[Preliminary information (Datasets, graphs) 3](#_Toc42947276)

[SemCor 3](#_Toc42947277)

[SemCor stats 3](#_Toc42947278)

[Graph 3](#_Toc42947279)

[Meetings 4](#_Toc42947280)

[IA 4](#_Toc42947281)

[IA, PSK, KF 4](#_Toc42947282)

[Meeting with IA, 12/05 5](#_Toc42947283)

[Multi-Sense LM. Part 0 6](#_Toc42947284)

[Structured prediction – select from K globals 12](#_Toc42947285)

[Design 12](#_Toc42947286)

[Mini-experiment – overfit on fragment of SemCor 13](#_Toc42947287)

[Graph check 14](#_Toc42947288)

[Senses.SelectK retrieval and selection check 15](#_Toc42947289)

[Modifications to SelectK 17](#_Toc42947290)

[Bugs & co 18](#_Toc42947291)

[Multi-sense LM. Part 1 20](#_Toc42947292)

[Gradient and loss 20](#_Toc42947293)

[GRUbase – debug on mini-experiments 21](#_Toc42947294)

[Overfitting base GRU on SemCor globals 28](#_Toc42947295)

[Preliminary experiment – GRUbase2 w/senses on SemCor 30](#_Toc42947296)

[Model Verification and Modification 32](#_Toc42947297)

[The loss when operating on senses 32](#_Toc42947298)

[Considerations 33](#_Toc42947299)

[Graph retrieval 34](#_Toc42947300)

[Boundaries of graph node types 34](#_Toc42947301)

[Analysis… 34](#_Toc42947302)

[Senses’ +1 displacement 36](#_Toc42947303)

[Adding lemmatization for the global node 37](#_Toc42947304)

[Structured prediction: SelectK – version 1 38](#_Toc42947305)

[Model 38](#_Toc42947306)

[Mini-Experiment 1 – Overfit on fragment of SemCor 39](#_Toc42947307)

[Losses 39](#_Toc42947308)

[Mini-experiment 1 – Analysis of the predictions 40](#_Toc42947309)

[Mini-Experiment 2 – Overfit on fragment of SemCor 41](#_Toc42947310)

[Losses 41](#_Toc42947311)

[Mini-experiment 2 – Analysis of the predictions 42](#_Toc42947312)

[Mini-experiment 3 44](#_Toc42947313)

[Losses 44](#_Toc42947314)

[Mini-experiment 3 – Analysis of the predictions 44](#_Toc42947315)

[Adding global-to-sense edges from the SLC 46](#_Toc42947316)

[Analysis of predictions – why do we have no senses for ‘act’, SensePerplexity=nan, etc. 46](#_Toc42947317)

[Mini-experiment 4 48](#_Toc42947318)

[Losses 48](#_Toc42947319)

[Mini-experiment 4 – Analysis of the predictions 48](#_Toc42947320)

[Mini-experiment 5 50](#_Toc42947321)

[Losses 50](#_Toc42947322)

[Mini-experiment 5 – Analysis of the predictions 51](#_Toc42947323)

[Experiment 1 – SelectK1 on SemCor 52](#_Toc42947324)

[Multi-sense LM. Part 2 54](#_Toc42947325)

[Checking & Debugging 54](#_Toc42947326)

[Baseline 0: 2 GRUs 55](#_Toc42947327)

[Model 55](#_Toc42947328)

[Mini-experiment: overfit on fragment of SemCor 56](#_Toc42947329)

[Experiment: GRUs on SemCor 56](#_Toc42947330)

[Experiment: GRUs on SemCor – version 2 57](#_Toc42947331)

[SelectK5 58](#_Toc42947332)

[Model 58](#_Toc42947333)

[Mini-experiment: overfit on fragment of SemCor 59](#_Toc42947334)

[Experiment: SelectK5 on SemCor 60](#_Toc42947335)

# Preliminary information (Datasets, graphs)

## SemCor

### SemCor stats

Note: currently I am also including punctuation in the vocabulary from SLC.

Training tokens: 646,038

Validation tokens: 80,760

Mini-dataset 1 : 180 tokens.

Mini-dataset 2: 63 tokens.

### Graph

We gather the WordNet data and build the graph using SemCor’s vocabulary.

Constructing X, matrix of node features

**X\_definitions**.shape=torch.Size([**25986**, 300])

**X\_examples**.shape=torch.Size([**26003**, 300])

**X\_senses**.shape=torch.Size([**25986**, 300])

**X\_globals**.shape=torch.Size([**21988**, 300])

Defining the edges: def, exs

def\_edges\_se.\_\_len\_\_()=25986

exs\_edges\_se.\_\_len\_\_()=26003

Defining the edges: sc

sc\_edges.\_\_len\_\_()=25986

sc\_edges\_with\_selfloops.\_\_len\_\_()=38611

This way, we also determine that there were 38611– 21988**= 16623** globals with no dictionary information, over a total of **21988 (75.6%).**

**syn\_edges**.\_\_len\_\_()=**26222**

**ant\_edges**.\_\_len\_\_()=**3780**

Data(edge\_index=[2, 120602], edge\_type=[**120602**], node\_types=[**99963**], num\_relations=[1], **x=[99963, 300]**)

# Meetings

## IA

* Alternative: simple selection from best globals, no self-attention, also works as estimate [the current simple selection overfits on fragment, but it does not give good results]
* Alternative: proceed in parallel with senses’ model and the selection from globals, and then make the probability derived from both, e.g. with a product
* Additional: when I have stable results, it would be relevant to re-run *everything* using Distil/Al-BERT’s pre-trained embreddings
* For the section meeting, I can present the whole project instead of focusing too much on what I am doing now – possibly also using the poster for the Innovation Foundation
  + It can be 10/11 to 20 minutes
  + I can also explain what different senses are, and show how SemCor works

## IA, PSK, KF

* Transformer-LM: plug in the filter that handles partial words / word completion.
* To improve performance on WikiText-2, I can use:
  + better/larger GRU
  + AWD-LSTM
  + Pre-trained embeddings from DistilBERT instead of FastText
* (How does Perplexity on Senses compare to standard Perplexity:  
  hp: the number of tokens that we compute PPL on. (e.g. PPL on text corpus of 2Mln will be different from PPL on text corpus of 103Mln words)  
  hp2: I could consider all the words that have a sense-specification on SemCor, and then compute the PPL over WikiText-2 only for those words – how does it compare with the standard PPL on WikiText-2?  
  note: PPL is difficult to compare between different datasets and tasks. Since we are in a new task, we can also just use our measure. (WSD uses F1-score))
* Implement structured prediction from globals to senses, as a baseline for the senses part of the task
* (Go on with self-attention for the senses task)

# Meeting with IA, 12/05

(Reordered) list:

* Rerun MultiSense Evaluation, using the new GRU in the architecture
  + select k=few eg. 5 globals, consider their senses, and choose among them.  
      
    question: I still need to open up to the senses’ logits; I would only get a distribution over the e.g. 20 senses from the k globals. Should I copy paste the logits over the senses’ logits, keeping everything at 0?   
    I can multiply the senses’logits (obtained, for instance, with a standard GRU side-architecture) per what I get from here. Everything not selected would be x0.
  + Compare with alternatives (e.g. Multi-sense alternatives)
* Standard Language Modeling:
  + AWD-LSTM implementation
  + Mogrifier LSTM – search for PyTorch implementation

# Multi-Sense LM. Part 0

**Baseline 1: 2 GRUs, shared first layer – bugs in the model were not solved**

**Model**

DataParallel(

(module): GRU\_base(

(maingru\_ls): ModuleList(

(0): GRU(300, 1150)

(1): GRU(1150, 1150)

(2): GRU(1150, 1150)

)

(gru\_senses): GRU(1150, 1150)

(linear2global): Linear(in\_features=1150, out\_features=21988, bias=True)

(linear2senses): Linear(in\_features=1150, out\_features=25986, bias=True))

)

Parameters:

('module.X', torch.Size([99963, 300]), torch.float32, True)

('module.select\_first\_indices', torch.Size([6]), torch.int64, False)

('module.embedding\_zeros', torch.Size([1, 300]), torch.float32, False)

('module.memory\_hn', torch.Size([3, 4, 1150]), torch.float32, False)

('module.memory\_hn\_senses', torch.Size([1, 4, 1150]), torch.float32, False)

('module.maingru\_ls.0.weight\_ih\_l0', torch.Size([3450, 300]), torch.float32, True)

('module.maingru\_ls.0.weight\_hh\_l0', torch.Size([3450, 1150]), torch.float32, True)

('module.maingru\_ls.0.bias\_ih\_l0', torch.Size([3450]), torch.float32, True)

('module.maingru\_ls.0.bias\_hh\_l0', torch.Size([3450]), torch.float32, True)

('module.maingru\_ls.1.weight\_ih\_l0', torch.Size([3450, 1150]), torch.float32, True)

('module.maingru\_ls.1.weight\_hh\_l0', torch.Size([3450, 1150]), torch.float32, True)

('module.maingru\_ls.1.bias\_ih\_l0', torch.Size([3450]), torch.float32, True)

('module.maingru\_ls.1.bias\_hh\_l0', torch.Size([3450]), torch.float32, True)

('module.maingru\_ls.2.weight\_ih\_l0', torch.Size([3450, 1150]), torch.float32, True)

('module.maingru\_ls.2.weight\_hh\_l0', torch.Size([3450, 1150]), torch.float32, True)

('module.maingru\_ls.2.bias\_ih\_l0', torch.Size([3450]), torch.float32, True)

('module.maingru\_ls.2.bias\_hh\_l0', torch.Size([3450]), torch.float32, True)

('module.gru\_senses.weight\_ih\_l0', torch.Size([3450, 1150]), torch.float32, True)

('module.gru\_senses.weight\_hh\_l0', torch.Size([3450, 1150]), torch.float32, True)

('module.gru\_senses.bias\_ih\_l0', torch.Size([3450]), torch.float32, True)

('module.gru\_senses.bias\_hh\_l0', torch.Size([3450]), torch.float32, True)

('module.linear2global.weight', torch.Size([21988, 1150]), torch.float32, True)

('module.linear2global.bias', torch.Size([21988]), torch.float32, True)

('module.linear2senses.weight', torch.Size([25986, 1150]), torch.float32, True)

('module.linear2senses.bias', torch.Size([25986]), torch.float32, True)

Number of trainable parameters=114.04M, where softmax=29.90M + 25.28M=55.19M, embeddings=29.99M, core = 28.86M

**Mini-experiment on fragment of SemCor**

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| For the globals:  **GRU** w/ 3 layers, 1150x3 (layer 1 shared).  Followed by **FF-NN** to logits | 1. The word embedding of the current global   (d=300) | batch\_size=4 |
| For the senses:  **1 shared layer** of the main GRU > **1-layer GRU** -> **FF-NN** to logits | TBPTT length=8 |
|  | learning rate=10^(**-3**) |

|  |  |  |
| --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |
| 1 | 4918.23 | 10418.61 |
| 2 | 105.08 | 134.04 |
| 3 | 77.19 | 68.57 |
| 4 | 65.73 | 56.92 |
| 5 | 66.21 | 54.71 |
| 10 | 62.92 | 51.64 |
| 30 | 62.92 | 19.82 |
| 50 | 56.17 | 4.42 |
| 75 | 37.24 | 1.44 |
| 100 | 23.13 | 1.1 |
| 150 | 9.02 | 1.03 |
| 200 | 3.57 | 1.01 |

**Experiment –GRUs w/ shared layer on SemCor (fault: not detaching the gru\_senses’ memory)**

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| For the globals:  **GRU** w/ 3 layers, 1150x3 (layer 1 shared).  Followed by **FF-NN** to logits | 1. The word embedding of the current global   (d=300) | batch\_size=40 |
| For the senses:  **1 shared layer** of the main GRU > **1-layer GRU** -> **FF-NN** to logits | TBPTT length=35 |
|  | learning rate=0.5\*10^(**-4**) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |  | *Validation perplexity on globals* | *Validation perplexity on senses* |
| 1 | 971.96 | 5169.47 |  | 490.31 | 835.0 |
| 2 | 540.3 | 3432.66 |  | 347.28 | **804.95** |
| 3 | 418.94 | 3129.2 |  | 315.93 | 831.75 |
| 4 | 373.93 | 2947.42 |  | 299.93 | 829.01 |
| 5 | 344.06 | 2821.13 |  | 289.18 | 886.89 |
| 6 | 321.49 | 2768.63 |  | 283.92 | 1034.11 |
| 7 | 302.41 | 3269.6 |  | 278.56 | 914.73 |
| 8 | 285.49 | 2665.07 |  | 270.99 | 1807.85 |
| 9 | 270.04 | 2528.11 |  | 267.79 | 1112.29 |
| 10 | 256.36 | 2502.85 |  | 262.02 | 970.6 |
| 11 | 243.94 | 2391.66 |  | 263.61 | 1200.81 |
| 12 | 232.49 | 2373.34 |  | 258.92 | 1399.38 |
| 13 | 221.78 | 2252.45 |  | 252.47 | 2179.9 |
| 14 | 211.73 | 2222.98 |  | 251.66 | 1309.55 |
| 15 | 202.54 | 2070.02 |  | 250.55 | 2772.74 |
| 16 | 193.66 | 1950.52 |  | 247.19 | 2650.16 |
| 17 | 185.68 | 1891.93 |  | 248.58 | 3095.87 |
| 18 | 177.93 | 1835.95 |  | 247.61 | 2480.83 |
| 19 | 170.97 | 1809.4 |  | 245.7 | 1232.44 |
| 20 | 164.49 | 1725.47 |  | 245.27 | 2917.69 |
| 21 | 158.3 | 1641.77 |  | 244.83 | 1680.78 |
| 22 | 152.17 | 1589.13 |  | 245.41 | 2118.22 |
| 23 | 146.72 | 1569.21 |  | 244.53 | 2307.36 |
| 24 | 140.69 | 1497.27 |  | 245.31 | 1977.77 |
| 25 | 134.79 | 1427.68 |  | 244.14 | 3117.87 |
| 26 | 129.49 | 1348.3 |  | 243.1 | 2807.55 |
| 27 | 124.34 | 1311.9 |  | **242.35** | 2739.16 |
| 28 | 119.55 | 1289.07 |  | 243.71 | 3153.67 |
| 29 | 115.16 | 1264.97 |  | 244.83 | 3012.75 |
| 30 | 110.88 | 1209.65 |  | 245.65 | 3195.36 |

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| For the globals:  **GRU** w/ 3 layers, 1150x3 (layer 1 shared).  Followed by **FF-NN** to logits | 1. The word embedding of the current global   (d=300) | batch\_size=40 |
| For the senses:  **1 shared layer** of the main GRU > **1-layer GRU** -> **FF-NN** to logits | TBPTT length=35 |
|  | **learning rate=10^(-5)** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |  | *Validation perplexity on globals* | *Validation perplexity on senses* |
| 1 | 1612.97 | 7651.04 |  | 522.04 | 1105.27 |
| 2 | 736.76 | 3579.03 |  | 510.6 | 1007.8 |
| 3 | 725.9 | 3486.44 |  | 510.74 | 983.15 |
| 4 | 723.02 | 3457.18 |  | 511.13 | **981.36** |
| 5 | 721.14 | 3443.05 |  | 508.15 | 991.13 |
| 6 | 720.06 | 3430.39 |  | 510.9 | 1001.09 |
| 7 | 719.66 | 3422.33 |  | 511.41 | 991.3 |
| 8 | 719.17 | 3410.85 |  | 508.73 | 1013.25 |
| 9 | 718.67 | 3410.02 |  | 511.87 | 1021.22 |
| 10 | 718.51 | 3407.2 |  | 508.25 | 1028.66 |
| 20 | 401.88 | 3258.57 |  | 309.43 | 1013.5 |
| 30 | 326.42 | 3116.9 |  | 274.1 | 1151.66 |
| 40 | 296.26 | 2955.15 |  | 267.23 | 1646.8 |
| 50 | 269.84 | 2696.74 |  | **263.68** | 1682.31 |
| 60 | 247.95 | 2453.59 |  | 264.04 | 2097.26 |

Observations:

a learning rate of 10^(-5) is too low. 10^(-4) is fast. Grid search suggests 0.5\*10^(-4)

I did not detach the senses’ memory!

