected per entity search. Those are called non-referring expressions, and they still belong to the set of referential expressions. Appositives, Predicative and Prenominal Noun Phrases, Expletives and Generics belong to that list of non-referring expressions. Below are listed some examples (non-referring expressions are in italics).

1. Appositives. Robert Koch Institute, *a German federal government agency and a research institute*, has issued a new press release.
2. Predicative and Prenominal NPs. the 38-year-old became *the country's first female president*
3. Expletives. *It* is frightening right now.
4. Generics. *He* likes apples. They are also good for health.

There are also some linguistic features that help researchers look for coreference if more direct features cannot evoke all possible referential expressions. Some of them include Number Agreement, Person Agreement, Gender or Noun Class Agreement, Case Agreement, some Reflexive pronouns, Recency, Verb Semantics, Definiteness, Semantic Compatibility, and so on. Keselj 2009 Denis and Baldridge 2008

Coreference resolution is a classification task and as any classification task it can be solved with the help of rule-based or learning-based algorithms. Supervised learning-based algorithms present our research interest in the current project report. Most of the current algorithms are mention-ranking, not pairwise-ranking. That is also the case for the paper of our interest written by Denis and Baldridge back in 2008. Denis and Baldridge 2008

They created and applied a mention-ranking algorithm with specialized models for referential expressions to ACE corpus. They reported very optimistic results:

“Evaluated on the ACE data sets, training the model on the train texts, and applying the classifier to the dev-test texts, the model achieves an overall accuracy score of 80.8 percent, compared to a baseline of 59. percent predicting the majority class (“discourse-old”).” Denis and Baldridge 2008

Based on their preceding work, our main interest was whether one can boost the performance of such a classification system with linguistic features by making some of the linguistic features even more fine-grained. For instance, Denis and Baldridge openly discuss in their paper on specialized models for referential expressions about the potential research that could be done by developing a model for demonstrative referential expressions. Also we were interested whether there are any disadvantages, or on the contrary advantages, that are associated with the choice of one or another corpus.

The following subsection addresses the notion of specialized model and referential expressions further. Moreover, we make the description of our research topic in the next section.

1.1 Specialized models for referential expressions

In literature there exists different methods on how to boost coreference resolution systems, including the use of linguistically motivated models. Already most early research was insisting on the idea that linguistic form indicates the status of the corresponding referent in the discourse model. Ariel 1988

For example, when the speaker uses a particular linguistic form, it is associated with a particular level of activation in the addressee’s discourse model. That is not a strict system, but rather a continuum; the above mentioned Ariel Mira developed a model for referential (or accessibility) hierarchy (see below). Ariel 1988

Zero pronouns » Pronouns » Demonstrative pronouns » Demonstrative NPs » Short PNs » Definite descriptions » Full PNs » Full PNs + appositive

In this model for accessibility hierarchy, one can clearly see that different referential expressions, which use a particular linguistic form (e.g. demonstrative pronouns), behave differently when they are used in coreference. For instance, The more on the left the linguistic forms are, the more salient they are, which means the entities they refer to are easier to access from the point the view of the speaker. In more detail, it is known that usage of pronouns requires usage of the referent somewhere close in the text, while proper names at the opposite end of the continuum can be used without previous mentions. This significant difference in behavior between the referential expressions was validated by corpus studies concerned with distribution of different types of elements Ariel 1988.

Pascal Denis and Jason Baldridge developed a strategy to use specialized models for different types of referential expressions. One of the main question for them was how to set types. Some researches like (Ng 2005) created types on the lexical basis (identical words that have the same grammatical function like pronouns he, she, they). This works for closed list categories such as pronouns, but it is unsuitable for proper names or definite descriptions. Therefore, Denis and Baldridge suggested a new strategy “to use different specialized models for different referential expressions” Denis and Baldridge 2008

Some authors chose to use linguistic information to improve the performance of the existing machine learning techniques. As many linguistic features encode some aspects of salience such as distance between mentions, syntactic context, etc, they are very likely to get different sets of parameters for different anaphora types.(Denis and Baldridge 2008)

Denis and Baldridge stated: “This might mean that better parameters are likely to be learned in the context of different models. While the above studies focus primarily on salience, there are of

course other dimensions according to which anaphors differ in their resolution preferences" (Denis and Baldridge 2008).

As they continue "Thus, the resolution of lexical expressions like definite descriptions and proper names is likely to benefit from the inclusion of features that compare the strings of the anaphor and the candidate antecedent (e.g., string matching) and features that identify particular syntactic configurations like appositive structures. This type of information is however much less likely to help in the resolution of pronominal forms. The problem is that, within a single model, such features are likely to receive strong parameters (due to the fact that they are good predictors for lexical anaphors) in a way that might eventually hurt pronominal resolutions." (Denis and Baldridge 2008).

2 Problem description

With our work we decided to close one of the gaps in implementation of Ariel's theory by Denis and Baldridge. They did not cover all the referential types, thus not getting a fully complete image of the whole potential of specialized models. In other words, There was no separate specialized model for demonstrative pronouns and noun phrases. Those were not considered separately and only assessed via specialized model "others" because demonstrative pronouns were very rare in the ACE corpus.

The first initial research topic was to replicate a specialized coreference model for demonstrative pronouns and demonstrative phrases. Nevertheless, it was soon clear that would be out of the question due to a lack of the original code and also because that would not be pertinent as a first topic for exploring coreference since it is much more complex than necessary. Nevertheless, we decided to keep the idea of working with different categories of referential expressions.

