英文マイニングが拓く TMStudio の新たな活用ステージ

株式会社数理システム 岩本 圭介

1. はじめに

(株)数理システム のテキストマイニングツール Text Mining Studio (TMStudio) は、お蔭様で現在約 200 のサイトにおけるユーザ様によってご利用いただいている製品となりました。その利用の分野も大きな広がりを見せており、使い勝手 と 自由度の高い分析 とを両立させるという開発コンセプトを受け入れていただいた結果であると感謝しております。

そんな中で、非常に大きなご要望として頂いておりました TMStudio の 英語対応版 がこの度リリースの運びとなりました。本公演では、TMStudio の英語対応の話題を中心に、その他、新機能や TMStudio のアドオンモジュールについてご紹介させて頂きます。

2. Text Mining Studio 英語アドオン

TMStudio の英語対応版は『Text Mining Studio 英語アドオン』として提供いたします。TMStudio 本体に対するアドオンモジュールという位置付けであり、TMStudio が利用可能な環境上に追加でインストールすることにより、表 1 の追加機能が利用できるようになります。

表 1 英語アドオンによる追加機能

機能	内容
分かち書き	日本語と英語の選択が可能。 日本語選択時はこれまでと同様の挙動、英語選択時は 英語の文法に則った品詞の付与と係り受けの抽出を行う。
辞書	ユーザ辞書の利用により、連語 のまとめ上げが可能。
分析機能	英語の文法に則って生成された分かち書き結果をもとに、 TMStudio の全分析機能が利用可能。

英文テキストを取り込んで分かち書き結果を行った後は、これまでの TMStudio の操作と同様の感覚で英文テキストに対してマイニングを行っていただけます。

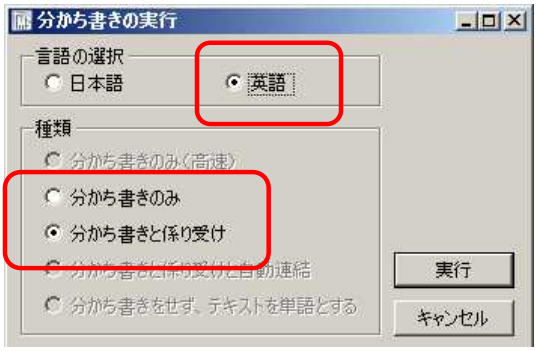

図 1 分かち書きの実行

英語アドオンをインストールすると、まず図 1 のように 分かち書き の段階で日本語・英語が選択可能になります。

英語の場合は、品詞を付与するのみの「分かち書き」モード、加えて構文を解析し係り受けを抽出する「分かち書きと係り受け」モードが選択できます。

このようにして分かち書き結果を作成したのちは、英文テキストを対象に TMStudio 上の各種機能をそのまま利用することができます。まず、係り受け頻度解析により係り受けを集計した結果を図 2 に示します。英文の各単語に対して「名詞」「動詞」といった品詞が正しく付与された上で、「何が・何を－どうした」といった単語間の意味の繋がりを、日本語での利用と同様に抽出することができます。

