Capturing Context using Attention Mechanisms and NeuraBase.

Kishor J, Neuramatix *13th February 2020*

Abstract

Capturing context in NLP is typically very computationally costly and complex. It typically relies on complex RNNs or CNNs that utilize an encoder and decoder architecture that has to be executed sequentially. Even state of the art transformer models are extremely complex and rely on parallelization to achieve efficiency. NeuraBase can be used to create a similar multi-layered attention grasping mechanism to capture the information that these more complex architectures are able to get.

1. Introduction

1.1 Sequences

Sequences are everywhere, everything that occurs in the world does so in some sequence. Because of this all of the information that comes through our senses is also received in that manner. Sequence modeling is vital to processing any form of input stream. Particularly for NLP and more broadly pattern recognition which is the basis of AI (Vinyals et al., 2014).

1.2 Attention

Attention can be seen as the amount of time that an input in a sequence is relevant for. This is particularly important in sequence modeling as there can be significant relationships between inputs that are not adjacent to each other. It is important to capture these relationships as they contain important information.

In NLP and text processing attention is about extracting and factoring in context while reading a sentence. In any complex NLP task such as Machine Translation, Text Summarization, etc. it is important to capture the context of a word. For example the word 'bank' can have a variety of meanings depending on the context. It could be a river bank or a financial bank. Information like this is necessary for translation as in different languages there may be different words for these two meanings. It is also valuable in text summarization as it's important to ensure that this kind of information is not lost when summarizing information.

1.3 Modeling Attention

Identifying these relationships can be computationally expensive and tricky (Vaswani et al., 2017). Typically they are done using recurrent language models and encoder-decoder architectures (Vaswani et al., 2017). These models and architectures often do computations linking the different positions of the inputs hidden state as a function of a previous hidden state (Vaswani et al., 2017). This process is sequential and is not able to take advantage of parallelization (Vaswani et al., 2017). An alternative method to this as described by Vaswani, is to have multiple layers where each layer is responsible for relationships that are a certain number of steps away (Vaswani et al., 2017). In Vaswani's approach each layer comprises a complex transformer model. They use bitwise operations to do what they call 'Scaled Dot-Product Attention'. By doing this they are able to get probabilities for the relationships between words in that layer. They then combine these by performing 'Multi Headed Attention' functions on them resulting in a single value.

2. Model

This report aims to implement a similar solution using the NeuraBase framework (Hercus, 2009). This is done by producing an output that is able to link words that are related to each other in some meaningful way. Be it to a noun that is

often associated with it. For example, Engineering and Industry or President and United States, or to verbs such as Engineering to build or President to govern. This is used to create a 'dictionary' association between words of sorts. Using this 'dictionary' context and can be extracted from an input sequence.

2.1 Pre-processing

As a start, a first pass of the data was made, flagging high frequency words that were unlikely to mean anything and simply dilute our output without contributing anything meaningful.

To do this the first approach used, was to simply tag and disregard the top 50 ~ 150 words. It was found that disregarding more than this resulted in the loss of some meaning and less than this resulted in too many meaningless words remaining in our output.

Then a second approach was used which involved preprocessing the data removing any words and punctuation that was not relevant. This was done in python using the NLTK (NLTK, 2018) and other preprocessing libraries available. Specifically by replacing stopwords with a specific token.

Furthermore lemmatization (Plisson, Lavrac and Mladenic, n.d.) and noun identification (spaCY, 2017) were considered, but these approaches often resulted in unusable data.

2.2 Data

The first attempt began by using books as the primary source for training data. They were sourced from the Gutenberg Project (Project Gutenberg, 2020). But it was found that the relationships observed were often very specific to books, often relating directly to characters. Along with this many nouns were linked to fictitious places that were present in the books.

As a solution to this data from Wikipedia (Wikipedia.org, 2020) was used. It was found that the relationships extracted were much better as they were not limited to locations and characters in the books. Overall it produced output that had knowledge on more topics and that when analysed gave more meaningful associations. In the end the NeuraBase was trained on about 550mb of Wikipedia data that had already been cleaned and processed in approximately 20 ~ 30mins. It was found that the more data used the better the results were.