In order to keep working with those categories, it was determined to assess how well a state-of-the-art coreference model (neuralcoref) *Neuralcoref 4.0* n.d. performs on those different categories for referential expressions, which are described below by using almost the whole GUM corpus (GUM - The Georgetown University Multilayer Corpus) Zeldes 2017. We first assess the general performance of neuralcoref with different coreference metrics, namely mention detection, MUC, B³, and two versions of CEAF (mention based and entity based), provided by Pradhan et al. 2014, which is also the official implementation of the revised coreference scorer used for CoNLL-2011/2012 shared tasks on coreference resolution. Finally, we assess how well neuralcoref performs on each individual category for referential expressions by making a coreference analysis based on mention errors. The following sections will explain in more detail each of the steps taken to carry out this analysis.

3 Categories for coreferential expressions

Here are the categories considered by Pascal Denis and Jason Baldridge: (i) third person pronouns, (ii) speech pronouns (first and second person pronouns), (iii) proper names, (iv) definite descriptions, and (v) others (i.e., all expressions that don't fall into the previous categories).

Our research is based on examining those 5 categories and the new category "demonstrative pronouns and noun phrases". To sum up, here are all the categories we are going to consider: (i) third person pronouns, (ii) speech pronouns, (iii) proper names, (iv) definite descriptions, (v) demonstrative pronouns and noun phrases, (vi) others (i.e., all expressions that don't fall into the previous categories).

One can also argue the addition of more concrete referential expressions might improve the performance of the specialized coreference model, but this topic is out of the scope of this paper. Below there are listed the numbers in total and in percentages of the types of anaphors by corpus, first the ACE corpus that was used by Denis and Baldridge and secondly the GUM corpus, which is the corpus where we tested our hypothesis.

Table 1: Distribution of different anaphors in ACE corpus, category others omitted

Type	Train corpus	Test Corpus
3rd pronoun	4, 389	1, 093
speech pronoun	2, 178	610
proper names	7, 868	1, 532
definite NPs	3, 124	796
Total	19, 322	4, 599

Table 2: Percent of different anaphors in ACE corpus, category others omitted

Type	Train corpus	Test Corpus
3rd pronoun	22.7 percent	23.7 percent
speech pronoun	11.2 percent	13.2 percent
proper names	40.7 percent	33.3 percent
definite NPs	16.1 percent	17.3 percent

Table 3: Distribution of different anaphors in GUM corpus, category others omitted

Type	Gold Corpus
3rd pronoun	882
speech pronoun	812
proper names	3,190
definite NPs	953
demonstratives	586
Total	6, 423

Table 4: Percent of different anaphors in GUM corpus, category others omitted

Type	Gold Corpus
3rd pronoun	13.7 percent
speech pronoun	12.6 percent
proper names	49.6 percent
definite NPs	14.8 percent
demonstratives	9.1 percent

As GUM corpus is much smaller than ACE corpus, it is clear that it will have less anaphors. However, not all differences in the distribution of categories are not to be explained this way. The rest of discrepancies comes from differences in texts in the corpora (ACE contains of weblogs, broadcast news, newsgroups, broadcast conversations, which have also less formal texts than GUM corpus containing interviews, news, travel guides, how-to guides, academic writing, biographies and fiction) and probably some differences or mistakes in our labeling queries.

As already mentioned we could not enrich the categories with apposition, acronym and semantic compatibility features. Also, labeling category "others" stays still a complex task.

4 Neuralcoref

Several existing implementations of coreference resolution were considered. There were some important restrictions on which system to choose. First of all, we decided that we will not implement the system from scratch but rather use an existing system. Secondly, it had to be a supervised learning system, because future comparison of melting ranking and specialized models in spirit of Denis and Baldridge seems only sensible this way. For example, we chose not to consider the unsupervised and thus depending less on the domain of the train corpus system of Bejan and

Harabagiu 2010.

The system we chose to use as a coreference resolution classifier is called Neuralcoref. It was developed based on the scientific work of Clark and Manning on deep reinforcement learning for mention-ranking coreference models. Clark and Manning 2016a Clark and Manning 2016b

NeuralCoref is in fact simply a pipeline extension for spaCy 2.1+ . The software is able to annotate and resolve coreference clusters, for that Neuralcoref system uses a neural network. *Neuralcoref* 4.0 n.d.

NeuralCoref is ready-to-use extension, which is integrated in spaCy's NLP pipeline. Moreover, it is extensible to new training datasets. NeuralCoref is written in Python/Cython and comes with a pre-trained statistical model for English. Nevertheless, since it is a pipeline extension for spaCy, it can access language models in spaCy for different languages, which is the path we took. *Neuralcoref* 4.0 n.d.

Moreover, since it is an extension to the spacy pipeline, it inherits all the properties that creating a spacy document entails, such as tokens, span, vocabulary, among others.

Another important fact to highlight is that Neuralcoref has also a set of parameters that allow the user to have more control on how Neuralcoref performs. This is the set of parameters that can be adjusted to one's needs:

1. greedyness: how greedy the model is about making coreference decisions
2. max_dist: how many mentions back to look.
3. max_dist_match: The system will consider linking the current mention to a preceding one further than max_dist away if they share a noun or proper noun.
4. blacklist: whether the system resolves or not first and second person pronouns.
5. store_scores: store the scores for the coreference in annotations.
6. conv_dict: conversion dictionary to use for the embedding of rare words.

4.1 Spacy models

Since Neuralcoref is an extension for the pipeline in spaCy, we decided to use two of the three available language models for English, namely en_core_web_sm and en_core_web_lg.

