PRESS RELEASE

報道関係者各位

2019 年 3 月 22 日
株式会社ワークスアプリケーションズ

国語研とワークスの産学共同研究成果＜第一弾＞「複数粒度の分割結果に基づく日本語単語分散表現」モデルを公開

～国内最大規模の日本語データを複数の単語単位で同時学習―実用的なモデルを構築～

株式会社ワークスアプリケーションズ（本社：東京都港区、代表取締役最高経営責任者：牧野正幸、以下 ワークス）の AI 研究機関であるワークス徳島人工知能 NLP 研究所は、大学共同利用機関法人人間文化研究機構国立国語研究所（以下 国語研）と産官の共同研究を実施し、国語研が保持する国内最大規模の日本語データベース「国語研日本語ウェブコーパス（NWJC）」[i]と、ワークス徳島人工知能 NLP 研究所の形態素解析器「Sudachi」[ii]を用いて学習した、実用的な単語分散表現モデルを新たに開発いたしました。このたび、第一弾では「複数粒度の分割結果に基づく日本語単語分散表現」モデルを商用利用可能なオープンデータとして無償公開[iii]いたしましたのでお知らせします。

大学共同利用機関法人 人間文化研究機構
国立国語研究所
National Institute for Japanese Language and Linguistics
NINJAL

WORKS APPLICATIONS
徳島人工知能ＮＬＰ研究所

本単語分散表現モデルを活用することで、コンピュータによる日本語の処理能力を向上させ、企業内に眠る様々なデータの解析、活用を促進します。更には、言語資源として広く公開することで、研究機関や技術者が手軽に高度な言語処理を実現できるようになるため、自然言語処理研究の推進に貢献できると考えます。

国立国語研究所コーパス開発センターの前川喜久雄センター長は、次のように述べています。
「国立国語研究所とワークスアプリケーションズとの共同研究により、国語研日本語ウェブコーパスに基づく新しい語彙資源が整備されました。国語研短単位がカバーしていないより長い単位の語を含む分散表現のオープンデータが学術界・産業界で活用され、言語学・言語処理研究の一助となることを大いに期待します。」

ワークスはマーケットリーダーとして、このような技術還元を通じて、企業のデジタルトランスフォーメーションの実現を支援してまいります。

研究成果

特長

- **国内最大 258 億語規模のコーパスにて学習を実施**
- **人名や地名、ブランド名、企業名、サービス名等の固有表現の語を大量に増強**
- **語の内部構造を考慮して類似度や相関度を学習することで高性能化を実現**

■本件に関するお問い合わせ先
TEL：03-6229-1210　FAX：03-6229-1211　Email：pr@worksap.co.jp
株式会社ワークスアプリケーションズ
広報（担当：金田、山際）

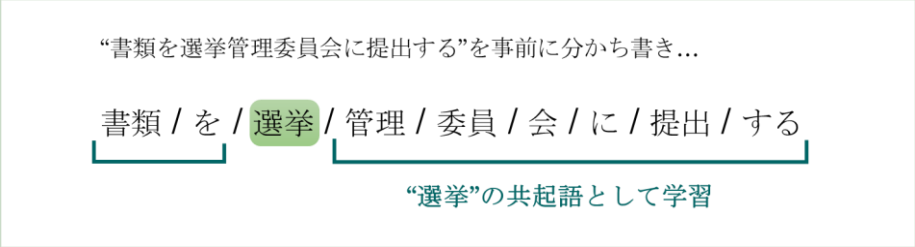
既存の日本語単語分散表現では固有表現のような長い単位の収録数が少なく、また語の内部構造を考慮できていない。

「Sudachi」を用いて複数の粒度で分割したコーパスを同時に学習することで、固有表現の収録数を大幅に拡大。加えて、内部の構造語との類似性を計算することで、単語分散表現モデルの実用性を向上させた。

参考：従来の単語分散表現における学習法および課題

0. 従来の単一的な学習

語を分割し、語の同時出現率（共起[iv]）を機械学習の技術を使って学習する。



この際、既存の日本語単語分散表現の問題点に、以下の 2 点が挙げられる。

ⅰ．固有表現や複合語のような長い語が登録されておらず、細分化されてしまい、全体の精度に影響を及ぼす。意味を持つ長い語を認識できることは応用時に有効となる。

例）
国立 / 競技 / 場、　イタリア / 料理 、　南 / アメリカ / 州

ⅱ．複数の粒度で分割したデータをもとに学習しておらず、長い語の内部にある単語同士の関連性が無視される。

例）
“ビール” ⟷ “クラフトビール”、　ウェアラブル” ⟷ “ウェアラブルデバイス”
“シャーロック” ⟷ “シャーロック・ホームズ”、　“デザイナー” ⟷ “グラフィックデザイナー”
“ベナン” ⟷ “ベナン共和国”、　“ワークス” ⟷ “ワークスアプリケーションズ”

