[Reflections and the next steps – 16/10 and onwards 2](#_Toc23928834)

[Development – 17/10 and onwards 5](#_Toc23928835)

[Step 0 : Vocabulary 6](#_Toc23928836)

[Step 1 : Retrieving and storing KB & Dictionary data for the Vocabulary 8](#_Toc23928837)

[How to deal with plural and inflected forms? 10](#_Toc23928838)

[Using the Vocabulary to retrieve KB data. 11](#_Toc23928839)

[Phrases 12](#_Toc23928840)

[Step 2: Corpus-based word embeddings 13](#_Toc23928841)

[Embeddings dimension 13](#_Toc23928842)

[From DistilBERT 14](#_Toc23928843)

[Step 2b: Definitions and Examples 300d (D)BERT-based embeddings 15](#_Toc23928844)

# Reflections and the next steps – 16/10 and onwards

Add-on: are there any related papers incoming from EMNLP 2019?

* “Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations” by Christian Hadiwinoto, Hwee Tou Ng and Wee Chung Gan
* “Challenging Supervised Word Sense Disambiguation with Lexical-Semantic Combinations” by Marco Maru, Federico Scozzafava, Federico Martelli and Roberto Navigli
* “Towards Zero-shot Language Modelling” by Edoardo Maria Ponti, Ivan Vulić, Ryan Cotterell, Roi Reichart and Anna Korhonen
* “MultiFiT: Efficient Multi-lingual Language Model Fine-tuning” y Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger and Jeremy Howard
* “GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge” by Luyao Huang, Chi Sun, Xipeng Qiu and Xuanjing Huang
* **Read** references on multi-sense embeddings & language models  
  90%+ done.   
  A quick overview:
  + J. Resinger & R.J. Mooney 2010: represent the words & contexts with TF-IDF / χ-square features, and then senses = centroids of the context clusters
  + Eric H. Huang et al. 2012: local+global(document) FF-NN architecture. Single-prototype phase: Quasi-LM, I see the whole document, but the objective is a score on the next word.  
    Then, cluster the contexts with spherical k-means.
  + X. Chen et al. 2014:   
    Single-prototype phase: Skip-gram  
    Senses from WordNet: definitions and examples form ‘glosses’; we take the average of the candidate vectors in glosses to get the sense embeddings.  
    WSD is done using the cosine-sim between context window and sense.
  + S. Rothe & H. Schutze 2015: AutoExtend  
    Embeddings for synsets and lemmas, using an AutoEncoder. Exploits the linearity assumption of a synset being constructed as the sum of lemmas.  
    Method: word embeddings (possibly pre-trained) -> encode -> synset embeddings -> decode -> re-create word embeddings
  + A. Neelakantan et al., 2015: MSSG, Multi-Sense Skip-Gram model  
    Single-prototype phase: initial global (Skip-Gram) word vectors.  
    2 variants. One with a fixed number of senses, the other that determines the senses based on the locations of the context embeddings (NP-MSSG)
  + S.K. Jauhar et al., 2015: Interacts with an ontology, uses context-independent sense proportions, considering the senses as latent variables
  + J. Li & D. Jurafsky 2015:   
    Uses CRP, Chinese Restaurant Process to create and populate new senses with contexts.  
    WSD can be either greedy or expectation.
  + H. Shi et al., 2016: Exploration of Pseudo Multi-Sense
  + D.Kartsaklis et al., 2018: Mapping text to KB entities  
    Only relevant here because it uses a (Q,K,V) attention mechanism on a fixed number of vectors (k=3) to choose the senses (with soft-attention) and update them.
  + M.T. Pilehvar et al., 2017: Given a text T, turn it into a graph (S,E) of all possible Senses + relation Edges.  
    At each iteration: select the sense s\_ij with the highest degree in the graph.
  + I. Iacobacci and R. Navigli, 2019: LSTMembed  
    1. Obtain a sense-labeled corpus, using the Babelfly instrument   
    2. BiLSTM over the training text > 1 layer > Training objective: minimize the cosine distance from a pre-trained embedding
  + A. Panigrahi et al., 2019: Word2Sense: interpretable embeddings  
    The dimensions have meaning:   
    A sense is a concept, represented by a set of similar words in the vocabulary. Like topics, senses are distributions over the vocabulary.   
    k senses = k dimensions; a word w will have a score on each sense.  
    (3000 -> JS-merge into ¾ -> truncate to 75 eventually)
  + S. Kumar et al., 2019: EWISE  
    Encode context with BiLSTM, then project…   
    Training objective: minimize the cosine distance from the pre-existing sense embeddings, that were created independently, either from pre-trained models, or from KB (e.g. WordNet) resources
* **Read** on evaluation measures for multi-sense embeddings & language models  
    