**Experiment –GRUs w/ shared layer on SemCor**

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| For the globals:  **GRU** w/ 3 layers, 1150x3 (layer 1 shared).  Followed by **FF-NN** to logits | 1. The word embedding of the current global   (d=300) | batch\_size=40 |
| For the senses:  **1 shared layer** of the main GRU > **1-layer GRU** -> **FF-NN** to logits | TBPTT length=35 |
|  | learning rate=**0.5\*10^(-4)** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |  | *Validation perplexity on globals* | *Validation perplexity on senses* |
| 1 | 974.01 | 5184.52 |  | 495.82 | 835.31 |
| 2 | 543.19 | 3431.93 |  | 343.02 | **809.3** |
| 3 | 420.57 | 3115.14 |  | 316.4 | 830.73 |
| 4 | 375.15 | 2931.76 |  | 300.37 | 849.11 |
| 5 | 345.35 | 2806.64 |  | 289.39 | 919.22 |
| 6 | 322.07 | 2755.25 |  | 283.68 | 1344.15 |
| 7 | 302.68 | 2966.38 |  | 275.61 | 914.12 |
| 8 | 285.5 | 2683.41 |  | 269.84 | 1144.84 |
| 9 | 269.57 | 2614.12 |  | 266.23 | 1091.96 |
| 10 | 255.34 | 2522.42 |  | 260.59 | 1562.29 |
| 11 | 242.45 | 2525.91 |  | 262.01 | 1201.69 |
| 12 | 230.17 | 2335.81 |  | 255.8 | 3216.74 |
| 13 | 218.73 | 2203.31 |  | 249.31 | 1609.48 |
| 14 | 208.14 | 2198.41 |  | 248.36 | 2093.41 |
| 15 | 198.53 | 2077.28 |  | 247.43 | 1538.86 |
| 16 | 189.53 | 2036.01 |  | 242.89 | 2184.96 |
| 17 | 181.32 | 2029.97 |  | 243.98 | 1695.24 |
| 18 | 173.23 | 1921.61 |  | 241.94 | 2423.03 |
| 19 | 165.53 | 1774.17 |  | 241.25 | 2554.37 |
| 20 | 158.48 | 1748.05 |  | 241.11 | 1825.42 |
| 21 | 152.1 | 1682.55 |  | 240.31 | 2533.01 |
| 22 | 146.39 | 1696.8 |  | 239.08 | 1760.2 |
| 23 | 141.34 | 1531.6 |  | 239.15 | 2551.63 |
| 24 | 136.85 | 1468.94 |  | 240.69 | 3231.04 |
| 25 | 132.03 | 1453.37 |  | 239.0 | 2714.26 |
| 26 | 126.83 | 1346.67 |  | 238.2 | 3285.9 |
| 27 | 122.56 | 1297.54 |  | **237.42** | 2934.18 |
| 28 | 118.1 | 1255.5 |  | 238.0 | 2678.2 |
| 29 | 113.91 | 1216.16 |  | 239.22 | 3284.78 |
| 30 |  |  |  |  |  |

**Baseline 2 – 2GRUs (x3 and x2), no shared layers**

**Model**

DataParallel(

(module): GRU\_base2(

(maingru\_ls): ModuleList(

(0): GRU(300, 1150)

(1): GRU(1150, 1150)

(2): GRU(1150, 1150)

)

(gru\_senses): GRU(300, 1150, num\_layers=2)

(linear2global): Linear(in\_features=1150, out\_features=21988, bias=True)

(linear2senses): Linear(in\_features=1150, out\_features=25986, bias=True)

)

)

Number of trainable parameters=119.05M, where softmax=29.90M + 25.28M=55.19M, embeddings=29.99M, core = 33.87M

**Variant**: Including the sense node-state among the input signals:

DataParallel(

(module): GRU\_base2(

(maingru\_ls): ModuleList(

(0): GRU(600, 1150)

(1): GRU(1150, 1150)

(2): GRU(1150, 1150)

)

(gru\_senses): GRU(600, 1150, num\_layers=2)

(gat\_senses): GATConv(300, 75, heads=4)

(linear2global): Linear(in\_features=1150, out\_features=21988, bias=True)

(linear2senses): Linear(in\_features=1150, out\_features=25986, bias=True)

)

)

Number of trainable parameters=121.21M, where core=36.03M

**Experiment – 2 GRUs on SemCor**

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| For the globals:  **GRU** w/ 3 layers, 1150x3 (layer 1 shared).  Followed by **FF-NN** to logits | 1. The word embedding of the current global   (d=300) | batch\_size=40  TBPTT length=35 |
| For the senses:  **GRU** w/ 2 layers -> **FF-NN** to logits | learning rate=**0.5\*10^(-4)** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |  | *Validation perplexity on globals* | *Validation perplexity on senses* |
| 1 | 967.94 | 5149.14 |  | 494.29 | 828.62 |
| 2 | 527.79 | 3413.73 |  | 344.56 | **809.02** |
| 3 | 415.68 | 3060.8 |  | 314.78 | 876.36 |
| 4 | 373.02 | 2898.12 |  | 299.13 | 960.85 |
| 5 | 343.53 | 2741.66 |  | 286.95 | 1717.02 |
| 6 | 319.55 | 2697.53 |  | 279.82 | 906.0 |
| 7 | 298.75 | 2779.89 |  | 272.17 | 1035.97 |
| 8 | 280.14 | 2865.53 |  | 263.82 | 1036.54 |
| 9 | 263.37 | 2729.85 |  | 259.78 | 831.5 |
| 10 | 248.22 | 2986.64 |  | 253.18 | 1417.8 |
| 11 | 234.43 | 2479.41 |  | 252.78 | 1582.18 |
| 12 | 221.68 | 2397.1 |  | 246.32 | 1551.16 |
| 13 | 209.56 | 2695.38 |  | 239.0 | 881.0 |
| 14 | 198.55 | 2565.65 |  | 237.06 | 1204.42 |
| 15 | 188.63 | 2354.02 |  | 235.9 | 1339.91 |
| 16 | 179.54 | 2433.03 |  | 232.66 | 1072.53 |
| 17 | 171.34 | 2335.96 |  | 233.08 | 2045.94 |
| 18 | 163.51 | 2525.17 |  | 232.08 | 1157.81 |
| 19 | 156.38 | 2600.38 |  | 231.01 | 1704.28 |
| 20 | 149.69 | 2189.19 |  | 231.73 | 4222.49 |
| 21 | 143.11 | 2137.6 |  | 230.99 | 3757.52 |
| 22 | 136.94 | 2096.51 |  | 229.7 | 2110.82 |
| 23 | 131.19 | 1997.07 |  | **228.25** | 2498.45 |
| 24 | 125.67 | 1997.78 |  | 229.65 | 2904.29 |
| 25 | 120.53 | 1954.38 |  | 229.96 | 2206.7 |
| 26 | 115.94 | 1863.62 |  | 230.31 | 3183.38 |
| 27 | 111.5 | 1911.42 |  | 231.0 | 2748.2 |
| 28 | 107.35 | 1749.79 |  | 233.52 | 2585.29 |
| 29 | 103.34 | 1804.05 |  | 234.73 | 2605.26 |
| 30 | 99.47 | 1669.6 |  | 236.99 | 3488.63 |

The Valid-PPL on globals and senses is (228, 809). Better than the architecture that shares a layer, that reaches (237, 809). So there is no reason to share the first layer between the GRUs.

## Structured prediction – select from K globals

### Design

Select k=5 globals, consider their senses, and choose among them.  
  
I still need to open up to the senses’ logits.

I get a distribution over the e.g. 20 senses from the k=5 globals.

I can multiply the senses’logits (obtained, for instance, with a standard GRU side-architecture) per what I get from here. Everything not selected would be x0.

Possibility: an alternative version may include the sense input.

For the sake of speed when retrieving the senses of the selected k globals:

we pass the graph\_area\_matrix as a parameter, to get the indices of the neighbouring nodes (since we are starting from globals, they will be either senses through the *sc* edges or other globals throught *synonyms/antonyms*.). We filter the neighbours to keep only the nodes whose index is in the senses’ range.

We obtain batch\_size\*sequence\_len\*k (e.g 4x8x5=160) tensors of variable size, containing the indices of the senses.

For every tensor containing the senses, we can apply X.index\_select(…) and get the sense embeddings.

We should get a probability distribution over these senses.

Idea #1: do not retrieve the embeddings. Assign 1 to the selected senses’indices and 0 to all others. Multiply per the softmax from the “other line”(the GRU). Then possibly scale up the probabilities that survived the filter so that they sum up to 1.

Problem: the nll\_loss works with log\_softmax, not with softmax.

Hypothesis a): assign 10^(-10) to the all the not-selected senses, instead of exactly 0 that would break the softmax and nll\_loss numerically. Then, rescale the softmax values that came from the relevant logits so that they sum up to 1.

**Hypothesis b)**: mask out the logits we don’t care for, and apply the softmax over the selected senses. Then, we will have to “make space” for all the 10^(-10) values, so we will have to subtract a small quantity δ from the selected values that we computed.

DataParallel(

(module): SelectK(

(maingru\_ls): ModuleList(

(0): GRU(300, 1150)

(1): GRU(1150, 1150)

(2): GRU(1150, 1150)

)

(gru\_senses): GRU(300, 1150, num\_layers=2)

(linear2global): Linear(in\_features=1150, out\_features=21988, bias=True)

(linear2senses): Linear(in\_features=1150, out\_features=25986, bias=True)

)

)

Number of parameters: 119.05M, where core=33.87M

### Mini-experiment – overfit on fragment of SemCor

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| For the globals:  **GRU** w/ 3 layers, 1150x3 (layer 1 shared).  Followed by **FF-NN** to logits | 1. The word embedding of the current global   (d=300) | batch\_size=40  TBPTT length=35 |
| For the senses:  **GRU** w/ 2 layers -> **FF-NN** to logits  - > applying the softmax of the senses of the k most likely globals | learning rate=**0.5\*10^(-4)** |

I do not manage to overfit. Sense loss stuck at 13.82, sense perplexity at 10^(6).

Incidentally, I set all the senses that do not belong to the k globals to 10^(-6).

When we are at the start, the first predicted globals are random, and do not provide a useful indication.

However, in later stages of the training they will. Here, the possibility of a correct sense prediction depends on the quality of the prediction of globals.

Let us use k=5 and wait until the globals’ part of the model starts overfitting…

We have an error due to the retrieval of globals and indices.

If my current most likely global is ‘cotton’, the senses I retrieve should not contain ‘factor.n.2’.

AD.get\_node\_data(self.grapharea\_matrix\_lil, 8977, self.N, features\_mask=(**True**,**False**,**False**))

-> tensor([ 8977, 56951, 83334, 32288, 8972, 8973, 8974, 8976, 8975, 35033,

34527])

We should remember that we are retrieving nodes 2 steps away, in a maximum number of 32.

Through the following edges, we can get the following node types:

sc -> senses; sc + def -> definitions; sc + ex -> examples; syn, ant -> globals; syn/ant + sc -> senses.

Since: EP.get\_globalword\_fromindex(32288-self.last\_idx\_senses) = ‘factor’ {global}

we are retrieving its senses.

However, the synonyms of ‘cotton’ , in the processed\_synonyms.h5 archive, are:

6386 cotton\_fiber

6387 cotton\_wool

6388 cotton\_plant

It is time to apply a sanity check to the graph, and the nodes we retrieve from the grapharea\_matrix.

## Graph check

Node: 8977 in the vocabulary of globals (cotton)

GA.get\_indices\_area\_toinclude(graph\_dataobject.edge\_index, graph\_dataobject.edge\_type, 8977, area\_size=32, max\_hops=1)

We get: ([8977(n: index of the global itself here), 56951, 83334, 32288],)

Remember that the X matrix has the sections:

**senses**: (0, 25986). **globals**: (25986, 47974). definitions: (47974, 73960). examples: (73960, 120602)

56951-47974=definition 8977. factor.v.03 consider as relevant when making a decision

83334-73960=example 9374. factor.v.03 You must factor in the recent developments

32288-25986=global 6302. factor 35

Why am I retrieving ‘factor’? What is the mistake here?

Let us add self.last\_idx\_senses to the node number I am retrieving…

This time,

GA.get\_indices\_area\_toinclude(graph\_dataobject.edge\_index, graph\_dataobject.edge\_type, 8977+25986{=34963}, 32, 1)

node\_indices\_ls, all\_edges\_retrieved\_ls =

([**34963**, 5673, 5672, 5671, 5674, *5670*], [57659, 57660, 57661, 57662, 57663])

**34963-**25986: global 8977 (‘cotton’)

Senses: 5670 to 5674: ﻿costume.v.02, cotton.n.01, cotton.n.02, cotton.n.03, cotton.n.04.

*Note*: it should be 5675, cotton.v.01 not 5670-costume.v.02. Must add 1. When I create the graph, maybe?

From the indices\_table.db, I get:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| word\_sense | vocab\_index | start\_defs | end\_defs | start\_examples | end\_examples |
| cotton.n.01 | 5670 | 5670 | 5671 | 5821 | 5821 |
| cotton.n.02 | 5671 | 5671 | 5672 | 5821 | 5821 |
| cotton.n.03 | 5672 | 5672 | 5673 | 5821 | 5821 |
| cotton.n.04 | 5673 | 5673 | 5674 | 5821 | 5821 |
| cotton.v.01 | 5674 | 5674 | 5675 | 5821 | 5822 |

When I execute get\_node\_edges, I get:

<class 'list'>: [(34963, 5673, 2), (34963, 5672, 2), (34963, 5671, 2), (34963, 5674, 2), (34963, 5670, 2)]

Let us review the most likely global and the senses that we retrieve for it in the SelectK architecture.

### Senses.SelectK retrieval and selection check

In Epoch 1, in the very first batch, at the start of the training process on mini-fragment:

(the initial most likely globals depend on the random initialization of the main GRU)

**INFO : sample: 0; k\_globals=['microorganisms']**

*INFO : sample: 1; k\_globals=['unscrupulous']*

…

**INFO : sample: 16; k\_globals=['shrink']**

*INFO : sample: 17; k\_globals=['unscrupulous']*

…

**# sample 0 is missing, since the plural form ‘microorganisms’ has no senses**.

*INFO : Sample: 1; selected\_senses=['unscrupulous.a.01']*

…

**INFO : Sample: 16; selected\_senses=['shrink.v.03', 'shrink.v.05', 'shrink.v.04', 'reduce.v.15', 'reduce.v.02', 'reduce.v.14', 'reduce.v.11', 'reduce.v.20', 'reduce.v.09', 'reduce.v.08', 'reduce.v.05', 'reduce.v.04', 'reduce.v.03', 'reduce.v.01', 'reduce.v.13', 'reduce.v.06', 'reduce.v.18']**

*INFO : Sample: 17; selected\_senses=['unscrupulous.a.01']*

From Epoch 2 to epoch… 161 and counting (globals train-PPL @161: 25.97, descending…) … epoch :

The most likely global is either ‘”’,‘the’ or ‘<unk>’ that have no senses.

I do not manage to get relevant globals to be the first one.

I use the verbose log of the predictions every 10 epochs, thus training faster (on GPUs) for 300 epochs.

Training, end of epoch 289. Global step n.1734. Time = 1180.27.

