Ukrainian Catholic University

Faculty of Applied Sciences

Data Science Master Programme

# Gendered pronoun resolution

## Machine Learning project report

*Authors:*
Dmytro Babenko
Maksym Opirskyi
(contributed equally, order chosen alphabetically)

April 23, 2019

**Abstract**

We are going to investigate several possible solution for gendered pronoun problem. One of them based on simple machine learning approach using One-vs-the-rest (OvR) multiclass/multilabel strategy. Furthermore, we are going to solve this problem using more classical natural language processing algorithm for anaphora and coreference resolution.

**Keywords:** Pronoun resolution · One-vs-the-rest (OvR) multiclass/multilabel strategy

# 1 Introduction

In computational linguistics, coreference resolution is a well-studied problem in discourse. To derive the correct interpretation of a text, or even to estimate the relative importance of various mentioned subjects, pronouns and other referring expressions must be connected to the right individuals. Algorithms intended to resolve coreferences commonly look first for the nearest preceding individual that is compatible with the referring expression. For example, she might attach to a preceding expression such as the woman or Anne, but not to Bill. Pronouns such as himself have much stricter constraints. Algorithms for resolving coreference tend to have accuracy in the 75 % range. As with many linguistic tasks, there is a tradeoff between precision and recall.[1]

In this report, we are going to compare several possible solutions based on machine learning (OvR multiclass strategy) and natural language processing algorithms.

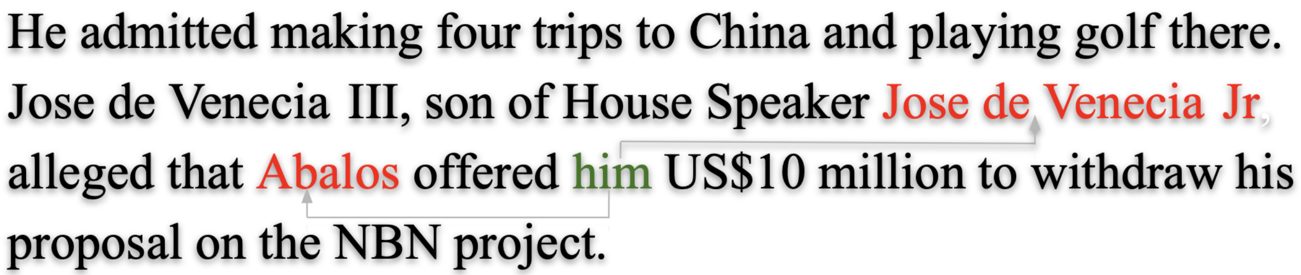
# 2 Importance of the problem

## 2.1 Motivation

Pronoun resolution is a part of coreference resolution, the task of pairing an expression to its referring entity. This is an important task for natural language understanding, and the resolution of ambiguous pronouns is a longstanding challenge. Obtaining effective reference resolution algorithms would give a huge boost to other NLP fields such as machine translation, sentiment analysis, paraphrase detection, summarization, etc.

## 2.2 Problem formulation

We have one or several sentences where two different subjects figure. There is also one pronoun which is related to one of these two subjects.



He admitted making four trips to China and playing golf there. Jose de Venecia III, son of House Speaker Jose de Venecia Jr, alleged that Abalos offered him US$10 million to withdraw his proposal on the NBN project.

Figure 1: The example of pronoun problem.

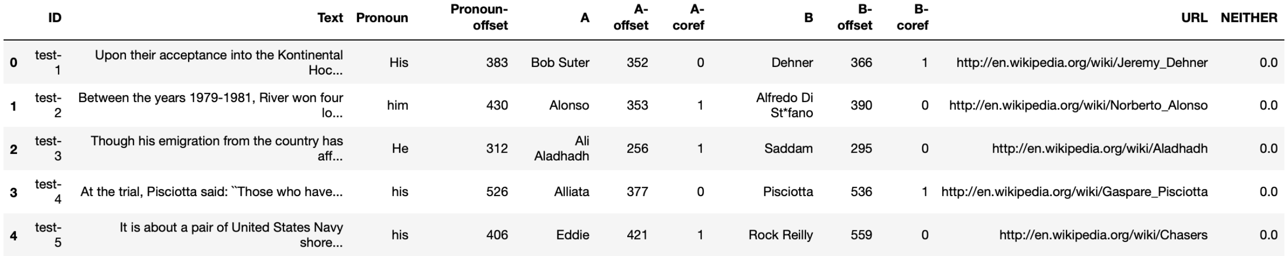
As we can see from Figure 1, we have a sentence and it has two persons *Jose de Venecia Jr.* and *Abalos.*

Additionaly, the sentence has pronoun *him* and we should define to which person this pronoun is related. This is what pronoun resolution solves.

# 3 Data

## 3.1 Collected data

As this project is a part of Kaggle competitions, there are availbale labeled training set on GAP Dataset Github Repo. This link was provided by Kaggle, because unlike many Kaggle challenges, this competition does not provide an explicit labeled training set.



