『統語・意味解析情報付き日本語コーパス』を使用した単語の分散表現獲得におけるデータ拡張の展望

岸山 健

平成 29 年 4 月 16 日

1 はじめに

本発表[1]では構文解析情報を利用した単語の分散表現獲得の手法と結果を報告し、「NPCMJ (統語・意味解析情報付き日本語コーパス)[2] を用いた、word2vec[3] による単語の分散表現におけるデータ拡張」を展望として述べる。

まず先行研究として単語の分散表現について述べ、次に機械学習における学習データの扱い、特にデータ拡張についてを述べる。その後、分散表現における問題の解決策として、構文解析情報を利用したデータ拡張の方法と結果を述べ、最後に NPCMJ を利用することにより、どうデータ拡張の問題が解決できるかを展望として述べたい。

2 先行研究

2.1 単語の分散表現

単語の統語と意味に関わる分散表現 (word embeddings) を獲得する手法として word2vec を挙げる。ここでは word2vec の理論の解説は省くが[4]、重要な情報は 3 つある。 1 つ目は、word2vec ではテキスト中の単語を入力として、周辺の単語を出力させるニューラルネットを形成させるということである。 2 つ目は、獲得されたニューラルネットによる出力自体には一切興味がなく、より重要なのはニューラルネットが形成された時に学習された重みベクトルが、入力の単語の意味を反映しているという点である。そして 3 つ目は、学習時に「周囲の何単語を予測するモデルを形成するか」などの値を hyper-parameter として指定する必要がある点である。

word2vec が話題になった理由の一つは、形成されたニューラルネットの重みベクトル自体が、「入力として与えられた単語」の分散表現として機能している点と、またその分散表現がベクトル加法に対応しているという点である (Mikolov et al. 2013)。例えば、「王」-「男」+「女」=「女王」のような計算ができる。特に 2 つ目の特徴が意味するところは、ベクトルが意味的な関係を表現しているという点であり、応用分野としてはレコメンドシステム[5] や機械翻訳 (Mikolov, Le, and Sutskever 2013)、評判分析 (加藤和平, 大島考範, and 二宮崇 2015)、感情分類 (黒崎優太 and 高木友博 2015) など複数の分野が挙げられる。

[1] 本稿は東北大学で行った発表に修正を加えたものです。余分に話しすぎた word2vec の理論を削り、話の根本にあるデータ拡張に関して追記いたしました。

[2] http://npcmj.ninjal.ac.jp/

[3] Apache License 2.0 で公開されています。 (code.google.com/archive/p/word2vec/)

[4] 個人のページですが、以下の解説が非常に丁寧で分かりやすかったです。 http://tkengo.github.io/blog/2016/05/09/understand-how-to-learn-word2vec/

[5] https://www.slideshare.net/recruitcojp/ss-56150629

2.2 word2vec の hyper-parameter

word2vec は入力として与えた単語から周囲の単語を予測するニューラルネットを形成させ、その重みベクトルを分散表現として得ることを目的としてるが、ニューラルネットを構築する上では hyper-parameter を決めなくてはならない。 hyper-parameter とは、ニューラルネットで獲得されるベクトルの重みの様なパラメターではなく、ニューラルネットで学習させる際にユーザーが指定するパラメターのことである。例えば、学習をどの程度で更新させるかや、重みベクトルの次元数、予測させる文脈長などが挙げられる。以下では予測させる文脈長をしているする window に焦点を当てる。

文は可変長であり、制限はない。したがって、文の長さは上限のないカウントデータということになる。例えば、今回の使用した wikipedia のサンプルデータは 165,739 文であった。下に x 軸を 1 文の形態素 [6] の数とし、y 軸を出現数として示した。

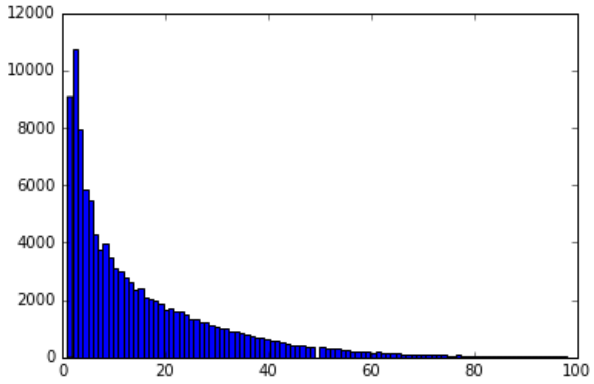


図 1: x 軸を 1 文の形態素の数とし、y 軸を出現数として示す

なお、データのばらつきを示す分散の値は 264.81 で、一文の中にある形態素数の平均は 17.04 であった。ここで問題となるのは、このデータに対して、`window = 8` という広く使われているパラメターは妥当なのか、という点である。日本語はいわゆる head final 言語であり文の構造上、文が長くなれば長くなるほど主部と述部の距離が遠くなる [7]。そこで、図 1 の様な平均 17.04 形態素 (≈ 単語) で分散の値は 264.81 のデータを周辺 8 単語で予測させるよりも、文を整形する方法 (つまり画像処理におけるデータ拡張のような処理) を検討した。