The first model, en_core_web_sm, is trained on the Ontonotes corpus (written text (blogs, news, comments)). It is an English multi-task CNN that is able to assign context-specific token vectors, POS tags, dependency parse and named entities. *spacy models* n.d.

The second Spacy model, en_core_web_lg, is trained on the same corpus, but it comes with GloVe vectors trained on Common Crawl. Because of that the model can assign not only context-specific token vectors, POS tags, dependency parse and named entities, but also word vectors (685k keys, 685k unique vectors in 300 dimensions). That allows the model to take care of more semantic information. *spacy models* n.d.

Even if the language models are trained on the same corpus, they have relevant differences on their capacities and therefore, we believe, will perform differently, even if just slightly.

5 Corpora

5.1 Training corpus

With the choice of neuralcoref system we automatically opted for the usage of training corpus called OntoNotes 5.0., which is the corpus on which the spaCy models were trained. The data for the corpus comes from the telephone conversations, newswire, newsgroups, broadcast news, broadcast conversation and weblogs.

According to the corpus authors the data was labeled with structural information and some semantic information (word sense linked to an ontology and coreference). Weischedel et al. 2011

ACE corpus is a part of OntoNotes 5.0 release, so there is some continuity in our research. The data for broadcast news belongs to a partial selection that used the previously annotated documents by the LDC as part of the ACE program. Weischedel et al. 2011

5.2 Evaluation and test corpus

It was important that the test and evaluation corpus are similar to OntoNotes 5.0, as Neuralcoref tool was trained on it, and we did not aim to retrain on another corpus.

We considered several publicly available corpora including WikiCoref (An English Coreference-annotated Corpus of Wikipedia Articles), The ECB+ Corpus (extension to the EventCorefBank), ACE Corpus and GUM Corpus. Those were the corpora that we were able to identify for our purposes. To the best of our understanding, the corpus coreference annotation layer should have a similar domain to Ontonotes, preferably be available in CoNLL-U format and have enough linguistically annotated layers (not less than ACE Corpus) as well as be freely available. We decided to use the GUM corpus (GUM - The Georgetown University Multilayer Corpus).

It has in total 130 files divided between different genres of topics, which are:

- interviews
- news
- travel guides
- how to guides
- academic writing
- biographies
- fiction
- online forum discussions

We decided not to use the online forum discussions part and simply use the rest of the corpus, because that part was not ready to access as compared to the rest of the corpus. We also decide to keep these genre categories for a more detailed analysis of the results.

The reasons behind this choice were the following:

- the domain of the corpus is much closer to the domain of OntoNotes 5.0 corpus than domain of all other above-mentioned corpora
- it is a freely available corpus (ACE corpus turned out to not be available freely)
- GUM is annotated in many formats, including CoNLL-U
- GUM has enough annotation layers and linguistic features labeled

5.3 Labeling categories

As already mentioned, we chose a different train and test corpora from the one used in the paper (ACE corpus).

The authors of the GUM corpus state on the front page of corpus repository that the easiest way to search in the corpus is using ANNIS. ANNIS stands for ANNotation of Information Structure and offers the interface for multi-layer linguistically annotated corpora.

Annis comes with its own query language called AQL . For example, `infstat="giv" ->coref[type="coref"] entity="person"` provides a list of given persons that are referred to by a coreferent non-anaphor. Here is a complete list of all the operations that are possible within an Annis Search Query:

Complete List of ANNIS Operators:

- direct precedence
- indirect precedence
- direct dominance
- identical coverage
- inclusion
- overlap
- left aligned
- right aligned
- directly near
- indirectly near
- labeled direct pointing relation
- labeled indirect pointing relation
- left-most child
- right-most child
- common parent node
- arity
- length
- root

Following the ANNIS Guidelines and documentation of the paper we started our task of searching for categories and labeling them. Two tables below show differences in two systems, the clarification comes below the tables.

As the chosen GUM corpus has different to ACE corpus annotation layers, we had to choose different from ACE study criteria for searching for a category. Still mostly we tried to stick to the paradigm presented in the paper. They used linguistic form, context, distance, morphosyntactic agreement, semantic compatibility, string similarity, apposition and acronyms. GUM corpus did not support semantic compatibility (no Wordnet information in the corpus, but one of our models supported wordnet vectors), neither apposition nor acronyms are supported in GUM.

As one can see from the table we tried to stick closely to the system presented in the paper , but had to omit three features (semantic compatibility, apposition and acronyms).

Table 5: ACE corpus

Features/Types	3P	12P	PN	Derf-NP	Other
Ling.form	Yes	Yes	Yes	Yes	Yes
Context	Yes	Yes	Yes	Yes	Yes
Distance	Yes	Yes	Yes	Yes	Yes
Agreement	Yes	Yes	Yes	Yes	Yes
Sem.compat.	Yes	Yes	Yes	Yes	Yes
Str.sim.	Yes	Yes	Yes	Yes	Yes
Apposition	No	No	Yes	Yes	No
Acronym	No	No	Yes	No	No

Table 6: GUM corpus

Features/Types	3P	12P	PN	Derf-NP	Demonst	Other
Ling.form	Yes	Yes	Yes	Yes	Yes	Yes
Context	Yes	Yes	Yes	Yes	Yes	Yes
Distance	Yes	Yes	Yes	Yes	Yes	Yes
Agreement	Yes	Yes	Yes	Yes	Yes	Yes
Sem.compat.	No	No	No	No	No	No
Str.sim.	Yes	Yes	Yes	Yes	Yes	Yes
Apposition	No	No	No	No	No	No
Acronym	No	No	No	No	No	No

Table 7: Features used for ANNIS search

Feature	Used Ones
Linguistic Form	Definiteness, ref, NNP, NNPS, NN, NNS, PRP
Context	POS of the preceding or following token
Distance	Distance between tokens
Morphological agreement	Gender, number, person
String similarity	Regular expressions

We also had to slightly alter Distance and Context measurements. We believe that it did not have a significant influence on the results of the queries, because the alterations were solely due to technical (or annotational) differences in measuring of the same linguistic phenomena.

