ISSN 2186-5647

Tranformer のための Word Embedding 手法の提案

日大生産工　○張　宏毅　　日大生産工　山内　ゆかり

1. まえがき

自然言語処理タスク（NLP）においてRecurrent Neural Network（RNN）とConvolutional Neural Network（CNN）が深層学習としてよく使われている。RNNは入力がSequenceの時に最適ではあるが、逐次的に単語を処理する仕組みでトレーニングをする時に並列処理ができないというのも事実である。CNNはトレーニングをする時に並列処理されやすいが、一方で入力が長くなれば精度が落ちるというデメリットがある。そこでAshish VaswaniらはTransformer[1]というAttention機構のみを用いたモデルを提案した。Transformerとは、RNNやCNNなどを一切使わずにAttentionだけを使うことで、入力と出力の内容の広範囲な依存関係を捉えられるモデルである。

Transformerに入力する時、自然言語をWord Embeddingの形にする必要がある。従来の方法では、Word2vec[2]がある。しかし、Word2Vecには大規模のトレーニングデータが必要で、小規模のデータセットをWord2Vecで言語の多様性を十分に捉えられないため、一部の単語や文脈が未学習のままになることがあり、質が低下する可能性がある。

そこで、上記の問題点を解決するために新たなWord Embedding手法を提案する。

2. 従来研究

2.1 Transformer

Transformerは、EncoderとDecoderという2つの部分で構成されている。Encoderの役割は、Transformerに入力されたデータを機械が処理できる形式に変換する。例えば、言語翻訳のための学習では、ドイツ語などで書かれた文章が入力され、Encoderによって数値のベクトルに変換される。この処理には、Word Embeddingが必要とされている。Decoderの役割は、Encoderによって変換されたデータを受け取り、処理内容に応じて別の形式へ変換する。例えば、ドイツ語から英語への翻訳を行う際は、数値のベクトルに変換されたドイツ語の文章を英語の文章へと変換する。

Transformerの仕組みが発表された論文では、EncoderとDecoderの層がそれぞれ6つずつ用意されていた。Transformerのイメージは図1に示す。

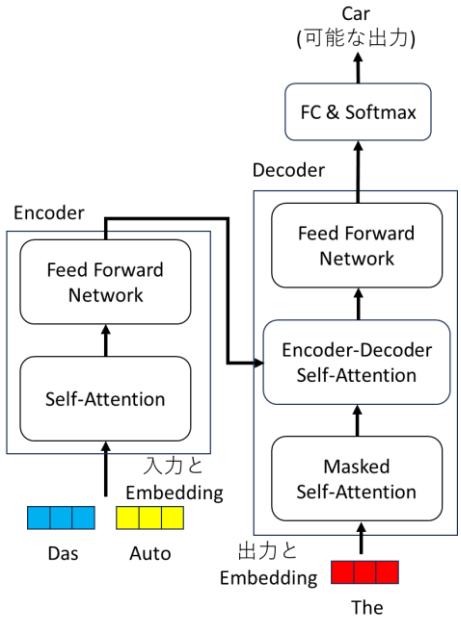

図 1.　Transformer

2.1.1 Encoder

Encoderは6つ同じ構造の層で構成されている。各層は Multi-Head Attention層と全結合層のサブ層で構成されている。それぞれのサブ層の後には残差結合と正規化関数であるLayer Normalizationがある。

2.1.2 Decoder

Decoderも同じく同じ構造の6層で構成されていている。但し、Decoderの各層はサブ層の間にEncoderの出力を受け取るMulti-Head Attention層も追加されている。そして、Decoderの最初のサブ層も書いてある通り Masked Multi-Head Attentionになっている。

2.1.3 Attention機構

Attentionの役割を簡単に言うと、ある単語の意味を理解する時に、どの単語に注目すれば良いかを表すスコアのことである。例えば英語の「it」を翻訳する場合、その単語だけでは翻訳できないだろう。「it」を含む文章中のどの単語にどれだけ注目すればいいのかいうスコアを表すのがAttentionである。

