4H3-GS-11d-02

The 35th Annual Conference of the Japanese Society for Artificial Intelligence, 2021

ニューロンを活性化させるテキストに基づく ニューラル自然言語処理モデルの解析手法

A Method for Inspecting Neural NLP Models Based on Text That Activates Neurons

大葉 大輔 [*1] Daisuke Oba　吉永 直樹 [*2] Naoki Yoshinaga　豊田 正史 [*2] Masashi Toyoda

[*1]東京大学 The University of Tokyo　[*2]東京大学 生産技術研究所 Institute of Industrial Science, the University of Tokyo

In this study, as an approach to provide a deeper insight into neural Natural Language Processing (NLP) models, we propose a method for analyzing the roles of individual neurons, which is the finest component of the models, by observing sentences which strongly activate individual neurons. Our method retrieves the sentences from the massive corpora, and abstracts the sentences for interpretation using data-mining techniques. In the experiments, we demonstrate that our method can give thoughtful insights into what linguistic aspects each neuron of a given model captures as well as how multiple neurons relate to each other.

1. はじめに

ニューラルネットワークベースの自然言語処理モデルは，目的タスクに最適化された内部ベクトル表現またはニューロン[*1]を学習する能力によって，様々な自然言語タスクで成功を収めている．その一方で，モデルが実際にどのような言語現象を学習するのかを理解することは困難であるため，モデルおよびその構成要素（例: 注意機構）が学習する言語現象を明らかにする取り組みが行われている (§ 2.)．しかし，その多くは調査の対象とするモデルの構成要素および言語現象ごとに特化したトップダウンなアプローチとなっている．結果として，これら既存の方法は，対象の言語現象とモデルの要素に関して深い洞察与える一方で，多様なモデルの構成要素や広範な言語現象を対象とした調査を行うことが難しい．

では，ニューラル自然言語処理モデルを分析する一般的な方法はないだろうか？本稿では，モデルの最も細かな構成要素である “ニューロン” に注目し，そのニューロンが捉える多様な言語現象を明らかにする統一的な方法論を提案する (§ 3.)．提案手法は，言語注釈（例: 品詞タグ）とメタデータ（例: 著者）で拡張された大規模テキストコーパスを用いて，ターゲットモデルの各ニューロンを活性化する文集合を抽出し，頻出パターンマイニングやクラスタリングによる抽象化によって，各ニューロンが捉える言語現象を明らかにする．加えて，各ニューロンに紐づく文集合や言語現象に基づいて，複数ニューロン間の関係についても深い洞察を与える．

実験では BERT (base-uncased) [Devlin 2019] を例に，BERT のニューロンが特定の単語やフレーズだけでなくドメインや感情極性，さらにはデータのバイアスといった非明示的な現象を捉えていることを示す (§ 4.2)．更に，同様の言語現象を捉えているニューロン群の存在を示す (§ 4.3)．

本研究の貢献は以下の通りである．

- ニューラル自然言語処理モデルの各ニューロンが捉える言語現象をデータに基づき明らかにする方法論を提案する．
- BERT の各ニューロンが明示的または暗黙的に捉えている多様な言語現象を明らかにする．
- BERT の同じレベルの層に存在するニューロン，または複数の層の同一次元に存在するニューロンの集合が同様の言語現象を捉えていることを示す．

連絡先: 東京大学大学院 情報理工学系研究科
〒 153-8505 東京都目黒区駒場 4-6-1 生産研 Ee-505
E-mail: oba@tkl.iis.u-tokyo.ac.jp

