Assignment #4

CS 486 - Fall 2016

Hanumanth Kumar Jayakumar - 20527136

Changqi Du - 20508921

Difei Zhang - 20530592

Jichen Zhao - 20453860

December 3, 2016

Generated Text Samples & Evaluation

In this document, we will see some samples of text generated after enriching our natural language resources. We will also evaluate these texts informally, based on their fluency and interestingness.

Wehrmacht 18th Army

With max graph distance: 1

> Wehrmacht 18th Army was an army. It was led by Georg Von Kuchler.
> It was a participant of the Siege of Leningrad.

With max graph distance: 2

> Wehrmacht 18th Army was an army. An army is a kind of Unit.
> It consists of at least two corps.
> Wehrmacht 18th Army was led by Georg Von Kuchler.
> Georg Von Kuchler was a general. He served Germany.
> Wehrmacht 18th Army was a participant of the Siege of Leningrad.
> The Siege of Leningrad was an engagement. It took place in Leningrad.
> It was defended by the Allies. It was attacked by the Axis powers.
> It was won by the Allies.
> Wehrmacht 16th Army and Wehrmacht 18th Army were participants of it.
> It ended on 1944-01-27T23:59:59Z. It started on 1941-09-08T00:00:00Z.

For this army, we see that with a max graph distance of 1, we have gramatically accurate text. It is fluent, albeit with slightly short sentences. With a max graph distance of 2, we have more interesting text. It is also gramatically accurate for the most part. While it is fluent, it also suffers from the same problem of sentences which are too short. In spoken English, we tend to combine short sentences using conjunctions, and their absence can be felt in the above two passages of text.

Siege of Leningrad

With max graph distance: 1

> The Siege of Leningrad was an engagement. It took place in Leningrad.
> It was defended by the Allies. It was attacked by the Axis powers.
> It was won by the Allies.
> Wehrmacht 16th Army and Wehrmacht 18th Army were participants of it.
> It ended on 1944-01-27T23:59:59Z. It started on 1941-09-08T00:00:00Z.

For this engagement, we see that we have quite rich information with a max graph distance of 1. The text is also gramatically accurate and fluent, albeit with slightly short sentences. Since we have some interesting information here, we will not include the sample of graph distance 2. The fluency of the text can be improved by combining the start and end dates into one sentence, as well as the powers involved. The dates used here are too detailed as well, and we should be able to shorten them into dates used in daily life.

Army

With max graph distance: 1

```
An army is a kind of Unit.
It consists of at least two corps.
It was led by only generals.
```

This is an example of sample text generated for a class rather than an individual. It is not completely gramatically accurate as evidenced by the last sentence. This is one of the problems we faced - how would we remove constraints from the class in its generated text. It is interesting in the sense that it conveys some information about an army by virtue of a brief description.

Battalion

With max graph distance: 1

```
Battalion isn't a kind of corps, division, squad,
platoon, company, and brigade.
It is a kind of Unit. It was led by exactly one Lieutenant Colonel.
It consists of at least four and at most six companies.
```

This is another example of sample text generated for a class rather than an individual. We see a few issues with this text. Firstly, it lists all the classes that it is disjoint with. This conveys too much information which is not required. Humans would generally assume that a battalion is not a kind of any of the other units, unless told otherwise. This brings up an interesting point in Natural Language Generation. We also see the specification that it was led by 'exactly one' lieutenant colonel. This again follows from the above, where we would say 'a' instead of exactly one in spoken or written English. The last sentence conveys some interesting information about this text, it gives us a range of companies which make up a battalion. This could be reworded as 'A battalion consists of between four to six companies'.

From the above sample texts, we can infer that our natural resources enable better text generation for instances, as compared to classes. In all cases, the generated text is gramatically accurate for the most part. One of the shortcomings is that the setences generated are too short. Thus, the fluency can be improved by combining short sentences using conjuctions, as we would do in day to day use.

In terms of interestingness, we observe that some samples provide interesting information with a max graph distance of 1, while others require a graph distance of 2 to produce interesting information. We believe there is a direct correlation between the number of natural language resources we add or improve and the interestingness of the text. To test this out, we compared two samples of text with a few NL names unlinked, and then linked. We observed that the latter produced more interesting text.

Overall, we have some decent natural language resources implemented. As always, this can be improved by further enriching our ontology, as well as our Lexicon entries, NL names, and sentence plans. In general, Natural Language Generation is hard, but we can solve subsets of problems with current techniques.