お知らせ
2019年4月18日
ヤフー株式会社

分散表現の学習時間を短縮化するAI技術「yskip」をOSSとして公開

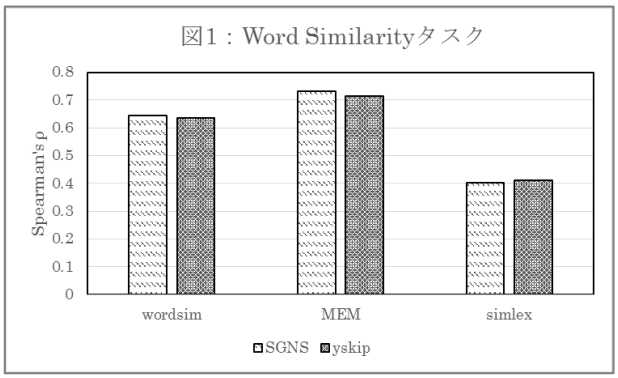
～ 増加し続けるビッグデータを追加分のみ機械学習し、学習効率を高める ～

ヤフー株式会社（以下、Yahoo! JAPAN）は、AI・自然言語処理技術の一つである分散表現の学習時間を短縮化する技術「yskip」を、4月18日よりオープンソースソフトウェア（OSS）として公開しました。

分散表現とは、単語をベクトルで表現する自然言語処理領域のAI技術の一つです。この技術を用いると、大量のテキストデータからさまざまな単語の関係性を機械学習し、単語同士の意味の相違を機械的に推定できます。Yahoo! JAPANでは、ユーザーの興味関心情報と記事や広告のマッチングの裏側の技術として活用しています。
分散表現は、より大量のテキストデータを用いて学習することで、精度向上が期待されます。日々増加し、また新たなトレンドも生まれてくる「検索キーワード」や「SNSのつぶやき」など、インターネットサービス上のテキストデータを活用する場合は、分散表現の学習モデルを頻繁に更新することが求められます。その際には、新しいデータだけでなく、既に学習したデータもあわせて一から学習しなおす必要があり、その都度学習時間がかかるなど非効率でした。

このような課題を受け、新しいデータのみの学習で、全データで学習する場合と比べ学習時間を短縮化しつつ同等精度を維持する分散表現技術「yskip」を開発し、4月18日よりGitHub上にOSSとして公開しました。Yahoo! JAPANでは、Twitterに投稿されたつぶやきを検索できる「リアルタイム検索」の裏側で「yskip」を用い、サービスの質の向上に役立てています。OSSとして公開することで、今後は広くAIエンジニア、研究者の方々にご利用いただき、サービス開発や研究開発の効率化にご活用いただけます。
なお「yskip」は、代表的な分散表現学習法であるskip-gram model with negative sampling（以下、SGNS）を拡張した技術です。

「yskip」と従来の学習法SGNSを、分散表現の精度を測定するために使われている5種のデータセットで検証したところ、同等精度で学習可能であることが実証されました（図1・2）。



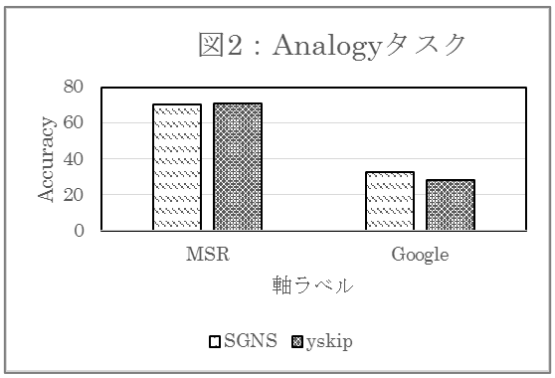


図1・図2：Word Similarityタスクのデータセット3種類（WordSim353，MEN，SimLex999）、Analogyタスクのデータセット2種類（GoogleデータとMSRデータ）を用いた結果。

詳細については、2017年9月に開催された、自然言語処理領域で権威のある国際会議（トップカンファレンス）「EMNLP2017」にて、論文として発表しています。また、開発者による技術解説記事を、Yahoo! JAPAN Tech Blogでも公開しています。

本技術は、導入後の特許侵害の発生リスクをおさえ、安心して利用いただくため、特許権を取得しています。研究用途だけでなく商業用途も含め、その特許権のライセンスを無償提供する形で、OSSとして公開しました。OSS公開を通じて、本技術のさらなる利便性向上を図り、データサイエンス領域の研究者・エンジニアコミュニティへ貢献したい考えです。

Yahoo! JAPANは今後も、強みである「ビッグデータ」をいかした先端研究・開発の推進を通じて、一人ひとりのユーザーにあった情報を提案し、最上級のおもてなしを実現するサービス提供を目指してまいります。

■GitHub 内の「yskip」公開ページURL
https://github.com/yahoojapan/yskip

■開発者による技術解説（Yahoo! JAPAN Tech Blog）
https://techblog.yahoo.co.jp/oss/yskip/

■論文の参照先URL
http://aclweb.org/anthology/D17-1037