2.2 Word Embedding

コンピュータは自然言語を読み取ることができず、数値しか処理できないことはよく知られ

A Proposal of Word Embedding Method for Transformer

Hongyi ZHANG and Yukari YAMAUCHI

ているため、自然言語は何らかの形で数値に変換する必要がある。 Word Embeddingは、1つの自然言語の単語を数値にマッピングする方法である。

Word Embeddingは、NLPにおける言語モデリングと表現学習技術の総称である。概念的には、単語の数の次元を持つ高次元空間を、より低次元の連続ベクトル空間に埋め込むことを指している。各単語やフレーズは実数フィールド上のベクトルとしてマッピングされる。

2.3 Word2vec

Word2vecは、2013年にTomas Mikolovらが提案したWord Embeddingの一種である。

Word2vecのアルゴリズムは、軽量化なニューラルネットワークモデルを用いて大規模なテキストコーパスから単語の関連付けを学習する。一旦学習されると、このモデルは同義語を検出したり、部分的な文に追加の単語を提案したりすることができる。その名の通り、Word2vecは各単語をベクトルと呼ばれる特定の数値リストで表現する。ベクトルは、単語の意味的および構文的な特質を捉えるように注意深く選択される。

Word2vecのニューラルネットワーク部分は、CBOWモデルとSkip-gramモデルの2つアルゴリズムがある。CBOWモデルは、単語w_tに関連する文脈が知られている場合、単語w_tを予測するモデルである。Skip-gramモデルは、単語w_tを知られている場合、単語w_tに関連する文脈を予測するモデルである。

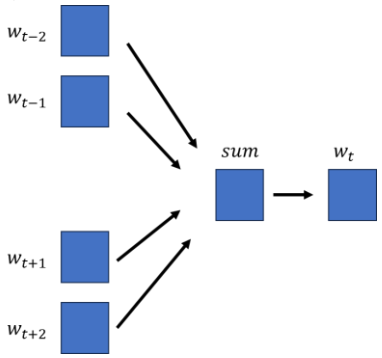

図 2. CBOW モデル

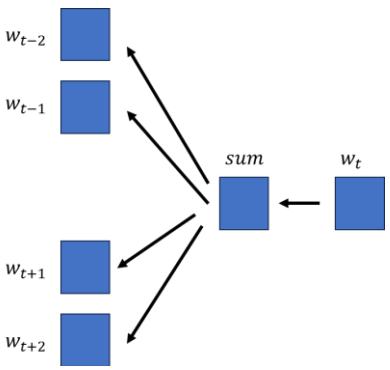

図 3. Skip-gram モデル

3. 提案手法

私は、Word2vec と異なるクラスタリング手法で構築された Word Embedding 方法を提案する。この提案手法は、入力データのトポロジーを保存しながらデータをマッピングする。これにより、関連性の高いデータが空間上でも近くに配置される。そしてこのクラスタリング手法は教師なし学習アルゴリズムであり、ラベルなしデータにも適用できる汎用性のあるかつ小規模なデータセットでも精度のあるアルゴリズムである。

4. 実験および検討

従来のWord2vecと提案手法と同じデータセットを用いて、Word Embeddingをする。得られたWord EmbeddingをTransformerの入力として入力させて、STS-B[3]テストで精度を確かめる。そして最後に提案手法が得られたWord Embeddingを従来のWord2vecに入力させることで、2つの手法を結合し再びSTS-Bテストをする。

5. まとめ

従来使われているWord Embeddingの手法であるWord2vecは大規模なデータセットでの学習が必要とされており、小規模なデータセットに向いていないというデメリットがある。私は従来のWord2vecと異なるクラスタリング手法を提案し、小規模なデータセットでもトレーニングできるWord Embeddingのアルゴリズムを提案する。

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, “Attention Is All You Need”, Proc arXiv (2017)

[2] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean.” Efficient Estimation of Word Representations in Vector Space”, arXiv(2013)

[3] STSbenchmark - stswiki, http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark, （参照2023-10-12）