ミニ特集「情報学と自然言語処理」

単語埋め込み手法の発展と応用

笹野 遼平・武田 浩一
（価値創造研究センター／知能システム学専攻）

はじめに

情報学の多くの分野と同じく、自然言語処理分野においても近年ニューラルネットに基づく様々な手法が開発され、既存のタスクの精度が大幅に向上したり、従来の技術では難しいとされていたシステムが実現したりしています。その中でも、単語の埋め込み手法は、ニューラルネットを用いることで大幅な発展を遂げた技術の代表例です。本稿では、近年の単語埋め込み手法の発展と、我々、名古屋大学大学院情報学研究科武田・笹野研究室における単語埋め込みを用いた取り組みについて紹介します。

単語埋め込み手法の発展

word2vecの登場

自然言語処理分野においてニューラルネットに基づく手法の有効性が認識されるきっかけの1つは2013年にMikolovにより発表されたword2vec[1]です。word2vecには、2層の単純なニューラルネットワークで構成されるCBOWとskip-gramという2つのモデルが実装されており、大量のテキストを与えると、そこに含まれる単語の意味を表現する単語埋め込み（word embedding）と呼ばれる固定長のベクトルを得ることができます。意味の近い単語の埋め込み表現は類似しているという性質があり、2つのベクトルがなす角の余弦を計算することにより単語間の類似度を計算することができます。また、word2vecが注目を集めた要因に、2つの単語のベクトルの差が2つの単語の関係を表すという性質があり、たとえば、“king”のベクトルから“man”のベクトルを引いたベクトルに“woman”のベクトルを足したベクトルは、“queen”と類似したベクトルになることが知られています。

様々な単語埋め込み獲得法

その後も、GloVe[2]やfastText[3]など新たな単語埋め込み手法が登場しています。GloVeは2014年にスタンフォード大学のグループにより発表された手法で、word2vecでは各単語ごとに周辺に出現する単語を確率的に扱うことで単語埋め込みを獲得するのに対し、GloVeではコーパス全体から得られる単語間の共起行列を用いて単語埋め込みを獲得します。fastTextは2015年にFacebook AI Researchから発表されたツールで、非常に高速に動作し、単語の部分文字列（subword）を考慮することで共通する形態素を含む単語間の類似性を捉えられるという特徴があります。これらの単語の埋め込み表現は単語の意味を固定長の密ベクトルで表現することから、自然言語を入力とするニューラルネットに基づくシステムにおいて広く利用されています。しかし、基本的に1つの単語に1つの

ベクトルを割り当てることから、多義語や文脈によって大きく意味が変化する語の意味を扱うのには不向きという問題がありました。

文脈を考慮した単語埋め込み

そこで、文脈による単語の意味の違いを扱えるよう、近年、ELMo[4]やBERT[5]をはじめとした、文脈を考慮した単語埋め込み（文脈化単語埋め込み）生成手法が提案され、言語理解タスク等の精度を大幅に向上させています。このうちELMoは2017年にAllen Institute for AIから発表されたモデルで、2層のLSTM（Long Short-Term Memory）言語モデルをベースとしています。LSTM言語モデルは、時系列を考慮したニューラルネットであるRNN（Recurrent Neural Network）を改良したもので、従来のRNNの問題であった勾配消失問題を解消し、長期的な依存関係の学習が可能なモデルです。ELMoでは、順方向と逆方向のLSTMを別々に学習させたネットワークの中間層を重み付けし結合することで、前後の文脈を考慮した埋め込み表現を生成します。一方、BERTは2018年にGoogleから発表されたモデルで、self-attention機構を採用したTransformerと呼ばれる構造をベースにしています。与えられた文から無作為に削除（マスク）された単語を予測するマスク単語予測と、与えられた次の文が本来の文であるかどうかを判定する次文判定という2つのタスクで事前訓練を行い、目的とするタスクでfine-tuningすることで、各タスクに適した文脈化単語埋め込みを生成します。