INFO : Perplexity: Globals perplexity=1.08 Sense perplexity=1383.87

Training, end of epoch 290. Global step n.1740. Time = 1182.03. The training losses are:

INFO : Perplexity: Globals perplexity=1.1 Sense perplexity=1979.8

|  |  |  |
| --- | --- | --- |
| Globals predicted | Senses predicted | Comment |
| Label: the next global is: said(from 1)  INFO : Label: the next sense is: state.v.01(from 22232)  INFO : The top- 5 predicted globals are:  INFO : Word: said ; probability = 92.4%  INFO : Word: further ; probability = 2.03%  INFO : Word: to ; probability = 1.34%  INFO : Word: thanks ; probability = 0.72%  INFO : Word: the ; probability = 0.54%  INFO : The top- 5 predicted senses are:  INFO : |  | Since “said” has no senses to select, we do not find any.  (We find only “say” in the vocabulary) |
| Label: the next global is: Friday(from 2)  INFO : Label: the next sense is: None(from -1)  INFO : The top- 5 predicted globals are:  INFO : Word: Friday ; probability = 82.86%  INFO : Word: it ; probability = 3.96%  INFO : Word: conducted ; probability = 2.76%  INFO : Word: , ; probability = 1.7%  INFO : Word: Only ; probability = 1.64%  INFO : |  |  |
| Label: the next global is: an(from 3)  INFO : Label: the next sense is: None(from -1)  INFO : The top- 5 predicted globals are:  INFO : Word: an ; probability = 91.28%  INFO : Word: the ; probability = 4.16%  INFO : Word: which ; probability = 1.6%  INFO : Word: . ; probability = 0.66%  INFO : Word: evidence ; probability = 0.36%  INFO : |  |  |
| Label: the next global is: investigation(from 4)  INFO : Label: the next sense is: probe.n.01(from 17882)  INFO : The top- 5 predicted globals are:  INFO : Word: investigation ; probability = 89.53%  INFO : Word: considering ; probability = 1.04%  INFO : Word: a ; probability = 0.97%  INFO : Word: won ; probability = 0.88%  INFO : Word: Only ; probability = 0.82% | INFO : The top- 5 predicted senses are:  INFO : Sense: probe.n.01 ; probability = 11.23%  INFO : Sense: investigate.v.01 ; probability = 10.91%  INFO : Sense: investigation.n.02 ; probability = 10.91%  INFO : Sense: investigate.v.02 ; probability = 10.86%  INFO : Sense: probe.n.03 ; probability = 10.82% |  |
| of Atlanta s |  |  |
| Label: the next global is: recent(from 8)  INFO : Label: the next sense is: late.s.03(from 13363)  INFO : The top- 5 predicted globals are:  INFO : Word: recent ; probability = 80.44%  INFO : Word: charge ; probability = 5.42%  INFO : Word: registration ; probability = 3.93%  INFO : Word: act ; probability = 1.94%  INFO : Word: handful ; probability = 1.29% | INFO : The top- 5 predicted senses are:  INFO : Sense: late.a.01 ; probability = 12.44%  INFO : Sense: late.r.03 ; probability = 12.3%  INFO : Sense: late.a.06 ; probability = 12.21%  INFO : Sense: late.s.03 ; probability = 12.18%  INFO : Sense: late.r.01 ; probability = 12.17% | we do not manage to focus on the correct sense… we are just picking the senses of the most likely global. |
|  |  |  |

Training epoch n.295: Perplexity: Globals perplexity=1.09 Sense perplexity=1712.68

Training epoch n.296: Perplexity: Globals perplexity=1.08 Sense perplexity=1776.37

Training epoch n.297: Perplexity: Globals perplexity=1.07 Sense perplexity=1786.04

Training epoch n.298: Perplexity: Globals perplexity=1.06 Sense perplexity=2161.16

Training epoch n.299: Perplexity: Globals perplexity=1.06 Sense perplexity=1197.16

Training epoch n.300: Perplexity: Globals perplexity=1.08 Sense perplexity=1712.68

### Modifications to SelectK

When the globals have already gone into overfit on the fragment, the sense perplexity is oscillating around 1100 and 2100, but it’s still very high.

We can try several modifications.

#1: Since “said” has no senses to select, we do not find any, even if there is a sense label – here, state.v.01

We can lemmatize ‘said’ (into ‘say’), and get the senses again.

#2: It may happen that even the lemmatized form has no senses or does not change. This happens with phrases, like ‘full of’.

We may decide to send a tensor of full-zeros, and ignore that sense label. This may or may not require modyfing the condition for ignoring the label, from a tensor with [0] to a full-zeros… it does not, because we do not explicitly ignore that label, we just exploit the fact that a uniform full-zero logsoftmax gives no gradient in the nll\_loss.

Working example of lemmatization:

sample\_k\_indices=[32671] -> [‘arches’] -> [‘arch’(6687+25986=32673)] -> we get sense\_neighbours\_t = tensor([ 1135, 1136, 1134, 1131, 1130, 1133, 1129, 1132, 25588])

The corresponding senses we extract are:

0 = {str} 'arch.s.03'

1 = {str} 'arch.v.01'

2 = {str} 'arch.s.02'

3 = {str} 'arch.n.03'

4 = {str} 'arch.n.02'

5 = {str} 'arch.s.01'

6 = {str} 'arch.n.01'

7 = {str} 'arch.n.04'

8 = {str} 'wicked.a.01'

We are picking, erroneously, one of arch’s syonyms:

arch.n.01 arch

arch.n.02 arch

arch.n.03 arch

arch.n.03 archway

arch.n.04 arch

arch.s.01 patronize

arch.s.01 patronise

arch.s.01 arch

arch.s.01 condescend

arch.s.02 arch

arch.s.03 pixilated

arch.s.03 puckish

arch.s.03 prankish

arch.s.03 arch

arch.s.03 wicked

arch.s.03 impish

arch.s.03 implike

arch.s.03 mischievous

arch.v.01 arch

arch.v.01 arc

arch.v.01 curve

Where does 25588 come from? It’s on the 2nd hop.

In order to avoid pulling the senses of synonyms (and antonyms), we need to use a 1-hop graph and graph area matrix for this particular use case.

Later on I will send the grapharea\_matrix with 1 hop specifically for this purpose while using the 2-hops version for the model input. For now, I can just use that one as the current grapharea matrix of the experiment

### Bugs & co

Current problems:

segfault only on GPU

git version

The segfault:

lib/python3.6/site-packages/tables/.libs/**libhdf5**-933c8d2d.so.103.0.0

In particular, we get:

Thread 38 "python3" received signal SIGSEGV, Segmentation fault.

[Switching to Thread 0x7ffd7b7ad700 (LWP 25678)]

0x00007fff53a8aab8 in H5C\_protect ()

from /home/andrealk3/venvs/torch15/lib/python3.6/site-packages/tables/.libs/libhdf5-933c8d2d.so.103.0.0

Problem:

we moved the model to DataParallel. HDF5 does not reliably allow for multithread access, or multiprocess for that matter.

New error: corrupted double-linked list.

Hypothesis 1:

“Some googling told me that apparently this one is due to a kernel debugging option:

CONFIG\_DEBUG\_STACK\_USAGE

and that the message is generally benign.”

Hypothesis 2:

“The two likely causes I can see are:

1) Writing into a block after it is freed.

2) A buffer overrun in a memory block into an adjacent freed block.”

On Cheetah we do not get this error. It is due to DataParallel-ism and replication.

Parallel reading on HDF5 strikes again. From the stacktrace:

k\_globals\_words = [EP.get\_globalword\_fromindex(global\_relative\_idx) for global\_relative\_idx in k\_globals\_relative\_indices]

File "/home/andrealk3/ALMA\_PhD/Code/TASK1\_Gbwe/GNN/ExplorePredictions.py", line 19, in get\_globalword\_fromindex

globals\_vocabulary\_df = pd.read\_hdf(globals\_vocabulary\_fpath, mode='r')

…’

tables.exceptions.HDF5ExtError: Problems reading records.

# Multi-sense LM. Part 1

## Gradient and loss

Whenever I have no sense labels, I should make it so that there is no gradient sent back in the model.

**Note**: maybe the train\_dataloader should use the lemmatizer to find senses of word like ‘said’? No, there is no need for that, either the training dataset has a sense label or it does not.

loss\_global = tfunc.nll\_loss(predictions\_globals, batch\_labels\_globals)

loss\_global = tensor(9.9997, grad\_fn=<NllLossBackward>)

<NllLossBackward object at 0x1a898ae190>

predictions\_globals.shape= torch.Size([32, 21988])

predictions\_globals[0] .shape= torch.Size([21988])

.grad\_fn= <SelectBackward object at 0x1a8b0f8e50>

.requires\_grad=True

Let us now examine:

batch\_labels\_senses=

tensor([ -1, 17882, -1, 13363, 17809, 17913, 8606, -1, -1, 13063,

9122, 22232, 23451, 8298, -1, 10898, 16289, 16237, 14922, 8131,

6745, 17606, 23521, 13952, 14322, 8131, 5081, -1, -1, -1,

23451, 13063])

loss\_sense = tensor(10.1642, grad\_fn=<NllLossBackward>)

predictions\_senses.shape= torch.Size([32, 25986])

predictions\_senses[0].shape= torch.Size([25986])

.grad\_fn= <SelectBackward object at 0x1a898ae750>

.requires\_grad=True

I may have to modify the functions:

compute\_model\_loss(model,batch\_input, batch\_labels, verbose=**False**)

and

compute\_sense\_loss(predictions\_senses, batch\_labels\_senses)

In torch.nn.functional.nll\_loss():

**ignore\_index** ([*int*](https://docs.python.org/3/library/functions.html#int)*, optional*) – Specifies a target value that is ignored and does not contribute to the input gradient. When size\_average is True, the loss is averaged over non-ignored targets.

Default: -100

Now loss\_sense is: tensor(10.1725, grad\_fn=<PermuteBackward>)

## GRUbase – debug on mini-experiments

Do not add any specific mechanism. Just use another GRU for the prediction on the senses’ logits&log\_softmax.

INFO : Model:

INFO : DataParallel(

(module): GRU\_base2(

(main\_gru): GRU(300, 1150, num\_layers=3)

(gru\_senses): GRU(300, 1150, num\_layers=2)

(linear2global): Linear(in\_features=1150, out\_features=21988, bias=True)

(linear2senses): Linear(in\_features=1150, out\_features=25986, bias=True)

)

)

p.s. I should also find a reason for using only 2 layers.

Maybe because not all the words have a sense label and thus I have fewer samples.

It may also be opportune to find a source for another set of hyperparameters of the architecture.

INFO : Parameters:

INFO : ('module.X', torch.Size([99963, 300]), torch.float32, True)

('module.select\_first\_indices', torch.Size([6]), torch.float32, False)

('module.embedding\_zeros', torch.Size([1, 300]), torch.float32, False)

('module.memory\_hn', torch.Size([3, 4, 1150]), torch.float32, False)

('module.memory\_hn\_senses', torch.Size([2, 4, 1150]), torch.float32, False)

('module.main\_gru.weight\_ih\_l0', torch.Size([3450, 300]), torch.float32, True)

('module.main\_gru.weight\_hh\_l0', torch.Size([3450, 1150]), torch.float32, True)

('module.main\_gru.bias\_ih\_l0', torch.Size([3450]), torch.float32, True)

('module.main\_gru.bias\_hh\_l0', torch.Size([3450]), torch.float32, True)

('module.main\_gru.weight\_ih\_l1', torch.Size([3450, 1150]), torch.float32, True)

('module.main\_gru.weight\_hh\_l1', torch.Size([3450, 1150]), torch.float32, True)

('module.main\_gru.bias\_ih\_l1', torch.Size([3450]), torch.float32, True)

('module.main\_gru.bias\_hh\_l1', torch.Size([3450]), torch.float32, True)

('module.main\_gru.weight\_ih\_l2', torch.Size([3450, 1150]), torch.float32, True)

('module.main\_gru.weight\_hh\_l2', torch.Size([3450, 1150]), torch.float32, True)

('module.main\_gru.bias\_ih\_l2', torch.Size([3450]), torch.float32, True)

('module.main\_gru.bias\_hh\_l2', torch.Size([3450]), torch.float32, True)

('module.gru\_senses.weight\_ih\_l0', torch.Size([3450, 300]), torch.float32, True)

('module.gru\_senses.weight\_hh\_l0', torch.Size([3450, 1150]), torch.float32, True)

('module.gru\_senses.bias\_ih\_l0', torch.Size([3450]), torch.float32, True)

('module.gru\_senses.bias\_hh\_l0', torch.Size([3450]), torch.float32, True)

('module.gru\_senses.weight\_ih\_l1', torch.Size([3450, 1150]), torch.float32, True)

('module.gru\_senses.weight\_hh\_l1', torch.Size([3450, 1150]), torch.float32, True)

('module.gru\_senses.bias\_ih\_l1', torch.Size([3450]), torch.float32, True)

('module.gru\_senses.bias\_hh\_l1', torch.Size([3450]), torch.float32, True)

('module.linear2global.weight', torch.Size([21988, 1150]), torch.float32, True)

('module.linear2global.bias', torch.Size([21988]), torch.float32, True)

('module.linear2senses.weight', torch.Size([25986, 1150]), torch.float32, True)

('module.linear2senses.bias', torch.Size([25986]), torch.float32, True)

INFO : Number of trainable parameters=119.05M

**Mini-experiments – Overfit on fragment of SemCor**

INFO : Hyperparameters: \_batchPerSeqlen32\_area32\_lr0.0001\_epochs300

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| model: GRUbase2 | 1) The word embedding of the current global (d=300) | batch\_size=4 |
| Globals’ prediction: **Main GRU** with 3 layers (1150 x3) | TBPTT length=8 |
| Followed by linear2Globals **FF-NN** | learning rate=10^(-4) |
| Senses’ prediction: **GRU** with 2 layers (1150x2),  Followed by linear2Senses **FF-NN** |  | grapharea=32, hops=2 |

|  |  |  |
| --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |
| 1 | 21007.48 | 25634.85 |
| 2 | 16496.35 | 23663.83 |
| 3 | 3434.95 | 19598.76 |
| 4 | 390.72 | 6683.16 |
| 5 | 206.94 | 335.28 |
| 6 | 105.66 | 108.93 |
| 7 | 77.35 | 73.31 |
| 8 | 69.2 | 61.01 |
| 9 | 62.86 | 58.51 |
| 10 | 65.38 | 52.86 |
| 30 | 59.68 | 47.63 |
| 50 | 59.34 | 47.45 |
| 75 | 59.16 | 47.26 |
| 100 | 59.17 | 47.1 |
| 125 | 59.08 | 43.13 |
| 150 | 59.1 | 17.43 |
| 175 | 59.09 | 1.49 |
| 200 | 59.05 | 1.1 |
| 225 | 58.72 | 1.04 |
| 250 | 21.44 | 1.03 |
| 275 | 1.38 | 1.02 |
| 300 | 1.11 | 1.01 |

**Important question:** given the way they are built, the 2 GRUs should be effectively independent – apart from moving the embeddings of words&nodes.

Is the fact that both move the embeddings enough to justify a degree of mutual influence?

We review how the main GRU alone overfits on the globals of the fragment of SemCor.

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| model: GRUbase2 | 1) The word embedding of the current global (d=300) | batch\_size=4 |
| Globals’ prediction: **Main GRU** with 3 layers (1150 x3) | TBPTT length=8 |
| Followed by linear2Globals **FF-NN** | learning rate=10^(-4) |
|  |  | grapharea=32, hops=2 |

|  |  |
| --- | --- |
| *Epoch* | *Training perplexity on globals* |
| 1 | 21184.4 |
| 2 | 16478.93 |
| 3 | 3339.3 |
| 4 | 359.31 |
| 5 | 192.21 |
| 6 | 100.5 |
| 7 | 74.33 |
| 8 | 68.74 |
| 9 | 61.69 |
| 10 | 64.21 |
| 30 | 59.63 |
| 50 | 59.37 |
| 75 | 59.34 |
| 100 | 59.2 |
| 150 | 59.1 |
| 200 | 59.12 |
| 250 | 59.09 |
| 300 | 58.38 |

This time we had a slightly better random initialization.

It never stops “hovering” at ~59, where it’s predicting the most common tokens - ", the, <unk>, of, . .

Let us try this again with a higher learning rate.

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| model: GRUbase2 | 1) The word embedding of the current global (d=300) | batch\_size=4 |
| Globals’ prediction: **Main GRU** with 3 layers (1150 x3) | TBPTT length=8 |
| Followed by linear2Globals **FF-NN** | learning rate=**10^(-3)** |
|  |  | grapharea=32, hops=2 |

|  |  |
| --- | --- |
| *Epoch* | *Training perplexity on globals* |
| 1 | 4839.46 |
| 2 | 98.55 |
| 3 | 73.22 |
| 4 | 65.45 |
| 5 | 65.05 |
| 6 | 61.82 |
| 7 | 57.81 |
| 8 | 60.41 |
| 9 | 58.0 |
| 10 | 63.08 |
| 30 | 63.02 |
| 50 | 48.98 |
| 75 | 36.56 |
| 100 | 27.55 |
| 150 | 19.68 |
| 200 | 14.46 |
| 250 | 11.29 |
| 300 | 8.68 |

Why do I not manage to overfit on SemCor?