  In the works I consulted (most of which were mentioned in the previous section), the evaluation measures that were used most often and/or most promisingly were:
  + SCWS (Stanford's Contextual Word Similarities).  
    Better than basic word similarity tasks such as WS-353 or RG-65; it also allows us to apply the WSD mechanism
  + {Downstream tasks: PoS-tagging, SentEval}
  + Downstream task: STS, Semantic Textual Similarity (STS12, STS13, STS14, STS15, STS16, STS-B), and also the SICK relatedness (SICK-R) and SICK-E (where the classes are ‘entailment’, ‘contradiction’, and ‘neutral’)
  + Word Sense Disambiguation: in particular, SemEval 2013 and 2015
  + Finally, Downstream task: Language model  
    WikiText-2 as a small, pocket-sized development set.  
    WikiText-103 as a larger dataset. We should obtain comparable/acceptable perplexity.

Note: it is also opportune to study the impact of dimensionality   
(e.g. 300, 384, 768)

* **A baseline** Multi-Sense LM  
  What models are going to compare our results with?   
  With different models, depending on the selected task. For some of them, When operating on well-known datasets, we do not need to train & implement anything, we can just compare result scores.  
  It is also opportune to consider models that are not too unorthodox (e.g. Pseudo Multi-sense, or interpretable Word2Sense) and not too old (Huang et al. 2012, FF-NN + context-based) .  
  Hypothesis regarding the models:
  + X. Chen et al. 2014. It uses Skip-gram > candidate words in the glosses > vectors for senses.  
    It is an approach relatively close to mine, but without any GNN.
  + A. Neelakantan et al., 2015: MSSG, Multi-Sense Skip-Gram. Here the senses are obtained from the contexts. A generic Multi-sense instrument, examined often and good for comparisons.
  + Transformer-XL, 2019. It already has perplexity values on WT-2 and more importantly WT-103. I even introduced it as a baseline (and a future feature) for Write-Assistant.
  + Possibility: The (pre-trained, unmodified) BERT.  
      
    We mention again here the (candidate) evaluation tasks:
  + SCWS for word similarity.(low priority)
  + **STS (Semantic Textual Similarity on sentences)**
  + Language Modeling perplexity, on WT-2 and WT-103
  + WSD

Given the evaluation tasks, there are already results, there should be no need to train anything anew.

# Development – 17/10 and onwards

* **Develop** the Graph-Based Multi-Sense Word Embeddings:
  + Step 0: define a vocabulary
  + Observation: our dictionary-resources system can only deal directly with base, non-inflected forms of words (e.g. only ‘plant’, not ‘plants’ or ‘planted’)
    - Lemmatization for inflected forms when training, to include them in the processing
    - We can worry about this later, in the next phase. Several alternative solutions are possible:
      * Inflected forms can be adjusted using a standard corpus-based approach
      * We can add singular / plural / past edges to the graph
      * We may also concatenate dict-vecs for base forms, and context-based vecs for inflected forms
    - **Different inflected forms distinct for each sense are confirmed as valid. I can recognize “moved to Denmark “ vs. “moved a cup” by operating on a sense annotated corpus.**
  + Preliminary distributional embeddings from training corpus, in order to provide for words without a dictionary definition (for, of, etc.). This also initializes the 'move.global' vectors.
  + Placing in the multi-dimensional space the definitions, and the examples. Sentence embedding based on BERT (average over the word tokens of the 2nd-to-last layer)
  + What about the embedding dimension? I can try:  
    - the original bert-base-uncased size (768)
  + **\* there must be a Bert-small with ~300 dim., and also UlmFit**   
    - PPA+PCA+PPA from V.Raunak et al. 2019 (768 / 2 = 384, or even to 300)  
    - PCA to 300
  + Consider the senses (move.v.1, ..., move.n.3) from the KB sources. How to initialize them? Alternatives:   
    1) on move.global, plus a random perturbation  
    2) Average of all definitions and examples for that particular sense
  + Build Graph Neural Network (Graph Attention Net, or DCGCN)
  + *Language Modeling objective on the training set  
    What if I predict the right word, but the wrong sense?.  
    Look into sense-annotated corpuses / Use example sentences (a held-out portion) / Look into sense-annotation instruments that are not this one itself*
  + Evaluate (see previous discussion)
  + The evaluation should be comparable with:  
    - SoA models on downstream tasks (eg. LSTMs for LMs on WT-2, we should not be too far off)  
    - contextual Transformer embeddings.  
    - And also old Multi-sense embeddings (MSSG, C.R.P.)

