CSC485 Homework Assignment 2: Write-up

Edwin Chacko

Student Number: 1009149716 *UTORid:* chackoed
edwin.chacko@mail.utoronto.ca

0 Warming up with WordNet and NLTK

(a) Deepest synset.

```
Deepest synset: rock_hind.n.01
Definition: found around rocky coasts or on reefs
Maximum Depth: 19

Depths on each hypernym path:
Depth: 15
Depth: 19
```

1 The Lesk algorithm & word2vec

(d) The reason that extension improves the accuracy. We are now considering more definitions and examples for each sense – the hyponyms, holonyms, and meronyms. Adding hyponyms allow for specific words from a greater topic. holonyms Holoynms are greater topics that the target belongs to. These two allow for more similar words to be used in the two sets, leading to greater potential for overlapping – and thus a greater score. For example, a sentence such as "The wheels on the bus go round and round.", bus provides sense for wheel to be the circular object rather than the verb. This is an example of a holonym. Similarly, in the same sentence, if the target is bus, wheels helps inform that this is the vehicle bus not the computer port bus. This is meronym.

The premise here is that we are now given more words that define the sense and can use these additional words with the context to better gauge the best sense. Overlap will count the signature words that appear in context and now we have more words to work with. Leading to greater accuracy.

(g) Compare how well `lesk_cos_oneside` performs compared to `lesk_cos`. In running the tests, `lesk_cos` gets an accuracy of 37.9% and `lesk_cos_oneside` gets an accuracy of 44.2%, a significant increase. The only change is that we are not using words unique to the signature in our calculations. This helps by filtering out irrelevant words in the signature that don't contribute to the correlation with context. In these dimensions, the context will be orthogonal to the signature suggesting that they are dissimilar. However, when removing these extraneous dimensions, it allows a more meaningful comparison, especially if there are many such orthogonal dimensions. By filtering out noise, `lesk_cos_oneside` focuses on similarity of the intersection of signature with the context as opposed to the holistic similarity of the signature and context.

Example: "The bank was flooded after the heavy rains."
Senses:

- Financial Institution: Words like "money," "account," "loan."
- Riverbank: Words like "shore," "water," "river," "flood."

Filtering for words that are only in the context, we get none of the additional financial words and for riverbank, we get flood. This leads to one dimension of similarity preferred with riverbank over the financial bank. This is what `lesk_cos_oneside` achieves. In contrast, if we use `lesk_cos`, there is one dimension of similarity, but both have 3 dimensions of orthogonality. This suggests a murkier, less clear distinction.

Looking at an example from the test cases: `lesk_cos`:
Synset('head.v.09') study after study – the most recent from the brookings institution – tells us that the best schools are those that are free of outside interference and are governed by a powerful head .

`lesk_cos_oneside`:
Synset('head.n.01') study after study – the most recent from the brookings institution – tells us that the best schools are those that are free of outside interference and are governed by a powerful head .

The meaning of Synset('head.v.09') is a verb for someone leading, while Synset('head.n.01') is a noun referring to the leader of an organization. It is obvious to us that the second one is more appropriate, and it seems that the oneside model was able to pick up on this distinction.

(h) The impact of using binary vectors. Using cosine similarity with binary vectors, when taking the dot product, the only terms that appear are shared terms, like regular cosine similarity. However, now we aren't concerned with the magnitude of the common word appearances, rather just that they exist. In the denominator, we have the count of unique terms in each set multiplied together. This gives the some proportion of shared elements relative to the overall set sizes.

(i) Report your scores in your written submission.

- mfs: 41.6%
- lesk: 39.4%

- lesk_ext: 45.8%
- lesk_cos: 38.1%
- lesk_cos_onesided: 44.1%
- lesk_w2v: 47.3%

(j) How lowercase change the result.

1. mfs: 41.6% – (No change)
2. lesk: 37.9% – (-1.5)%
3. lesk_ext: 46.1% – (+0.3%)
4. lesk_cos: 36.4% – (-1.7%)
5. lesk_cos_onesided: 44.0% – (-0.1%)
6. lesk_w2v: 46.6% – (-0.7%)

It appears that lowercasing leads to negative performance overall. This can be due to loss of sense in cases like “Bank” versus “bank” (financial instituition versus riverbank) or “Turkey” the country versus “turkey” the animal. Additionally, it can cause words that would not have been overlapping, since they were capitalized, to now overlap and increase noise. However, intuition makes me think that these would be counterbalanced by words that were previously capitalized now being found, but it seems that this not the case.

2 Word sense disambiguation with BERT

(a) The reason of using contextual information. Yes, context is indeed necessary for accurate WSD. Word order-invariant methods would not be able to disambiguate precisely that, sentences where the word order is provides important information to the sense of a word. An example is “He almost always tell the truth” versus “He always almost tell the truth.”. The first sentence means he is honest while the second means he struggles with honesty. Static methods, without dependencies, semantic roles, or parts of speech, would not be able to disambiguate these two sentences as they consider the set of words and will not care about the order. Contextual methods are able to use the difference between “almost always” and “always almost”. Another example is “People flooded the bank”. Here, just taking the raw words, word2vec will assume bank to be a riverbank and not a financial bank since flooded often occurs with riverbank. Taking context, we can see that people were doing the flooding which does not occur in a riverbank flooding, helping distinguish these senses.

(c) Sorting the corpus by length. Ordering by size means similar sized sentences will be in the same batch, requiring a similar time to complete the batch. The time to complete a batch is governed by the longest element, so if there is a random distribution of sentence lengths, the smaller ones will have to "wait". When sorting by length, we ensure that each batch is far more similar in size minimizing redundant calculations when there is a long sentence that is still being computed while the others are just sitting there. Padding comes into play as the input is padded to the model input size.