¨The correct prediction is not the first one, as it should be. Examples:

Label: the next global is: Georgia(from 69)

INFO : Label: the next sense is: None(from -1)

INFO : The top- 5 predicted globals are:

INFO : Word: relative ; probability = 8.64%

INFO : Word: Georgia ; probability = 6.33%

INFO : Word: produced ; probability = 6.27%

INFO : Word: of ; probability = 6.2%

INFO : Word: primary election ; probability = 5.32%

INFO :

-----

Label: the next global is: s(from 5197)

INFO : Label: the next sense is: None(from -1)

INFO : The top- 5 predicted globals are:

INFO : Word: . ; probability = 15.55%

INFO : Word: in ; probability = 8.96%

INFO : Word: , ; probability = 7.06%

INFO : Word: irregularities ; probability = 6.73%

INFO : Word: s ; probability = 6.41%

INFO :

-----

Label: the next global is: registration(from 70)

INFO : Label: the next sense is: registration.n.01(from 19059)

INFO : The top- 5 predicted globals are:

INFO : Word: been ; probability = 7.61%

INFO : Word: over-all ; probability = 7.54%

INFO : Word: recent ; probability = 6.98%

INFO : Word: primary election ; probability = 6.82%

INFO : Word: registration ; probability = 6.58%

INFO :

-----

Let us try to decrease the dimensions of the network: for the 2nd and 3rd layer, hidden\_units=600 (from 1150)

**Mini-experiment – smaller network**

DataParallel(

(module): GRU\_base2(

(main\_gru): GRU(300, 600, num\_layers=3)

(linear2global): Linear(in\_features=600, out\_features=21988, bias=True)

)

)

INFO : Parameters:

INFO : ('module.X', torch.Size([99963, 300]), torch.float32, True)

('module.select\_first\_indices', torch.Size([6]), torch.float32, False)

('module.embedding\_zeros', torch.Size([1, 300]), torch.float32, False)

('module.memory\_hn', torch.Size([3, 4, 600]), torch.float32, False)

('module.memory\_hn\_senses', torch.Size([2, 4, 600]), torch.float32, False)

('module.main\_gru.weight\_ih\_l0', torch.Size([1800, 300]), torch.float32, True)

('module.main\_gru.weight\_hh\_l0', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.bias\_ih\_l0', torch.Size([1800]), torch.float32, True)

('module.main\_gru.bias\_hh\_l0', torch.Size([1800]), torch.float32, True)

('module.main\_gru.weight\_ih\_l1', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.weight\_hh\_l1', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.bias\_ih\_l1', torch.Size([1800]), torch.float32, True)

('module.main\_gru.bias\_hh\_l1', torch.Size([1800]), torch.float32, True)

('module.main\_gru.weight\_ih\_l2', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.weight\_hh\_l2', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.bias\_ih\_l2', torch.Size([1800]), torch.float32, True)

('module.main\_gru.bias\_hh\_l2', torch.Size([1800]), torch.float32, True)

('module.linear2global.weight', torch.Size([21988, 600]), torch.float32, True)

('module.linear2global.bias', torch.Size([21988]), torch.float32, True)

INFO : Number of trainable parameters=49.15M

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| model: GRUbase2 | 1) The word embedding of the current global (d=300) | batch\_size=4 |
| Globals’ prediction: **Main GRU** with 3 layers (hd=600) | TBPTT length=8 |
| Followed by linear2Globals **FF-NN** | learning rate=**10^(-3)** |
|  |  | grapharea=32, hops=2 |

|  |  |
| --- | --- |
| *Epoch* | *Training perplexity on globals* |
| 1 | 8470.51 |
| 2 | 168.78 |
| 3 | 75.62 |
| 4 | 63.13 |
| 5 | 63.36 |
| 6 | 61.18 |
| 7 | 56.09 |
| 8 | 58.49 |
| 9 | 56.59 |
| 10 | 61.41 |
| 30 | 61.02 |
| 50 | 61.01 |
| 75 | 44.54 |
| 100 | 30.0 |
| 150 | 14.0 |
| 200 | 8.75 |
| 250 | 5.52 |
| 300 | 3.99 |

This time we have the correct prediction as among the most likely – not always the most likely one, but we can surmise it would become such with ~50 or 100 epochs.

Label: the next global is: It(from 78)

INFO : Label: the next sense is: None(from -1)

INFO : The top- 5 predicted globals are:

INFO : Word: The ; probability = 18.3%

INFO : Word: " ; probability = 13.33%

INFO : **Word: It ; probability = 10.97%**

INFO : Word: to ; probability = 10.87%

INFO : Word: that ; probability = 8.55%

INFO :

-----

Label: the next global is: recommended(from 79)

INFO : Label: the next sense is: recommend.v.01(from 18880)

INFO : The top- 5 predicted globals are:

INFO : Word: the ; probability = 18.07%

INFO : Word: investigate ; probability = 17.02%

INFO : **Word: recommended ; probability = 15.33%**

INFO : Word: term ; probability = 9.32%

INFO : Word: city ; probability = 8.91%

INFO :

-----

Label: the next global is: that(from 14)

INFO : Label: the next sense is: None(from -1)

INFO : The top- 5 predicted globals are:

INFO : Wo**rd: that ; probability = 37.19%**

INFO : Word: the ; probability = 11.67%

INFO : Word: evidence ; probability = 5.46%

INFO : Word: conducted ; probability = 5.08%

INFO : Word: reports ; probability = 4.72%

INFO :

-----

Label: the next global is: Fulton(from 44)

INFO : Label: the next sense is: location.n.01(from 13952)

INFO : The top- 5 predicted globals are:

INFO : **Word: Fulton ; probability = 22.13%**

INFO : Word: the ; probability = 13.96%

INFO : Word: investigate ; probability = 11.38%

INFO : Word: recommended ; probability = 10.7%

INFO : Word: city ; probability = 6.97%

INFO :

-----

Label: the next global is: legislators(from 80)

INFO : Label: the next sense is: legislator.n.01(from 13543)

INFO : The top- 5 predicted globals are:

INFO : **Word: legislators ; probability = 56.71%**

INFO : Word: <unk> ; probability = 18.73%

INFO : Word: Fulton ; probability = 10.65%

INFO : Word: , ; probability = 2.35%

INFO : Word: jury ; probability = 2.15%

INFO :

-----

Label: the next global is: act(from 81)

INFO : Label: the next sense is: act.v.01(from 260)

INFO : The top- 5 predicted globals are:

INFO : **Word: act ; probability = 37.36%**

INFO : Word: find ; probability = 13.51%

INFO : Word: conducted ; probability = 6.28%

INFO : Word: The ; probability = 4.85%

INFO : Word: of this ; probability = 4.45%

It is easier to overfit a model with fewer parameters, because it has less inertia: it takes fewer epochs for it to move from a “predict the most common token” local minimum to predicting the current token at each step.

Let us now try to overfit on globals + senses.

DataParallel(

(module): GRU\_base2(

(main\_gru): GRU(300, 600, num\_layers=3)

(gru\_senses): GRU(300, 600, num\_layers=3)

(linear2global): Linear(in\_features=600, out\_features=21988, bias=True)

(linear2senses): Linear(in\_features=600, out\_features=25986, bias=True)

)

)

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| model: GRUbase2 | 1) The word embedding of the current global (d=300) | batch\_size=4 |
| Globals’ prediction: **Main GRU** with 3 layers (hd=600) | TBPTT length=8 |
| Followed by linear2Globals **FF-NN** | learning rate=10^(-3) |
| Senses’ prediction: **GRU** with 3 layers (hd=600),  Followed by linear2Senses **FF-NN** |  | grapharea=32, hops=2 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Train.PPL on globals (without sense prediction)* | *Training perplexity on senses* |
| 1 | 8902.51 | *8470.51* | 19133.89 |
| 2 | 173.04 | *168.78* | 304.25 |
| 3 | **76.14** | ***75.62*** | 67.81 |
| 4 | 62.71 | *63.13* | 55.04 |
| 5 | 63.94 | *63.36* | 50.21 |
| 10 | 61.41 | *61.18* | 49.39 |
| 30 | 61.01 | *56.09* | 49.12 |
| 50 | **61.01** | ***58.49*** | 49.13 |
| 75 | **43.16** | ***56.59*** | 48.61 |
| 100 | 23.94 | *61.41* | 32.57 |
| 125 | 16.34 | *61.02* | 21.9 |
| 150 | 11.62 | *61.01* | 13.04 |
| 175 | 8.91 | *44.54* | 6.18 |
| 200 | 7.18 | *30.0* | 2.86 |
| 225 | 6.19 | *14.0* | 1.56 |
| 250 | 4.85 | *8.75* | 1.2 |
| 275 | 3.95 | *5.52* | 1.08 |
| 300 |  | *3.99* |  |

For reproducibility, from now on we will use torch.manual\_seed(seed)

### Overfitting base GRU on SemCor globals

DataParallel(

(module): GRU\_base2(

(main\_gru): GRU(300, 600, num\_layers=3)

(linear2global): Linear(in\_features=600, out\_features=21988, bias=True)

)

)

INFO : Parameters:

INFO : ('module.X', torch.Size([99963, 300]), torch.float32, True)

('module.select\_first\_indices', torch.Size([6]), torch.float32, False)

('module.embedding\_zeros', torch.Size([1, 300]), torch.float32, False)

('module.memory\_hn', torch.Size([3, 4, 600]), torch.float32, False)

('module.memory\_hn\_senses', torch.Size([2, 4, 600]), torch.float32, False)

('module.main\_gru.weight\_ih\_l0', torch.Size([1800, 300]), torch.float32, True)

('module.main\_gru.weight\_hh\_l0', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.bias\_ih\_l0', torch.Size([1800]), torch.float32, True)

('module.main\_gru.bias\_hh\_l0', torch.Size([1800]), torch.float32, True)

('module.main\_gru.weight\_ih\_l1', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.weight\_hh\_l1', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.bias\_ih\_l1', torch.Size([1800]), torch.float32, True)

('module.main\_gru.bias\_hh\_l1', torch.Size([1800]), torch.float32, True)

('module.main\_gru.weight\_ih\_l2', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.weight\_hh\_l2', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.bias\_ih\_l2', torch.Size([1800]), torch.float32, True)

('module.main\_gru.bias\_hh\_l2', torch.Size([1800]), torch.float32, True)

('module.linear2global.weight', torch.Size([21988, 600]), torch.float32, True)

('module.linear2global.bias', torch.Size([21988]), torch.float32, True)

INFO : Number of trainable parameters=49.15M

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| model: GRUbase2 | 1) The word embedding of the current global (d=300) | batch\_size=4 |
| Globals’ prediction: **Main GRU** with 3 layers (hd=600) | TBPTT length=8 |
| Followed by linear2Globals **FF-NN** | learning rate=**10^(-3)** |
|  |  | grapharea=32, hops=2 |

|  |  |
| --- | --- |
| *Epoch* | *Training perplexity on globals* |
| 1 | 8908.56 |
| 2 | 163.35 |
| 3 | 76.22 |
| 4 | 62.64 |
| 5 | 62.95 |
| 10 | 61.31 |
| 30 | 61.03 |
| 50 | 61.04 |
| 75 | 44.11 |
| 100 | 37.3 |
| 150 | 19.39 |
| 200 | 12.73 |
| 250 | 7.89 |
| 300 | 4.8 |
| 400 | 1.89 |

Double-checking that setting the PyTorch manual random seed brings the same results in the experiment: 8908.56, 163.35, 76.22,… yes, confirmed.

**Experiment II**:

We set the word embeddings matrix X as requires\_grad=False in the GRUbase2 model.

We try again to overfit on the globals-only task with the GRU.

Then, we will add the senses’ prediction part of the model, and since there are no other elements in common apart from X, we expect the globals’ prediction to be unchanged.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Epoch* | *Training perplexity on globals* |  | *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |
| 1 | 8914.2 |  | 1 | 9238.8 | 16922.44 |
| 2 | 163.32 |  | 2 | 175.25 | 277.27 |
| 3 | 76.21 |  | 3 | 75.28 | 66.57 |
| 10 | 61.31 |  | 10 | 61.3 | 49.36 |
| 50 | 61.04 |  | 50 | 59.56 | 49.14 |
| 100 | 47.01 |  | 100 | 31.19 | 29.46 |
| 200 | 18.6 |  | 200 | 12.87 | 9.4 |
| 300 | 7.16 |  | 300 | ... |  |
| 400 | 5.11 |  | 400 | … |  |
|  |  |  | 500 | 3.13 | 1.18 |

It seems that we do have different results.

This may be due to loss=loss\_global+loss\_sense.

It can still be correct. However, we should **verify** that after a label (global=any, sense=-1) we do not modify the gru\_senses.

### Preliminary experiment – GRUbase2 w/senses on SemCor

INFO : Hyperparameters: \_batchPerSeqlen32\_area32\_lr5e-05\_epochs100

INFO : Model:

INFO : DataParallel(

(module): GRU\_base2(

(main\_gru): GRU(300, 600, num\_layers=3)

(gru\_senses): GRU(300, 600, num\_layers=3)

(linear2global): Linear(in\_features=600, out\_features=21988, bias=True)

(linear2senses): Linear(in\_features=600, out\_features=25986, bias=True)

)

)

INFO : Parameters:

INFO : ('module.X', torch.Size([99963, 300]), torch.float32, True)

('module.select\_first\_indices', torch.Size([6]), torch.float32, False)

('module.embedding\_zeros', torch.Size([1, 300]), torch.float32, False)

('module.memory\_hn', torch.Size([3, 40, 600]), torch.float32, False)

('module.memory\_hn\_senses', torch.Size([3, 40, 600]), torch.float32, False)

('module.main\_gru.weight\_ih\_l0', torch.Size([1800, 300]), torch.float32, True)

('module.main\_gru.weight\_hh\_l0', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.bias\_ih\_l0', torch.Size([1800]), torch.float32, True)

('module.main\_gru.bias\_hh\_l0', torch.Size([1800]), torch.float32, True)

('module.main\_gru.weight\_ih\_l1', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.weight\_hh\_l1', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.bias\_ih\_l1', torch.Size([1800]), torch.float32, True)

('module.main\_gru.bias\_hh\_l1', torch.Size([1800]), torch.float32, True)

('module.main\_gru.weight\_ih\_l2', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.weight\_hh\_l2', torch.Size([1800, 600]), torch.float32, True)

('module.main\_gru.bias\_ih\_l2', torch.Size([1800]), torch.float32, True)

('module.main\_gru.bias\_hh\_l2', torch.Size([1800]), torch.float32, True)

('module.gru\_senses.weight\_ih\_l0', torch.Size([1800, 300]), torch.float32, True)

('module.gru\_senses.weight\_hh\_l0', torch.Size([1800, 600]), torch.float32, True)

('module.gru\_senses.bias\_ih\_l0', torch.Size([1800]), torch.float32, True)

('module.gru\_senses.bias\_hh\_l0', torch.Size([1800]), torch.float32, True)

('module.gru\_senses.weight\_ih\_l1', torch.Size([1800, 600]), torch.float32, True)

('module.gru\_senses.weight\_hh\_l1', torch.Size([1800, 600]), torch.float32, True)

('module.gru\_senses.bias\_ih\_l1', torch.Size([1800]), torch.float32, True)

('module.gru\_senses.bias\_hh\_l1', torch.Size([1800]), torch.float32, True)

('module.gru\_senses.weight\_ih\_l2', torch.Size([1800, 600]), torch.float32, True)

('module.gru\_senses.weight\_hh\_l2', torch.Size([1800, 600]), torch.float32, True)

('module.gru\_senses.bias\_ih\_l2', torch.Size([1800]), torch.float32, True)

('module.gru\_senses.bias\_hh\_l2', torch.Size([1800]), torch.float32, True)

('module.linear2global.weight', torch.Size([21988, 600]), torch.float32, True)

('module.linear2global.bias', torch.Size([21988]), torch.float32, True)

('module.linear2senses.weight', torch.Size([25986, 600]), torch.float32, True)

('module.linear2senses.bias', torch.Size([25986]), torch.float32, True)

INFO : Number of trainable parameters=70.73M

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| model: GRUbase2 | 1) The word embedding of the current global (d=300) | batch\_size=40 |
| Globals’ prediction: **Main GRU** with 3 layers (hd=600) | TBPTT length=35 |
| Followed by linear2Globals **FF-NN** | learning rate=5\*10^(-5) |
| Senses’ prediction: **GRU** with 3 layers (hd=600),  Followed by linear2Senses **FF-NN** |  | grapharea=32, hops=2 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* | *Validation PPL on globals* | *Validation PPL on senses* |
| 1 | 1117.21 | 5521.64 | 513.59 | 897.91 |
| 2 | 738.79 | 3522.27 | 506.68 | 849.49 |
| 3 | 617.83 | 3365.4 | 395.15 | **826.48** |
| 4 | 472.55 | 3233.66 | 338.72 | 827.52 |
| 5 | 419.23 | 3032.93 | 313.37 | 835.21 |
| 6 |  |  |  |  |
| 7 |  |  |  |  |
| 8 |  |  |  |  |
| 9 |  |  |  |  |
| 10 |  |  |  |  |
| 11 |  |  |  |  |
| 12 |  |  |  |  |
| 13 |  |  |  |  |
| 14 |  |  |  |  |
| 15 |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

(note: for SemCor’s 650K words, compared to WT-2’s 2Mln words, it may be opportune to increase the learning rate from 5e-05 to 1e-04)

We do not manage to learn the senses, but the validation PPL has a much more reasonable trend – using the ignore\_index flag on the torch.nn.functional.nll\_loss fixed the computation of the senses’ loss.

The inability to learn can be explained by the ‘snapshot’ way the gru\_senses works, as explained elsewhere: it does not read the whole text word-by-word, but only sees the locations that have a sense label. This means that the GRU memory does not work the way it’s supposed to.