*1　内部ベクトル表現の任意の次元の値を指す．

2. 関連研究

これまで，ニューラルネットの “層” [Hewitt 2019, Liu 2019], “注意機構の重み” [Kovaleva 2019, Clark 2019, Brunner 2020], 及び “モデル全体” [Petroni 2019, Broscheit 2019, Roberts 2020] などのモデルの部分構造が捉える言語現象を調べる研究が行われてきた．これらの研究では，分析するモデルがどのような言語現象（例: 構文情報 [Jawahar 2019, Bolkubasi 2016, Miaschi 2020] および意味論的知識 [Tenney 2019a, Ettinger 2020]）を捉えているか仮説を立てた上で，評価分類器の訓練や評価データセットの構築を行なってその仮説を検証している．このようなトップダウンなアプローチは，モデルのどの要素が注目に値し，更にどのような言語現象を捉えていそうなのか “あたり” がついていない場合には，手探りの状態となり効率が悪い．

本研究が焦点を当てる，モデルの最小構成要素である “ニューロン” が捉える言語現象を，ニューロンの値と入力文の対応に基づいて分析することも行われている [Karpathy 2015, Shi 2016]．これらの研究では，要素あたりごく少数の入力文に焦点を当てて分析しているため，対象とする言語現象は限定される．さらに，モデルの部分構造を分析の対象とした研究と同様に，特定の言語現象に特化して分析を行なっており，事前に分析対象とする言語現象に “あたり” をつける必要がある．

我々の研究と最も関係する研究として，勾配上昇に基づきニューロンを強力に活性化する入力文を生成する研究が行われている [Poerner 2018]．しかし，この方法が生成する入力文は事前に定めた長さの文に制限されるだけでなく，生成する入力文ごとに学習率やアニーリング温度等のハイパーパラメータの調整を要するため，モデル内の大量のニューロンに対して，それらを活性化する文を求める際にはコスト面で問題となる．

3. 方法論

関連研究で述べたように，注目するモデルの構成要素と，その要素が捉える言語現象について仮説を立てて検証する既存の分析手法は，モデルが膨大な数のニューロンから構成されるこ

とや，言語現象が多岐に渡ることを考えると，効率が悪いものとなっている．そこで，本節では，モデルの最小構成要素である各ニューロンが強く反応するテキストを大量に収集し，データマイニング技術を用いて得られたテキストを抽象化することで，モデルがその細部でどのような言語現象を捉えているのかを分析する方法論を提案する．

提案手法は，BERT の事前訓練に使用される BookCorpus [Zhu 2015] のような大規模テキストコーパスを前提に，まず，各テキストを基礎解析やメタデータを用いて多角的な分析を補助する情報を付与する (**Step 1**)．このようにして得られたコーパスを分析対象のモデルに入力し，各ニューロンが最も活性化する文集合を獲得する (**Step 2**)．獲得した文集合を頻出パターンマイニング等のデータマイニング技術により抽象化することで，各ニューロンが特徴付ける多様な言語現象を明らかにする (**Step 3**)．さらに，獲得したテキストおよび抽象化した言語現象に基づいてニューロン間の関係を分析する (**Step 4**)．以降では，提案手法の詳細を述べる．

Step 1. テキストコーパスのデータ拡張:
大規模なテキストコーパスを用意し，品詞解析や構文解析を実行することにより得られる基本的な言語注釈情報 (例: 品詞タグや構文木) を各文に関連づける．また，テキストの表層に現れない情報として，コーパスに付与されているドメインやトピックといったメタデータ (表 1) があれば合わせて活用する．

Step 2. 各ニューロンを強く活性化させるテキストの獲得:
Step 1 で得られたコーパスから各ニューロンが強く反応する文を獲得する．基本的には，Step 1 のコーパスの各文を分析対象のモデルに入力として与え，各文に対する各ニューロンの値を計算する．全入力文に対する各ニューロンの反応を保持するのは空間計算量的に非現実的であるため，優先度付きキューを用いて，各ニューロンが最も反応するテキスト K 件をシングルパスで抽出する．