## Step 0 : Vocabulary

Our vocabulary must be useful for Language Modeling.

* Word-level or subword level? Word level.  
  BERT uses internally subword level, but we employ it only to produce sentence embeddings for defs & examples.
* Keep punctuation or remove it? Keep it. Later on, possibly, a version for Write Assistant may remove it.  
  Punctuation should be processed as separate tokens.
* Considering the end task, obviously: keep stopwords.
* Lower-case or not?  
  An interesting possibility is *truecasing*: lower-case only the words at the beginning of a sentence, thus leaving acronyms (USA, CAT) and proper names (Marco Polo) with the correct starting letter.
* Numbers: keep all vs. replace all with <num> vs. keep only 1-digit and 4-digits   
  The last alternative appears to me as the most appropriate
* Non-latin words: if a word contains a character that is non-latin and not even in the Greek alphabet, such as “マギカ”, we replace the word with the <unk> token.

Example line from WikiText-2 (training set):

Father <unk> joined the observatory in 1889 after it was upgraded with more modern instruments . The same year , a second tower was built some 400 metres ( **1 @,@ 300** ft ) away from the main Gregorian Tower , overlooking the Vatican Gardens behind St. Peter 's Basilica on the **south @-@ west** border . It was built to a diameter of 17 metres ( 56 ft ) with a lower wall thickness of **4 @.@ 5 metres** ( 15 ft ) , which could bear the load of a 13 inch photographic <unk> , newly procured from Paris . ………’

Note: It is also necessary to reconvert some of the symbols found in the WikiText datasets, in particular:

‘@-@’ 4 ‘@.@’ 5 metres

Ampersand symbols taken away. Sentence-level tokenization:

'Father <unk> joined the observatory in 1889 after it was upgraded with more modern instruments .',

"The same year , a second tower was built some 400 metres ( **1,300** ft ) away from the main Gregorian Tower , overlooking the Vatican Gardens behind St. Peter 's Basilica on the **south-west** border .",

'It was built to a diameter of 17 metres ( 56 ft ) with a lower wall thickness of **4.5 metres** ( 15 ft ) , which could bear the load of a 13 inch photographic <unk> , newly procured from Paris .'

Without starting and ending spaces (the effect is seen elsewhere, 2 sentences before here), and lowercasing the first word in a sentence …

Tokens:

'father', '<', 'unk', '>', 'joined', 'the', 'observatory', 'in', '1889', 'after', …

Problem: I am decomposing the Unknown token, ‘<unk>’. …

Recap of the steps to process the text of a line in order to build a vocabulary that will be appropriate for a Language Model downstream task:

1. It is necessary to reconvert some of the symbols found in the WikiText datasets, in particular: ‘@-@’, ‘@.@’ etc.
2. nltk's tokenizer - sentence tokenizer
3. Eliminate outer spaces, and true-case the sentences.
4. Split into word tokens, and reunite the lls
5. Recreate <unk> from '<','unk','>'
6. There are word-tokens with non-Latin and non-Greek characters, such as: 匹, 枚, マギカ , etc. We replace those words with <unk>
7. Numbers. All decimals, together with all numbers that are not 1 digit (basic) or 4 digits (years), get replaced with <num>

Later on, these steps could be used to re-write the corpus and save it to disk for other uses.

The vocabulary should be created at the start of everything. Before retrieving the data from the KB sources.

With min\_count = 1, I get a vocabulary of 32,156 from WikiText-2.

Dataset statistics: |V| = 33,278 (close enough)

Emergent problem:

The titles and subsection-titles of the Wikipedia articles are presented in-between ‘=’ signs.

= Valkyria Chronicles III =

…

= = Gameplay = =

…

We thus add a regex check in the vocabulary processing. to remove the ‘ = ‘ segments from the vocabulary creation process.

