News Release

2018 年 11 月 6 日
TIS 株式会社

TIS、自然言語処理・機械学習向け データ作成ツール「doccano」を OSS で公開
～　自然言語処理・機械学習を行うためのデータ作成を容易にし、企業システムにおける活用を促す　～

ＴＩＳインテックグループの TIS 株式会社(本社:東京都新宿区､代表取締役会長兼社長:桑野　徹､以下 TIS）は、自然言語処理・機械学習向けのデータ作成ツール(アノテーションツール)「doccano（ドッカーノ)」をオープンソースソフトウェア（OSS）として公開することを発表します。

「doccano」公開ページ：
https://github.com/chakki-works/doccano

「doccano」は、自然言語処理・機械学習に使われるラベル付きデータ（教師データ）の作成を容易にするツール(アノテーションツール)です。

テキスト分類、系列ラベリング※1、系列変換※2 という 3 つの基本的なタスクで使用するデータを作成することができます。セットアップが容易であり、英語以外に日本語にも対応しています。

※1：系列ラベリングは、文中の人名や地名を特定するといったタスクのこと
※2：系列変換は、要約/翻訳といったタスクのこと

＜「doccano」の利用画面イメージ＞

＜「doccano」の活用例＞

「doccano」を利用することで、作成に特に手間がかかる系列ラベリングのデータを簡単に作成

できます。テキスト分類や系列変換は、Excel などの帳票ツールでも作成が可能ですが、系列ラベリングでは文字/単語単位でデータを作る必要があるため、帳票ツールのみでは作成が困難です。系列ラベリングに「doccano」を活用すれば、対象の単語を選択し、ボタン、またはショートカットキーを押すだけでラベル付けが可能です。

これまで手間だった自然言語処理・機械学習向けのデータ作成を容易にすることで、機械学習・自然言語処理を適用できる業務を拡大できます。

TIS は「doccano」を OSS として公開し、より多くのフィードバックを得ることで機能を改善し、データ作成業務を効率化することで機械学習・自然言語処理の市場の拡大を目指します。

■「doccano」公開の背景

機械学習・自然言語処理の研究・開発を行うためには教師データが欠かせません。教師データとは問題と解答をセットにしたデータであり、機械学習モデルに与えることで正しい答えを学習させることができます。しかし、教師データの作成には非常に手間がかかるという課題があります。TIS が公開した機械学習で感情解析を行うためのデータセット「chABSA-dataset」(チャブサ・データセット) においても、作成には多くの手間がかかりました。そこで、その経験を元に今回「doccano」を開発し OSS として公開しました。

「doccano」を利用することで、機械学習・自然言語処理に用いるデータの作成が容易になります。一方で、ラベルの定義を明確にするといった、データ作成における本質的な難しさのサポートにはまだ改善の余地があります(これは感情解析であれば、どんな場合をネガティブ・ポジティブと判断するかなどです)。TIS では「doccano」を OSS として公開し、より多くのフィードバックを得ることでツールの改善に活かしていきます。

■TIS の自然言語処理・機械学習への取り組みと今後の展開

データが増え続ける中で重要なデータの見逃しは許されない、といったビジネス課題を解決すべく、TIS では「観点要約」の研究開発を進めています。

「観点要約」は、単純な要約ではなく、情報の見るべきポイント(観点)を指定することで、情報の粒度をそろえて提示する、また同じ観点で異なる時期に書かれた文書、また別々の人によって書かれた文書などを比較する、といったことを可能にします。

これまで TIS では「観点要約」の研究開発の一環として「chABSA-dataset」を公開し、現在はテキストと数値とが混在した会計文書を対象に適用検証を進めています。

TIS では、こうした研究開発活動をオープンな姿勢で行っています。「doccano」を含めた自然言語処理の研究に関連して開発したソフトウェアは、以下のページで OSS として公開しています。

「doccano」公開ページ：
https://github.com/chakki-works/doccano

■「chABSA-dataset」について

「chABSA-dataset」は上場企業の有価証券報告書(2016 年度)をベースに作成されたデータセットです。各文に対してネガティブ・ポジティブの感情分類だけでなく、「何が」ネガティブ・ポジティブなのかという観点を表す情報が含まれています。こうした観点単位の感情分類を機械学習

モデルに学習させることで、より高度な感情解析が実現できます。
「chABSA-dataset」の詳細については、データセット公開ページ、また付随する論文をご参照ください。

「chABSA-dataset」データセット公開ページ：
https://github.com/chakki-works/chABSA-dataset

ＴＩＳ株式会社について

ＴＩＳインテックグループの TIS は、SI・受託開発に加え、データセンターやクラウドなどサービス型の IT ソリューションを多数用意しています。同時に、中国・ASEAN 地域を中心としたグローバルサポート体制も整え、金融、製造、流通/サービス、公共、通信など様々な業界で 3000 社以上のビジネスパートナーとして、お客様の事業の成長に貢献しています。詳細は以下をご参照ください。http://www.tis.co.jp/

ＴＩＳインテックグループについて

ＴＩＳインテックグループはグループ会社約 60 社、2 万人が一体となって、それぞれの強みを活かし、日本国内および海外の金融・製造・サービス・公共など多くのお客さまのビジネスを支えるＩＴサービスをご提供します。

※ 記載されている会社名、製品名は、各社の登録商標または商標です。
※ 記載されている情報は、発表日現在のものです。最新の情報とは異なる場合がありますのでご了承ください。

【本件に関するお問合わせ先】

◆報道関係からのお問い合わせ先
TIS 株式会社　企画本部　コーポレートコミュニケーション部　浄土寺/橋田
TEL：03-5337-4232　E-mail：tis_pr@ml.tis.co.jp

◆本件に関するお問い合わせ先
TIS 株式会社　戦略技術センター　担当：久保/油谷
TEL：03-5909-4501　E-mail：info-stc@ml.tis.co.jp