2.3 機械学習における学習データの扱い

データの整形に関して、機械学習におけるデータ拡張について述べる。画像処理の分野では学習の際にデータ拡張 (data augmentation) を行い、データの量を増やし学習させるモデルの頑強性を高めている (斎藤康毅 2016)。データ拡張においては、画像を回転させたり、歪めたり、ノイズを加えたり、コントラストを調整したりすることによってデータ量を増やしている。例えば、Python で実装された、TensorFlow または Theano 上で実行可能な高水準のニューラルネットワークライブラリである Keras[8] にも、画像データの拡張を簡単に行う ImageDataGenerator が標準で実装されている。

一方で言語データに手を加えてデータ拡張を行った研究は頻繁には見られない。著者推定のタスクなどにおいては、文の区切り方、句読点の使い方などの情報が特徴として学習される (石田基広 2008)。した

[6] 形態素と単語は別ですが便宜上、形態素を単語として今回は扱いました。

[7] 例えば、「**仙台市** は、宮城県の中部に位置する同県の県庁所在地かつ**政令指定都市である** 。」のような文と、 “**Sendai** is **the capital city** of Miyagi Prefecture, Japan, and the largest city in the Tōhoku region, and the second largest city north of Tokyo.” のような文を比較した場合のことを示しています。

[8] https://keras.io/ja/

がってデータ拡張が見られない理由は、画像処理の様な形でデータの変形を行うことが学習のための特徴を損なうことに繋がりかねないからだと考えられる。

しかし話者推定のタスクならば問題となりうるが、word2vec は単語の周辺を予測して分散表現を獲得する手法であるため、必要なのは文の構造ではなく文における単語の分布であると仮定する。すると、「仙台市は、宮城県の中部に位置する同県の県庁所在地かつ政令指定都市である。」のような文があった場合に、「仙台市は政令指定都市である。」や「仙台市は、宮城県の中部に位置する同県の県庁所在地。」、またトレースの情報を考慮して「同県は宮城県の中部に位置する。」のような文を作ることで、複数のメリットが得られる。

1 つは学習データの文の長さの平均と分散が下がることである。単文の長さが特定の分布に従うと仮定すると、複文構造を持った文の長さは、「いくつの複文を持つか」というパラメターに規定される。複文を単文に分けることで、文の長さは「複文の数」というパラメターに依存しなくなり、特定のパラメーターに従う確率分布 [9] に沿うと予想される。同時に文が分割されるため、文の長さの平均も下がり、window の指定の指標となる。

もう 1 つは利用できるデータの数が増えることである。学習させる対象によってはデータ量が部分的に不足する場合がある (黒崎優太 and 高木友博 2015)。そうした不足を複文を単文に分けデータ拡張を行うことで、 単純な “トリック” ではあるが、問題の軽減はできると考えられる。

3 構文解析情報を利用したデータ拡張と分散表現の学習

3.1 方法

今回の発表ではテキストデータに対するデータ拡張 (data augmentation) として、CaboCha による形態素解析を行い、 word2vec により分散表現の学習を試みた。

方法としては構文解析器である CaboCha (拓 and 裕治 2002) を使用し、計算機に処理しやすいフォーマットで出力させる [10]。フォーマットでは各チャンクに `* 0 5D 0/1 1.062087` のような記述があるが、意味するところは「指定されているチャンクはインデックスが 0 であり、そのチャンクはインデックスが 5 であるチャンクにかかっている」というものである。このフォーマットを利用し、それぞれのチャンクがどのチャンクにかかっているかを参照し、テキストデータを拡張した。

そしてデータ拡張の終了したデータに対して word2vec による分散表現の学習を試みた。パラメターである window　の値は 8 を指定した。精度の比較の手法に関しては形態論情報付きデータを使用するもの (正幸 and 照晃 2017) や、実際の文書分類課題の精度の向上を考慮するものがあり、今後参考にする予定ではあるが、今回は不本意ながらデータ拡張有りの条件と無しの条件を主観的に比較した。