For example, one we accessed most linguistic forms via category of morphological layer "Definiteness" instead of looking for definite descriptions or we could use ref layer labeled for coreference type, information status and coreference edge type annotation (anaphora, cataphora, apposition, coreference),

6 Coreference from Neuralcoref

In order to use the chosen coreference evaluation metrics module, it is important that the key file (annotated file) and the response file (file with predicted coreference chains) have the same tokenization system. As a pipeline from spaCy, Neuralcoref receives the tokenizers and parser from the en_core_web_sm and en_core_web_lg language models. Nevertheless, this produced a different tokenization than the GUM annotated corpus. For this reason, a tokenization map was passed to the spaCy pipeline so that the Neuralcoref output would have the same exact tokenization as the annotated GUM corpus.

One must highlight another important fact, which is that Neuralcoref has different parameters that can be changed. For instance the parameter max_dist, which tells the model how many mentions

back to look when looking for a possible antecedent for the current mention. We decided not to change any parameters and use the default parameters.

7 Evaluation metrics

We used the standard metrics that are described in Luo and Pradhan 2016 and used for evaluation in Denis and Baldrige 2008.

This coreference metrics evaluation module that we use Pradhan et al. 2014 takes three arguments; namely: metric, key file and response file, where key file stands for gold standard file and response file is the file one gets after applying the system. Coreference links happen between same tokens in key and response files. Missing coreference links are the links that are found in key entities, but not in the response entities. Luo and Pradhan 2016

Along with NLP tasks that can be evaluated via standard mention based evaluation metrics there are NLP areas like coreference resolution which metric should be able to handle false-alarm and missing mentions. Below we quickly summarize what metrics are suggested for coreference resolution tasks. Luo and Pradhan 2016

MUC F-Measure

This measure is computed by measuring the common coreference links between the key and the response. It counts how many key links are missing in the response. Luo and Pradhan 2016

B-Cubed F-Measure

This measure is very similar to MUC F-Measure, but it relies differently from MUC F-measure on mentions instead of using links connections. Luo and Pradhan 2016

Constrained Entity-Aligned F-Measure (CEAF)

This measure was suggested in order to solve one of the issues associated with B-Cubed F-Measure. The B-Cubed measure allows an entity to be credited multiple times. The Constrained Entity-Aligned F-Measure (CEAF) fixes the problem by introducing one-to-one alignment. Moreover, only aligned entities are a part of the final score.

There exist two forms of this measure: mention-based CEAF (CEAF_m) and entity-based CEAF (CEAF_e).

8 Results

8.1 Models evaluation results

The first thing that one immediately sees is that recall is greatly higher than precision. That stays so for all possible evaluation metrics that we used, that also does not change much for different models. As we know high precision and low recall means that both classifiers recognize samples correctly, but misses many of them.

It depends on the classification task whether high precision or high recall is more important. In case of our task we had hoped to get a higher recall in order to cover as many coreference cases as possible. Nevertheless, the precision values look optimistic in general for every single genre and coreference metric.

Table 8: Evaluation results en_core_web_lg, first part

	mentions			muc			bcub		
	recall	precision	F1	recall	precision	F1	recall	precision	F1
1-1 academic	8.98	80.66	15.95	11.76	62.75	19.34	5.05	66.77	9.25
1-1 bio	22.9	92.16	36.31	39.34	84.72	53.28	9.89	78.54	17.26
1-1 fiction	29.9	92.3	44.64	39.8	80.63	52.28	15.99	73.89	25.89
1-1 interview	23	88.76	36.1	24.31	72.83	35.88	11.54	72.21	19.47
1-1 news	23.89	82.19	35.93	28.22	65.93	38.47	12.58	64.78	20.66
1-1 voyage	16	81.45	26.59	23.87	63.13	34.32	7.79	61.77	13.71
1-1 whow	18.36	82	29.21	17.82	64.69	27	7.61	66.97	13.48
1-1 MEAN	20.43	85.65	32.1	26.45	70.67	37.22	10.06	69.28	17.1
1-1									

Table 9: Evaluation results en_core_web_lg, second part

	ceafm			ceafe			MEAN		
	recall	precision	F1	recall	precision	F1	recall	precision	F1
1-1 academic	7.3	65.87	12.97	3.52	51.75	6.53	12.81	65.56	7.32
1-1 bio	14.9	59.31	23.56	4.02	46.33	7.32	27.55	72.21	18.21
1-1 fiction	20.77	65.17	31.14	7.86	53.22	13.59	33.51	73.04	22.86
1-1 interview	16.31	63.4	25.61	8.55	49.5	14.42	26.3	69.34	16.74
1-1 news	17.84	63.98	27.1	8.5	50.96	14.27	27.29	65.57	18.21
1-1 voyage	11.63	59.07	19.33	4.41	47.78	8	20.39	62.64	12.74
1-1 whow	11.85	56.2	19.13	6.71	41.89	11.38	20.04	62.35	12.47
1-1 MEAN	14.37	61.86	22.69	6.22	48.78	10.79	23.98	67.25	15.51
1-1									

After presenting the first model results we are presenting the results of the second model. We actually expected the first model (en_core_web_lg) to perform better as it had the advantage of utilising word vectors. That did not turn out to be so, the mean F1 score of the first model (en_core_web_lg) is 15.51, the mean F1 score of the second model (en_core_web_sm) is 15.19.

