News Release

TIS INTEC Group | Go Beyond

2018 年 4 月 10 日
TIS 株式会社

TIS、機械学習で感情解析を行うための データセット「chABSA-dataset」を無償公開

～　観点を指示した自然言語処理により、文章の要約作成や図表化を目指す　～

ＴＩＳインテックグループの TIS 株式会社（本社：東京都新宿区、代表取締役社長：桑野 徹、以下 TIS）は、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）を、無償公開することを発表します。

「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、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、より高度な解析が実現できます。

＜「chABSA-dataset」に収録されているデータのイメージ＞

通常の感情解析用のデータセット

商品Aの売上が上がった。	○
商品Bについては、コストが上がった。	×
・ ・ ・	

chABSA-dataset(観点単位のデータセット)

商品Aの売上が上がった。	商品A#売上	○
商品Bについては、コストが上がった。	商品B#コスト	×
・ ・ ・	「何が」の情報も付与	

「chABSA-dataset」を利用した感情解析では、例えば、「商品 A の売上が上がった」という文について単にポジティブ、というだけでなく、「商品 A」の「売上」が「上がった」(=ポジティブ)である、ということが判断できます。こうした解析結果は、以下のように表としてまとめることが可能となります。

＜観点単位の感情分類結果を表にした場合のイメージ＞

	売上	利益	数量	価格	コスト
商品A					
商品B					
商品C					
商品D					

この表では、緑の色が濃いほどポジティブ、灰色の色が濃いほどネガティブであることを示しています。図中では、「商品 A」の「売上」についてポジティブな表現がされ、「商品 B」の「コスト」についてはネガティブな表現がされている、といった解析結果をまとめたイメージになります。このように、「chABSA-dataset」を活用することで、機械学習による高度な解析が可能になり

ます。
　また、今回公開する「chABSA-dataset」は、上場企業の有価証券報告書をベースとしているため、機械学習による企業分析に活用することも可能です。

　「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

■「chABSA-dataset」公開の背景

　TIS では、機械学習・自然言語処理を用いた業務の生産性向上について研究・開発を行っています。その取り組みの一つとして、機械学習・自然言語処理を用いて観点に沿って情報をまとめる「観点要約」に取り組んでいます。
　「観点要約」とは、例えば議事録であれば決定事項や Todo といった特定の「観点」に沿い文書をまとめることです。文章から情報を抽出・要約する際には、まとめられた文書が“どれだけ短いか”という点より“必要な情報が抜けていないか”という点が重視されます。機械学習・自然言語処理によって、“指定されたポイントを押さえて情報をまとめる”ということを実現するには「観点要約」が欠かせない技術になります。

　今回公開した「chABSA-dataset」は、この「観点要約」の研究の一貫で作成されたものです。「chABSA-dataset」を利用することで、「何が」良い評価・悪い評価なのかを判断する機械学習モデルの開発が可能になります。こうしたモデルは、将来的にはマーケティングデータに対し“商品のどういった点が評価され、どういった点が不満に思われているのか”などの分析に役立ちます。また、各商品を同じ観点で評価することが可能になるため、商品間の評価の比較を行う際にも活用が期待できます。

　TIS では、同様の研究を行う研究者にも活用をしてもらい、その知見を交換することを目的に「chABSA-dataset」を無償公開します。

■TIS の自然言語処理・機械学習への取り組みと今後の展開

　TIS では、2017 年 4 月に AI・ロボット分野における専門組織「AI サービス事業部」を立ち上げ、機械学習・自然言語処理などを中心に AI に関する技術・知識と、長年のシステム構築・運用の実績で培った企業の業務プロセス・システムの理解を組み合せ、課題解決に向けた AI 活用の各種ソリューション・サービスを提供しています。

　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」では、ユーザーの指示する様々な「観点」を理解し、それに沿い文書をまとめることが必要になります。こうした柔軟な解析を実現するためには、自然言語処理における「転移学習」※1 が有力な技術であると TIS では考えています。
　「観点要約」以外でも、「転移学習」を用い少量のデータでカスタマイズ可能な自然言語処理の機能を今後開発していく予定です。

　今回「chABSA-dataset」を無償公開したように、TIS では研究開発活動をオープンな姿勢で行

っています。データセットだけでなく、自然言語処理の研究に際して開発したソフトウェアは、以下のページでオープンソースとして公開しています。

https://github.com/chakki-works

※1:「転移学習」とは、あるタスクを行うために学習させた機械学習モデルを、別のタスクを行えるよう少ないデータで「転移」させる技術です。「転移学習」は会社における社員の配置転換と似ています。配置転換は、ある部署で優秀な人材は、他の部署へ異動しても短い時間で適応し成果が出せることを期待します。同様に、機械学習モデルでも既にあるタスクで優秀なモデルであれば、別のタスクでも少ないデータで高い精度が出せるという期待が「転移学習」の背景にあります。
　様々な観点を機械学習モデルで一から学習するのは難しいですが、「転移学習」を利用することである観点を十分に学習させたモデル、あるいは別の自然言語のタスクで優秀なモデルを、少ないデータで指示された観点に「転移」させることが可能と考えています。

■「chABSA-dataset」の利用について

「chABSA-dataset」の利用を希望する方は、下記のページからダウンロードが可能です。
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