質問応答と言語理解

質問応答は1999年にTREC-8国際会議で評価タスクとして導入されて以来、情報検索の発展した手法として研究されてきました。2011年にIBMが開発した質問応答システムWatsonが、TVクイズ番組Jeopardy!で人間のチャンピオン2名に勝利したことで商用化が進み、その後がん治療の選択肢を医師に提示する医療応用などにつながりました。質問応答手法は、質問に対して解答を生成するための情報源の特定（主に情報検索）と、解候補の抽出と順位付けに分離でき、後者は提示された質問と情報をもとに解答を生成する機械読解（machine comprehension）というタスクとして一般化されています。機械読解は、与えられた情報を知識として蓄積・再利用するための「言語理解」の機能を実現する鍵となる重要なタスクであると認識されています。前述の文脈化単語埋め込みがこのような機械読解や言語理解に関する9種類のタスクを評価するベンチマークGLUE[6]の精度を大きく向上させることが知られています。

研究室での取り組み

単語ベクトル空間における単語の分布分析

単語ベクトル空間において、特定の意味クラスに属する単語集合がどのように分布しているかの分析に取り組みました[7]。具体的には、特定の意味クラスに属する単語集合の一部を与え、その意味クラスに属する単語（正例）とそれ以外の単語（負例）を識別するにはどのようにモデル化を行えば良いか調査し、正例の中心ベクトルからの距離や混合ガウス分布に基づくモデル化では十分な精度を得ることはできず、高い精度の実現には負例も考慮した識別学習に基づくモデル化が必要であること

を明らかにしました。さらに、各単語の対象クラスにおける典型度合のモデル化にも取り組み、正例と負例の識別には識別学習に基づくモデルが適していたのに対し、典型度合のモデル化には負例と正例それぞれの中心ベクトルの差分（off-set）ベクトルに基づく単純なモデルが適していることを明らかにしました。

文脈化埋め込みを用いた慣用句判定

慣用句として使用される表現の中には「足を洗う」などのように、慣用句として用いられる場合と文字通りの意味で用いられる場合の両方があるものが存在します。このような表現の用法を正しく判定できるよう、慣用句を構成する動詞と名詞（「足を洗う」の場合は「洗う」と「足」」）のBERTにより得られた文脈化単語埋め込みを使った判定モデルを構築し、その有効性を示しました[8]。さらに、慣用句を構成する動詞と名詞の平均的な埋め込み表現も考慮することで判定精度が向上することを明らかにしました。

文脈化埋め込みを用いたフレーム知識の自動獲得

文脈化埋め込みを用い、言語理解のための基本的な知識の１つであるフレーム知識を自動獲得する研究にも取り組んでいます[9]。フレームとはある語の意味を理解するために必要となる背景知識をまとめたもので、テキスト中に出現した述語は何らかのフレームを喚起するとされます。たとえば、人手で整備された代表的なフレーム知識であるFrameNetでは、“eat”や“consume”が喚起するIngestionフレームや、“sleep”や“nap”が喚起するSleepフレームなど1200あまりのフレームが定義されており、出現する述語ごとに喚起するフレームを人手で付与したコーパスとともに公開されています。多くの動詞は文脈に応じて異なるフレームを喚起するため、フレーム知識をコーパスから自動獲得しようとした場合、その第一歩として各述語が喚起するフレームの違いを認識する必要があります。我々の研究室では現在、文脈化埋め込みを活用し、テキスト中の述語が喚起するフレームの違いを認識する研究に取り組んでいます。

参考資料

[1] https://github.com/tmikolov/word2vec

[2] https://github.com/stanfordnlp/GloVe

[3] https://github.com/facebookresearch/fastText

[4] https://allennlp.org/elmo

[5] https://github.com/google-research/bert

[6] https://gluebenchmark.com/

[7] Ryohei Sasano, Anna Korhonen: Investigating Word-Class Distributions in Word Vector Spaces, （ACL 2022）

[8] Ryosuke Takahashi, Ryohei Sasano, Koichi Takeda: Leveraging Three Types of Embeddings from Masked Language Models in Idiom Token Classification（*SEM 2022）

[9] Kosuke Yamada, Ryohei Sasano, Koichi Takeda: Semantic Frame Induction using Masked Word Embeddings and Two-Step Clustering（ACL-IJCNLP 2021）

https://tamatebako.i.nagoya-u.ac.jp/4256/