Table 10: Evaluation results en_core_web_sm, first part

	mentions			muc			bcub		
	recall	precision	F1	recall	precision	F1	recall	precision	F1
1-1 academic	9.24	85.4	16.4	11.58	64.74	19.13	5.13	69.38	9.43
1-1 bio	23.05	91.09	36.4	39.01	83.09	52.66	9.89	77.15	17.19
1-1 fiction	28.87	87.89	42.86	37.96	76.67	49.76	14.32	70.84	23.47
1-1 interview	22.57	86.74	35.33	23.82	71.49	35.12	11.31	70.81	19.04
1-1 news	23.08	78.49	34.83	27.92	64.79	38.2	12.09	62.92	19.94
1-1 voyage	16.29	82.85	27.05	24.61	65.04	35.39	8.07	63.97	14.22
1-1 whow	17.83	79.57	28.36	16.94	61.63	25.69	7.19	64.02	12.8
1-1 MEAN	20.13	84.58	31.6	25.98	69.64	36.56	9.71	68.44	16.58
1-1									

Table 11: Evaluation results en_core_web_sm, second part

	ceafm			ceafe			MEAN		
	recall	precision	F1	recall	precision	F1	recall	precision	F1
1-1 academic	7.45	70.23	13.24	3.72	54.37	6.87	13.01	68.82	7.42
1-1 bio	15.02	58.21	23.61	4.2	46.26	7.61	27.49	71.16	18.23
1-1 fiction	18.91	59.03	28.27	7.31	47.88	12.53	31.38	68.46	21.47
1-1 interview	15.94	62.13	25	8.47	48.25	14.26	25.75	67.88	16.42
1-1 news	17.3	61.94	26.43	8.19	49.02	13.81	26.64	63.43	17.72
1-1 voyage	11.99	61.01	19.92	4.54	50.37	8.26	20.97	64.65	13.1
1-1 whow	11.36	54.03	18.34	6.43	40.39	10.9	19.22	59.93	11.95
1-1 MEAN	14	60.94	22.12	6.12	48.08	10.61	23.49	66.34	15.19
1-1									

8.1.1 Referential expressions

Now that we looked at the general coreference evaluation metrics, we also decided to work with the categories for referential expressions. This is the reason why, apart from labeling categories with the ANNIS tool from the GUM corpus, we also decided to label referential expressions ourselves by using different methods. For instance, closed lists for the third person pronouns, speech pronouns, regular expressions for definite noun phrases, demonstrative pronouns and noun phrases and finally, we decided to create a list of proper names by extracting them from the GUM corpus, where they had the NNP tag from the Penn Treebank. Below there are tables that summarizes in percentages the quantities of different categories for referential expressions that we could find.

Table 12: Gold standard GUM: quantities of referential expressions by category given in percentages

	key						
	speech	third	demonstrative	definite	proper	other	SUM
1-1 academic	2.87	3.6	2.66	21.18	8.95	60.74	4358
1-1 bio	0.76	13.67	1.29	15.71	20.28	48.29	5660
1-1 fiction	18.34	18.99	2.61	17.29	6.14	36.62	4672
1-1 interview	16.94	9.73	3.73	16.85	10.72	42.03	4616
1-1 news	1.55	7.83	1.65	22.06	19.28	47.62	3807
1-1 voyage	2.99	5.26	1.48	22.99	20.11	47.17	3784
1-1 whow	14.96	9.77	3.27	17.18	1.48	53.33	4063
1-1 SUM	8.36	10.21	2.37	18.75	12.52	47.79	30960
1-1							

Below you can find the quantities of referential expressions by category given in percentages for the en_core_web_sm and en_core_web_lg Spacy models:

Table 13: en_core_web_sm Model

	resp-sm						
	speech	third	demonstrative	definite	proper	other	SUM
1-1 academic	10.65	26.52	5.22	27.17	13.91	16.52	460
1-1 bio	0.78	51.78	0.78	10.2	28.39	8.06	1402
1-1 fiction	9.27	50.49	1.04	9.53	10.37	19.31	1543
1-1 interview	18.08	30.29	3.18	16.82	17.07	14.56	1195
1-1 news	0.27	25.32	1.72	27.4	31.83	13.47	1106
1-1 voyage	0	23.11	2.7	27.97	33.65	12.57	740
1-1 whow	1.06	39.74	2.81	30.6	1.64	24.15	853
1-1 SUM	5.9	38.07	2.08	19	19.74	15.19	7299
1-1							

Table 14: en_core_web_lg Model

	resp-lg						
	speech	third	demonstrative	definite	proper	other	SUM
1-1 academic	12.05	26.79	4.91	25	12.95	18.3	448
1-1 bio	0.66	53.1	0.51	9.91	28.7	7.14	1373
1-1 fiction	9.58	52.21	1.19	11.17	11.9	13.95	1513
1-1 interview	18.11	31	2.95	16.93	17.78	13.23	1187
1-1 news	0.35	24.33	1.82	27.63	31.97	13.9	1151
1-1 voyage	0	22.91	2.7	28.03	34.64	11.73	742
1-1 whow	1.41	39.95	2.35	32.55	1.06	22.68	851
1-1 SUM	6.04	38.5	1.97	19.56	20.33	13.6	7265
1-1							

8.2 Referential expressions results

Here are the final and most interesting for our research question and hypothesis results.