Step 3. 各ニューロンを強く活性化させるテキストの抽象化:
Step 2 で各ニューロンに対して獲得した文集合を観察することで，各ニューロンが捉える様々な言語現象に関する洞察が得られると期待できるが，膨大な文集合からそれらに共通する言語現象を人手で読み解くことは難しい．そこで，頻出パターンマイニングにより文が含む情報を抽象化することで，文集合に共通して現れる特徴的な言語現象を明らかにする．

具体的には，まず PrefixSpan [Han 2001] の適用により頻出する (skip) n-gram を観察する．各ニューロンに特有の n-gram を収集するために，コーパス全体から計算された各 n-gram の頻度を用いた相対頻度を参照する．また，事前に各文に紐づけた言語注釈を活用することで，例えば FREQT [Asai 2004] を用いて頻出する部分構文構造を獲得することや，メタデータの分布を用いて表層には表われない文の特性を獲得することが考えられる．更に，クラスタリングを用いて，各ニューロンが特徴付ける典型的な文および言語現象を確認する．

Step 4. 解析対象とするモデルの構成要素の抽象化:
これまでに獲得した各ニューロンを強く活性化させるテキスト集合およびその抽象化により得られた言語現象をニューロン横断的に分析する．各ニューロンに対応するテキスト集合や言語現象をニューロン間で比較することや，与えられたテキストに対するニューロンの反応パターンを k-means クラスタリングすることにより，ニューロン間での役割の類似および相異について分析する．また，上述する情報に基づいて学習データの異なるモデル間でニューロンを比較する．

表 1: 実験で用いたデータセット．各データのドメインと利用可能なメタデータを示す．すべて英語のテキストで構成されている．

データセット	文数	ドメイン	メタデータ
BookCorpus [Zhu 2015]	40M	book	author
English Wikipedia*2	40M	wiki.	entity
Sentiment140 [Go 2009]	2M	twitter	sentiment
IMDB [Maas 2011]	0.3M	review	sentiment
20Newsgroups [Ken 1995]	0.2M	news	topic
Reuters*3	39K	news	category
Total	82M	-	-

4. 実験

BERT [Devlin 2019] を対象に，提案手法を使用することで，先行研究が明らかにした様々な BERT の振る舞いを再確認できることを示すとともに，これまでに明らかとなっていない洞察が得られることを示す．

4.1 設定

モデル　本論文では解析対象として事前訓練済み BERT（base-uncased，12 層，768 次元の隠れ状態）*4 を使用する．以降，各層の隠れ状態の各次元をニューロンとして扱う（合計で 12×768 ニューロン）．BERT では入力文の各トークンに対して隠れ状態が計算されるため，入力文に対するニューロンの活性度合いは隠れ状態の値のトークン平均とする．

データ　さまざまなドメインからなる 6 つの英語テキストコーパスを使用する．事前に各コーパスを文分割器*5 を使用して文単位に分割し，記号*6 の繰り返しを 1 つの記号に正規化する (例: +++ → +)．BERT の入力形式に合わせるために FastTokenizer*7 を使用して各文をトークン化する．このとき 3 トークン未満の文は除外する．コーパス内の各文にメタデータを関連づけることに加えて，品詞解析器*8 を用いて品詞タグを，さらには構造解析器 [Manning 2014] を用いて句構造を各文に注釈付けする．表 1にデータセット統計を示す．

4.2 BERT の各ニューロンが反応するテキストの抽象化

各ニューロンを最も活性化させる 10K 文を頻出パターンマイニングにより抽象化する．10K 文に平均して最も高いスコアを与えるニューロンの中からランダムに選択したニューロン (N_1, N_2, N_3) についてのマイニング結果を表 2に示す．表から，BERT のニューロンが捉える様々な言語現象を読み取ることができる．例えば N_1 は国際的な話題を想起させる単語 (例: *polulation, Olympics, World*) によって活性化され，N_2 はカジュアルな名詞 (例: *lol, haha, cuz, ya*) によって活性化されている．事前に紐付けた言語注釈の観点では，前置詞 (IN) の後に基数 (CD) が出現することや，複数の名詞句 (NP) で構成される名詞句を N_1 が捉えていることなどが分かる．こういった語順や品詞，さらには構文構造などの基礎的な言語現象を BERT が捉えていることは既に報告されている [Tenney 2019a, Kunz 2020]．本研究は，モデルの基本的な単位 (ニューロン) が捉える，これら多様な言語現象を統一的に明らかにしている点で先行研究と大きく異なる．