Executing our pre-processing, we now get:

Vocabulary created, after processing 2,089,445 tokens [statistics= 2,088,628]

32013 rows x 2 columns – (33,278)

{What about normalizing acronyms? Not necessary, especially if we wish to be able to face datasets that have not been pre-processed in the same way}

I decide to set min\_count=2 on the training set of WikiText-2,

finally obtaining 31628 words.

## Step 1 : Retrieving and storing KB & Dictionary data for the Vocabulary

Reviewing the mini-vocabulary data that we extracted from the KBs with the current code.

Observation: Denominated antonyms have repetitions.

Such as:

move verb.7 refrain

move verb.7 refrain

move verb.7 refrain

move verb.3 stay

light adj.10 heavy

light adj.10 heavy

etc…

* Double-check their creation

Denominated definitions and Denominated examples are both ok.

Denominated synonyms contain duplicates as well.

0 wide adj.4 wide-eyed

0 wide adj.4 wide-eyed

1 wide adv.1 widely

0 wide adj.4 wide-eyed

1 wide adv.1 widely

2 wide adv.2 astray

0 wide adj.4 wide-eyed

1 wide adv.1 widely

2 wide adv.2 astray

…

The Processed Synonyms and Processed Antonyms, that contain the bn\_id as they are before the sense denomination step, already contain the duplicates.

Synonyms and Antonyms (i.e. just after retrievel, before processing) contain no duplicates.

* Reason found: when extending the dataframe, I add it repeatedly, including again the previous elements.  
  It can be seen in the indices, for instance: 0,1,01,2,0,1,2,3,4,0,1,2,3,4,5.

Status: Adjusted.

We get back to reviewing the Input Data that we collected and prepared.

* Definitions:
  + denominated\_definitions.h5, table with fields: word, sense, definitions.
    - 2 light adj.2 having colors relatıvely near white.  
      3 light adj.2 (used of color) having a relatively small amount of coloring agent
    - 6 light adj.1 low in degree or quantity or number (e.g. of rain, snow, accent).
    - 7 light adj.1 not great in degree or quantity or number
    - 9 light noun.1 a small, reusable, handheld device for creating fire.
  + vocabulary\_table.sql, also used for the examples.  
    The database contains:   
    word, sense, vocab\_index, start\_defs, end\_defs, start\_examples, end\_examples  
    The last 4 fields are used to select the appropriate rows in the matrix of embeddings for definitions / examples.
  + vectorized\_definitions.npy
* Examples:
  + denominated\_examples.h5
    - 2 light adj.4 a light load
    - 3 light adj.4 my bag was much lighter once i had dropped off the books.
    - 20 light adj.1 a light sentence
    - …
    - 25 light adj.1 light smoke from the chimney
    - 47 light verb.1 we lit the fire to get some heat.
  + vocabulary\_table.sql, as stated
  + vectorized\_examples.npy
* Synonyms:
  + denominated\_synonyms:
    - move noun.1 motileness
    - move noun.1 freeswimmer
    - move noun.1 motility
    - move noun.2 movement
    - move noun.2 motion
    - light verb.1 kindle
    - light verb.1 ignite
    - light verb.2 illumine
    - light verb.2 illuminate
    - light verb.2 illume
    - light noun.2 sunlight
* Antonyms:
  + denominated\_antonyms.h5
    - wide adj.1 narrow
    - plant verb.1 abolish
    - move verb.2 stay

### How to deal with plural and inflected forms?

Example: what about ‘moves’ (that can be either the verb or the plural), and ‘moved’?

‘moved’ on WordNet :

Redirects to the verb,

1. (v) travel, go, move, locomote

2. (v) move, displace

…

and to one adjective sense:

(adj) moved, affected, stirred, touched

‘moves’ brings to the main page for ‘move’, with 5 noun senses and 16 verb senses.

‘moved’ in BabelNet: using BN.retrieve\_DESA(‘moved')

Out[7]:

({'bn:00096594a': ['Being excited or provoked to the expression of an emotion']},

{'bn:00096594a': ['Too moved to speak',

"Very touched by the stranger's kindness"]},

{'bn:00096594a': ['moved', 'affected', 'stirred', 'touched']},

{'bn:00096594a': ['unmoved']})

‘moves’ gives nothing,   
BN.retrieve\_DESA('moves')

Out[8]: ({}, {}, {}, {})

#### Hypotheses

1. Different inflected forms distinct for each sense are confirmed as valid. I can recognize “moved to Denmark “ vs. “moved a cup” by operating on a sense annotated corpus.
2. The Inflected form should be different for each sense, but the transformation verb-to-past or noun-to-plural should not change. The ideal would be to learn a transformation… although it would probably be simpler (and also, not requiring a sense-annotated corpus) to register the translation between ‘move.global’ and ‘moves.global’, and apply this translation starting from each sense.