3.2 結果

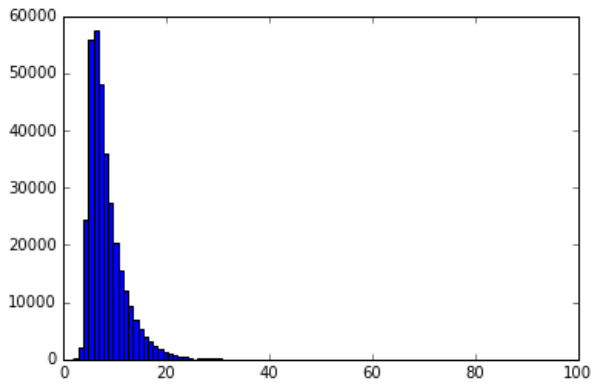
3.2.1 データの分布

以下にデータ拡張の結果の報告をする。データ拡張を行った結果、165,739 文のデータ (分散=264.81, 平均=17.04) は 381,762 文になり、分散は 15.23、形態素数の平均は 8.11 となった。以下にデータ拡張を行っ

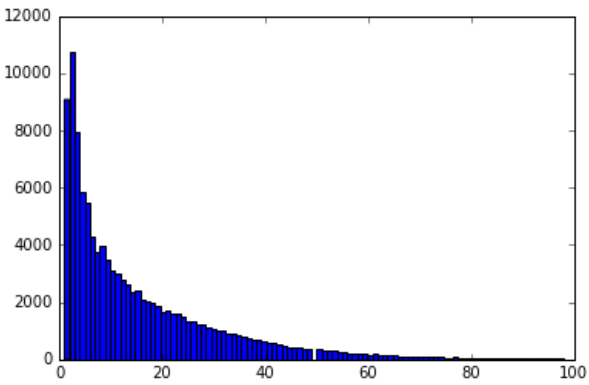
[9] おそらく離散のカウントデータであるため、ポアソン分布に従うと考えられる。
[10] https://taku910.github.io/cabocha/

た文の長さの分布と、データ拡張を行っていない文の長さの分布を示す。図で y 軸の値が異なることを留意されたいが、図に示されている通り文の単語数の分散は低くなっている。

Figure: x 軸を 1 文の形態素の数とし、y 軸を出現数として示す。



(a) データ拡張処理あり



(b) データ拡張処理なし (再掲)

3.2.2 学習の結果

学習されたモデルの評価の参考に、複数の単語に対してコサイン類似度から推定した近いベクトルをもつ単語を以下に示す。一つの単語に対して 2 つの列があり、左の列がデータ拡張した学習データに対する結果であり、右の列がデータ拡張を行っていない学習データに対する結果である。単語の選択が恣意的であることは確かであるが、もともとのデータを拡張した文は予測の当てはまりが良いように思える [11]。

表 2: 単語に対する学習結果（左：データ拡張有, 右：無）

言語		将棋		九州		漫画	
形態	文法	オセロ	パニック	東北	方面	ゴラク	マンガ
音韻	体系	ワーキン	遊戯	北陸	市内	タッチ	執筆
共通	共通	囲碁	湘南	中部	四国	貸本	作家
文法	HTML	てきぱき	テリー	中央	北陸	ギャグ	投稿
生物	定義	永田	α	四国	地区	少女	劇画

4 『統語・意味解析情報付き日本語コーパス』活用の展望

最後に本発表の本題に移る。本題は、NPCMJ (統語・意味解析情報付き日本語コーパス) を用いて、データ拡張をどう改善できるかという点である。 NPCMJ は公開中のコーパスで、文中の語句間の完全な統語的・意味的リンク付けを行うことにより，文法情報の根幹をなす依存関係 (dependency) の抽出を可能にする (アラステア・バトラー et al. 2015)。このコーパスに記述された情報を使用すれば、データ拡張の精度の向上が期待できる。

以下に wikipedia から「言語」を説明した一文を引用する。引用した文にデータ拡張を施した例と、コーパスの情報を活用した場合の期待するデータ拡張の例を下に挙げる。

[11] 今回はデータの評価に関して深く検討できなかったので、今後の課題とさせていただきます。

> 言語は、人間が用いる意志伝達手段であり、社会集団内で形成習得され、意志を相互に伝達することコミュニケーションや、抽象的な思考を可能にし、結果として人間の社会的活動や文化的活動を支えている。