∗2　12/20/2020 ver. https://dumps.wikimedia.org/enwiki
∗3　http://kdd.ics.uci.edu/databases/reuters21578
∗4　https://github.com/huggingface/pytorch-pretrained-BERT
∗5　nltk sentence tokenizer ver. 3.2.4
∗6　!@#$%^&∗()_+-=[]{};:?,.
∗7　https://github.com/huggingface/tokenizers
∗8　https://www.logos.ic.i.u-tokyo.ac.jp/ tsuruoka/lapos

表 2: BERT の各ニューロンを活性化させるテキスト集合から抽象化された言語現象の例.

N_1	11 層目の 309 番目の隠れ状態
名詞	% Summer population Olympics August
動詞	competed make held received According
形容詞	total old full-time equivalent federal
単語系列	(competed Summer Olympics)
品詞系列	(IN DT CD), (IN DT IN), (IN CD CD)
部分句構造	(NP (NP) (PP (NP))), (NP (NP) (PP (IN)))
品詞の分布	名詞: 35.2%, 動詞: 8.6%, 形容詞: 4.3%
平均系列長	15.9
N_2	6 層目の 200 番目の隠れ状態
名詞	lol haha LOL twitter im cuz Twitter u ya
動詞	'm got im miss lol think love know want
形容詞	twitter u im sad new good old sick last
単語系列	(Br Open J), (Open Sci J)
品詞系列	(NN NN NN), (JJ NN NN), (PRP VBP NN)
部分句構造	(ROOT (NP (PRP)) (VP)), (ROOT (NP) (VP (NP)))
品詞の分布	名詞: 29.7%, 動詞: 18.3%, 形容詞: 6.2%
平均系列長	7.5
N_3	8 層目の 270 番目の隠れ状態
名詞	fun part life movie kind thing moment
動詞	's going 'm felt fucking feel looked
形容詞	first much ready funny amazing best
単語系列	(is going be), (was going be)
品詞系列	(DT JJ NN), (PRP DT NN), (PRP RB JJ)
部分句構造	(ROOT (NP (PRP) (NP))), (ROOT (NP) (VP (VBD)))
品詞の分布	名詞: 16.2%, 動詞: 16.5%, 形容詞: 11.2%
平均系列長	11.4

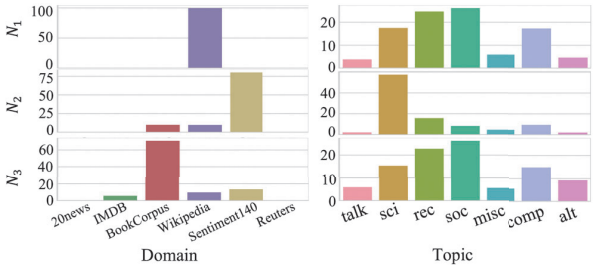

図 1: 各ニューロンを活性化させるテキスト集合におけるメタデータの分布. Topic に関しては, 関連する 20Newsgroups データセットのみから収集したテキスト集合について分布を計算した.

図 1に各ニューロンに対応する文集合に紐づけられたメタデータの分布を示す. 図より, 3 つのニューロンが捉えるテキストには現れない情報がわかる. 例えば N_1 は Wikipedia ドメインに, また N_2 は 20Newsgroup の *sci* (science) ドメインに強く反応する. 興味深いことに, 一種の分類器のように特定のクラスに対してより強力に反応するニューロンが存在した. 例えば 12 層目の 104 番目のニューロンを活性化させる文の 89.1% は正の感情極性を持ち, また 11 層目の 131 番目のニューロンを活性化させる文の 61.7%は同じ著者のものであった. これらは, BERT が教師なしでタスクに関する知識を獲得しているという先行研究の報告 [Petroni 2019, Roberts 2020] と一致する.