# Model Verification and Modification

## The loss when operating on senses

We launch GRUbase2 on a fragment of SemCor, and examine what happens when we compute the loss and after we

call loss.backward()

We set batch\_size=1, sequence\_length=1…

batch\_labels = tensor([[ 2, -1]])

Shape of the batch\_input tensor: torch.Size([1, 1, 1150])

The model, at the start of the forward:

(objective: the weights of the gru\_senses should *not* change)

self.memory\_hn.shape=(3,1,600)

At the start, it’s a tensor made entirely of zeros.

self.memory\_hn.nonzero()=tensor([],…)

The same for self.memory\_hn\_senses. Shape=(3,1,600), .nonzero()=tensor([])

We examine

self.main\_gru.weight\_ih\_l1.mean()=tensor(-4.3736e-05, grad\_fn=<MeanBackward0>)

self.main\_gru.weight\_hh\_l2.mean()=tensor(-8.4784e-06, grad\_fn=<MeanBackward0>)

, that we expect to change, and

self.gru\_senses.weight\_ih\_l1.mean()=tensor(-1.2207e-05, grad\_fn=<MeanBackward0>)

self.gru\_senses.weight\_hh\_l2.mean()=tensor(-2.3060e-05, grad\_fn=<MeanBackward0>)

that we expect to remain the same.

x\_indices\_g= tensor([ 1, 47975, 34783, 0, 32616])

edge\_index\_g= tensor([[1, 2, 2, 2, 2, 4, 2],

[0, 3, 0, 2, 2, 2, 4]])

edge\_type\_g= tensor([0, 2, 2, 3, 3, 3, 3])

x\_indices\_s= tensor([22232, 70206, 96175, 96176, 96177, 96178, 26158, 22228, 22229, 22230, 22227, 22231, 36926, 26852, 28113, 34096, 29324, 29292, 37538, 26756,

30394, 29242

edge\_index\_s= tensor([[ 1, 2, 3, 4, 5, 6, 6, 6, 6, 6, 6, 17, 6, 19, 6, 15, 6, 6,

12, 6, 6, 6, 6, 6, 6, 6, 18, 6, 16, 6, 20, 6, 17, 6, 6, 6,

6, 6, 6, 6, 13, 6, 14, 21, 6],

[ 0, 0, 0, 0, 0, 10, 7, 8, 9, 11, 0, 6, 17, 6, 19, 6, 15, 12,

6, 6, 6, 6, 6, 6, 6, 18, 6, 16, 6, 20, 6, 17, 6, 6, 6, 6,

6, 6, 6, 13, 6, 14, 6, 6, 21]])

edge\_type\_s= tensor([0, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,

3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3])

We compute the logits for both globals and senses, but only the predictions\_globals should encounter a meaningful label in the next step…

loss\_global= tensor(10.0134, grad\_fn=<NllLossBackward>)

Then, we have:

loss\_sense = tfunc.nll\_loss(predictions\_senses, batch\_labels\_senses, ignore\_index=-1)

with predictions\_senses.shape=(1,25986) and batch\_labels\_senses=tensor([-1])

loss\_sense= tensor(0., grad\_fn=<NllLossBackward>)

Then, in the training loop we execute:

batch\_sense\_tokens = (batch\_labels.t()[1][batch\_labels.t()[1]!=-1].shape[0])  
sum\_epoch\_loss\_sense = sum\_epoch\_loss\_sense + loss\_sense.item() \* batch\_sense\_tokens

epoch\_senselabeled\_tokens = epoch\_senselabeled\_tokens + batch\_sense\_tokens

And

loss = loss\_global + loss\_sense

loss.backward()

Now,

model.main\_gru.weight\_ih\_l1.mean()=tensor(-4.3115e-05, grad\_fn=<MeanBackward0>)

(from -4.3736e-05)

model.gru\_senses.weight\_ih\_l1.mean()=tensor(-1.2207e-05, grad\_fn=<MeanBackward0>)

(from -1.2207e-05) # unchanged, as expected.

### Considerations

The gru\_senses does not “see” anything related to the processing of the text word-after-word, as instead the main\_gru does.

It sees the current “snapshot” of input signals (whether that be the word embedding alone or input from the graph). gru\_senses adjusts itself and then waits “dormant” for a number of words, until the next sense label comes along.

Thus, gru\_senses has no understanding of the text / of the language. It proceeds as follows:

Label: the next global is: <unk>(from 21987) (from: Fulton\_County\_Grand\_Jury)

Label: the next sense is: group.n.01(from 10898)

INFO : batch\_labels=tensor([[21987, 10898]])

INFO : loss\_sense=tensor(10.1451, grad\_fn=<NllLossBackward>)

Label: the next global is: said(from 1)

INFO : Label: the next sense is: state.v.01(from 22232)

INFO : batch\_labels=tensor([[ 1, 22232]])

INFO : loss\_sense=tensor(10.1698, grad\_fn=<NllLossBackward>)

Label: the next global is: Friday(from 2)

INFO : Label: the next sense is: None(from -1)

INFO : batch\_labels=tensor([[ 2, -1]])

INFO : loss\_sense=tensor(0., grad\_fn=<NllLossBackward>)

Label: the next global is: an(from 3)

INFO : Label: the next sense is: None(from -1)

INFO : batch\_labels=tensor([[ 3, -1]])

INFO : loss\_sense=tensor(0., grad\_fn=<NllLossBackward>)

Label: the next global is: investigation(from 4)

INFO : Label: the next sense is: probe.n.01(from 17882)

INFO : batch\_labels=tensor([[ 4, 17882]])

INFO : loss\_sense=tensor(10.2071, grad\_fn=<NllLossBackward>)

## Graph retrieval

### Boundaries of graph node types

(n: node types’ boundaries are at: 25986, 47974, 73960)

**Senses: (0, 25986). Globals: (25986, 47974). Definitions: (47974, 73960). Examples: (73960, end)**

### Analysis…

Let us examine:

x\_indices\_g= tensor([ 1, 47975, 34783, 0, 32616])

# Likely mistake: the 1st global has not been made a relative index to get the correct row from the X matrix.

x\_indices\_s= tensor([22232, 70206, 96175, 96176, 96177, 96178, 26158, 22228, 22229, 22230, 22227, 22231, 36926, 26852, 28113, 34096, 29324, 29292, 37538, 26756,

30394, 29242])

When I am starting to read the SemCor fragment, how do I retrieve nodes?

The DataLoader calls:

get\_forwardinput\_forelement(global\_idx, sense\_idx, grapharea\_matrix, area\_size)

with:

global\_idx=0, sense\_idx=-1, grapharea\_matrix=…, area\_size=32

**if** (sense\_idx == -1): we use all torch.zeros(…).

Let us see the globals:

area\_x\_indices\_global, edge\_index\_global, edge\_type\_global = AD.get\_node\_data(grapharea\_matrix, global\_idx, area\_size)

Inside that function:

*# Accessing sparse matrix. Everything was shifted +1, so now: we ignore 0 ; we shift -1; we get the data*nodes\_ls =list(map(**lambda** value: value - 1, filter(**lambda** num: num != 0, grapharea\_matrix[i, 0:k].todense().tolist()[0])))

and nodes\_ls=<class 'list'>: [0.0, 47974.0, 34783.0, 1.0, 32616.0].

Which nodes did we get from the matrix?

0: 0th sense

47974: 47974-47974=0th definition

34783: 34783-25986= 8797th global

1: 1st sense

32616: 32616-25986=6630th global

Let us review what we get from the graph, instead of the graph matrix. Starting node: 0.

GA.get\_indices\_area\_toinclude(graph\_dataobject.edge\_index, graph\_dataobject.edge\_type, node\_index=0, area\_size=32, max\_hops=1):

nodes\_queue\_at\_current\_level=[0]

node\_edges=<class 'list'>: [(47974, 0, 0), (34783, 0, 2)]

nodes\_queue\_at\_next\_hop=we add 47974 and 34783

In the end, node\_indices\_ls, all\_edges\_retrieved\_ls=([0, 47974, 34783], [0, 51989])

Issue #1: if we are extracting the 1st global, we should not operate on index 0, but on the index (start\_of\_globals+0).

Examining: NI.get\_tokens\_tpls(…)

When we are in the dataloader’s \_\_getitem\_\_(self, index):

current\_token\_tpl=<class 'tuple'>: (0, -1)

next\_token\_tpl=<class 'tuple'>: (21987, 10898)

So the 2nd is correct, but the 1st is not.

What happens when token\_dict={'surface\_form': 'said', 'lemma': 'say', 'pos': 'VBD', 'wn16\_key': 'say%2:32:00::', 'wn30\_key': 'say%2:32:00::'}?

wordnet\_sense=state.v.01 ; From querying the SQL indices\_table.db, we get: sense\_index\_queryresult=<class 'tuple'>: (22232,)

For the global,

global\_absolute\_index = Utils.select\_from\_hdf5(globals\_vocabulary\_h5, **'vocabulary'**, [**'word'**], [word]).index[0]

I took out the addition of the last\_sense\_index…

global\_index = global\_absolute\_index *# + last\_idx\_senses; do not add this to globals, or we go beyond the n\_classes*

For the label purpose, the consideration is correct. However, it is necessary to add that term for the purpose of retrieving graph data and rows of X.

Therefore, in the current version, global\_index=1.

next\_token\_tpl=<class 'tuple'>: (1, 22232)

Whereas the current\_token\_tpl was correct: <class 'tuple'>: (21987, 10898)

Then we go on:

{'surface\_form': 'Friday', 'lemma': 'friday', 'pos': 'NNP', 'wn16\_key': 'friday%1:28:00::', 'wn30\_key': 'friday%1:28:00::'}

The sense is friday.n.01… The global\_absolute\_index is 2.

So we have:

current\_token\_tpl=<class 'tuple'>: (1, 22232) , next\_token\_tpl=<class 'tuple'>: (2, -1)

the current tuple, in the meantime, leads us to get\_node\_data with i=1.

nodes=tensor([ 1(sense!), 47975(def), 34783(global), 0(sense), 32616(global)])

edgeindex.T=tensor([[1, 0],

[2, 3], [2, 0],

[2, 2], [2, 2], [4, 2], [2, 4]])

edgetype=tensor([0, 2, 2, 3, 3, 3, 3]) (edge types are: 0=defs, 1=examples, 2=SenseChildren, 3=synonyms, 4=antonyms). (n: considering that here we may be from the point of view of the sense=1, instead of the global 25987)

What are the nodes we got?

“1”: It should have been the global: said, but this is sense n.1=ab.n.04

47975: 47975-47974= definition n. 1= from ab.n.04, the blood group whose red cells carry both the A and B antigens

34783: 34783-25986=8797th global= 8797 word=ab frequency=3

0: Sense n.0 = ab.n.02

32616: 32616-25986= 6630th global= 6630 word=abdominal frequency=3

While the token\_tuple that is also used for the labels and can use the “absolute” global index, the retrieval of data from the graph must use the **relative global index** (i.e. + last\_sense\_idx).

If I retrieve what I should, i.e. the node index 25986+1, from the graph:

nodes, edges=([25987], [77976])

The only node is the global n.1 in the vocabulary of globals:

“said”.

Since it’s not “say”, it does not have any connection in the graph. We should apply lemmatization before attempting node retrieval.

“say” is the global n. 2127.

If we execute get\_indices\_area\_toinclude(edge\_index, edge\_type, node\_index*=25986+2127=28113*, area\_size, max\_hops):

Using the node itself (during the 1st iteration, at hop=1)

node\_edges=<class 'list'>: [

(28113, 20173, 2),

(28113, 20178, 2),

(28113, 20177, 2),

(28113, 20175, 2),

(28113, 20176, 2),

(28113, 20174, 2),

(26631, 28113, 3), (28113, 26631, 3),

(28113, 26158, 3), (26158, 28113, 3),

(28113, 29596, 3), (29596, 28113, 3)

(28113, 43985, 3), (43985, 28113, 3),

(28113, 26599, 3), (26599, 28113, 3),

(28113, 28113, 3)]

Senses: say.n.01, say.v.07, say.v.08, say.v.09, say.v.10, say.v.11

***~~Note~~***~~: I am still with a displacement of 1. This displacement could be due to <unk>… I will have to check sense 20173 and what I retrieve from it…~~

Global synonyms – to find them, subtract 25968:

[645:order, 172:state, 3610:suppose, 17999:aforesaid, 613:read]

Moreover, now that we are able to make the correct request to the graph, I decide to operate with 1 hop of distance as the starting hyperparameter.

### Senses’ +1 displacement

Lastly, we review the senses’ retrieval:

when we have global n.1 (‘said’), we will operate with:

area\_x\_indices\_sense: tensor([

22232: sense ~~state.n.06 (or so we get from consulting indices\_table.sql)~~ state v.01,

70206: 22232nd definition (state.v.01 – express in words),

96175: 22215th example (state.v.01 – He said that he wanted to marry her),

96176: 22216th example (state.v.01 – tell me what is bothering you),

96177: “” state your opinion, 96178: “” state your name,

26158: 172nd global “state”,

22228: sense – from consulting: state.n.02,

22229: sense – from consulting: state.n.03,

22230: sense – from consulting: state.n.04,

22227: sense – from consulting: state.n.01,

22231: sense – from consulting: state.n.06,

36926: 10940th global “province”,

26852: 866th global “tell”, # I still have 2 hops in the graph area in this particular experiment, so it’s: sense->global->synonym.

28113: 2127th global “say”,

34096: 8110th global “express”,

etc. 29324, 29292, 37538, 26756, 30394, 29242])

This may be just due to the row-counting in indices\_table.sql.

The rows that we visualize in the table go from 1 to 25986.

When my sense-node index is 0, I actually refer to row 1 in the table.

(I adjust the meaning of the indices above)

## Adding lemmatization for the global node

If we encounter a word like ‘said’, it would be informative to include the graph-input-signal from the global node ‘say’.

# Structured prediction: SelectK – version 1

## Model

The prediction of senses should build upon the standard-LM prediction of globals.

Select the senses of the *k* most likely globals. In particular, we select their logits from the senses’ GRU.

We apply the softmax over them alone. Then, we remove a small amount of probability mass to give ε=10^(-6)

INFO : Model:

INFO : DataParallel( (module): SelectK(

(main\_gru): GRU(300, 1150, num\_layers=3)

(gru\_senses): GRU(300, 575, num\_layers=3)

(linear2global): Linear(in\_features=1150, out\_features=21988, bias=True)

(linear2senses): Linear(in\_features=575, out\_features=25986, bias=True)

) )

INFO : Parameters:

INFO : ('module.X', torch.Size([99963, 300]), torch.float32, True)

('module.select\_first\_indices', torch.Size([6]), torch.float32, False)

('module.embedding\_zeros', torch.Size([1, 300]), torch.float32, False)

('module.memory\_hn', torch.Size([3, 4, 1150]), torch.float32, False)

('module.memory\_hn\_senses', torch.Size([3, 4, 575]), torch.float32, False)

('module.main\_gru.weight\_ih\_l0', torch.Size([3450, 300]), torch.float32, True)

('module.main\_gru.weight\_hh\_l0', torch.Size([3450, 1150]), torch.float32, True)

('module.main\_gru.bias\_ih\_l0', torch.Size([3450]), torch.float32, True)

('module.main\_gru.bias\_hh\_l0', torch.Size([3450]), torch.float32, True)

('module.main\_gru.weight\_ih\_l1', torch.Size([3450, 1150]), torch.float32, True)

('module.main\_gru.weight\_hh\_l1', torch.Size([3450, 1150]), torch.float32, True)

('module.main\_gru.bias\_ih\_l1', torch.Size([3450]), torch.float32, True)

('module.main\_gru.bias\_hh\_l1', torch.Size([3450]), torch.float32, True)

('module.main\_gru.weight\_ih\_l2', torch.Size([3450, 1150]), torch.float32, True)

('module.main\_gru.weight\_hh\_l2', torch.Size([3450, 1150]), torch.float32, True)

('module.main\_gru.bias\_ih\_l2', torch.Size([3450]), torch.float32, True)

('module.main\_gru.bias\_hh\_l2', torch.Size([3450]), torch.float32, True)

('module.gru\_senses.weight\_ih\_l0', torch.Size([1725, 300]), torch.float32, True)

('module.gru\_senses.weight\_hh\_l0', torch.Size([1725, 575]), torch.float32, True)

('module.gru\_senses.bias\_ih\_l0', torch.Size([1725]), torch.float32, True)

('module.gru\_senses.bias\_hh\_l0', torch.Size([1725]), torch.float32, True)

('module.gru\_senses.weight\_ih\_l1', torch.Size([1725, 575]), torch.float32, True)

('module.gru\_senses.weight\_hh\_l1', torch.Size([1725, 575]), torch.float32, True)

('module.gru\_senses.bias\_ih\_l1', torch.Size([1725]), torch.float32, True)