Table 15: Mention errors by categories

			academic	bio	fiction	interview	news	voyage	whow
sm	speech	r	41.7	8.17	22.49	24.57	1.92	0	1.45
		p	75	20	83.33	72.64	9.52	0	21.05
		f1	50.79	11.14	29.5	34.61	3.17	0	2.63
	third	r	69.7	93.34	85.65	76.35	84.57	78.79	72.85
		p	88.11	99.24	99.31	93.4	91.5	93.46	89.48
		f1	76.9	96.09	91.65	82.62	87.43	83.81	79.3
	demonstrative	r	16.19	7	6.78	15.2	11.04	23.7	20.23
		p	55	33.33	37.5	52.44	28.57	62.75	60.09
		f1	23.83	11.37	11.29	21.51	15.24	33.89	27.99
	definite	r	11.44	12.97	17.11	24	27.21	20.09	31.99
		p	83.17	79.52	87.03	88.27	74.43	82.74	83.3
		f1	19.77	21.2	27.89	36.97	39.14	31.45	44.04
	proper	r	9	31.96	41.47	32.45	35.58	33.39	11.96
		p	37.54	91.96	72.65	90.82	73	91.33	21.05
		f1	14.39	45.82	50.73	45.49	45.56	47.37	14.54
	other	r	2.06	2.49	9.69	5.82	5.67	2.74	6.36
		p	53.92	55.04	61.82	64.16	60.87	45.82	67.02
		f1	3.9	4.71	15.9	10.09	10.11	5.1	11.34
lg	speech	r	43.79	6.17	22.57	24.54	2.88	0	2.34
		p	75	15	83.33	73.33	14.29	0	26.32
		f1	53.16	8.29	29.7	34.8	4.76	0	4
	third	r	69.09	93.51	87.05	78.09	89.22	78.26	72.69
		p	88.28	98.99	98.86	93.84	96.34	92.86	89.58
		f1	76.45	96.05	92.27	83.66	92.21	83.29	79.17
	demonstrative	r	13.69	4.12	13.92	14.21	12.63	23.01	13.04
		p	48.75	22.5	48.41	51.57	28.57	54.9	54.82
		f1	20.26	6.89	18.84	20.36	16.67	31.92	20.71
	definite	r	10.6	12.54	19.27	24.1	27.63	19.2	34.39
		p	79.53	81.79	87.52	90.29	77.75	78.67	85.67
		f1	18.32	20.48	30.59	37.32	39.86	30.06	46.4
	proper	r	8.35	31.84	44.04	33.36	38.53	33.62	7.89
		p	40.38	91.26	68.33	91.24	76.21	89.68	10.53
		f1	13.15	45.76	51.41	46.52	48.58	47.51	8.77
	other	r	2.12	2.29	9.44	5.94	6.17	2.35	6.76
		p	62.26	61.44	75.22	71.69	63.06	37.52	71.51
		f1	4.07	4.37	16.39	10.43	10.92	4.37	12.11

9 Discussion and error analysis

The first issue we want to address are the changes we had to perform to the initial plan due to some technical issues. As already mentioned we wanted to assess how two (not specialized) spaCy models perform on correctly identifying how many of coreference entities belong to any of our 6 categories. We were not able to do that because we failed at the task of correctly converting ANNIS category files to the neuralcoref system. Still theoretically this task remains one of the goals one can set for the future as we keep thinking that this stays the better option.

We had to assess models' performance using the regular expressions we created extra regular expressions. It might sound as if it ruins our and whole concept of (Denis and Baldrige 2008). After all they invented the whole construct of labeling categories not only with regular expressions. Even though that is true we actually could partially save some of the linguistic features due to the fact that spaCy model train on some linguistic features. One of the models even saves some semantic features.

When one looks closely at the differences in the results of ANNIS GUM distribution and Gold Standard GUM distribution one definitely sees some differences. We omitted category "other" results from the results table, because they happened to turn out not satisfactory and containing some tokens that do not belong to the query.

That explains the big portion of the other category in Neuralcoref search results: the query we gave to Neuralcoref was even less precise. Category other was the least represented category in the paper Denis and Baldrige 2008, it contained only 1763 mentions, similar to speech pronouns. As already mentioned, neuralcoref alone could not assess some linguistic features (e.g. Definiteness), which clarifies some differences from ANNIS in results. That would be very interesting to analyze which categories (real categories or categories-to-be) landed in others category file. That might also explain low recall of our model.

Another important issue to address is whether it makes sense to make the list of linguistic features and referential expressions further and further. We believe that should really depend on the aims of the research and on whether the corpus size and the distribution sizes of planned categories are big enough and are not too disproportional. We mentioned some of the potential categories (indefinite noun phrases, clitics, zero anaphora, etc) and listed some additional linguistic features in the introduction.

Moreover, one can speak with caution about elimination of some of the categories provided they are poorly represented. Even category other does not actually add any important information to the specialized model and can only serve as a statistical tool for evaluation of the research.

Only after completing the evaluation process we came across an interesting system, which proposes a working mechanism of the order of linguistic criteria application. They apply them from highest to lowest precision. Lee et al. 2013 You can find the scheme they used in the References and tables section.

"For example, the first (highest precision) sieve links first-person pronouns inside a quotation with the speaker of a quotation, and the tenth sieve (i.e., low precision but high recall) implements generic pronominal coreference resolution." (Lee et al. 2013)

Additionally, looking at the general metrics for coreference evaluation results, one can notice that both models have a similar performance by having low recall and high precision, which means, they weren't able to identify a considerable amount of mentions and coreference chains in the Gold standard corpus. However, the ones they do identify as mentions and in coreference chains tend to be the right ones. One could further investigate whether changing the parameters from Neuralcoref will produce better results. Also, it seems there is a difference in performance when the genre of the text is taken into account. For instance, both `en_core_web_lg` and `en_core_web_sm` perform

better with fiction genre where their F1 scores are 22.86 and 21.47 respectively.