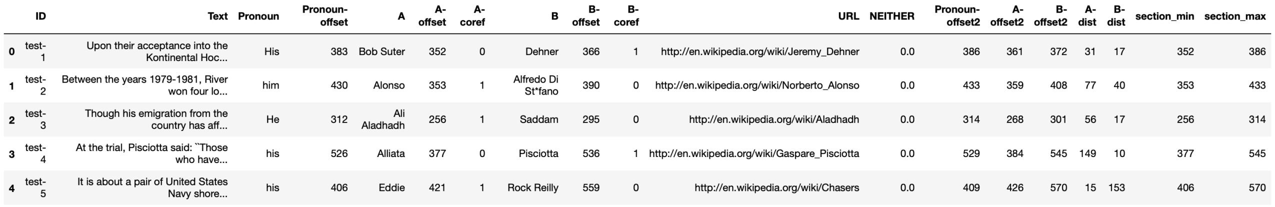
| | ID | Text | Pronoun | Pronoun-offset | A | A-offset | A-coref | B | B-offset | B-coref | URL | NEITHER |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | test-1 | Upon their acceptance into the Kontinental Hoc... | His | 383 | Bob Suter | 352 | 0 | Dehner | 366 | 1 | http://en.wikipedia.org/wiki/Jeremy_Dehner | 0.0 |
| 1 | test-2 | Between the years 1979-1981, River won four lo... | him | 430 | Alonso | 353 | 1 | Alfredo Di St*fano | 390 | 0 | http://en.wikipedia.org/wiki/Norberto_Alonso | 0.0 |
| 2 | test-3 | Though his emigration from the country has aff... | He | 312 | Ali Aladhadh | 256 | 1 | Saddam | 295 | 0 | http://en.wikipedia.org/wiki/Aladhadh | 0.0 |
| 3 | test-4 | At the trial, Pisciotta said: ``Those who have... | his | 526 | Alliata | 377 | 0 | Pisciotta | 536 | 1 | http://en.wikipedia.org/wiki/Gaspare_Pisciotta | 0.0 |
| 4 | test-5 | It is about a pair of United States Navy shore... | his | 406 | Eddie | 421 | 1 | Rock Reilly | 559 | 0 | http://en.wikipedia.org/wiki/Chasers | 0.0 |

Figure 2: Available datasets.

From Figure 2 above, we can see that this set has 11 columns: ID, Text, Pronoun, Pronoun-offset, A, A-offset, A-coref, B, B-offset, B-coref and URL.

- ID - id of sample item
- Text - the main text (one or couple sentences), where all necessary info contains.
- Pronoun - the pronoun in the text which should be related to A or B subject
- Pronoun-offset - start position of the pronoun in the text
- A - first subject in the text
- A-offset - start position of the subject A in the text
- A-coref - boolean value which shows whether considered pronoun related to subject A
- B - second subject in the text
- B-offset - start position of the subject B in the text
- B-coref - boolean value which shows whether considered pronoun related to subject B
- URL - link to the site where this text was found

The example of extended dataset is described below on the Figure 3.



| | ID | Text | Pronoun | Pronoun-offset | A | A-offset | A-coref | B | B-offset | B-coref | URL | NEITHER | Pronoun-offset2 | A-offset2 | B-offset2 | A-dist | B-dist | section_min | section_max |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | test-1 | Upon their acceptance into the Kontinental Hoc... | His | 383 | Bob Suter | 352 | 0 | Dehner | 366 | 1 | http://en.wikipedia.org/wiki/Jeremy_Dehner | 0.0 | 386 | 361 | 372 | 31 | 17 | 352 | 386 |
| 1 | test-2 | Between the years 1979-1981, River won four lo... | him | 430 | Alonso | 353 | 1 | Alfredo Di St*fano | 390 | 0 | http://en.wikipedia.org/wiki/Norberto_Alonso | 0.0 | 433 | 359 | 408 | 77 | 40 | 353 | 433 |
| 2 | test-3 | Though his emigration from the country has aff... | He | 312 | Ali Aladhadh | 256 | 1 | Saddam | 295 | 0 | http://en.wikipedia.org/wiki/Aladhadh | 0.0 | 314 | 268 | 301 | 56 | 17 | 256 | 314 |
| 3 | test-4 | At the trial, Pisciotta said: ``Those who have... | his | 526 | Alliata | 377 | 0 | Pisciotta | 536 | 1 | http://en.wikipedia.org/wiki/Gaspare_Pisciotta | 0.0 | 529 | 384 | 545 | 149 | 10 | 377 | 545 |
| 4 | test-5 | It is about a pair of United States Navy shore... | his | 406 | Eddie | 421 | 1 | Rock Reilly | 559 | 0 | http://en.wikipedia.org/wiki/Chasers | 0.0 | 409 | 426 | 570 | 15 | 153 | 406 | 570 |

Figure 3: The example of extended dataset.