We can introduce new edge types in the GNN: Plural (for Noun senses), VerbDeclination (for Verb senses).

### Using the Vocabulary to retrieve KB data.

We extracted a vocabulary of 31628 words from WikiText-2, setting min\_count=2.

Problem: BabelNet currently allows me to send 5000 queries a day.

It will be necessary to:

* insert an Append setting to all operations in PrepareInput.prepare(), or even make it the default setting and add a reset() command.
* Correction: the Append mode has to start all the way back in GetInputData.

Then:

Track the number of BabelNet requests in GetInputData, or just pick a random number like 3000?

Tracking is better. A global variable in Utils can do it.

If I use this choice… then I must save in a txt file the vocabulary index I processed last.

And then, start the processing from the next one.

This is now implemented. Current threshold before stopping and saving the index : 4950.

“Retrieving Multisense data for word: III”…

Some words have no definitions/data in any of the KB sources.

We must recognize this and move on, before we attempt to access a dictionary field that does not exist.

Retrieving Multisense data for word: 2011

BN\_request\_sender.requests\_counter= 1

({}, {}, {}, {})

({}, {})

{}

// When all dictionaries are empty, we should move on, without attempting to postprocess and compute BERT embeddings for the definitions. In fact, it would be opportune to exclude such terms from the vocabulary\_chunk sent to postprocessing.

Retrieving Multisense data for word: for

BN\_request\_sender.requests\_counter= 1

({}, {}, {}, {})

({}, {})

{}

Retrieving Multisense data for word: Raven

BN\_request\_sender.requests\_counter= 1

({'bn:00086765v': ['Eat greedily'], 'bn:00091936v': ['Prey on or hunt for', 'To prey on or hunt for.'], 'bn:00022930n': ['Large black bird with a straight bill and long wedge-shaped tail', 'A raven is one of several larger-bodied species of the genus Corvus.', 'The common raven, also known as the northern raven, is a large all-black passerine bird.', 'A bird that is usually black, very rarely white', 'A large black bird, similar to the crow, but larger.'], 'bn:00092393v': ['Feed greedily'], 'bn:00092392v': ['Obtain or seize by violence']}, {'bn:00086765v': ['He devoured three sandwiches'…},

{'bn:00086765v': ['devour', 'guttle', 'raven', 'pig', 'devour', 'guttle'], 'bn:00091936v': ['raven', 'prey', 'predate', 'predate', 'prey', 'raven', 'predate'], 'bn:00022930n': ['raven', 'corvus\_corax', 'common\_raven', 'raven', 'common\_raven', 'corvus\_corax', 'raven', 'raven', 'common\_raven', 'northern\_raven', 'raven', 'common\_raven', 'corbie', 'northern\_raven', 'raven', 'ravens', 'canary\_islands\_raven', 'common\_raven', 'corbies', 'corvus\_corax', 'corvus\_corax\_varius', 'corvus\_tingitanus', 'northern\_raven', 'northern\_raven', 'raen', 'western\_raven', 'western\_raven', 'corbie', 'corbie', 'raven', 'ravin', 'raven'…)

{}

Observation: Raven appears to be in Camelcase, despite our truecasing operation on the vocabulary.

Where does it appear in the training set?

Calamity Raven, in the game Valkyria Chronichles III; also, Big Raven.

This highlights another issue:

### Phrases

As of now, we have no phrase aggregation at all in the vocabulary over the training set.

“New York” would be processed as “New” and “York”, just like “Calamity” “Raven”.

The gensim instrument for phrases may come in handy.

Note: In BabelNet, ‘New\_York’ redirects to New York and is still ok.

In WordNet, ‘New York’ is ok, ‘New\_York’ has no correspondences.

What about the parameter values for gensim phrases?

Default: min\_count=5, threshold=10.0

We consider that even in the training corpus of WikiText-2 we get:

United States – 783 results

New York – 509 results

The parameters can be proportional to the training corpus’ size. Hypothesis:

min\_freq = cubic root of total\_num\_tokens // 2 (e.g. cbrt(2.2\*106) / 2 = 65)

phrases\_score\_threshold = min\_freq \* 2 (130)

It is probably better to just train them based on the phrases we get.

