

Short Homework 6

Coreference Resolution

November 13, 2018; due November 20, 2018 (11:59pm)

In this homework, you will be solving the problem of pronominal coreference resolution. You will try to answer the following question: *given a set of candidate mentions in a document, which is a correct [antecedent](#) for each pronoun?* You'll find the following files on bCourses in `Files/SHW6`:

- `shw6.py`: A file with starter code
- `train.coref.data.txt`: training dataset
- `dev.coref.data.txt`: a dataset to help evaluate your model's performance
- `test.coref.data.txt`: the test dataset for us to evaluate your model's performance

Your training data will look like this:

document_id	sentence_word_id	word	pos	mention_ids	entity_ids
0910	0	Humphrey	NNP	1	0910:9
0910	1	Bogart	NNP	1	0910:9
0910	2	is	VBZ		
0910	3	married	VRN		
0910	4	to	TO		
0910	5	Lauren	NNP	2	0910:10
0910	6	Bacall	NNP	2	0910:10
0910	7	.	.		
0910	0	Bogart	NNP	3	0910:9
0910	1	starred	VBD		
0910	2	in	IN		
0910	3	the	DT	4	0910:14
0910	4	Maltese	JJ	4	0910:14
0910	5	Falcon	NN	4	0910:14
0910	6	.	.		
0910	7	He	PRP	5	0910:9
0910	8	also	RB		
0910	9	starred	VBD		
0910	10	in	IN		
0910	11	Casablanca	NP	6	0910:18

The training and development data sets will be in the format above. In the test data set, the **entity_ids** column will be missing.

The **mention_ids** column here specifies the identity of each unique mention in the document (here, [Humphrey Bogart] is mention 1, while the second [Bogart] is a separate mention 3). In the training data, **entity_ids** specifies the entity in the real world to which those mentions refer; coreferent mentions are those that point to the same entity_ids. Here, mention_ids 1, 3 and 5 are all listed as referring to entity 0910:9 and are hence coreferent with each other.

Your task will be to identify the correct antecedent for each pronoun with a mention_ids —i.e., every word for which the POS tag is **PRP** or **PRP\$**, and is marked as a valid mention. You will submit your predictions for the test data: for each pronoun, select the single best mention_ids it corefers with within the same document. Note that several

possible antecedent mentions might be correct solutions; here, for example, selecting 1 ([Humphrey Bogart] as a coreferent mention of “He” is correct, as is selecting 3 ([Bogart]); all refer to the same `entity_ids`. (Again, you’ll know the true `entity_ids` for training and development data, but not for test data.) The evaluation function `check_accuracy` will accept as correct any (or multiple) `mention_ids` as long as they map to the correct `entity_ids`.

For the test file you submit, you should replace the `mention_ids` for the pronoun with your prediction for the correct antecedent mention. A valid response (in which the original `mention_ids` for “He” is replaced with “3” as a choice denoting “Bogart” as the antecedent) could look like this:

document_id	sentence_word_id	word	pos	mention_ids
0910	0	Humphrey	NNP	1
0910	1	Bogart	NNP	1
0910	2	is	VBZ	
0910	3	married	VRN	
0910	4	to	TO	
0910	5	Lauren	NNP	2
0910	6	Bacall	NNP	2
0910	7	.	.	
0910	0	Bogart	NNP	3
0910	1	starred	VBD	
0910	2	in	IN	
0910	3	the	DT	4
0910	4	Maltese	JJ	4
0910	5	Falcon	NN	4
0910	6	.	.	
0910	7	He	PRP	3
0910	8	also	RB	
0910	9	starred	VBD	
0910	10	in	IN	
0910	11	Casablanca	NP	6

In this assignment, you are free to use any resources or libraries that you wish, except those that explicitly perform coreference resolution (`spacy` in particular is a good python library for NLP: <https://spacy.io>).

1 Improving the baseline

a.) We have implemented the `naive_resolver` function. This function predicts that every pronoun refers to the antecedent that is closest to it. This very naive implementation achieves a development accuracy of ~0.53. Your goal is to come up with a method that gains a higher accuracy than this method on the **development** set. Write your code in `shw6.py`. Check your accuracy with the `check_accuracy` function. Include the output of `check_accuracy` function for the development set in the write-up.

b.) Your improved model may use, but not be limited to, refined heuristics, the Hobbs algorithm, logistic regression based classifier using your designed features or any other method. Some approaches have been discussed in the [slides](#). You can use any library/data to achieve this task except a library that explicitly performs coreference resolution. In your write-up, explain your method along with the relevant justification/rationale for the model design.

c.) After creating your model using the training and development data (use the dev data to guard against overfitting), make your predictions on the test set. Your submission is a text file `test_predictions.txt` with a similar structure to `test.coref.data.txt`, except with pronoun `mention_ids` changed to map to the appropriate `mention_ids` of the predicted antecedent. We will test this submission and check if it does better than the naive implementation on the **test** data. A model that makes an overall improvement of at least 8% to the baseline will get full credit.

2 Deliverables

Submit `writeup.pdf`, `shw6.py`, `test_predictions.txt` on gradescope.