## 3.2 Preprocessed data

Firstly, we tried to train our model based on this original dataset, and we got log loss 0.9198. After that, we constructed new features based on available ones:

- Pronoun-offset2 - end position of pronoun in the text
- A-offset2 - end position of subject A in the text
- B-offset2 - end position of subject B in the text
- A-dis - distance between subject A and pronoun
- B-dis - distance between subject B and pronoun
- section_min - minimum of Pronoun-offset, A-offset, B-offset
- section_max - maximum of Pronoun-offset2, A-offset2, B-offset2

Adding these features reduced log loss to 0.9127.

## 3.3 Linguistic features

Obviously, the feature matrix, which was constructed before cannot give us good accuracy for natural language understanding task. As a result, we acquire additional features using nlp algorithm. We created the Stanford Dependencies (SD) representation [3] for text (one or couple sentences). After that, we know what each subject in the text has Penn Treebank tags [4]. We found top five tags which occur the most often for subject A and B. There are the next tags:

- conj - conjunct
- dobj - direct object
- poss - possession modifier
- pobj - object of a preposition
- nsubj - nominal subject

Then, there was created additional 10 columns which show how many times specific tag occur for specific subject. Using this extended feature matrix we got log loss 0.7692.

# 4 Modeling

## 4.1 Baseline model

To start with, we tried a classical machine learning approach to solve a problem. Since we deal with supervised multiclass classification problem, the common thing is, having the text, to construct features that are able to grasp references between candidates and target pronoun as well as possible. These are typically combination of distance features (e. g. distance between pronoun and candidate, in characters or in words) and linguistic features like syntactic dependencies, gender or number agreement. The features are fed into classifier, the most common is decision tree (ref, ref, ref). We experimented with various classifiers like decision trees, logistic regression, random forest and XGBoost. They all showed similar and quite unpleasant result. Random forest classifier turned out to be the best with a very little difference to other classifiers.

## 4.2 BERT embeddings

The approach described above is often not sufficiently effective, since it depends on manually constructed features which makes it unable to generalize well and hard to maintain. Moreover, the text possesses much greater amount of information which can be used to learn relations between words. With rise of deep learning words could be represented as vectors that carry information about semantic dependencies [6]. Each word is embedded in a vector space such that vectors representing words that have similar semantic meaning are close in this space. This allows to obtain more useful information from the given data, comparing to the limited knowledge that give hand-crafted features. Furthermore, such an approach significantly reduces the need to create features manually, which can be a time consuming end error prone process.

Having the vector representations of words in each training sample one of the possible solutions to coreference problem would be to feed embeddings of candidate subjects and a pronoun into a classifier. To create word embeddings we used BERT which is a state of the art model for general-purpose ”language understanding” [bert ref]. BERT is capable of extracting contextual representations in a deeply bidirectional manner - meaning that it considers both left-side and right-side contexts of a word.

## 4.3 Neural models

Apart from simple classifiers that did not get a satisfactory result we used multi layer perceptron with one hidden layer. We also had the idea that due to the ordered nature of the embeddings (permuting the entries in a word vector would eliminate all its usefulness) convolution layers would help to increase the accuracy of the prediction. So we used one dimensional convolutional neural network.

## 4.4 Vision for improvements (iteration plans)

As an improvement we are going to try to remove non-anaphoric pronominal references. Some contestants reported that their models did bad on the cases where neither of suggested candidates for coreferencing was correct, i. e. when the pronoun was an empty referent. Having this in mind, we can preprocess our data by removing e. g. mentions that are not agreeing in gender.
Another possible approach is to use deep learning. On Kaggle, it is very popular to use BERT for this task. BERT is a neural network capable of generating word embeddings, which could then serve as features in aforementioned classification model. Combining BERT with simple classification tools like we used before, will certainly improve overall performance of the complete model. Moreover, we consider using LSTM model which probably gives us better accuracy.

# 5 Evaluation

Currently, we trained our model using One-vs-the-rest (OvR) multiclass/multilabel strategy. In this case we got log loss 0.7692 which is better than we trained without linguistic features.

# 6 Conclusions

To summarize, we trained our model based on available dataset and got several results. We extended feature matrix by addition of linguistic features and got better result than before. As a next steps, we are going to try more sophisticated neural based approaches which hopefully give us better accuracy. Experiments done so far on kaggle public kernel [5].

# References

[1] https://en.wikipedia.org/wiki/Coreference#Coreference_resolution

[2] https://github.com/google-research-datasets/gap-coreference

[3] https://nlp.stanford.edu/software/dependencies_manual.pdf

[4] https://www.sketchengine.eu/modified-penn-treebank-tagset/

[5] https://www.kaggle.com/dmytrobabenko/simple-ml-model

[6] Rhea Sukthanker, Soujanya Poria, Erik Cambria, Ramkumar Thirunavukarasu: Anaphora and Coreference Resolution: A Review

@articledevlin2018bert, title=BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, author=Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina, journal=arXiv preprint arXiv:1810.04805, year=2018