Semantic Web Technology – Assignment 1

Leon F.A. Wetzel
s3284174
l.f.a.wetzel@student.rug.nl

September 14, 2020

1 Introduction

In this report, we take a closer look at two systems designed for the task of named entity linking. We use different English corpora to evaluate the performance of both systems and to determine if they perform well or not. Altough there are a number of systems capable of named entity linking, we only take a look at **TextRazor** and **OpenTapioca** as this fits in the scope of the actual assignment. As describing both models does not fall in the scope of this assignment, we refer to the documentation of TextRazor (Crayston, 2020) and OpenTapioca (Delpeuch, 2019) for more information about their respective history and architecture.

This report covers several aspects of these systems. We take a look at the data which we will use for the evaluation, we discover more about the metrics used for the evaluation and we take an in-depth look at the actual annotations that were generated by these systems.

2 Data

Our test data for evaluating the entity linking systems is fairly simple. We use two English news articles which will be annotated by both systems independently. The first news article is a sample from McGee (2020), which covers the situation in Britain regarding new, controversial legislation for Brexit. The second news article is a sample from McKeever (2020) and explains more about upcoming COVID-19 vaccins. Both articles have been picked based on presence of named entities and differing topic.

Let's take a look at the corpora. Our sample from McGee (2020) contains the following text:

" Ahead of a very crucial round of talks between London and Brussels over the future trading relationship between the UK and the European Union, the British government made a startling admission: That it would be prepared to break the terms of an international treaty. The threat was relatively technical – over an aspect of the withdrawal agreement that allowed the UK to leave the EU at the end of January – but the admission by a government minister in the House of Commons sent shockwaves through diplomatic circles and raised questions about whether the UK can be trusted on the world stage. "

We can already spot several named entities by hand. The sample from McKeever (2020) contains the following text:

“ More than 150 coronavirus vaccines are in development across the world—and hopes are high to bring one to market in record time to ease the global crisis. Several efforts are underway to help make that possible, including the U.S. government’s Operation Warp Speed initiative, which has pledged $10 billion and aims to develop and deliver 300 million doses of a safe, effective coronavirus vaccine by January 2021. The World Health Organization is also coordinating global efforts to develop a vaccine, with an eye toward delivering two billion doses by the end of 2021. The candidates, like all vaccines, essentially aim to instruct the immune system to mount a defense, which is sometimes stronger than what would be provided through natural infection and comes with fewer health consequences. ”

A similar observation counts for McKeever (2020): we can already spot a number of named entities. We can now feed the articles to both TextRazor and OpenTapioca.

3 Using and evaluating the models

Our way of working comprises of manually entering the samples from McGee (2020) and McKeever (2020) in both TextRazor and OpenTapioca. We can then compare the output from both systems. We measure model performances by taking a look at precision (equation 1) and recall (equation 2), which take as input observed entities and real entities. Observed entities are entities that have been detected by a system, real entities are entities that are real-world entities.

$$precision = \frac{TP}{TP + FP} \quad (1) \qquad recall = \frac{TP}{TP + FN} \quad (2)$$

3.1 TextRazor

We start our journey with TextRazor. We head to the web interface by visiting `www.textrazor.com/demo`. A demonstration sample has already been given, but we will use our own corpora instead. We enter our samples from McGee (2020) and McKeever (2020) and TextRazor returns us the annotations as shown in figures 3.1 and 3.1.

Figure 1: Generated annotations by TextRazor for McGee (2020).

We can see that - next to annotating the corpus - TextRazor splits its output per sentence, presumably for sake of readability. Our first observation is that TextRazor has performed decently in underlining what relevant named entities are. We displayed the results of the performance of TextRazor in tables 2 and 3. We can observe that TextRazor has found all real entities in the sample of McGee (2020), but that it was rather generous in observing entities as well.