2.3 Weights

The pre-processed data is then processed on a sentence by sentence basis at the various attention levels to form a NeuraBase with 6 networks in it. Each network is responsible for relationships 1 - 6 input steps away. It was found that by using different attention profiles different sorts of relationships were favoured.

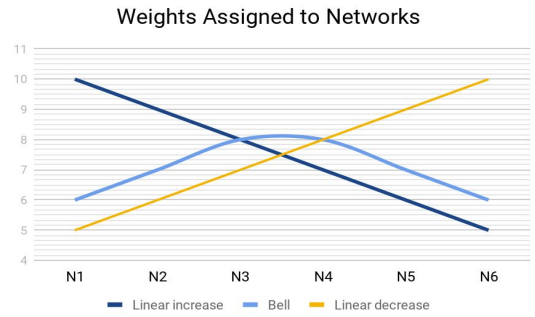

Figure 1. Illustrates the weights assigned to each network.

When a linearly decreasing profile was used, it was found that the results were skewed to words that occured in close proximity but often belong together anyway. A good example of this is shown in Table 1. Where it can be seen that using the linearly declining attention profile results in names of presidents and that using a linearly increasing profile these relationships are avoided and relationships that are related to the role of a president are prioritised. In order to balance both these results a bell profile was used.

Results for the word 'President' under different attention profiles

N1 [10,9,8,7,6,5] N6		N1 [6,7,8,8,7,6] N6		N1 [5,6,7,8,9,10] N6	
Word : PRESIDENT		Word : PRESIDENT		Word : PRESIDENT	
Associate	Score	Associate	Score	Associate	Score
UNITED	9030	UNITED	9149	STATES	9227
STATES	7828	STATES	8955	UNITED	8430
REPUBLIC	4734	REPUBLIC	5027	REPUBLIC	4941
GEORGE	4547	GEORGE	2889	STATE	3150
JOHN	3820	STATE	2617	COUNCIL	2595
FRANKLIN	2649	COUNCIL	2545	GEORGE	2518
REAGAN	2620	JOHN	2543	NATIONAL	2429
ROOSEVELT	2535	NATIONAL	2248	VICE	2321
OBAMA	2514	VICE	2237	JOHN	2315

Table 1. Top 10 results for the word 'president', under different weights.