Finally, looking at the mention errors by category, one can notice some difference. For instance, one can see that the referential expression category with better results is the third person pronoun. That could be due to different factors.

One could argue, third person pronouns are almost exclusively used in coreference and that they almost refer to some entity except in expressions such as 'it is raining'. Due to the fact that they are almost used in coreference, have a specific form and are widely and often used, one could argue the model has learned to recognize it and deal with it in a more efficient way. One could compare the third person referential expression results with ones of the proper name category in the academic genre.

One can notice that the models have a worse performance when dealing with proper names in academic writing and that might be due to the fact that proper names have different forms compared to the closed list that third person pronouns represent, one can also argue that proper names (especially those that are not well known such as the ones from academia) do not appear as often in corpora so that the model learns to recognize them.

From the results, one can conclude that both models perform in a similar way when compared to one another, but they perform differently when dealing with different referential expression categories and even text genre. Therefore, coreference evaluation can benefit from specialized models for different referential categories.

As a final note, one could assess the validity of this research by following the same process and analyze all the GUM corpus by all the already mentioned referential categories except the demonstrative categories by assessing it under the category "others" in order to have a much clearer perspective. In this manner one could get a firmer perspective whether the model Neuralcoref performed better with or without that category. Nevertheless, that was out of the scope of this paper.

10 Annexed Figures and Tables

Table 16: Penntreebank pos tags

Tag	Description
CC	Coordinating conjunction
DT	Determiner
EX	Existential there
FW	Foreign word
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNP	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
POS	Possessive ending
PRP	Personal pronoun
PRPS	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VNG	Verb, gerund or present participle
VCN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WPS	Possessive wh-pronoun
WRB	Wh Adverb

Figure 1: Ontonotes

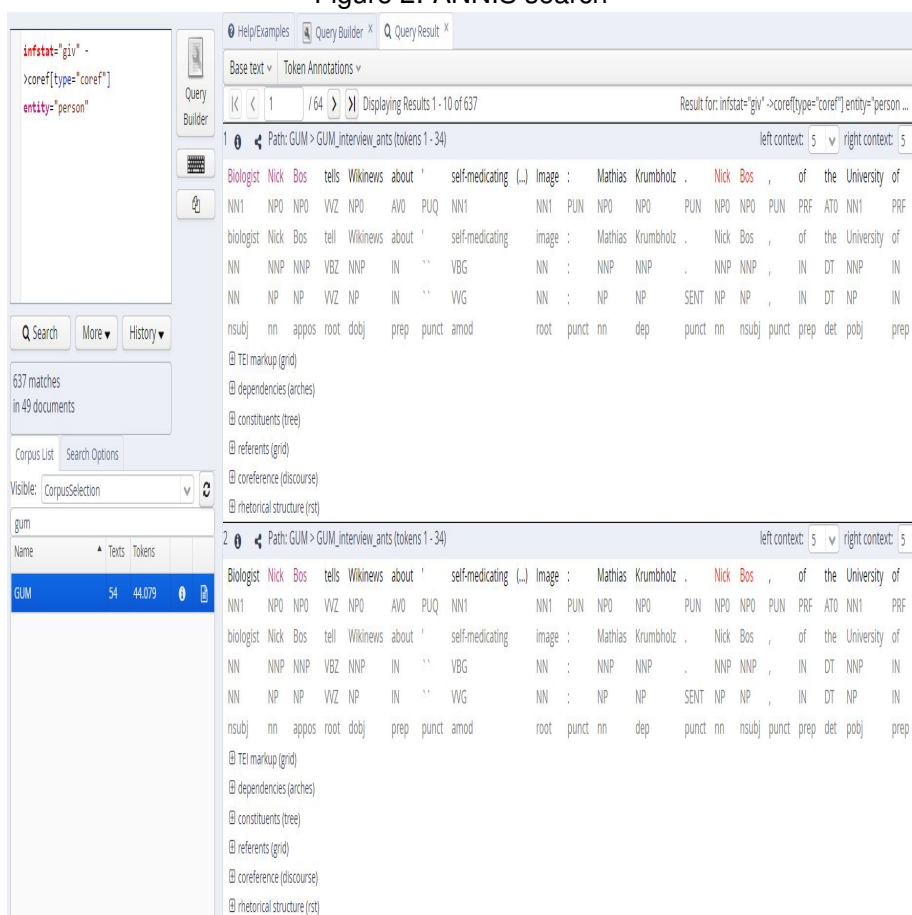
```

1 -----
2
3 Plain sentence:
4 -----
5     I ground the rye on number 6 click-LRB-out of 8-RRB-in my Champion Juicer grinder.
6
7 Treebanked sentence:
8 -----
9     I ground the rye on number 6 click -LRB- out of 8 -RRB- in my Champion Juicer grinder .
10
11 Tree:
12 -----
13     (TOP (S (NP-SBJ (PRP I))
14              (VP (VBD ground)
15                   (NP (DT the)
16                        (NN rye))
17                   (PP-MNR (IN on)
18                           (NP (NP (NML (NN number)
19                                   (CD 6))
20                                   (NN click))
21                                   (-LRB- -LRB-)
22                                   (PP (IN out)
23                                       (PP (IN of)
24                                           (NP (CD 8))))
25                                       (-RRB- -RRB-)))
26                           (PP-LOC (IN in)
27                                   (NP (PRP$ my)
28                                       (NML (NNP Champion)
29                                           (NNP Juicer))
30                                           (NN grinder))))
31                                   (. .)))
32
33 Leaves:
34 -----
35     0 I
36     1 ground
37     2 sense: ground-v.6
38     2 the
39     3 rye
40     4 on
41     5 number

