PRESS RELEASE

報道関係者各位

2020年10月9日
株式会社ワークスアプリケーションズ

国内最大規模の日本語言語処理資源「SudachiDict」および「chiVe」をOpen Data on AWSで公開開始

ー自然言語処理技術で日本語の曖昧さを吸収し、さらに便利でオープンなシステムへー

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：井上直樹、以下ワークス）は、この度、2020年10月9日(金)より、ワークス徳島人工知能NLP研究所（以下、同研究所）が開発した国内最大規模の日本語言語処理資源「SudachiDict」および「chiVe」をOpen Data on AWSにて公開を開始したことをお知らせいたします。
以前より本製品を商用利用可能なライセンスにて無償公開していましたが、大規模な言語資源であるためデータサイズが非常に大きく、取り扱いが難しいというご意見をいただいておりました。今回、Open Date on AWS上で公開することで、さらに使いやすく、オープンにご利用いただけるようになります。

公開先URL：https://registry.opendata.aws/sudachi/

（徳島県ご当地キャラとコラボレーションした「AIすだちくん」）

●「SudachiDict」とは

「SudachiDict」は専門家の手によりメンテナンスされた290万語以上の語彙を収録する高品質な日本語自然言語処理のための辞書です。
SyncThought社、リクルート社など複数企業の製品サービスで「Sudachi Dict」は活用されています。

特徴

(1) １つの語に対する複数の分割情報を付与
　日本語処理に必要な語の区切りは必ずしも一意ではありません。「SudachiDict」ではさまざまな利用シーンにあうよう3種類の区切りを用意しているため、用途に応じて区切り方を選べます。

A単位	ワークス/徳島/人工/知能/NLP/研究/所
B単位	ワークス/徳島/人工/知能/NLP/研究所
C単位	ワークス徳島人工知能NLP研究所

(2) すべての語彙に表記正規化情報を付与
　日本語では同じ語がさまざまな表記で書かれることがあります。これらの表記を正規化することにより同一のものとして統一的に扱うことができます。

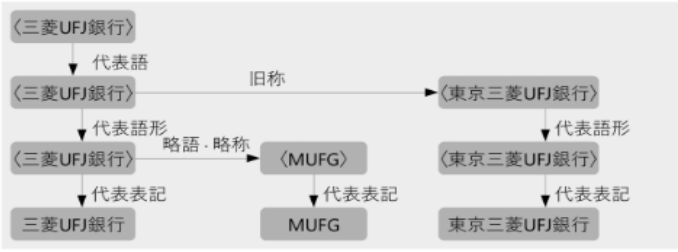
パターン	例
文字種の違い	向日葵−ひまわり−ヒマワリ
漢字の違い（異体字、代用表記、慣用表記）	芸術−藝術、驚歎−驚嘆、徳用−得用
送り仮名の違い	受け付け−受付け−受付
外来語の表記違い	コミュニティー−コミュニティ−コミニティー−communitiy
誤用	シミュレーション−シュミレーション
くだけた言い方	～ちゃあ − ～ては

(3) 約60,000語に同義関係を詳細化した同義語情報を付与
　SudachiDictでは収録語に同義語の情報を付与しており、全文検索を始めさまざまな用途に利用できます。また同義語の関係を精密に記述するために階層化された詳細な同義関係を導入しています。

■本件に関するお問い合わせ先
TEL：03-6229-1200　FAX：03-6229-1201　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：池内

関係	例
同義語彙素	支払い/勘定
旧称	三菱東京UFJ銀行/三菱UFJ銀行
対訳	日本/Japan
別称	社会保障･税番号制度/マイナンバー制度
誤用	おもむろ/突然
略語･略称	流行性感冒/流感
翻字	インフルエンザ/influenza
異表記	子供/子ども
誤表記	シミュレーション/シュミレーション

例）三菱UFJ銀行



(4) 継続的な語彙の拡充・整備

言葉は日々変化しています。実用的な日本語処理のためには新語や言葉の新しい使われ方に追随する必要があります。SudachiDictでは継続的に語彙の拡充・整備を続け、常に最新の辞書を提供していきます。

●「chiVe（チャイブ）」とは

chiVeと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が保持する「国語研日本語ウェブコーパス（NWJC）」をSudachiDictを用いて解析させ、word2vec（自然言語処理技術の一つで大量のテキストデータを解析し、各単語の意味をベクトル表現（分散表現）する手法）により学習した、大規模な単語分散表現リソースです。

