InterSystems IRIS NLP Japanese

Version 2.0

InterSystems IRIS NLP Japanese

はじめに

InterSystems IRIS Natural Language Processing （NLP）は、非構造化データから本質的な情報を得ることによりお客様のソリューションの価値を高めていくことができる、インターシステムズ社独自のテキスト探索（text exploration）技術です。IRIS NLP のテクノロジーは「ボトムアップ」形式のテキスト解析により、文章そのものから意味のある「エンティティ」を見出していくため、辞書やオントロジーを予め定義する、といった面倒で限界のある作業は必要ありません。

IRIS NLP は非構造化データ・セット内のすべての文を、言語構造だけを基に文法解析します。定義した辞書内の特定の単語の存在を探すのではなく、筆者が書いた原文そのものからセマンティックな（意味のある）「エンティティ」を見つけ出します。また、これら「エンティティ」の文脈から「パス」を見極めてエンティティ同士の関連性を示したり、Proximity（近接性）や Dominance（優位性）などのエンティティ関連メトリクスを算定したりします。

事前知識を不要とするこの「ボトムアップ」形式の解析は、IRIS NLP の「**オープンエンド型の発見**」手法（open-ended discovery）と呼ばれています。

InterSystems IRIS NLP Japanese の仕組み

文を単語単位に切り出すという従来の日本語形態素解析ツールとは異なり、IRIS NLP Japanese は一単語以上から成る「エンティティ」を検出します。

例: **米利上げ**が接近し、**世界経済減速の懸念**も広がるなかで、**外国人投資家**は**慎重**。

ハイライトされている部分が、IRIS NLP Japanese が検出した「エンティティ」です。
従来の形態素解析ツールでは、下のように単語単位でこれらのエンティティをさらに切り分けていました。

米利上げ	=> 米／利上げ
世界経済減速の懸念	=> 世界／経済／減速／の／懸念
外国人投資家	=> 外国／人／投資／家
慎重	=> 慎重

「エンティティ」を個々の単語としてではなく、単語クラスターとして扱うことにより、文すべてを読み直すことなく、エンティティおよびそこに含まれる単語それぞれの意味が明確になります。この特徴により、従来のツールに比べてテキスト解析と探索の価値と精度が高まります。上の例では、単語レベルまで分かち書きされると「米」という字が「お米」を意味しているのか「米国」を指しているのかが不明になってしまいますが、「米利上げ」というエンティティからは一目で理解できます。また、エンティティの一部分を基に「Similar Entity」（類似エンティティ）を引き出すことも簡単です。例えば、同じデータ・セットに存在する「米利上げ」と「早期の米利上げ観測」という2個のエンティティは「Similar Entity」と認識されます。
IRIS NLP Japanese はさらに、独自の「Entity Vector」アルゴリズムを利用して、各文の文法構造だけを基にセマンティックな「パス」を特定し、エンティティ同士の関連性やその繋がりの強さを示します。前述の例文の「パス」は下のようになります。

外国人投資家　慎重　世界経済減速の懸念　米利上げ

文書内の「エンティティ」と「パス」の情報を基に、IRIS NLP では「Dominance」、「Proximity」などの追加メトリクスを計算します。「Dominance」は、特定のエンティティの文書内での優位性を表し、長い文章の要点を押さえるのに役立ちます。「Proximity」は、文書内で同一の「パス」に存在するエンティティ同士の関連性（近接性）を示します。

従来のソフトウェア・システムは、膨大な構造化データを集約しグラフや表にすることで、より良い決断に導いていましたが、非構造化データは活かされない状態でした。IRIS NLP を利用することで、非構造化データから「エンティティ」、「パス」、および関連メトリクスという形での構造を容易に得られるようになります。このように構造化されたデータは、従来使用してきた構造化データとともに「全てのデータ」（all the data）となり、これを様々なインターフェースやアルゴリズムに取り込むことで、お客様の意思決定を、より有益なものとなるよう支援します。

お客様事例

IRIS NLP 独自の「オープンエンド型の発見」手法は、世界各国のお客様により実証されています。以下は、非構造化データの「コンテンツベース・プロファイリング」の例です。

> ある大手製薬会社では、治験対象となる患者コホートを特定するために IRIS NLP が使用されました。対象疾患に対する危険因子情報はコード・システムには存在しなかったため、個々の患者に関する非構造化データが非常に重要となりました。従来の単語単位の技法では言葉の微妙な区別が読み取れませんでしたが（例：「metformin allergy （メトフォミンアレルギー）」 vs. 「metformin （メトフォミン）」）、IRIS NLP の「ボトムアップ」形式で、カルテの自由記入欄などから効率的に目的の概念を選び、患者コホート対象者選択のためのコンテンツベース・ルールを構築することができました。

他にも「データ探索」、「トレンド分析」、「情報抽出」など、様々なエリアで IRIS NLP によって検出された「エンティティ」の付加価値を活かしたユースケースがあります。[1]

おわりに

日本語の特徴の一つとして、ヨーロッパ言語のように単語境界に明確な指標（空白文字）がないことが挙げられます。このため、従来の典型的な形態素解析では、文内でテキストの最小単位を求めることに焦点を当てることで、一般テキスト解析ツールの「Bag of Words」モデルに当てはめる形をとってきました。IRIS NLP の「ボトムアップ」アプローチは、反対にテキストの最大単位をエンティティとして認識するため、著者が執筆時に書こうとした内容を理解する上での、より確かな基盤となります。このことは日本語に限らず、世界各国のお客様により実証されています。

また、この**オープンエンド型の発見**手法は自然言語そのものの特徴を利用しており、特定の業界や専門用語に依存しません。IRIS NLP はインターシステムズ社のデータプラットフォームの**組み込み**テクノロジーとして提供されており、そこに構築されたアプリケーションやサービスならば、どのアプリケーションやサービスからもご利用が可能なのは、**多様な活用事例**からもご理解いただける通りです。

---

[1] こちらもご参照ください:
http://www.intersystems.com/wp-content/uploads/sites/6/Use_Cases_for_Unstructured_Data_20141205-1.pdf

インターシステムズジャパン株式会社

〒160-0023
東京都新宿区西新宿 6-10-1
日土地西新宿ビル 15F
Tel: 03-5321-6200

InterSystems.com/jp/