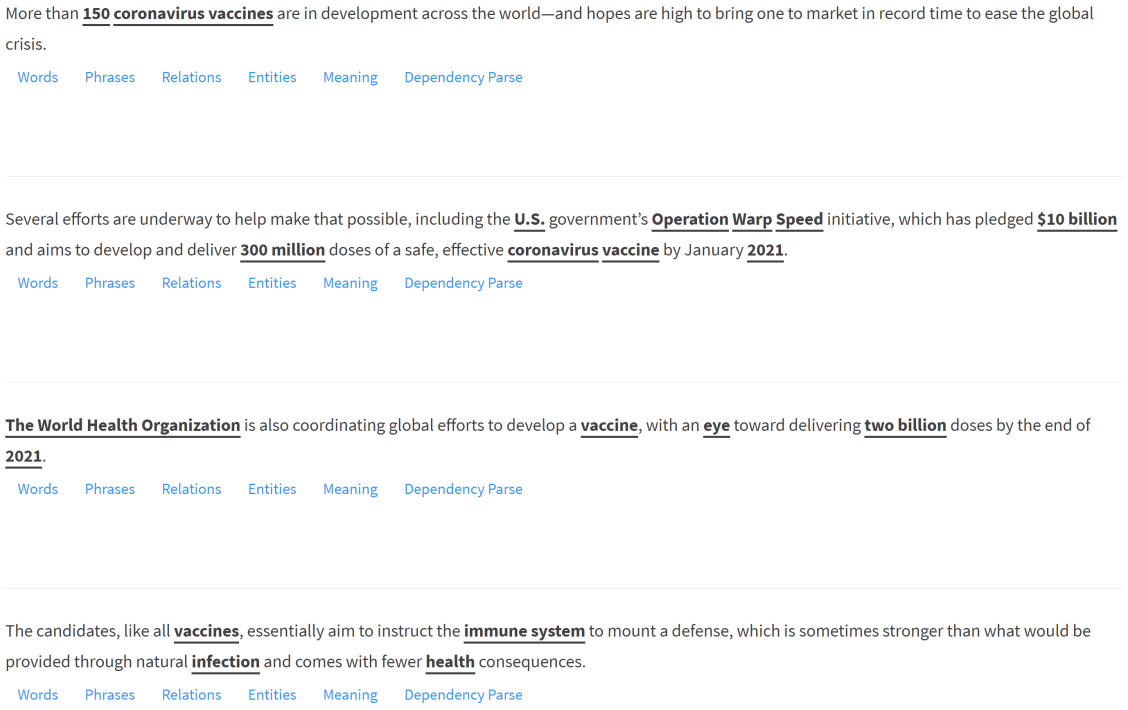

Figure 2: Generated annotations by TextRazor for McKeever (2020).

3.2 OpenTapioca

The web interface of OpenTapioca (`www.opentapioca.org`) has a very minimalistic and simple design when compared to TextRazor, When it comes to functionality, they are nearly identical. Once again, a demonstration sample is already provided and we use our own samples instead.

⟦Ahead⟧ of a very crucial round of talks between ⟦London⟧ and ⟦Brussels⟧ over the future trading relationship between the ⟦UK⟧ and the ⟦European Union⟧, the British government made a startling admission: ⟦That⟧ it would be prepared to break the terms of an international treaty.
⟦The⟧ threat was relatively technical -- over an aspect of the withdrawal agreement that allowed the ⟦UK⟧ to leave the ⟦EU⟧ at the end of ⟦January⟧ -- but the admission by a government ⟦minister⟧ in the ⟦House of Commons⟧ sent shockwaves through diplomatic circles and raised questions about whether the ⟦UK⟧ ⟦can⟧ be trusted on the world stage.

Figure 3: Generated annotations by OpenTapioca for McGee (2020).

⟦More⟧ than 150 coronavirus vaccines are in development across the world—and hopes are ⟦high⟧ to bring one to market in record time to ease the ⟦global⟧ crisis. ⟦Several⟧ efforts are underway to help make that possible, including ⟦the U.S.⟧ government's ⟦Operation Warp Speed⟧ initiative, which has pledged $10 billion and aims to develop and deliver 300 million doses of a safe, effective coronavirus vaccine by ⟦January⟧ 2021. ⟦The World⟧⟦World Health Organization⟧ is also coordinating ⟦global⟧ efforts to develop a vaccine, with an eye toward delivering two billion doses by the end of 2021. ⟦The⟧ candidates, like all vaccines, essentially aim to instruct the immune system to mount a defense, which is sometimes stronger than ⟦what⟧ would be provided through natural infection and ⟦comes⟧ with fewer health consequences.