特徴

(1) 国内最大258億語規模のコーパスにて学習を実施

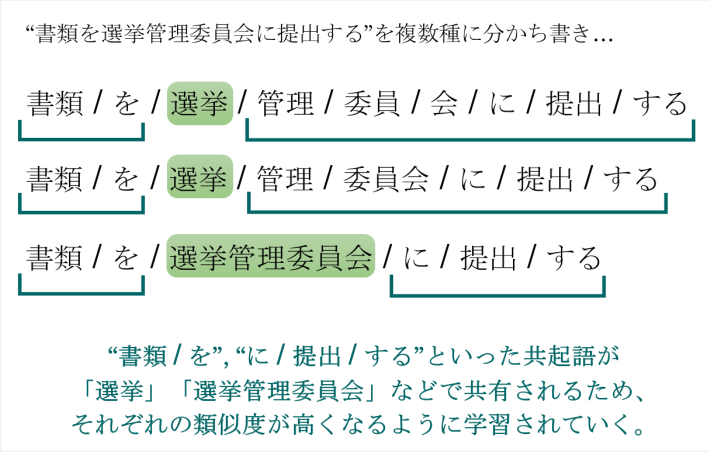
国立国語研究所による超大規模なコーパス「NWJC」を利用して学習しています。これは、ウェブ上の様々な情報源から作成された日本語のテキストデータセットです。分散表現の学習においてはデータ量が重要なファクターとなることが知られています。今回この超大規模コーパスを利用することにより、小・中規模なデータによるリソースに比べて、更に有益なものになることが想定されます。

(2) 人名や地名、ブランド名、企業名などの固有表現を大量に追加

「SudachiDict」は290万以上の語彙を含み、その中には多くの新語も存在します。この類をみない高品質で大規模な辞書を利用することで、既存の辞書ではカバーできなかった幅広い固有表現な複合語に対する分散表現を学習できます。

(3) 複数分割情報を活用することにより、語の内部構造を考慮した高性能化を実現

「SudachiDict」の分割情報による解析結果により、同じ文を違った粒度で出力できます。この結果を活用し、同じテキストから獲得した複数の単位での語とその文脈を学習時に入力として利用しました。これにより、固有表現や複合語といった長い表現と、その内部の語との類似度が高くなるといった傾向のある分散表現の学習が可能です。



■本件に関するお問い合わせ先
TEL：03-6229-1200　FAX：03-6229-1201　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：池内

●「SudachiDict」および「chiVe」の活用方法

「SudachiDict」は、同研究所がオープンソースソフトウェアで公開する日本語形態素解析エンジン「Sudachi※1」と「SucdachiPy※2」を利用することで、辞書内に付与されている様々な情報を効率的に利用できます※3。また、全文検索エンジン「Elasticsearch」から利用可能なプラグインも公開しています※4。

「chiVe」と「SudachiPy」は、多言語対応自然言語処理フレームワークである「spaCy※5」、日本語自然言語処理オープンソースライブラリ「GiNZA※6」からも利用可能です。

豊富な語彙を収録する「SudachiDict」および「chiVe」を活用することによりコンピュータによる日本語の処理を向上させ、企業内に眠る様々なデータの解析、活用の実現を促進します。

●Open Data on AWSとは

Open Data on AWSは、公益価値のある一般公開用データをAWSがホストするスポンサーシップ・プログラムです。認可されたデータセットは2年間、AWSでホスティングされ、その保存とデータ転送の費用はAWSが負担します。AWS上でデータを一般公開することで、データを多くの人々に共有でき、AWS内外からの利便性が高まります。また、登録されている全てのデータセットは、レジストリで管理され、一覧・検索することができ、より多くの人が情報を発見することが期待されます。
詳細URL：https://aws.amazon.com/jp/opendata/

なお、Amazon Web Services ブログにおいて、今回Open Date on AWS上で公開する「SudachiDict」および「chiVe」についてご紹介いただいています。

●ワークス徳島人工知能NLP研究所

2017年2月にワークスが徳島県に開設した自然言語処理（NLP、 Natural Language Processing）に特化した研究機関です。

ワークスが開発するERPパッケージソフト「HUE」は企業内に蓄積されるオペレーションログを機械学習のトレーニングデータとして活用しています。

本研究所では「HUE」に蓄積されるオペレーションログをより有効活用し、よりユーザーニーズに即したAI機能を実用化するために自然言語処理を活用した研究開発を進めています。

◆ワークス徳島人工知能NLP研究所では共に働く仲間を募集しています。
詳細URL:https://job.axol.jp/vb/c/worksap/job/detail/cGrkDH85Yx0o-

※1 https://github.com/WorksApplications/Sudachi
※2 https://github.com/WorksApplications/SudachiPy
※3 https://github.com/WorksApplications/Sudachi/blob/develop/README.md
※4 https://github.com/WorksApplications/elasticsearch-sudachi
※5 https://spacy.io/
※6 https://megagonlabs.github.io/ginza/

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

■本件に関するお問い合わせ先
TEL：03-6229-1200　FAX：03-6229-1201　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：池内