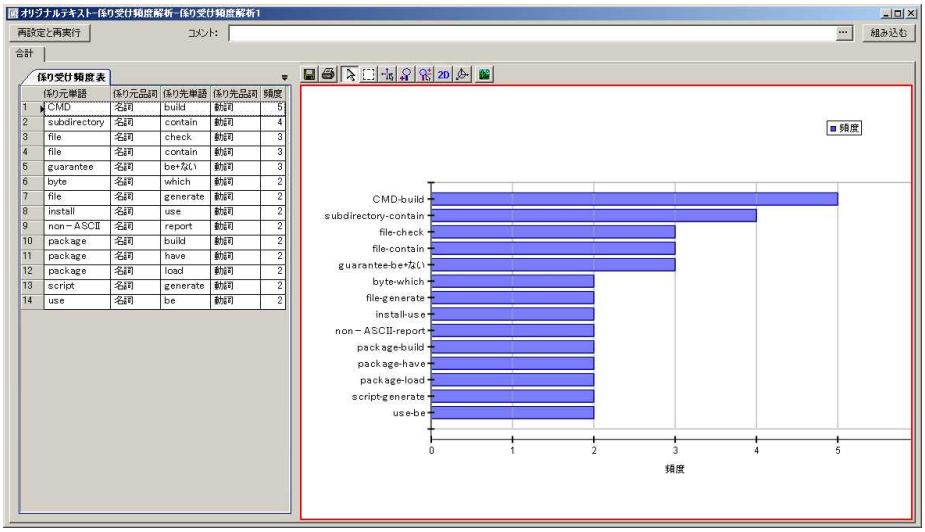

図 2 係り受け頻度解析

TMStudio の分かち書きでは『述語属性』として「否定・可能・不可能・要望・疑問・容易・困難・過度」といった、いわば「その語がどのようなニュアンスで発せられたか」といった情報を単語に付与します。これは、分析上大いに利用価値があり、これを用いて要望表現のみを取り出したい、また不満表現のみを抽出したい、といった抽出が非常に気軽に行えるようになっています。英語での分かち書き時には、現バージョンではまず「否定」の明示的な抽出にのみ対応いたします。ただ、no, never, not のような明示的な否定語を抽

出するのみではなく、文脈的に否定表現とみられる各種の言い回しもサポートします。

また、英文において考慮しなければならない 連語 については、図3 のようにユーザ辞書を用いてまとめ上げていただくことで、複数の語の連接をあたかも 1 単語であるかのように扱うことができます。

図 3 連語のまとめ上げ

日本語と英語とで、分析パラメタ や 各種辞書ファイルの形式は共通です。TMStudio の上で、日・英テキストマイニングの資産共有を図ることができます。

3. TMStudio3.2 エンジンの利用

日本語解析エンジンの方もアップグレードされます。TMStudio バージョン 3.2 では、日本語解析エンジンとして

- TMStudio3.1 エンジン
- TMStudio3.2 エンジン

の 2 種類が選択可能となっていました。英語アドオン リリースと同時期の更新において、3.2 エンジン が修正されました。より高精度の単語の抽出、また仮名と漢字の統一といった 3.2 エンジンのメリットを、3.1 エンジン利用ユーザ様がより違和感なく移行のうえ感じていただける形となっております。図 4 に 3.2 エンジンでの分かち書き結果例を示します。是非その威力を実感いただけたらと存じます。

図 4 TMStudio3.2 エンジンによる分かち書き

4. TextCutter

図 5 TextCutter

TextCutter は、テキストを話題（トピック）毎に自動的に分割する、TMStudio のアドオンツールです。

一件の意見の中で様々な話題について触れているようなテキストについては、TextCutter を用いて分析者に興味ある話題の部分のみを抽出することによって、テキストマイニングの精度を大幅に向上させることができます。

トピックの定義は、TMStudio の各種分析機能を用いて作成することができます。

TMStudio 英語アドオンにより、TextCutter も英文に対するトピック分割に対応します。英文入力テキストに対して、トピックの付与と分割を行った結果を図 5 に示します。

5. 今後の Text Mining Studio

Text Mining Studio 英語アドオン 及び 修正された TMStudio3.2 エンジンを利用可能にするための更新ツール は 10 月末にリリースされました。

今後の機能追加に関して、以下のような内容を予定しております。

- グルーピング機能 大幅刷新
  - テキストの意味的なまとめ上げを行うため、同じ意見カテゴリに属する
    単語 や 係り受け表現 のグループを作成する機能がグルーピングになります。
    グループ作成において、よりきめ細かい条件指定を可能とすると共に、
    単語・係り受けを超えて「キーとなる文章」を与えてまとめ上げを可能とします。
- 英語アドオン 機能強化
  - 「否定」以外の 述語属性 についても、
    英文の文脈的な内容を考慮の上付与し、利用可能とします。
- ASP サービスとしての形態を提供
  - Web 経由で TMStudio の機能を利用可能といたします。