('module.gru\_senses.bias\_hh\_l1', torch.Size([1725]), torch.float32, True)

('module.gru\_senses.weight\_ih\_l2', torch.Size([1725, 575]), torch.float32, True)

('module.gru\_senses.weight\_hh\_l2', torch.Size([1725, 575]), torch.float32, True)

('module.gru\_senses.bias\_ih\_l2', torch.Size([1725]), torch.float32, True)

('module.gru\_senses.bias\_hh\_l2', torch.Size([1725]), torch.float32, True)

('module.linear2global.weight', torch.Size([21988, 1150]), torch.float32, True)

('module.linear2global.bias', torch.Size([21988]), torch.float32, True)

('module.linear2senses.weight', torch.Size([25986, 575]), torch.float32, True)

('module.linear2senses.bias', torch.Size([25986]), torch.float32, True)

INFO : Number of trainable parameters=96.64M

## Mini-Experiment 1 – Overfit on fragment of SemCor

### Losses

batchPerSeqlen32\_area32\_lr0.0005\_epochs400

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| model: SelectK | 1) The word embedding of the current global (d=300) | batch\_size=4 |
| Globals’ prediction: **Main GRU** with 3 layers (d=1150) | TBPTT length=8 |
| Followed by linear2Globals **FF-NN** | learning rate=5\* 10^(-4) |
| Senses’ prediction:  **Senses GRU** (L=3, d=575), apply softmax on the selected logits of the senses of the most likely ***k*** globals |  | grapharea=32, hops=2 |

|  |  |  |
| --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |
| 1 | 9026.6 | 1000000.32 |
| 2 | 178.7 | 1000000.46 |
| 3 | 81.98 | 854154.08 |
| 4 | 61.46 | 1000000.32 |
| 5 | 64.17 | 1000000.32 |
| 10 | 61.24 | 1000000.32 |
| 30 | 60.88 | 1000000.32 |
| 40 | 58.78 | 1000000.32 |
| 50 | 45.55 | 1000000.32 |
| 75 | 29.57 | 1000000.32 |
| 100 | 17.73 | 1000000.32 |
| 125 | 8.1 | 530171.45 |
| 150 | 3.8 | 14308.24 |
| 200 | 1.66 | 1378.08 |
| 250 | 1.23 | 616.07 |
| 300 | 1.1 | 616.07 |

### Mini-experiment 1 – Analysis of the predictions

|  |  |  |
| --- | --- | --- |
| Label: the next global is: jury(from 19)  INFO : Label: the next sense is: jury.n.01(from 13063) | INFO : The top- 5 predicted globals are:  INFO : Word: jury ; p=91.75%  INFO : Word: September ; p=5.21%  INFO : Word: <unk> ; p=1.17%  INFO : Word: term ; p=0.78%  INFO : Word: often ; p=0.25% | INFO : The top- 5 predicted senses are:  INFO : Sense: jury.n.02 ; p = 49.45%  INFO : Sense: jury.n.01 ; p = 47.95% |
| Label: the next global is: find(from 67)  INFO : Label: the next sense is: **rule.v.04**(from 19931) | INFO : The top- 5 predicted globals are:  INFO : Word: find ; p=93.61%  INFO : Word: interest ; p=1.1%  INFO : Word: relative ; p=0.96%  INFO : Word: inadequate ; p=0.58%  INFO : Word: took place ; p=0.44% | INFO : The top- 5 predicted senses are:  INFO : Sense: find.v.03 ; p = 16.9%  INFO : Sense: find.v.05 ; p = 16.86%  INFO : Sense: find.v.13 ; p = 16.21%  INFO : Sense: find.v.10 ; p = 16.06%  INFO : Sense: find.v.15 ; p = 16.05% |
| Label: the next global is: registration(from 70)  INFO : Label: the next sense is: registration.n.01(from 19059) | INFO : The top- 5 predicted globals are:  INFO : Word: registration ; p=93.85%  INFO : Word: recent ; p=1.64%  INFO : Word: of this ; p=1.47%  INFO : Word: had ; p=0.4%  INFO : Word: by ; p=0.34% | INFO : The top- 5 predicted senses are:  INFO : Sense: registration.n.04 ; p = 25.92%  INFO : Sense: registration.n.01 ; p = 24.63%  INFO : Sense: registration.n.02 ; p = 24.03%  INFO : Sense: registration.n.03 ; p = 22.83% |

Etc.

2 observations follow:

* As we see in *registration*, the model does not choose the correct sense among those of the predicted global.
* As wee see in *find+rule.v.04*, sometimes the sense of a word will not be one of those directly associated with the word (here the text is “the court did find”…)

If we were able to choose the correct sense among those of the predicted global…

I examine the computation graph, going backwards from predictions\_senses, to see where it breaks…

predictions\_senses has \_grad=None, \_grad\_fn=None

senses\_softmax : the same

logits\_sense.\_grad\_fn= <AddmmBackward object at 0x1a866d3ad0>

Moreover:

logits\_global = self.linear2global(main\_gru\_out)

logits\_globa.\_grad\_fn = <AddmmBackward object at 0x1a8ae6ba50>

k\_globals\_indices = logits\_global.sort(descending=**True**).indices[:, 0:self.k]

k\_globals\_indices.\_grad\_fn=None

I try to change how I pick the logits\_senses, let us try to use index\_select to maintain the gradient…

## Mini-Experiment 2 – Overfit on fragment of SemCor

### Losses

batchPerSeqlen32\_area32\_lr0.0005\_epochs400

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| Globals’ prediction: **Main GRU** with 3 layers (d=1150) | 1) The word embedding of the current global (d=300) | batch\_size=4  TBPTT length=8 |
| linear2Globals **FF-NN** | grapharea=32, hops=2 |
| **Senses GRU** (L=3, d=575) into linear2Senses **FF-NN**, apply softmax on the logits of the senses of the most likely ***k*** globals | learning rate=5\* 10^(-4) |

|  |  |  |
| --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |
| 1 | 9817.6 | 1000000.32 |
| 2 | 178.86 | 1000000.46 |
| 3 | 79.23 | 1000000.32 |
| 4 | 63.24 | 1000000.32 |
| 5 | 63.86 | 1000000.32 |
| 10 | 61.42 | 1000000.32 |
| 35 | 60.81 | 1000000.32 |
| 50 | 42.0 | 1000000.32 |
| 75 | 25.35 | 1000000.32 |
| 100 | 15.01 | 856529.48 |
| 125 | 7.66 | 155110.68 |
| 150 | 3.62 | 5586.52 |
| 200 | 1.51 | 733.13 |
| 225 | 1.27 | 617.06 |
| 250 | 1.16 | 617.06 |
| 300 | 1.08 | 617.06 |

### Mini-experiment 2 – Analysis of the predictions

|  |  |  |
| --- | --- | --- |
| Label: the next global is: jury(from 19)  INFO : Label: the next sense is: jury.n.01(from 13063) | INFO : The top- 5 predicted globals are:  INFO : Word: jury ; p=96.72%  INFO : Word: September ; p=1.73%  INFO : Word: <unk> ; p=0.29%  INFO : Word: end ; p=0.26%  INFO : Word: manner ; p=0.22% | INFO : The top- 5 predicted senses are:  INFO : Sense: jury.n.02 ; p = 48.72%  INFO : Sense: jury.n.01 ; p = 48.68% |
| Label: the next global is: find(from 67)  INFO : Label: the next sense is: rule.v.04(from 19931) | INFO : The top- 5 predicted globals are:  INFO : Word: find ; p=95.2%  INFO : Word: , ; p=1.75%  INFO : Word: Georgia ; p=0.58%  INFO : Word: ambiguous ; p=0.43%  INFO : Word: produced ; p=0.29% | INFO : The top- 5 predicted senses are:  INFO : Sense: find.v.13 ; p = 16.69%  INFO : Sense: find.v.05 ; p = 16.6%  INFO : Sense: find.v.01 ; p = 16.48%  INFO : Sense: find.v.15 ; p = 16.19%  INFO : Sense: find.v.03 ; p = 15.74% |
| Label: the next global is: registration(from 70)  INFO : Label: the next sense is: registration.n.01(from 19059) | INFO : The top- 5 predicted globals are:  INFO : Word: registration ; p=90.77%  INFO : Word: recent ; p=5.32%  INFO : Word: recommended ; p=0.83%  INFO : Word: in ; p=0.7%  INFO : Word: a ; p=0.57% | INFO : The top- 5 predicted senses are:  INFO : Sense: registration.n.02 ; p = 25.15%  INFO : Sense: registration.n.03 ; p = 24.6%  INFO : Sense: registration.n.01 ; p = 24.55%  INFO : Sense: registration.n.04 ; p = 23.1% |
| … |  |  |
| Label: the next global is: act(from 81)  INFO : Label: the next sense is: act.v.01(from 260) | INFO : The top- 5 predicted globals are:  INFO : Word: act ; p=95.13%  INFO : Word: relative ; p=1.86%  INFO : Word: October ; p=0.69%  INFO : Word: <unk> ; p=0.6%  INFO : Word: , ; p=0.28% | INFO : The top- 5 predicted senses are:  INFO : Sense: act.v.04 ; p = 7.72%  INFO : Sense: act.v.03 ; p = 7.6%  INFO : Sense: act.n.03 ; p = 7.58%  INFO : Sense: act.n.05 ; p = 7.56%  INFO : Sense: act.v.02 ; p = 7.55% |

Both problems spotted earlier remain:

**Some senses do not coincide with the global word.**

Examples:

|  |  |
| --- | --- |
| often | frequently.r.01 |
| outmoded | antique.s.02 |
| are | be.v.01 – we could find it if we lemmatized, but we retrieve INFO : Sense: are.n.01 ; p = 97.4%  This presents a case for lemmatizing always |
| find | rule.v.04 |

This could be solved by reading the training split of the sense-labeled corpus at graph creation, and connecting the global nodes to the sense nodes appropriately.

**We are unable to choose between the selected senses.**

We did not manage to reconnect the gradient to predictions\_senses yet.

logits\_selected\_senses.\_grad\_fn= <IndexSelectBackward object at 0x1a81070dd0>

…

softmax\_selected\_senses.\_grad\_fn=<SubBackward0 object at 0x1ae7219450>

I had to modify:

sample\_senses\_softmax.masked\_scatter\_(mask=i\_senseneighbours\_mask[s], source=softmax\_selected\_senses~~.data.clone())~~

**Temp – issue:**

RuntimeError: one of the variables needed for gradient computation has been modified by an inplace operation: [CUDABoolType [25986]] is at version 21; expected version 20 instead.

Warning: Error detected in torch::autograd::CopySlices. Traceback of forward call that caused the error:

result = self.forward(\*input, \*\*kwargs)

File "/home/andrealk3/ALMA\_PhD/Code/TASK1\_Gbwe/GNN/Models/Senses.py", line 234, in forward

senses\_softmax[s].masked\_scatter\_(mask=i\_senseneighbours\_mask[s], source=softmax\_selected\_senses)

(print\_stack at /pytorch/torch/csrc/autograd/python\_anomaly\_mode.cpp:60)

Traceback (most recent call last):

File "<stdin>", line 1, in <module>

File "/home/andrealk3/ALMA\_PhD/Code/TASK1\_Gbwe/GNN/Training.py", line 195, in training\_loop

loss.backward()

## Mini-experiment 3

### Losses

Hyperparameters: \_batchPerSeqlen32\_area32\_lr0.0005\_epochs300

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| Globals’ prediction: **Main GRU** with 3 layers (d=1150) | 1) The word embedding of the current global (d=300) | batch\_size=4  TBPTT length=8 |
| linear2Globals **FF-NN** | grapharea=32, hops=2 |
| **Senses GRU** (L=3, d=575) into linear2Senses **FF-NN**, apply softmax on the logits of the senses of the most likely ***k*** globals | learning rate=5\* 10^(-4) |

|  |  |  |
| --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |
| 1 | 9741.0 | 1000000.32 |
| 2 | 175.6 | 1000000.46 |
| 3 | 79.5 | 1000000.32 |
| 4 | 62.28 | 1000000.32 |
| 5 | 63.95 | 1000000.32 |
| 10 | 61.28 | 1000000.32 |
| 35 | 60.88 | 1000000.32 |
| 50 | 52.69 | 1000000.32 |
| 75 | 31.13 | 1000000.32 |
| 100 | 18.38 | 421930.79 |
| 125 | 10.65 | 355068.55 |
| 150 | 7.32 | 251424.08 |
| 200 | 3.26 | 3997.55 |
| 225 | 2.31 | 424.04 |
| 250 | 1.81 | 252.63 |
| 300 | 1.39 | 252.63 |

### Mini-experiment 3 – Analysis of the predictions

|  |  |  |
| --- | --- | --- |
| Label: the next global is: jury(from 19)  INFO : Label: the next sense is: jury.n.01(from 13063) | INFO : Word: jury ; p=70.15%  INFO : Word: September ; p=21.89%  INFO : Word: investigation ; p=1.17%  INFO : Word: <unk> ; p=1.11%  INFO : Word: any ; p=0.71% | INFO : Sense: jury.n.01 ; p = 98.7%  *This time we manage to pick jury.n.01, instead of being undecided between jury.n.01 and jury.n.02* |
| Label: the next global is: find(from 67)  INFO : Label: the next sense is: rule.v.04(from 19931) | INFO : Word: find ; p=80.47%  INFO : Word: irregularities ; p=4.93%  INFO : Word: or ; p=3.77%  INFO : Word: that ; p=3.33%  INFO : Word: said ; p=2.67% | INFO : Sense: find.v.10 ; p = 21.37%  INFO : Sense: find.v.05 ; p = 20.19%  INFO : Sense: find.v.15 ; p = 16.35%  INFO : Sense: find.v.01 ; p = 14.71%  INFO : Sense: find.v.03 ; p = 12.83%  *We select the senses of the global ‘find’, but we* *are unable to find rule.v.04. That issue must still be addressed* |
| Label: the next global is: registration(from 70)  INFO : Label: the next sense is: registration.n.01(from 19059) | INFO : Word: registration ; p=66.08%  INFO : Word: received ; p=6.73%  INFO : Word: to ; p=6.45%  INFO : Word: , ; p=3.85%  INFO : Word: The ; p=3.11% | INFO : Sense: registration.n.01 ; p = 99.35%  *Instead of being split at 25% x4 between n.01,…,n.04* |
| … |  |  |
| Label: the next global is: legislators(from 80)  INFO : Label: the next sense is: legislator.n.01(from 13543) | INFO : Word: legislators ; p=87.93%  INFO : Word: been ; p=1.96%  INFO : Word: of ; p=1.93%  INFO : Word: and ; p=1.78%  INFO : Word: <unk> ; p=1.55% | INFO : Sense: legislator.n.01 ; p = 97.4% |
| Label: the next global is: act(from 81)  INFO : Label: the next sense is: act.v.01(from 260) | INFO : Word: act ; p=73.69%  INFO : Word: deserves ; p=4.82%  INFO : Word: that ; p=3.08%  INFO : Word: no ; p=2.64%  INFO : Word: . ; p=2.36% | INFO : The top- 5 predicted senses are:  *We have lost the senses, the previous version managed to pick them giving them a p=~7.5%* |

## Adding global-to-sense edges from the SLC

We have seen that sometimes the correct sense attributed to a word is not among those immediately associated with that word in the dictionary.

e.g. find -> rule.v.04

When we create the graph, it is opportune to read the training split of the senses’ corpus and add the senseChildren edges that derive from all occurrences of   
global\_w1 -> sense\_from\_w2

\* done when the global word is *not* in the sense denomination, e.g. say->state.v.04

Checking that the new senseChildren edges that we added are correct:

This time the adjacent nodes for ‘say’ (relative global 28113) are:

(remember: Senses: (0, 25986). Globals: (25986, 47974). Definitions: (47974, 73960). Examples: (73960, end) )

[28113, global 2127, ‘say’

22232, sense ‘state.v.01’

20173, sense ‘say.n.01’

20178, say.v.11

20175, say.v.08

18711, read.v.02

20177, say.v.10

20176, say.v.09

524, ﻿aforesaid.s.01

20174, ﻿say.v.07

16073, order.v.02

22929, ﻿suppose.v.01

26631, global 645, ‘order’

43985, global 17999 ’aforesaid’

26158, global 172 ‘state’

29596, global 3610 ‘suppose’

26599] global 3613 ‘read

The connections are all pertinent. We have the old senses of ‘say’ plus the external senseChildren connections that we have read from the training split of SemCor.

The larger the SenseLabeledCorpus, the more of these connections we will include.