上に引用した文から生成されるデータは、以下のものである。引用した文のトップのレベルを切り取り、長かった文をある程度は短い文のデータの集合に分解している。

> 言語は、意志伝達手段であり、形成習得され、可能にし、結果として文化的活動を支えている。
> 人間が用いる。 用いる意志伝達手段であり、。 社会集団内で形成習得され、。意志を相互に伝達する。 伝達することコミュニケーションや、。抽象的な思考を。 人間の社会的活動や文化的活動を。ことコミュニケーションや、思考を可能にし、。

しかしながら、文の体をなしていないものが多く、ここに NPCMJ による改善点が見込まれる。例えば、仮に意味のコントロール関係を考慮したデータが得られれば、より分解された文の意味関係を明確にすることができる。元の文では `window = 8` というパラメターを用いても「言語」という入力からは「文化的活動」や「抽象的な思考」というコンテキストは予想できない。一方で、下のようにデータ拡張が行えれば、「言語」が示す内容は `window = 8` というパラメターで予測する範囲内に分布する。

> 言語は、意志伝達手段であり、形成習得され、可能にし、結果として文化的活動を支えている。
> 意志伝達手段を人間が用いる。 言語は社会集団内で形成習得され、。言語は意志を相互に伝達する。 伝達することコミュニケーションや、。思考は抽象的な。 人間の社会的活動や文化的活動を。言語はことコミュニケーションや、思考を可能にし、。

上の様なデータ拡張を可能にするために、コーパスを用いたアプローチは主に 2 つ考えられる。 1 つは単純に、既に公開されているデータの情報を利用して文を分解してデータ拡張を行うことである。 NPCMJ に与えられた統語的・意味的情報を付けを用いれば、より精度の高い文の再構築が期待できる。そして 3 つのモデル、つまりデータ拡張のないデータに基づき学習させたモデル、CaboCha による構文解析によるデータ拡張に基づき学習させたモデル、NPCMJ の情報によるデータ拡張に基づき学習させたモデルを比較することで、手法の有効性を測ることができる。問題としては、新規のデータに対応することができない点が挙げられる。もう 1 つは現段階では課題が多いが、新規のデータに対応するために平文から NPCMJ のデータを予測するモデルを立て、モデルから予測された結果をデータ拡張に反映させることである。

5 まとめ

本発表 [12] では構文解析情報を利用した単語の分散表現獲得の手法と結果を始めに報告した。 CaboCha により係り受けの情報を得て文を細かくすることで、文の長さの分散と平均を下げ、より word2vec の学習のパラメターを考慮した形にデータを整形し、データ拡張を行った。その結果、主観にしか基づいていない点があるが、学習されたモデルは改善されたようであった。そして NPCMJ を用いてデータ拡張を行う今後の展望、また課題として 2 つのアプローチ、直接情報を使用するアプローチと、平文から NPCMJ の情報を予測し、その情報をを利用するアプローチを述べた。

[12] 本稿は東北大学で行った発表に修正を加えたものです。余分に話しすぎた word2vec の理論を削り、話の根本にあるデータ拡張に関して追記いたしました。

参考文献

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. 2013. “Exploiting Similarities Among Languages for Machine Translation.” *CoRR* abs/1309.4168. http://arxiv.org/abs/1309.4168.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. “Distributed Representations of Words and Phrases and Their Compositionality.” In *Advances in Neural Information Processing Systems 26*, edited by C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, 3111–9. Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

アラステア・バトラー, 吉本啓, 岸本秀樹, and プラシャント・パルデシ. 2015. “統語・意味解析情報付き日本語コーパスの開発.” 言語処理学会第 *21* 回年次大会 発表論文集.

加藤和平, 大島考範, and 二宮崇. 2015. “Word2vec と深層学習を用いた大規模評判分析.” 言語処理学会第 *21* 回年次大会 発表論文集.

拓, and 松本 裕治. 2002. “チャンキングの段階適用による日本語係り受け解析” 43 (6): 1834–42.

斎藤康毅. 2016. ゼロから作る *Deep Learning* ——*Python* で学ぶディープラーニングの理論と実装. オライリージャパン.

正幸, and 岡 照晃. 2017. “Nwjc2vec:『国語研日本語ウェブコーパス』に基づく単語の分散表現データ.” 言語処理学会 第 *23* 回年次大会発表論文集.

石田基広. 2008. *R* によるテキストマイニング入門. 森北出版.

黒崎優太, and 高木友博. 2015. “Word2Vec を用いた顔文字の感情分類.” 言語処理学会 第 *21* 回年次大会発表論文集.