(e) Issue with an arbitrary sentence. For starters, token length of the sentences is bounded, but for an arbitrary sentence, it may exceed the max token length for BERT. Moreover, idioms like "break a leg" or "kick the bucket" have vastly different meanings from the words that compose it. Our methods have no way of correctly identifying the meaning behind such phrases. More generally, multi-word expressions in general may be difficult as it is not identified in lesk. Lastly, word not included in the models vocabulary or in the word2vec vocab will be difficult. These can be domain specific terms or new words that are not yet standardized.

3 Understanding transformers through causal tracing

(c) Causal tracing result plots for the prompt "The Eiffel Tower is located in the city of" with the output "Paris". Your answers and explanations here.

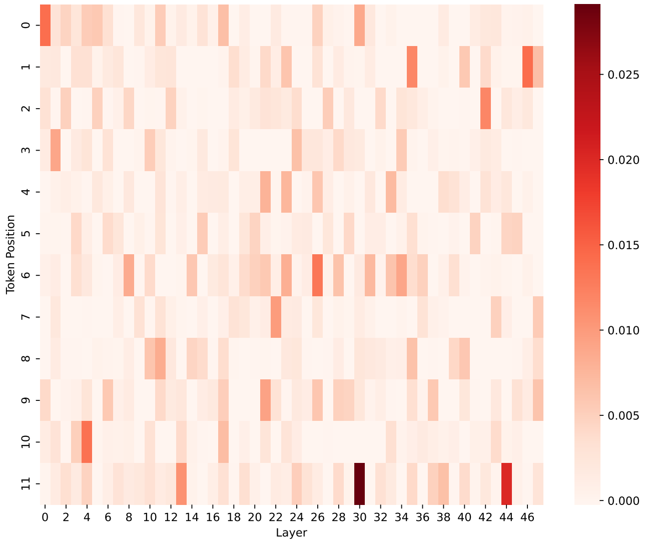

Figure 1: attn

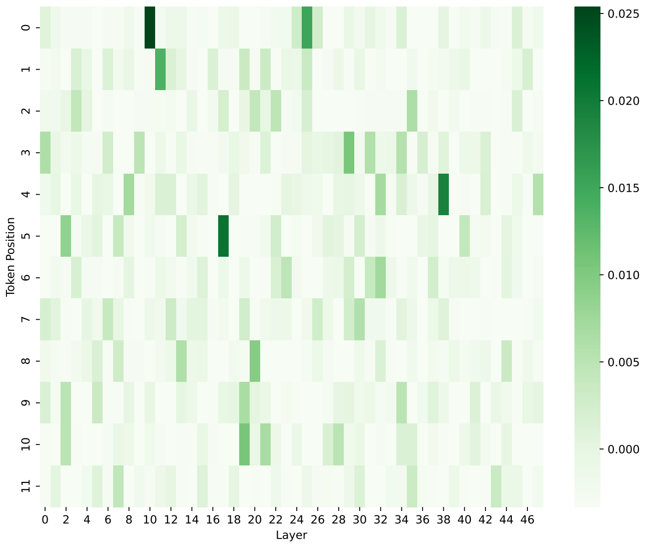

Figure 2: mlp

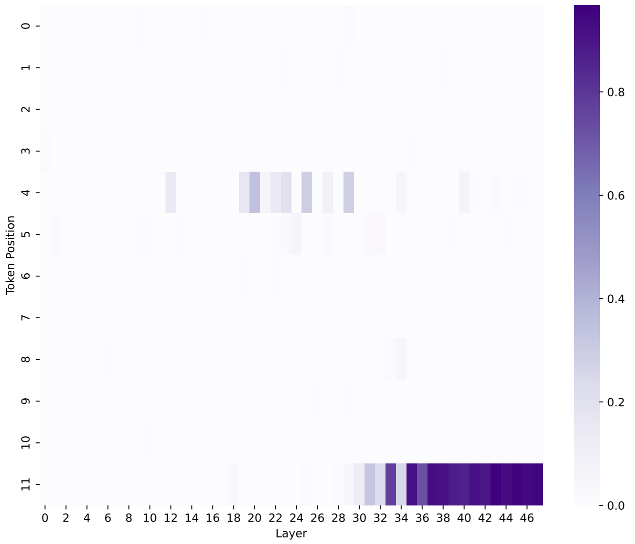

Figure 3: states

(d) Analyze the impact of model size. In causal tracing, patterns increase as model size increases, and is most visible in GPT-2 XL (1.5B parameters), and weak or inconsistent in smaller models like GPT-2 Small (124M). Larger models have more parameters, allowing them to store and retrieve more factual knowledge and better represent more complex relationships. Additionally, they can handle more nuanced prompts beter. Smaller models lack this capacity, leading to reduced causal tracing, as they cant create or maintin deep, relationships between tokens. This suggests that model size directly affects the reliability of factual retrieval and the ability to make meaningful connections.

(e) Identify prompt types. Factual prompts are those that lead to observable patterns. This is because we are looking to see where a fact is stored in the network. Examples include “The capital of France is ”, or “Canada gained independence in the year ”. The similarity between such plots is exactly that they have factual answers, requiring the model to draw upon memory.

(e) Identify prompt types that causal tracing pattern is absent. Some prompts I found that did not have any pattern were ambiguous ones where there were many ways to go about answering. Prompts I tested include, “What is next”, “ What is the weather today” and “The meaning of life is”. This shows that the ability to retrive information is dependent on knowing what info needs to be retrived.