For WikiText-2, min\_freq = 40 and phrases\_score\_threshold = 120 is appropriate.

At this point, there are 2 possibilities:

1. Fast & scalable, less accurate for the subsequent Language Model training & task: just add the phrases to our vocabulary.
2. More time-consuming, more precise: re-process the corpus, joining into a\_b all the bigram occurrences.

Choosing: (2)

Add-on:

And what about tri-gram expressions? Such as ‘Salt Lake City’ or ‘World War II’?

Make 2 passes for Phrases: one with min\_freq = 40 and phrases\_score\_threshold = 120, the other one with min\_freq = 40 and phrases\_score\_threshold = 150

Note: It would appear that the benefit from a 2nd pass with the given parameters is very limited (e.g. +30 phrases). It can be removed.

## Step 2: Corpus-based word embeddings

This is a preparatory phase: given the training corpus, I need to obtain old-style 1-to-1 word embeddings.

This is necessary for 2 reasons:

* The words like “for”, “and”, “of”, and so on that do not have any definition whether in BabelNet or in WordNet.
* To initialize the ‘global’ elements in our graph. They have 3 roles:
  + They are connected to the senses with the ‘*sc’ (senseChildren*) edges
  + They are connected to other ‘w.global’ nodes through the *‘syn’* and *‘ant’ (synonyms* and *antonyms)* edge types
  + They can be used as an initialization anchor for the senses (with a random perturbation)

There are several alternatives for it:

* BERT with the words in absolute isolation. Possibly, small-BERT / distilBert
* UlmFit.
* Skip-Gram.
* Fast-Text (should be slightly better than Skip-Gram, by empirical evidence)

And the dimensionality should be the same one used for definitions and examples, since they end up in the same embedding space.

For the corpus-based word embeddings, as a first choice I can use:

Fast-Text (wiki-news-300d-1M.vec.zip): 1 million word vectors trained on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset.

### Embeddings dimension

In the spirit of building a low-resource model, and also to have a direct comparison with some past Multi-Sense methods (e.g. MSSG by Neelakantan et al. 2015), it would be opportune to set the embeddings dimensionality to d=300 (at least as an alternative).

d=300 allows us to operate directly with the FastText pre-trained single-prototype word embeddings.

However, this poses a problem later on, when dealing with definitions and examples embeddings (as specified in the module EmbedWithBERT.py). The embedding dimension of bert-base-uncased, etc., and even of DistilBERT that uses one layer every two of the original model, is 768.

Idea for dimensionality reduction:

Add to BERT a last-layer that leads from 768 dimensions to 300, and train it on a Language Model task on the chosen training set. Then, to get the sentence embedding use the same approach considered previously: average each token, producing one 300-length vector.

I can either:

* add a last linear layer, and fine-tune the whole BERT or DistilBERT architecture on a Language Model
* or: get the pretrained 768-embedding of the definition/example, and just train the last additional layer (less accurate, but simpler and not requiring more resources than it is feasible)

{Emerging issue: in assign\_senses\_to\_word(word, input\_dbs, output\_dbs), there is a KeyError: 'No object named antonyms in the file' if the word had no antonyms…}

{Updated Tensorflow in Conda-MacOs to 2.0.0 to use SummaryWriter}

### From DistilBERT

Currently: implementing the Finetuning of DistilBERT on a standard Language Model task (on WikiText-2)

{RuntimeError: index out of range: Tried to access index 512 out of table with 511 rows

while training, at outputs = model(inputs, labels)

>

In TextDataset block\_size must be 510 not 512, because this is \*before\* adding [CLS] and [SEP], we would end up out of bounds at 514}

* Status: I have written the code that finetunes DistilBERT on a standard Language Model instead of a masked LM. In the next step, I will add a final linear layer that goes from token dimensionality 768 to 300.
* There were 2 main changes on DistilBERT:
  + Turn all the relevant script arguments into function parameters, and use the default values for the others (gradient accumulation, distributed-hardware training etc.)
  + Introduce the function mask\_last\_token(inputs, tokenizer). Instead of having a Masked Language Model, we train only to predict the last word in a sentence, in a sequential way.
  + ***Course corrections***: I can also state that, since my objective is the quality of representations for definitions and examples, I may as well predict a MLM.  
    My code can also be changed, I can go and use DistilBertForMaskedLM
* Question: does my modified code produce sensible results, comparable with the original examples?