2.4 Results

Here are some examples of the results achieved. Only the top 10 associations for each example were printed .

```
Word : presidential Freq: 5586
-- Associate : election Score : 9895
-- Associate : elections Score: 3032
-- Associate : candidate Score : 282
-- Associate : freedom Score : 1680
-- Associate : campaign Score : 1551
-- Associate: candidates Score: 1507
-- Associate : results Score : 1409
-- Associate : county Score : 1409
-- Associate : medal Score : 1327
-- Associate: nomination Score: 1213

Word : episode Freq: 5570
-- Associate : series Score : 3038
-- Associate : season Score : 1453
-- Associate: television Score: 1198
-- Associate : show Score : 846
-- Associate : aired Score : 772
-- Associate : tv Score : 750
-- Associate : first Score : 549
-- Associate : titled Score : 531

Word : american Freq: 94021
-- Associate : player Score : 54275
-- Associate : actor Score : 45649
-- Associate : actress Score : 32178
-- Associate: politician Score:25961
-- Associate : author Score : 24543
-- Associate : producer Score: 23334
-- Associate : singer Score : 19551
-- Associate : football Score: 18142
-- Associate : baseball Score: 18086

Word : humanitarian Freq: 728
-- Associate : aid Score : 424
-- Associate : law Score : 270
-- Associate : award Score : 222
-- Associate : assistance Score: 216
-- Associate : relief Score : 172
-- Associate : disaster Score : 127
-- Associate : missions Score : 126
-- Associate : work Score : 124
-- Associate : efforts Score : 124

Word : trademark Freq: 728
-- Associate : office Score : 254
-- Associate : name Score : 196
-- Associate:infringement Score: 132
-- Associate : owned Score : 102
-- Associate : company Score : 93
-- Associate : patent Score : 84
-- Associate : use Score : 65
-- Associate : brand Score : 65
-- Associate : used Score : 63
-- Associate : style Score : 60

Word : human Freq: 13886
-- Associate : rights Score : 15904
-- Associate : beings Score : 3030
-- Associate: development Score:1844
-- Associate : nature Score : 1700
-- Associate : body Score : 1669
-- Associate : life Score : 1593
-- Associate : watch Score : 1178
-- Associate : history Score : 1035
-- Associate: violations Score :1023
-- Associate : activity Score : 993

Word : bbc Freq: 4299
-- Associate : radio Score : 4595
-- Associate : news Score : 3424
-- Associate : day Score : 2645
-- Associate : world Score : 1724
-- Associate : series Score : 1652
-- Associate: television Score :1518
-- Associate : service Score : 966
-- Associate : one Score : 833
-- Associate : programme Score : 828
-- Associate : broadcast Score : 821

Word : chemical Freq: 4105
-- Associate : weapons Score : 1780
-- Associate: reactions Score : 1384
-- Associate : element Score : 1138
-- Associate : reaction Score : 909
-- Associate : elements Score : 848
-- Associate: engineering Score :817
-- Associate: properties Score : 736
-- Associate : symbol Score : 731
-- Associate: biological Score : 704
-- Associate : compounds Score : 628

Word : mathematics Freq: 4283
-- Associate : science Score : 1370
-- Associate : physics Score : 1317
-- Associate: university Score : 928
-- Associate : theory Score : 793
-- Associate : computer Score : 655
-- Associate: philosophy Score : 545
-- Associate : sciences Score : 471
-- Associate : logic Score : 461
-- Associate : set Score : 409
-- Associate: mathematical Score:408

Word : nuclear Freq: 6680
-- Associate : weapons Score : 6568
-- Associate : power Score : 3220
-- Associate : war Score : 1717
-- Associate : test Score : 1275
-- Associate : weapon Score : 1123
-- Associate : energy Score : 1092
-- Associate : reactors Score : 1033
-- Associate : plant Score : 1028
-- Associate : reactor Score : 948
-- Associate : program Score : 931
Word : collection Freq: 7194
-- Associate : stories Score : 2044
-- Associate : short Score : 1326
-- Associate : art Score : 998
-- Associate : poems Score : 874
-- Associate : essays Score : 868
-- Associate : library Score : 809
-- Associate : including Score : 805
-- Associate : works Score : 773
-- Associate: university Score : 683
-- Associate : published Score : 665

Word : health Freq: 7181
-- Associate : care Score : 3899
--Associate:organization Score: 1894
-- Associate : services Score : 1699
-- Associate : education Score :1170
-- Associate : problems Score : 1039
-- Associate : economic Score : 1007
-- Associate : status Score : 920
-- Associate : insurance Score : 838
-- Associate : social Score : 827
-- Associate : populace Score : 824

Word : commercial Freq: 7167
-- Associate : success Score : 3768
-- Associate : use Score : 1022
-- Associate : radio Score : 857
-- Associate : service Score : 758
-- Associate : failure Score : 697
-- Associate : stations Score : 664
-- Associate : television Score :591
-- Associate : production Score :543
-- Associate : banks Score : 509
-- Associate : center Score : 489

Word : committee Freq: 7123
-- Associate : council Score : 1223
-- Associate : party Score : 1111
-- Associate : members Score : 962
-- Associate : national Score : 843
-- Associate : communist Score : 790
-- Associate : privy Score : 685
--Associate: international Score:621
-- Associate : public Score : 575
-- Associate : chairman Score : 480
-- Associate : report Score : 468

Word : united Freq: 56399
-- Associate : states Score : 240294
-- Associate : kingdom Score : 47159
-- Associate : nations Score : 28275
```

```
-- Associate : canada Score : 6041    -- Associate : america Score : 4858    -- Associate : national Score : 4169
-- Associate : army Score : 5802      -- Associate : air Score : 4287
-- Associate : navy Score : 5405
```

2.5 Uses

Uses for these kinds of results can be wide and varied but fall into a few major categories. Often it is used in NLP but it can also be used in the wider field of analysis. While analyzing and extracting information from any input stream, knowledge of the context is required to make accurate predictions or decisions on the next course of action. There are many uses for this in NLP, two of those are translation and context extraction.