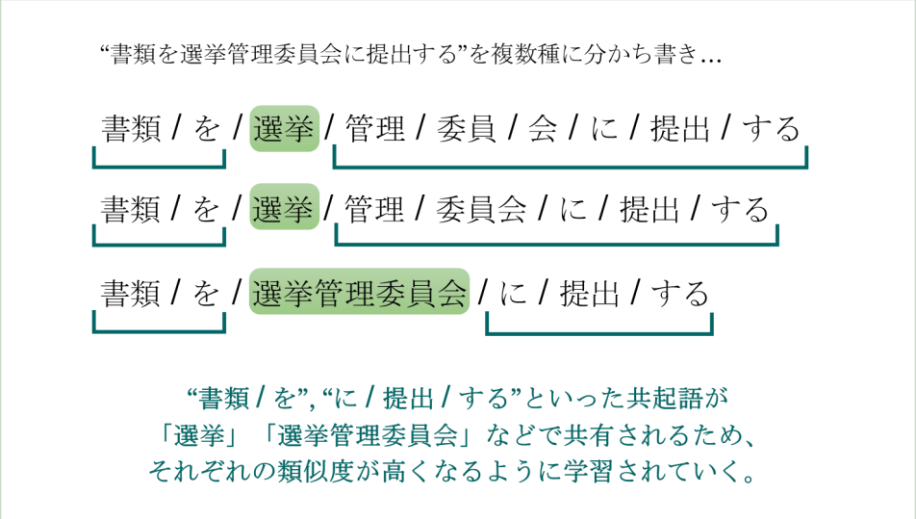
研究内容

1. 複数の単語単位で同時に学習

「Sudachi」の長・中・短単位の各分割モードで語を複数パターンに分割し、それぞれのパターンにおいて語の同時出現率（共起）を同時に計算して学習する。

■本件に関するお問い合わせ先
TEL：03-6229-1210　FAX：03-6229-1211　Email：pr@worksap.co.jp
株式会社ワークスアプリケーションズ
広報（担当：金田、山際）

学習データとして各単位の分割を同時に考慮することで、注目する語の周辺に分布する語が各分割単位で共有され類似度が高くなりやすい。



2. 比較実験・分析の結果

ⅰ. 単語間類似度（jwsan-1400）[v]

	類似度	関連度
単語ベクトルA	47.46	62.43
単語ベクトルB	52.85	61.55
単語ベクトルC	**60.09**	**66.03**
NWJC + Sudachi（短・中・長単位すべて）	**53.85**	**66.13**

ⅱ. 文書分類（livedoor-news）[vi]

	精度
単語ベクトルA	0.820 ± 0.0013
単語ベクトルB	**0.843 ± 0.0010**
単語ベクトルC	0.810 ± 0.0009
NWJC + Sudachi（短・中・長単位すべて）	**0.838 ± 0.0012**

リソースの公開先

「複数粒度の分割結果に基づく日本語単語分散表現」モデルは、以下の URL にて公開しています。

ワークス徳島人工知能 NLP 研究所 特設ページ
https://www.worksap.co.jp/nlp-activity/word-vector/

解説

「単語の分散表現」（単語のベクトル表現ともいう）とは、単語の周辺文脈から単語間の関連性や類似性を機械学習して高次元のベクトルとして表現（数値化）したものです。この単語分散表現モデルは、深層学習等の技術においてコンピュータが日本語を意味解析・意味理解する上で欠かせない自然言語処理技術の基礎技術です。

日本語の単語分散表現モデルは、「単語の区切り」が明示されないといった特殊性や、学習データの不足等が課題となり、実用化にいたる単語分散表現モデルの研究は英語などの言語に比較して後れ

■本件に関するお問い合わせ先
TEL：03-6229-1210　FAX：03-6229-1211　Email：pr@worksap.co.jp
株式会社ワークスアプリケーションズ
広報（担当：金田、山際）

をとっていました。

このたびの国語研とワークス徳島人工知能 NLP 研究所の共同研究により、国語研が有する国内最大規模の「国語研日本語ウェブコーパス」と、単語を複数の単位で分割可能な「Sudachi」を活用することで、高精度な単語分散表現モデルの構築を行っています。

株式会社ワークスアプリケーションズ Web サイト https://www.worksap.co.jp/

* 会社名、製品名等はそれぞれ各社の商標または登録商標です。
* 本リリースに掲載された内容は発表日現在のものであり、予告なく変更または撤回される場合があります。また、本リリースに掲載された予測や将来の見通し等に関する情報は不確実なものであり、実際に生じる結果と異なる場合がありますので、予めご了承ください。

---

[i] 「国語研日本語ウェブコーパス」とは、ウェブ上の日本語テキストから 100 億語を超える規模のサンプルを収集することで稀言語現象の言語学的、心理学的および情報処理的視点からの究明の可能性を開くことを目的に構築されています。

[ii] 形態素解析とは、テキストを語に分割し各種情報を付与する技術です。「Sudachi」は、ワークス徳島人工知能 NLP 研究所が開発した形態素解析ツールです。

[iii] ライセンス：Copyright (c) 2019 National Institute for Japanese Language and Linguistics, and Works Applications Co., Ltd. All rights reserved.
Apache License, Version 2.0 のライセンスの下で国立国語研究所と株式会社ワークスアプリケーションズによって提供されています。

[iv] 「共起」とは、自然言語処理の分野において、任意の文書や文の中に、ある文字列とある文字列が同時に出現することを指します。

[v] スピアマンの順位相関係数による評価. jwsan-1400 内のエントリは短単位語からなります。

[vi] 9 クラス分類のタスク 10 分割交差検証の結果、名詞の形態素のみ対象に分散表現の平均和を文書の特徴量とし, ロジスティク回帰による分類をしたものです。

■本件に関するお問い合わせ先
TEL：03-6229-1210　FAX：03-6229-1211　Email：pr@worksap.co.jp
株式会社ワークスアプリケーションズ
広報（担当：金田、山際）