### Analysis of predictions – why do we have no senses for ‘act’, SensePerplexity=nan, etc.

At Epoch 200, using the GPUs:

In the model’s forward(), k\_globals\_words=

<unk>, many of, The,

jury:

softmax\_selected\_senses = tensor([9.9997e-01, 3.2930e-05], device='cuda:0', grad\_fn=<SoftmaxBackward>)

said:

softmax\_selected\_senses = tensor([5.1017e-06, 7.6327e-06, 1.3423e-05, 5.6478e-06, 5.9111e-06, 6.8113e-06,

8.5246e-06, 8.5297e-06, 1.0177e-05, 9.9992e-01, 9.2005e-06],

device='cuda:0', grad\_fn=<SoftmaxBackward>)

ok, found the error.

When the probability of not-correct senses is < ε and we subtract ε, we eventually get something like

softmax\_selected\_senses = tensor([-0.0024, -0.0024, -0.0023, -0.0024, -0.0024, -0.0024, -0.0024, -0.0024,

-0.0024, 0.9976, -0.0024], device='cuda:0', grad\_fn=<SubBackward0>)

Which breaks the computation.

Let us change the epsilon (the small probability mass given to all the non-selected senses) from 10^(-6) to 10^(-8).

Moreover, as a safety measure, I could either:

1. set the epsilon to be the minimum between [10^(-8)] and [the smallest probability accorded to one of the softmax\_selected\_senses + 10^(-9)]
2. do not modify the selected\_senses\_softmax values < 10^(-8)

## Mini-experiment 4

Note: to have faster experiments while I debug the numerical computation on the senses\_softmax, I now use an even smaller fragment of SemCor (**MiniDataset 2**, with 63 tokens instead of 180).

### Losses

INFO : Hyperparameters: \_batchPerSeqlen32\_area32\_lr0.0001\_epochs300

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| Globals’ prediction: **Main GRU** with 3 layers (d=1150) | 1) The word embedding of the current global (d=300) | batch\_size=4  TBPTT length=8 |
| linear2Globals **FF-NN** | grapharea=32, hops=2 |
| **Senses GRU** (L=3, d=575) into linear2Senses **FF-NN**, apply softmax on the logits of the senses of the most likely ***k*** globals | learning rate=10^(-4) |

|  |  |  |
| --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |
| 1 | 21472.66 | 100000216.32 |
| 2 | 20215.56 | 100000216.32 |
| 3 | 18591.51 | 100000216.32 |
| 10 | 178.51 | 100000216.32 |
| 50 | 23.89 | 100000216.32 |
| 100 | 23.79 | 100000216.32 |
| 150 | 18.44 | 100000216.32 |
| 175 | 3.4 | 105.54 |
| 200 | 1.07 | 13.92 |
| 250 | 1.01 | 13.91 |
| 300 | 1.01 | 13.9 |

### Mini-experiment 4 – Analysis of the predictions

(standard correct predictions for globals and senses, with >99% probability for the label, are not reported)

|  |  |  |
| --- | --- | --- |
| Label: the next global is: investigation(from 4) |  | INFO : Label: the next sense is: probe.n.01(from 17882) |
| Label: the next global is: recent(from 8) |  | INFO : Label: the next sense is: late.s.03(from 13363) |
| Label: the next global is: primary election(from 9) |  | INFO : Label: the next sense is: primary.n.01(from 17809) |
| Label: the next global is: produced(from 10) |  | INFO : Label: the next sense is: produce.v.04(from 17913) |
| Label: the next global is: evidence(from 13) |  | INFO : Label: the next sense is: evidence.n.01(from 8606) |
| Label: the next global is: took place(from 17) |  | INFO : Label: the next sense is: happen.v.01(from 11119) |
| Label: the next global is: jury(from 19) |  | INFO : Label: the next sense is: jury.n.01(from 13063) |
| Label: the next global is: further(from 20) |  | INFO : Label: the next sense is: far.r.02(from 9122) |
| Label: the next global is: said(from 1) |  | INFO : Label: the next sense is: state.v.01(from 22232) |
| Label: the next global is: term(from 22) |  | INFO : Label: the next sense is: term.n.02(from 23451) |
| Label: the next global is: end(from 23) |  | INFO : Label: the next sense is: end.n.02(from 8298) |
| Label: the next global is: <unk>(from 21987) | *Probably the global is* Fulton\_County\_Grand\_Jury  *reported as <unk>* | INFO : Label: the next sense is: group.n.01(from 10898) |
| Label: the next global is: <unk>(from 21987) | *Probably the global is* City\_Executive\_Committee  *reported as <unk>* | INFO : Label: the next sense is: group.n.01(from 10898) |
| Label: the next global is: said(from 1) |  | INFO : Label: the next sense is: state.v.01(from 22232) |
| Label: the next global is: investigation(from 4) |  | INFO : Label: the next sense is: probe.n.01(from 17882) |
| … etc (all the others are correct) |  |  |

Some globals that appear only once in the training corpus are set to <unk>, and when we overfit on a fragment we have <unk> as the most likely global, that has no senses. So we can not guess the senses correctly with SelectK=1

Moreover, on a rerun I have:

Label: the next global is: produced(from 10)

INFO : Label: the next sense is: produce.v.04(from 17913)

INFO : The top- 5 predicted senses are:

INFO :

Why do I not manage to predict the senses here? Must debug.

This time

“Label: the next global is: said(from 1)

INFO : Label: the next sense is: state.v.01(from 22232)”

does not work…

k\_globals\_words = ‘said’

k\_globals\_lemmatized = ‘say’

lemmatized\_indices=[28113]

sense\_neighbours\_t = tensor([22232, 20173, 20178, 20175, 18711, 20177, 20176, 524, 20174, 16073, 22929], device='cuda:0')

softmax\_selected\_senses= tensor([1.1597e-05, 1.1795e-05, 1.4018e-05, 8.8305e-06, 1.0876e-05, 1.1523e-05, 1.0502e-05, 8.8967e-06, 1.1274e-05, 9.9989e-01, 1.3601e-05], device='cuda:0', grad\_fn=<SoftmaxBackward>)

quantity\_to\_subtract\_from\_selected = 2.3613636363636363e-05

**softmax\_selected\_senses** = tensor([-1.2016e-05, -1.1818e-05, -9.5958e-06, -1.4783e-05, -1.2737e-05, -1.2091e-05, -1.3112e-05, -1.4717e-05, -1.2340e-05, 9.9986e-01,

-1.0013e-05], device='cuda:0', grad\_fn=<SubBackward0>)

The newest experiment, after the adjustments, gets:

Perplexity: Globals perplexity=1.2 Sense perplexity=4.27

<unk>, said and produced still do not work.

When I try to apply the fix, I get:

RuntimeError: one of the variables needed for gradient computation has been modified by an inplace operation: [torch.cuda.FloatTensor [1]], which is output 0 of SoftmaxBackward, is at version 1; expected version 0 instead. Hint: enable anomaly detection to find the operation that failed to compute its gradient, with torch.autograd.set\_detect\_anomaly(True).

And what if I just didn’t subtract anything? Let the nll\_loss sort it out, even if the sum of the softmax values is \*slightly\* more than one (order of magnitude: epsilon x |vocab\_senses|, ~10^-8 x 10^5 =~ 10^-3).

Perplexity: Globals perplexity=1.01 Sense perplexity=4.27

Label: the next global is: said(from 1)

INFO : Label: the next sense is: state.v.01(from 22232)

INFO : Sense: state.v.01 ; p = 99.98%, now it works.

Now the only places where we do not manage to predict the senses are those where the global is <unk>.

## Mini-experiment 5

As the last modification to test before trying out the model on SemCor, we increase the number of globals we select the senses from, **k**, from 1 to 5.

### Losses

INFO : Hyperparameters: \_batchPerSeqlen32\_area32\_lr0.0001\_epochs300

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| Globals’ prediction: **Main GRU** with 3 layers (d=1150) | 1) The word embedding of the current global (d=300) | batch\_size=4  TBPTT length=8 |
| linear2Globals **FF-NN** | grapharea=32, hops=2 |
| **Senses GRU** (L=3, d=575) into linear2Senses **FF-NN**, apply softmax on the logits of the senses of the most likely ***k=5*** globals | learning rate=10^(-4) |

|  |  |  |
| --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |
| 1 | 21796.59 | 100000216.32 |
| 2 | 20597.76 | 5375446.1 |
| 3 | 19078.43 | 3381108.91 |
| 10 | 153.35 | 28035284.03 |
| 50 | 23.89 | 3075497.52 |
| 100 | 23.77 | 2489515.49 |
| 150 | 23.33 | 114444.77 |
| 175 | 18.07 | 14368.72 |
| 200 | 8.21 | 52.94 |
| 250 | 1.25 | 16.76 |
| 300 | 1.03 | 17.94 |

### Mini-experiment 5 – Analysis of the predictions

Mostly correct, apart from <unk>s as expected, with one exception:

Label: the next global is: investigation(from 4)

INFO : Label: the next sense is: probe.n.01(from 17882)

INFO : The top- 5 predicted globals are:

**INFO : Word: investigation ; p=96.87%**

INFO : Word: <unk> ; p=0.85%

***INFO : Word: jury ; p=0.52%***

INFO : Word: Atlanta ; p=0.39%

INFO : Word: term ; p=0.28%

INFO : The top- 5 predicted senses are:

INFO : Sense: jury.n.02 ; p = 93.08%

INFO : Sense: probe.n.01 ; p = 6.22%

And one partial exception:

Label: the next global is: produced(from 10)

INFO : Label: the next sense is: produce.v.04(from 17913)

INFO : The top- 5 predicted globals are:

**INFO : Word: produced ; p=96.67%**

INFO : Word: " ; p=0.86%

INFO : Word: any ; p=0.64%

INFO : Word: recent ; p=0.38%

INFO : Word: primary election ; p=0.38%

INFO : The top- 5 predicted senses are:

**INFO : Sense: produce.v.04 ; p = 64.3%**

INFO : Sense: any.r.01 ; p = 21.21%

INFO : Sense: recent.s.01 ; p = 6.0%

INFO : Sense: produce.v.03 ; p = 3.15%

INFO : Sense: produce.v.02 ; p = 1.75%

**Observations**: increasing k moves the “burden” of prediction from the globals, that have an easier task (as long as they manage to include the correct global among the k most likely ones) to the senses (since we have to discriminate among more senses).

It is probably worth to explore k=1(as a baseline – the senses of the most likely predicted global), k=3, k=5.

Moreover, alternative architectures for the senses’logits may help. We noted previously that the senses’part “activates” only when there is a sense label, and thus does not have a continuous view of the text. Maybe instead of 3layers with hd=575 we should have 1layer with 1150?

## Experiment 1 – SelectK1 on SemCor

INFO : Model:

INFO : DataParallel(

(module): SelectK(

(main\_gru): GRU(300, 1150, num\_layers=3)

(gru\_senses): GRU(300, 575, num\_layers=3)

(linear2global): Linear(in\_features=1150, out\_features=21988, bias=True)

(linear2senses): Linear(in\_features=575, out\_features=25986, bias=True)

)

)

INFO : Hyperparameters: \_batchPerSeqlen1400\_area32\_lr5e-05\_epochs100

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| Globals’ prediction: **Main GRU** with 3 layers (d=1150) | 1) The word embedding of the current global (d=300) | batch\_size=40  TBPTT length=35 |
| linear2Globals **FF-NN** | grapharea=32, hops=2 |
| **Senses GRU** (L=3, d=575) into linear2Senses **FF-NN**, apply softmax on the logits of the senses of the most likely ***k=1*** global | learning rate=5\*10^(-5) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* | *Validation PPL on globals* | *Validation PPL on senses* |
| 1 | 918.46 | 28057.44 | 434.44 | 25986.01 |
| 2 | 490.94 | 26994.04 | 314.98 | **25549.72** |
| 3 | 367.5 | 31471.46 | 268.2 | 47641.5 |
| 4 | 309.25 | 44004.55 | 243.77 | 66291.41 |
| 5 | 272.81 | 46505.83 | 228.62 | 74292.94 |
| 6 | 246.53 | 49694.05 | 219.81 | 78444.04 |
| 7 | 225.84 | 52122.73 | 212.56 | 81960.8 |
| 8 | 208.45 | 54768.2 | 206.16 | 84532.23 |
| 9 | 193.63 | 57472.16 | 203.44 | 86204.86 |
| 10 | 180.89 | 59758.27 | 199.58 | 88047.98 |
| 11 | 169.68 | 61785.0 | 201.01 | 87741.85 |
| 12 | 159.69 | 63496.94 | 197.92 | 90574.67 |
| 13 | 150.52 | 65106.49 | 193.72 | 93140.19 |
| 14 | 142.27 | 66584.56 | 193.75 | 94614.72 |
| 15 | 134.68 | 68381.98 | 194.04 | 99078.3 |
| 16 | 127.65 | 70024.39 | 192.0 | 102214.42 |
| 17 | 121.23 | 71601.27 | 193.27 | 105976.18 |
| 18 | 115.08 | 72829.32 | 192.83 | 112505.01 |
| 19 | 109.42 | 74349.62 | 192.25 | 119116.06 |
| 20 | 104.15 | 75866.42 | 192.37 | 127979.96 |
| 21 | 99.13 | 76963.22 | 192.07 | 131361.44 |
| 22 | 94.44 | 77852.04 | 191.46 | 139729.8 |
| 23 | 90.09 | 78860.22 | **190.51** | 141888.52 |
| 24 | 85.91 | 79598.94 | 192.02 | 150453.28 |
| 25 | 81.98 | 80649.1 | 192.48 | 156545.26 |

Validation perplexity on globals = 190.5

Validation perplexity on senses = 25549.7

Which is extremely high.

**Hypothesis:**

We only try to predict the senses from the most likely predicted global, k=1.

If the globals’ prediction is unable to choose the correct global, we are only adjusting the weights (and softmax probabilities) to choose among a set of senses that does not overlap with the correct solution.

**Follow-up:**

Set k=5 and verify what happens.

# Multi-sense LM. Part 2

## Checking & Debugging

Before doing anything else, trying architectures, experiments etc. it is necessary to verify that the node indices and retrieval are correct.

Let us examine the variables when training on the fragment of SemCor (Mini-dataset 2: 63 tokens)

Setting batch\_size=1 and seq\_len=1 for simplicity.

At step 1,

(input) = (global\_index, sense\_index)=(0, -1), corresponding to : (The, no\_sense)

(labels) = (global\_index, sense\_index)=(21987, 10898) (<unk>, group.n.01)

Regarding the batch input tensor and how it’s unpacked:

For example, operating with batch\_size=1 and seq\_len=2, we have

batch\_input.shape=torch.Size([1, 2, 1150])

Only 1 element in the batch size, so we have

sequences\_in\_the\_batch\_ls = [batchinput\_tensor]

Then,

padded\_sequence = padded\_sequence.squeeze() # so we obtain [2,1150]. We should probably only squeeze 1 dimension here.

Back on track:

x\_indices\_g = tensor([25986]) (corresponding to global n.0 + number of senses)

batch\_input\_signals.shape= torch.Size([1, 1, 300]) # no other input signals here

Then,

senses\_gru\_out, hidden\_s =self.gru\_senses(batch\_input\_signals, self.memory\_hn\_senses)

followed by logits\_sense = self.linear2senses(senses\_gru\_out), etc.

The most likely global is: n.19876, “collects”

It is a consequence of the random initialization of the GRU.

lemmatized indices=[33154], corresponding to the 7168th global ‘collect’

sense\_neighbours\_t = tensor([ 4593, 4590, 4591, 4594, 4595, 10373, 4592]), ‘collect.n.01’, … , ‘collect.n.05’, ‘gather.v.01’

It seems that now the retrieval of sense nodes is correct.

The next steps & experiments will be:

* Baseline GRU for the senses, no methods
* SelectK with k=5 (k=1 was executed after the modifications to the retrieval mechanism that gave us the current, correct version of the models)
* SelectK with k=10, or alternatively
* The Self-attention mechanism using the sense embeddings

## Baseline 0: 2 GRUs

### Model

1 GRU for the globals, 1 for the Senses. No specific mechanism or refinement for the Senses, and no connection between the 2 predictions (apart from the moveable embeddings for globals, senses, defs and examples).