Firstly translation, the associations that produced can be very useful in translations as words can have multiple meanings based on their context. With the attention model different contexts that a word represents are able to be captured. From here the context can be used to determine the appropriate word in the target language by comparing it to a similar structure in the target language (Clark et al., 2019).

Secondly keywords and context can be extracted from a sentence. To do this the associations for all the words in a sentence are found and the intersect of these sets are taken. This leaves us with a set of words that are related to all the words in the sentence. An implementation of this was made and an example of the output is provided below.

Processing : [river bank]

Words Found : [river] [bank]

Context :

valley : 826; north : 799; south : 779; **flows : 682**; near : 644; west : 631; east : 608; **lake : 592**; new : 530; **city : 516**; **thames : 453**; **water : 389**; also : 385; **border : 381**; area : 377; **tributary : 364**; part : 349; one : 303; along : 280; **sea : 276**; **bridge : 274**; northern : 271; southern : 269; western : 262; **system : 249**; **tributaries : 246**; known : 245; **region : 244**; two : 240; called : 238; boundary : 232; **bay : 227**; first : 226; major : 220; forms : 213; including : 208; county : 208;

```
park : 207; eastern : 198; named : 196; mountains : 196; states : 192;
state : 191; name : 190; central : 187; rivers : 187; town : 187;
modern : 184; great : 183; main : 182;
```

Output 1. The program was given 'river bank' as input and tasked with extracting the context.

Although the results above have some 'junk' like `[central]`, `[major]` and `[town]`. The most important characteristic to note is that although the word bank is present, there is no mention of finance. This is a good result as in the context above the word bank has no association with a financial bank. And so it is clear that in terms of context identification this is successful.

To produce output like this the score of any associations was normalised against the frequency of the word. Then only the top 50 words were displayed. The intersects of the sets were also taken in order of the occurrence of the word as the network only captured relations in a forward direction as such the 'filtering' should be done in a forward direction. Furthermore to prevent one word that had no associations with other words in the sentence from removing all the results, any word whose intersect results in an empty set was disregarded.

3. Conclusion

This report presented a concept of an attention model using the NeuraBase framework and subsequently presented a use for it via its implementation as a context extraction tool in NLP. It also discussed other uses for the attention model and explored some concepts on how they can be implemented.

This implementation using NeuraBase provides a less complex attention model that is parallelizable. As an NLP tool it can be improved with better preprocessing. Specifically by performing better lemmatization and noun recognition. The attempts made in this report to implement effective lemmatization and noun recognition were unsuccessful but the results achieved without employing these processes are promising. And it is the author's belief that correctly employing lemmatization and noun recognition in this model would result in stronger results.

4. References

Clark, K., Khandelwal, U., Levy, O. and Manning, C. (2019). What Does BERT Look At? An Analysis of BERT's Attention.

Hercus, R. (2009). Neural network with learning and expression capability. 20090119236.

NLTK. (2018). NLTK.org.

Plisson, J., Lavrac, N. and Mladenic, D. (n.d.). A Rule based Approach to Word Lemmatization.

Project Gutenberg. (2020). Project Gutenberg. [online] Available at: https://www.gutenberg.org/ [Accessed 10 Feb. 2020].

spaCY. (2017). spaCY.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I. (2017). Attention Is All You Need.

Vinyals, O., Kaiser, L., Koo, T., Petrov, S., Sutskever, I. and Hinton, G. (2014). Grammar as a Foreign Language.

Wikipedia.org. (2020). Wikipedia. [online] Available at: https://www.wikipedia.org/ [Accessed 13 Feb. 2020].