Figure 4: Generated annotations by OpenTapioca for McKeever (2020).

The output (see figures 3.2 and 3.2) is displayed on the same page, and named entities are *highlighted* by using brackets. A quick observation shows that OpenTapioca makes more directly visible mistakes when compared to TextRazor.

3.3 Overview

All results of our tests can be found in table 1. When comparing the models using their precision and recall values, we can accept that TextRazor outperforms OpenTapioca. When we look at the generated annotations by both TextRazor and OpenTapioca, we can see that they behave very differently from each other. For example, we can see that TextRazor is eager to also link numbers

used to describe amounts of money. Although this does not have to be marked wrong, it shows that the linking goes beyond the classical tagging of named entities.

Corpus	**Model**	**Precision**	**Recall**
McGee (2020)	TextRazor	0.7273	1.0
McGee (2020)	OpenTapioca	0.2857	0.5714
McKeever (2020)	TextRazor	0.75	0.5714
McKeever (2020)	OpenTapioca	0.1765	0.375

Table 1: Test results for TextRazor and OpenTapioca when applied to the McGee (2020) and McKeever (2020) samples.

The performances of OpenTapioca are disappointing when compared to TextRazor. The system sometimes looks confused when tagging named entities. OpenTapioca also seems to be too eager when tagging certain words, such as *Ahead* or *can*. The cause for this phenomenon can be found in that these words could possibly be synonyms for *real* entities; *can* appears to lead to the entity Canada according to OpenTapioca. The system extensively uses Wikidata for the entity linking, hence why the tagging is rather *aggressive* and strange when compared to TextRazor.

References

Toby Crayston. Textrazor, Sep 2020. URL `https://www.textrazor.com/`.

Antonin Delpeuch. Opentapioca: Lightweight entity linking for wikidata. *CoRR*, abs/1904.09131, 2019. URL `http://arxiv.org/abs/1904.09131`.

Luke McGee. Boris johnson's government is threatening to breach international law. it could backfire spectacularly, Sep 2020. URL `https://edition.cnn.com/2020/09/09/uk/boris-johnson-rule-of-law-brexit-intl-gbr/`.

Amy McKeever. Dozens of covid-19 vaccines are in development. here are the ones to follow., Sep 2020. URL `https://www.nationalgeographic.com/science/health-and-human-body/human-diseases/coronavirus-vaccine-tracker-how-they-work-latest-developments-cvd/`.

A Results of TextRazor

Entity	Observed entity?	Real entity?	Normalized entity
London	Yes	Yes	London
Brussels	Yes	Yes	Brussels
UK	Yes	Yes	United Kingdom
the European Union	Yes	Yes	European Union
the British Government	Yes	Yes	Government of the United Kingdom
international treaty	Yes	No	Treaty
Withdrawal agreement	Yes	No	Brexit withdrawal agreement
the UK to leave the EU	Yes	No	Brexit
January	Yes	Yes	2021-01-01T00:00:00.000+00:00
government minister	Yes	Yes	Minister (government)
House of Commons	Yes	Yes	House of Commons

Table 2: Performance report of TextRazor analyzing the McGee (2020) sample.

Entity	Observed entity?	Real entity?	Normalized entity
150	Yes	No	150
coronavirus vaccines	Yes	Yes	Coronavirus vaccine
world	No	Yes	N.A.
crisis	No	Yes	N.A.
U.S.	Yes	No	United States
U.S. government	No	Yes	N.A.
Operation Warp Speed	Yes	Yes	Operation Warp Speed
$10 billion	Yes	No	10000000000
300 million	Yes	No	300000000
2021	Yes	Yes	2021
The World Health Organization	Yes	Yes	World Health Organization
vaccine	Yes	Yes	Vaccine
eye	Yes	Yes	Eye
two billion	Yes	No	2000000000
immune system	Yes	Yes	Immune system
infection	Yes	Yes	Infection
health	Yes	Yes	Health

Table 3: Performance report of TextRazor analyzing the McKeever (2020) sample.