4.3 ニューロン横断的な分析

まず, 4.2節で抽象化した言語現象をニューロン間で比較する. 例として各ニューロンが反応する文の "平均長" を図 2に示す. 紙面の都合上, 一定範囲の (各層 301〜350 番目のニューロン) 結果を掲載した. 興味深いことに, 各層の同次元に位置するニューロンが同じような長さの文に強く反応していた. また "ドメイン" や "動詞の数" などへの反応に関しても同様の傾向が見られた. この分析結果は, 既に報告されている各層の役割 [Liu 2019, Hewitt 2019] とは別に, BERT の層の各次元には何らかの共通した役割があることを示唆する.

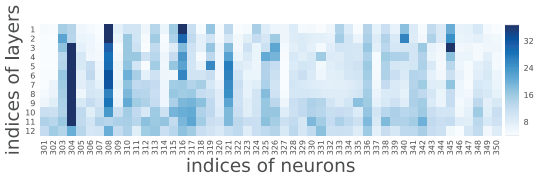

図 2: 収集したテキスト集合の平均文長.

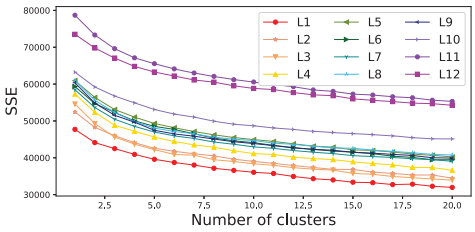

図 3: クラスタの数を変化させた時のクラスタ内誤差平方和 (SSE) の層 (L) ごとの結果.

次に, 各ニューロンを層ごとにまとめて見たときに, それらが捉える言語現象が層によってどれだけ多様なものであるかを分析する. 各ニューロンを最も活性化させる文を, BERT の [CLS] トークンの埋め込み $\in \mathbb{R}^{768}$ を使用してベクトル表現に符号化し, それらを層ごとに集約し k-means クラスタリングする. クラスタ数を変化させた際の損失 (SSE) を図 3に示す. 深い層ほどクラスタ数が小さい時の損失が大きいことから, 深い層に存在するニューロンほど, 集団で見たときに互いに多様な言語現象を捉えていることが読み取れる. これは, BERT が深い層で高度な言語現象を捉えているという主張 [Jawahar 2019, Tenney 2019b] に合う結果である.

最後に, 与えられた文集合に対するニューロンの活性化パターンをクラスタリングすることにより, 同様の言語現象を捉えているニューロン群を抽出する. 具体的には, 表 1のコーパスからサンプリングした 1K 文それぞれを, 各ニューロンの活性化スコアを各次元の値に対応させたベクトル表現とし, k-means クラスタリング ($k = 10$) を実行する. ニューロンの値の範囲は層に依存するため, 例として 6 番目の層のニューロンについて検証する. 図 4にクラスタリング結果を示す. 可視化には t-SNE [Van 2008] を使用しており, 各色は帰属クラスタを示している. 図より, ニューロンの活性化パターンには大まかな分類があることがわかる. この分析から, BERT のニューロンは集団として同一言語現象を捉えていることが分かった.

4.4 モデル横断的な分析

単一モデルを超えて, 追加学習前後での BERT のニューロンを分析する. 具体的には, 追加学習前後における, 与えられた文集合に対するニューロンの反応の変化を観察する.