INFO : Model:

INFO : DataParallel(

(module): RNN(

(dropout): Dropout(p=0, inplace=False)

(main\_rnn\_ls): ModuleList(

(0): GRU(300, 1024)

(1): GRU(1024, 1024)

(2): GRU(1024, 512)

)

(rnn\_senses): GRU(300, 1024, num\_layers=2)

(linear2global): Linear(in\_features=512, out\_features=21988, bias=True)

(linear2senses): Linear(in\_features=1024, out\_features=25986, bias=True)))

INFO : Parameters:

INFO : ('module.X', torch.Size([99963, 300]), torch.float32, True)

('module.select\_first\_indices', torch.Size([1024]), torch.float32, False)

('module.embedding\_zeros', torch.Size([1, 300]), torch.float32, False)

('module.memory\_hn', torch.Size([3, 4, 1024]), torch.float32, False)

('module.memory\_cn', torch.Size([3, 4, 1024]), torch.float32, False)

('module.memory\_hn\_senses', torch.Size([2, 4, 1024]), torch.float32, False)

('module.memory\_cn\_senses', torch.Size([2, 4, 1024]), torch.float32, False)

('module.main\_rnn\_ls.0.weight\_ih\_l0', torch.Size([3072, 300]), torch.float32, True)

('module.main\_rnn\_ls.0.weight\_hh\_l0', torch.Size([3072, 1024]), torch.float32, True)

('module.main\_rnn\_ls.0.bias\_ih\_l0', torch.Size([3072]), torch.float32, True)

('module.main\_rnn\_ls.0.bias\_hh\_l0', torch.Size([3072]), torch.float32, True)

('module.main\_rnn\_ls.1.weight\_ih\_l0', torch.Size([3072, 1024]), torch.float32, True)

('module.main\_rnn\_ls.1.weight\_hh\_l0', torch.Size([3072, 1024]), torch.float32, True)

('module.main\_rnn\_ls.1.bias\_ih\_l0', torch.Size([3072]), torch.float32, True)

('module.main\_rnn\_ls.1.bias\_hh\_l0', torch.Size([3072]), torch.float32, True)

('module.main\_rnn\_ls.2.weight\_ih\_l0', torch.Size([1536, 1024]), torch.float32, True)

('module.main\_rnn\_ls.2.weight\_hh\_l0', torch.Size([1536, 512]), torch.float32, True)

('module.main\_rnn\_ls.2.bias\_ih\_l0', torch.Size([1536]), torch.float32, True)

('module.main\_rnn\_ls.2.bias\_hh\_l0', torch.Size([1536]), torch.float32, True)

('module.rnn\_senses.weight\_ih\_l0', torch.Size([3072, 300]), torch.float32, True)

('module.rnn\_senses.weight\_hh\_l0', torch.Size([3072, 1024]), torch.float32, True)

('module.rnn\_senses.bias\_ih\_l0', torch.Size([3072]), torch.float32, True)

('module.rnn\_senses.bias\_hh\_l0', torch.Size([3072]), torch.float32, True)

('module.rnn\_senses.weight\_ih\_l1', torch.Size([3072, 1024]), torch.float32, True)

('module.rnn\_senses.weight\_hh\_l1', torch.Size([3072, 1024]), torch.float32, True)

('module.rnn\_senses.bias\_ih\_l1', torch.Size([3072]), torch.float32, True)

('module.rnn\_senses.bias\_hh\_l1', torch.Size([3072]), torch.float32, True)

('module.linear2global.weight', torch.Size([21988, 512]), torch.float32, True)

('module.linear2global.bias', torch.Size([21988]), torch.float32, True)

('module.linear2senses.weight', torch.Size([25986, 1024]), torch.float32, True)

('module.linear2senses.bias', torch.Size([25986]), torch.float32, True)

INFO : Number of trainable parameters=91.008M

### Mini-experiment: overfit on fragment of SemCor

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| Globals’ prediction: **Main GRU** with 3 layers (1024 > 1024 > 512) | 1) The word embedding of the current global (d=300) | batch\_size=4  TBPTT length=8 |
| linear2Globals **FF-NN** | grapharea=32, hops=1 |
| **Senses GRU** with 2 layers (1024 > 1024) | learning rate=10^(-3) |

INFO : Hyperparameters: \_batchPerSeqlen32\_area32\_lr0.0005\_epochs300

|  |  |  |
| --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |
| 1 | 19936.75 | 22948.16 |
| 2 | 2651.57 | 3397.39 |
| 3 | 288.8 | 20.79 |
| 10 | 25.17 | 11.21 |
| 50 | 23.61 | 1.54 |
| 100 | 20.34 | 1.05 |
| 150 | 17.46 | 1.02 |
| 200 | 15.21 | 1.01 |
| 250 | 14.22 | 1.01 |
| 300 | 13.64 | 1.01 |

This GRU architecture for the globals appears to be slower to learn to overfit than other alternatives.

Regardless, I will try mini-experiment hyperparameters again, but in the meantime I will start the standard experiment on the full SemCor

### Experiment: GRUs on SemCor

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| Globals’ prediction: **Main GRU** with 3 layers (1024 > 1024 > 512) | 1) The word embedding of the current global (d=300) | batch\_size=40  TBPTT length=35 |
| linear2Globals **FF-NN** | grapharea=32, hops=1 |
| **Senses GRU** with 2 layers (1024 > 1024) | learning rate=10^(-4) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* | *Validation perplexity on globals* | *Validation Perplexity on senses* |
| 1 | 960.38 | 4852.58 | 487.37 | **793.37** |
| 2 | 608.2 | 3104.06 | 372.77 | 805.0 |
| 3 | 437.28 | 2688.92 | 297.47 | 862.89 |
| 4 | 338.82 | 2532.54 | 250.72 | 1183.94 |
| 5 | 281.69 | 2409.66 | 226.53 | 1274.36 |
| 6 | 247.24 | 2358.39 | 214.41 | 891.52 |
| 7 | 222.84 | 2341.0 | 205.94 | 1434.78 |
| 8 | 203.59 | 2177.51 | 199.12 | 1464.13 |
| 9 | 187.78 | 2071.02 | 195.98 | 1649.94 |
| 10 | 174.29 | 1980.39 | 192.04 | 1799.48 |
| 11 | 162.52 | 1980.49 | 193.21 | 2270.78 |
| 12 | 152.06 | 1788.65 | 189.99 | 2408.59 |
| 13 | 142.42 | 1743.1 | 185.98 | 1681.45 |
| 14 | 133.8 | 1565.92 | **185.95** | 2171.4 |
| 15 | 125.8 | 1476.37 | 186.36 | 2240.74 |
| 16 | 118.53 | 1352.14 | 184.78 | 3159.46 |
| 17 | 111.91 | 1264.11 | 186.69 | 2708.47 |
| 18 | 105.64 | 1164.57 | 187.41 | 2474.85 |
| 19 | 100.05 | 1063.79 | 187.73 | 3647.71 |
| 20 | 95.04 | 983.45 | 189.34 | 4330.14 |
| … |  |  |  |  |
| 40 | 34.88 | 138.33 | 256.96 | 5288.73 |
| … |  |  |  |  |
| 60 | 12.98 | 19.67 | 463.47 | 9549.37 |

The GRUs eventually overfit on SemCor.

Best values for the Valid-PPL: on globals=185.9. on senses=793.4, reached immediately in Epoch 1.

### Experiment: GRUs on SemCor – version 2

This time we use the architecture 1024>1024>**1024**, that in the SelectK5 mini-experiment actually manages to overfit on a fragment for the globals.

Lower learning rate.

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| Globals’ prediction: **Main GRU** with 3 layers (1024 > 1024 > **1024**) | 1) The word embedding of the current global (d=300) | batch\_size=40  TBPTT length=35 |
| linear2Globals **FF-NN** | grapharea=32, hops=1 |
| **Senses GRU** with 2 layers (1024 > 1024) | learning rate=5\*10^(-5) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* | *Validation perplexity on globals* | *Validation Perplexity on senses* |
| 1 | 972.76 | 5214.45 | 466.09 | 842.91 |
| 2 | 527.05 | 3437.78 | 332.01 | 811.1 |
| 3 | 389.62 | 3110.54 | 279.08 | 817.96 |
| 4 | 324.65 | 2892.64 | 250.44 | 842.01 |
| 5 | 284.52 | 2780.03 | 233.17 | 856.07 |
| 6 | 256.73 | 2705.51 | 223.32 | 886.45 |
| 7 | 235.48 | 2650.44 | 215.0 | 1649.36 |
| 8 | 217.94 | 2793.9 | 207.86 | 978.71 |
| 9 | 203.1 | 2633.58 | 204.51 | 2335.16 |
| 10 | 190.3 | 2464.52 | 200.2 | 1864.83 |
| 11 | 178.99 | 2525.9 | 201.09 | 1082.45 |
| 12 | 168.89 | 2447.28 | 197.38 | 1165.9 |
| 13 | 159.62 | 2387.41 | 192.88 | 1689.99 |
| 14 | 151.26 | 2240.95 | 192.46 | 1418.81 |
| 15 | 143.58 | 2188.04 | 192.3 | 1972.79 |
| 16 | 136.46 | 2081.01 | 190.04 | 1957.07 |
| 17 | 129.97 | 2102.42 | 191.05 | 1755.35 |
| 18 | 123.73 | 1965.57 | 190.37 | 2243.94 |
| 19 | 118.02 | 1942.21 | 189.86 | 1798.67 |
| 20 | 112.7 | 1886.14 | 190.0 | 2460.83 |
| 21 | 107.6 | 1840.18 | 189.92 | 2157.65 |
| 22 | 102.81 | 1780.39 | 189.58 | 2446.24 |
| 23 | 98.37 | 1726.86 | 188.67 | 1950.77 |
| 24 | 94.12 | 1777.95 | 190.4 | 2151.06 |
| 25 | 90.06 | 1648.05 | 190.87 | 2376.13 |
| … |  |  |  |  |
| 40 | 48.05 | 876.12 | 212.67 | 3407.63 |
| … |  |  |  |  |
| 60 | 24.65 | 447.34 | 251.02 | 4499.36 |

## SelectK5

### Model

Idea: make the senses’ task dependent on the globals’ standard language model task, choosing among the senses of the most likely *k* globals.

We consider the globals’ logits (main GRU > FF-NN, standard language model), and the senses’ logits (senses’ GRU > FF-NN):

logits\_global = self.linear2global(main\_gru\_out)

logits\_sense = self.linear2senses(senses\_gru\_out)

Sort the logits of the globals, to see which globals are most likely in our LM prediction:

k\_globals\_indices = logits\_global.sort(descending=**True**).indices[:, 0:self.k]

sample\_k\_indices\_lls\_relative = k\_globals\_indices.tolist()

(’relative’ means ’in-vocabulary’, without considering X and the last sense idx)

For every sample *s*: (we are handling batch\_size x seq\_len elements)

k\_globals\_relative\_indices = sample\_k\_indices\_lls\_relative[s]

k\_globals\_words > k\_globals\_lemmatized

Take the most likely *k* globals for the current sample. We lemmatize them, so we turn ‘said’ into ‘say’, ‘collects’ into ‘collect’, ‘are’ into ‘be’, etc.

lemmatized\_indices

sense\_neighbours\_t = get\_senseneighbours\_of\_k\_globals(self, lemmatized\_indices)

Using the grapharea\_matrix for speed, we retrieve the neighbours of every most-likely global, and collect them in a tensor, keeping only the nodes with the index in the range of senses.

Next: from the sensesGRU>linear2Senses, we select the logits of the senses of the most likely k globals:

logits\_selected\_senses = sample\_logits\_senses.index\_select(dim=0, index=sense\_neighbours\_t)

and apply a softmax only over those:

softmax\_selected\_senses = tfunc.softmax(input=logits\_selected\_senses, dim=0)

Then, having defined a probability distribution over the whole vocabulary of senses with a fixed ε value for every sense (epsilon = 10 \*\* (-8)), we copy-paste the values from the softmax over the selected senses.

senses\_softmax[s].masked\_scatter\_(mask=i\_senseneighbours\_mask[s].data.clone(), source=softmax\_selected\_senses)

As a side effect, this means that the sum of the softmax over the whole vocabulary will be slightly more than 1, sum=1+(10^(-8)\*num\_senses) = 1+ 0.00256986.

Currently, the sum>1 does not seem to produce a significant error. It could be possible to subtract this probability mass from the most likely entry (or entries) in the selected senses. In this case, one needs to be careful not to push very unlikely selected entry to have a probability < 0, because this breaks the nll\_loss numerically.

Eventually, INFO : Model:

INFO : DataParallel(

(module): SelectK(

(dropout): Dropout(p=0, inplace=False)

(main\_rnn\_ls): ModuleList(

(0): GRU(300, 1024)

(1): GRU(1024, 1024)

(2): GRU(1024, 1024))

(gru\_senses): GRU(300, 1024, num\_layers=2)

(linear2global): Linear(in\_features=1024, out\_features=21988, bias=True)

(linear2senses): Linear(in\_features=1024, out\_features=25986, bias=True))

)

### Mini-experiment: overfit on fragment of SemCor

**Observation**:

The 1024 > 1024 > 512 architecture for the globals DOES NOT manage to overfit on a fragment. Instead, an architecture 1024 > 1024 > 1024 does.

Thus, we change the main\_gru – and above we repeat the experiments regarding the GRU-baseline.

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| Globals’ prediction: **Main GRU** with 3 layers (d=1024) | 1) The word embedding of the current global (d=300) | batch\_size=4  TBPTT length=8 |
| linear2Globals FF-NN | grapharea=32, hops=1 |
| **Senses GRU** (**L=2, d=1024**) into linear2Senses FF-NN, apply softmax on the logits of the senses of the most likely ***k=5*** globals | learning rate=10^(-3) |

INFO : Hyperparameters: \_batchPerSeqlen32\_area32\_lr0.001\_epochs400

|  |  |  |
| --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* |
| 1 | 19691.81 | 18904264.16 |
| 2 | 734.52 | 28635572.41 |
| 3 | 45.67 | 10122294.55 |
| 10 | 24.34 | 3791925.79 |
| 50 | 22.56 | 1295454.06 |
| 100 | 12.69 | 50190.08 |
| 150 | 9.52 | 83.24 |
| 200 | 6.2 | 7.28 |
| 250 | 4.06 | 4.85 |
| 300 | 2.83 | 8.53 |
| 350 | 1.78 | 1.13 |

### Experiment: SelectK5 on SemCor

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Input Signals** | **Hyperparameters** |
| Globals’ prediction: **Main GRU** with 3 layers (d=1024) | 1) The word embedding of the current global (d=300) | batch\_size=40  TBPTT length=35 |
| linear2Globals FF-NN | grapharea=32, hops=1 |
| **Senses GRU** (**L=2, d=**1024) into linear2Senses FF-NN, apply softmax on the logits of the senses of the most likely ***k=5*** globals | learning rate=5\*10^(-5) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Epoch* | *Training perplexity on globals* | *Training perplexity on senses* | *Validation perplexity on globals* | *Validation Perplexity on senses* |
| 1 | 966.81 | 31392.45 | 459.5 | 25986.01 |
| 2 | 534.55 | 112360.31 | 340.65 | 405029.77 |
| 3 | 406.77 | 2021427.64 | 288.92 | 1602356.61 |
| 4 | 335.86 | 5084236.43 | 255.78 | 1754059.22 |
| 5 | 291.11 | 4701992.56 | 236.08 | 1808811.46 |
| 6 | 260.5 | 4655555.52 | 225.17 | 1831024.41 |
| 7 | 237.87 | 4717672.25 | 216.66 | 1894365.78 |
| 8 | 219.72 | 4688206.98 | 209.8 | 2094054.77 |
| 9 | 204.64 | 4577951.25 | 206.46 | 2036662.93 |
| 10 | 191.8 | 4410303.11 | 202.24 | 2034214.03 |
| 11 | 180.6 | 4346401.35 | 203.27 | 1948670.31 |
| 12 | 170.73 | 4297053.32 | 199.64 | 1902468.18 |
| 13 | 161.65 | 4206292.45 | 194.88 | 1985912.05 |
| 14 | 153.5 | 4111808.46 | 194.5 | 1905634.53 |
| 15 | 146.01 | 4078699.91 | 194.63 | 1869208.05 |
| 16 | 139.06 | 4016920.82 | 192.49 | 1820520.83 |
| 17 | 132.74 | 3976029.51 | 193.47 | 1852512.54 |
| 18 |  |  |  |  |
| 19 |  |  |  |  |
| 20 |  |  |  |  |
| … |  |  |  |  |
| 40 |  |  |  |  |
| … |  |  |  |  |
| 60 |  |  |  |  |

Observations: the senses’ prediction is consistently incorrect, and given that we add an epsilon of 10^-8 to every sense that does not belong to the top k most likely globals, this means that the nll\_loss in those samples is ~10^8, and thus brings up the PPL on senses.

Idea: do not use the nll\_loss / PPL to evaluate the senses’ prediction when operating with structured prediction / SelectK.

Add the evaluation measures: “correct label” and “correct sense in the top-5 senses”.

After epoch 10, the Training and Valid PPL on senses decreases, so the model is learning something. However, we do not know how relevant this is until we consider the number of correct labels, given the problem with computing the Senses’ PPL in this architecture.