On **WikiText-2**, the SoA is:

Mogrifier LSTM + dynamic eval : Validation perplexity = 40.2, Test perplexity = 38.6

FRAGE + AWD-LSTM-MoS + dynamic eval : V.p.= 40.85 , T.p.= 39.14

On **WikiText-103**, the SoA is:

Transformer-XL + RMS dynamic eval : Valid.P. = 15.8, Test P.= 16.4

Running on the Lambda machine – waiting to see the results (if it is not too much for the hardware… and it is, if I attempt to finetune the whole model, even with DistilBERT. I must use the standard finetuning approach of adding a last layer and modifying only that one).

## Step 2b: Definitions and Examples 300d embeddings

(old) Objective:

Finetune DistilBERT, adding a last linear layer that goes from dimensionality 768 to 300, in order to use it to obtain 300-d sentence embeddings for Dictionay Definitions and Examples

{Error: outputs = model(inputs, labels)

… in \_\_call\_\_

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

**TypeError: forward() takes 2 positional arguments but 3 were given**

Let us try with:

outputs = model((inputs, labels))

* result = self.forward(\*input, \*\*kwargs)
  + attention\_mask = torch.ones\_like(input\_ids) # (bs, seq\_length)

TypeError: ones\_like() received an invalid combination of arguments - got (tuple), but expected one of:

\* (Tensor input, torch.dtype dtype, torch.layout layout, torch.device device, bool pin\_memory, bool requires\_grad)

\* (Tensor input, bool requires\_grad)

softmax() is not the right function, I should be looking for crossEntropyLoss() and similar…

}

* Hypothesis:
  + use distilBERT to encode the input window in: 768 x N dimensions
  + Operate with a learned linear layer, that goes 786d x N> 300d x N
  + Use the 300d x N embedding to predict Masked Language Model  
    (note: it’s complicated – it can not be only one layer, I need to keep the words coherent, and I should mask the same positions – there are 2 alternatives to solve this:
    - Modifying BertForMLM
    - Changing the layer, that will produce directly a sentence embedding with (768 x N) > 300. Must change the objective then?, from MLM
  + When the added linear layer has been optimized, apply the sequence:
    - 768 x N (even if we have only |s| words), distilBert embedding for all the words in the sentence
    - 300d x N, after applying the linear layer
    - Get the average over the N slots, in 300d

Observation:

The DistilBERT tokenizer does not include phrases – and it is named ‘distilbert-base-uncased’.

Therefore… I should exclude phrase pre-processing from the training corpus, or maybe use an alternative version of it without it.

Observation:

It is possible that even a dimensionality reduction would \*not\* mean that the Bert-originated sentence embeddings be in the same embedding space as the pre-trained 300d FastText word embeddings.

* **Hypothesis: AutoEncoder approach:**The objective is to create a 768 > 300 Neural Network that holds in 300 dimensions as much information as possible from the sentence embedding from (d)BERT, that was computed averaging over the last layer on (768 x N)  
  + Read a line (768 x |s|). Pad it to (768 x N).
  + Obtain the line embedding, averaging the last layer on (768 x N)
  + Apply linear layer: 768 > 300
  + Decoder part: try to reconstruct the window embedding of 768.

Writing…

{Runtime Error: CUDA out of memory…

torch.cuda.empty\_cache() necessary?

It helps, but the main issue was not using torch.no\_grad()}

Now the Autoencoder is made of Encoder + Decoder where each of them has 2 linear layers:

768 > 534 > 300 , and 300 > 534 > 768

Currently:

* read lines from WT-2 train
* pad or cut to a 256-tokens window
* obtain DistilBERT embeddings
* Train the AutoEncoder to encode in 300 dimensions and then reconstruct.  
  Optimized on the cosine-distance reconstruction loss on the WikiText-2 validation set. Use early-stopping.

**Hypothesis 2) Simple vector average from FastText’s pre-trained 300d vectors**

We recall the previous criticism:

“It is possible that even a dimensionality reduction would \*not\* mean that the Bert-originated sentence embeddings be in the same embedding space as the pre-trained 300d FastText word embeddings.”

This means there are 3 alternatives:

1. Use distilBERT for everything, including single words (taken from the vocabulary of the training set), to produce embeddings of d=768
2. Use distilBERT + encoding for everything, to get to 300d
3. Obtain the sentence embeddings for Definitions and Examples by averaging the FastText pre-trained 300d vectors