Reuters (表 1) を用いた文書分類タスクにより BERT, および一層の線形層を追加学習した. Epoch 数 5, 学習率 10^{-5}, Dropout 率 0.5, バッチサイズ 8 で, Adam [Kingma 2015] を用いて最適化を行なった. 追加学習前後のモデルに与えるテキス

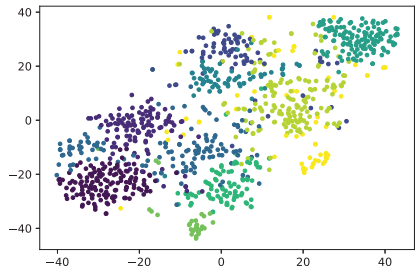

図 4: ニューロンの活性化パターンのクラスタリング結果.

表 3: 与えられたテキスト集合に対する，追加学習 (FT) 前後におけるモデルのニューロンの反応の変化.

	Used in FT	In-domain	Manhattan	Euclidean
Reuters (train)	✓	✓	**2161.4**	**34.8**
Reuters (test)		✓	1947.9	31.2
Sentiment140			1553.4	24.5

トには，追加学習に使用した Reuters の学習データ，Reuters のテストデータ，異なるドメインの Sentiment140 から 10,000 文ずつサンプリングした，異なる性質の 3 つのデータを用いた．

各文をモデルへ入力した時のニューロンの値からなるベクトル (768 × 12 = 9,216 次元) によって各文を表現し，追加学習前後における各文ベクトルの距離を計算し，平均したものを表 3に示す．表から，追加学習データに対する反応が特に大きく変化していること，及び同ドメインでも学習データ外のテキストへの変化は小さいことが分かる．この結果から，BERT のニューロンは今回の分類タスクを解くために，テキストの “ドメイン” を超えた学習データ特有の情報を学習していることが考えられる．それが例えば，より細かなドメインであるのか，またはデータのバイアス (例: 注釈バイアス [Oba 2019]) であるのかを明らかにすることは今後の課題としたい．

5. おわりに

ニューラル NLP モデルの細部が持つ役割を調べる際に，既存の方法論は分析対象とするモデルの部分構造とその部分構造が捉える言語現象に予めあたりをつけて分析を行うため，効率の面で課題があった．本稿では各ニューロンを強力に活性化させる大量のテキスト集合を獲得および解析することで，各ニューロンが持つ様々な役割を明らかにするボトムアップな方法論を提案した．実験では BERT (base-uncased) が実際にニューロンレベルで捉えている多様な言語現象を，テキストに暗黙的に存在するものまで示した．また，抽象化した言語現象やニューロンの活性化パターンの比較により，協調的に動作するニューロンの集団の存在を示唆した．特に BERT の各層の同一次元が層間で共通の役割を持っていることを示唆する分析結果は非常に興味深い．更に，追加学習を行なった BERT のニューロンの解析から，一種の過学習とも取れる学習データへのニューロンの選択的な応答を示した．

今後は，本稿で示唆された BERT の各層の同一次元のような注目に値するモデルの構成要素を引き続き特定し，それらを既存手法 (§ 2.) と絡めて深掘りする．また，学習データやタスクを変化させて，今回示唆されたニューロンの学習データに対する選択的応答を体系的に分析する．更に，モデルの核となるニューロンを特定し圧縮や推論の高速化への応用を模索する．

謝辞 本研究は，JST, CREST, JPMJCR19A4 の支援を受けたものです．

参考文献

[Asai 2004] Asai, T., Abe, K., Kawasoe, S., Sakamoto, H., Arimura, H., and Arikawa, S. Efficient substructure discovery from large semi-structured data In IEICE. Vol. 87. No. 12. pp. 2754–2763. (2004)

[Bolkubasi 2016] Bolukbasi, T., Chang, Kai-Wei, Zou, J., Saligrama, V., and Kalai, A. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In NIPS. pp. 4356–4364. (2016)

[Broscheit 2019] Broscheit, S., Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking. In CoNLL. pp. 677–685. (2019)