```

<https://www.overleaf.com/project/5e96d64c769a610001dda4dd>

Figure 2: ANNIS search



The screenshot displays the ANNIS search interface. On the left, the 'Query Builder' panel shows a query: `infstat="giv" ->coref[type="coref"] entity="person"`. Below this, it indicates '637 matches in 49 documents'. The 'Corpus List' shows 'gum' selected. The main search results area displays two results, both from the 'gum' corpus. The first result (1) is titled 'Path: GUM > GUM_interview_ants (tokens 1 - 34)' and shows a snippet of text with various annotations. The second result (2) is titled 'Path: GUM > GUM_interview_ants (tokens 1 - 34)' and shows a similar snippet. The interface includes navigation controls at the top and bottom of the results list, and a sidebar on the left for query building and corpus selection.

Query Builder

infstat="giv" ->coref[type="coref"] entity="person"

Q Search More History

637 matches in 49 documents

Corpus List Search Options

Visible: CorpusSelection

gum

Name Texts Tokens

GUM 54 44.073 0

Help/Examples Query Builder Query Result

Base text Token Annotations

1 0 Path: GUM > GUM_interview_ants (tokens 1 - 34) left context: 5 right context: 5

Biologist Nick Bos tells Wikinews about ' self-medicating (...) Image : Mathias Krumbholz . Nick Bos , of the University of

NN1 NPO NPO VVZ NPO AVO PUQ NN1 NN1 PUN NPO NPO PUN NPO NPO PUN PRF ATO NN1 PRF

biologist Nick Bos tell Wikinews about ' self-medicating image : Mathias Krumbholz . Nick Bos , of the University of

NN NNP NNP VBZ NNP IN '' VBG NN : NNP NNP . NNP NNP , IN DT NNP IN

NN NP NP VVZ NP IN '' VVG NN : NP NP SENT NP NP , IN DT NP IN

nsubj nn appos root dobj prep punct amod root punct nn dep punct nn nsubj punct prep det pobj prep

TEI markup (grid)

dependencies (arcs)

constituents (tree)

references (grid)

coreference (discourse)

rhetorical structure (rst)

2 0 Path: GUM > GUM_interview_ants (tokens 1 - 34) left context: 5 right context: 5

Biologist Nick Bos tells Wikinews about ' self-medicating (...) Image : Mathias Krumbholz . Nick Bos , of the University of

NN1 NPO NPO VVZ NPO AVO PUQ NN1 NN1 PUN NPO NPO PUN NPO NPO PUN PRF ATO NN1 PRF

biologist Nick Bos tell Wikinews about ' self-medicating image : Mathias Krumbholz . Nick Bos , of the University of

NN NNP NNP VBZ NNP IN '' VBG NN : NNP NNP . NNP NNP , IN DT NNP IN

NN NP NP VVZ NP IN '' VVG NN : NP NP SENT NP NP , IN DT NP IN

nsubj nn appos root dobj prep punct amod root punct nn dep punct nn nsubj punct prep det pobj prep

TEI markup (grid)

dependencies (arcs)

constituents (tree)

references (grid)

coreference (discourse)

rhetorical structure (rst)

References

- Ariel, Mira (1988). “Referring and accessibility”. In: *Journal of linguistics* 24.1, pp. 65–87.
- Bejan, Cosmin Adrian and Sanda Harabagiu (2010). “Unsupervised event coreference resolution with rich linguistic features”. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1412–1422.
- Clark, Kevin and Christopher D Manning (2016a). “Deep reinforcement learning for mention-ranking coreference models”. In: *arXiv preprint arXiv:1609.08667*.
- (2016b). “Improving coreference resolution by learning entity-level distributed representations”. In: *arXiv preprint arXiv:1606.01323*.
- Denis, Pascal and Jason Baldridge (2008). “Specialized models and ranking for coreference resolution”. In: *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 660–669.
- Keselj, Vlado (2009). *Speech and Language Processing Daniel Jurafsky and James H. Martin (Stanford University and University of Colorado at Boulder) Pearson Prentice Hall, 2009, xxxi+ 988 pp; hardbound, ISBN 978-0-13-187321-6*.
- Lee, Heeyoung et al. (2013). “Deterministic coreference resolution based on entity-centric, precision-ranked rules”. In: *Computational Linguistics* 39.4, pp. 885–916.
- Luo, Xiaoqiang and Sameer Pradhan (2016). “Evaluation metrics”. In: *Anaphora Resolution*. Springer, pp. 141–163.
- Neuralcoref 4.0* (n.d.). URL: <https://github.com/huggingface/neuralcoref>.
- Ng, Vincent (2005). “Supervised ranking for pronoun resolution: Some recent improvements”. In: Pradhan, Sameer et al. (June 2014). “Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 30–35. URL: <http://www.aclweb.org/anthology/P14-2006>.
- spacy models* (n.d.). URL: <https://spacy.io/usage/models>.
- Weischedel, Ralph et al. (2011). “OntoNotes Release 4.0”. In: *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.
- Zeldes, Amir (2017). “The GUM corpus: creating multilayer resources in the classroom”. In: *Language Resources and Evaluation* 51.3, pp. 581–612.