[Brunner 2020] Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., and Wattenhofer, R. On Identifiability in Transformers. In ICML. (2020)

[Clark 2019] Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. What Does BERT Look At? An Analysis of BERT's Attention. arXiv 1906.04341. (2019)

[Devlin 2019] Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT. pp. 4171-4186. (2019)

[Ettinger 2020] Ettinger, A. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. arXiv:1907.13528. (2020)

[Go 2009] Go, A., Bhayani, R., and Huang, L. Twitter sentiment classification using distant supervision. (2009)

[Han 2001] Han, J., Pei, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., and Hsu, M. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE. pp. 215. (2001)

[Hewitt 2019] Hewitt, J. and Manning, C. D. A Structural Probe for Finding Syntax in Word Representations. In NAACL. pp. 4129-4138. (2019)

[Jawahar 2019] Jawahar, G., Sagot, B., and Seddah, D. What Does BERT Learn about the Structure of Language? In ACL. pp. 3651-3657. (2019)

[Karpathy 2015] Karpathy, A., Johnson, J., and Fei-Fei, L. Visualizing and understanding recurrent networks. arXiv:1506.02078. (2015)

[Ken 1995] Ken, L. Newsweeder: Learning to filter netnews. In ICML. pp. 331-339. (1995)

[Kingma 2015] Kingma DP. and Ba, J. Adam: A Method for Stochastic Optimization. In CoRR. Vol. abs/1412.6980 (2015)

[Kovaleva 2019] Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. Revealing the Dark Secrets of BERT. In EMNLP-IJCNLP. pp. 4365-4374. (2019)

[Kunz 2020] Kunz, J. and Kuhlmann, M. Classifier Probes May Just Learn from Linear Context Features. In COLING. pp. 5136-5146. (2020)

[Liu 2019] Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. Linguistic Knowledge and Transferability of Contextual Representations. In NAACL-HLT. pp. 1073–1094. (2019)

[Maas 2011] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning Word Vectors for Sentiment Analysis. In NAACL-HLT. pp. 142–150. (2011)

[Manning 2014] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. pp. 55–60. (2014)

[Miaschi 2020] Miaschi, A., and Dell'Orletta, F. Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation. In RepL4NLP. pp. 110–119. (2020)

[Oba 2019] Oba, D., Yoshinaga, N., Sato, S., Akasaki, S., and Toyoda, M. Modeling Personal Biases in Language Use by Inducing Personalized Word Embeddings. In NAACL. pp. 2102–2108. (2019)

[Petroni 2019] Petroni, F., Rocktäschel, T., Miller, A. H., Lewis, P., Bakhtin, A., Wu, Y., and Riedel, S. Language Models as Knowledge Bases? In EMNLP. pp. 2463-–2473. (2019)

[Poerner 2018] Poerner, N., Roth, B., and Schütze, H. Interpretable Textual Neuron Representations for NLP. In EMNLP-BlackBoXNLP. pp. 325-327. (2018)

[Roberts 2020] Roberts, A., Raffel, C., and Shazeer, N. How Much Knowledge Can You Pack Into the Parameters of a Language Model? arXiv:2002.08910. (2020)

[Shi 2016] Shi, X., Knight, K., and Yuret, D. Why Neural Translations are the Right Length. In EMNLP. pp. 2278–2282. (2016)

[Tenney 2019a] Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., and Bowman, S. R., Das, D., and Pavlick, E. What do you learn from context? Probing for sentence structure in contextualized word representations. In ICLR. (2019)

[Tenney 2019b] Tenney, I., Das, D., and Pavlick, E. BERT Rediscovers the Classical NLP Pipeline. In ACL. pp. 4593–4601. (2019)

[Van 2008] Van der Maaten, L. and Hinton, G. Visualizing Data using t-SNE. In JMLR. vol. 9. pp. 2579–2605. (2008)

[Zhu 2015] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In ICCV. pp. 19-27. (2015)