InterSystems IRIS NLP Japanese の概要

Version 2.0

InterSystems IRIS NLP Japanese の概要

InterSystems IRIS Natural Language Processing （NLP）は、文章（非構造化データ）を意味のあるデータ項目（構造化データ）に変換することのできる、自然言語処理の技術です。筆者が書いた原文そのものから検出される「エンティティ」と、そのエンティティ同士の関連性「パス」がデータ項目となります。こうして構造化されたデータ項目は、一般のアプリケーションやアルゴリズムで活用できるようになり、ナビゲート、分析、その他の処理など、様々な用途で元の非構造化コンテンツを利用できます。

IRIS NLP Japanese

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。

従来の NLP ツールの多くは、各文を単語のシーケンスとして捉えていました。辞書、統計および構文解析のメカニズムを使用して、どの語群が意味のある単位かを検出するのが一般的です。InterSystems IRIS NLP はこれらのツールとは異なり、「ボトムアップ」形式のテキスト解析により、単語ではなく文法を基に、文章そのものから意味のある「エンティティ」を見出します。IRIS NLP のテクノロジーで本質的な情報を得る際には、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。こうして先入観なくして得られた「エンティティ」は、一般の統計やアルゴリズムで分析してナビゲートや探索シナリオを実装したり、固有表現抽出（NER）や品詞タグ付け（POS）など、他の NLP ツールに入力してそれらの処理精度を高めたりすることができます。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。

従来の形態素解析ツールによるテキスト解析

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。日本語を対象とした典型的な形態素解析ツールは、文を単語（言語で意味を持つ最小単位）に分割し、それぞれの品詞を判別することを目的に設計されています。文から切り出した単語が属する品詞を辞書で調べていき、結果得られた品詞の並びから文法的に正しい並びであるものを正解と見なすのが一般的で、主な手法には、規則による方法と確率的言語モデルを用いる方法があります。

このようなツールは、オープンソース、商用システムともにいくつか存在します。ここでは、一般の形態素解析ツールの典型的な分かち書き結果を紹介します。

例文： 文部科学副大臣には「ヤンキー先生」こと義家弘介元文科政務官が就任した。

ツールAによる分かち書き：

文部／科学／副／大臣／に／は／「／ヤンキー／先生／」／こと／義／家／弘／介／元／文科／政務／官／が／就任／し／た／。

ツールBによる分かち書き：

文部／科学／副／大臣／に／は／「／ヤンキー／先生／」／こと／義家／弘介／元／文科／政務／官／が／就任／し／た／。

このようなアプローチによる解析結果は、自然言語処理の特定分野では有益ですが、昨今の研究動向では統計を基としたテキスト解析よりも、セマンティック解析（言葉の意味解析）が主流となっています。

また、従来のオープンソース・ツールで利用できる形態素解析辞書はいくつか存在し、ここ10年以上にわたり現代日本語の書き言葉についての辞書は、妥当なレベルで提供されてきましたが、更新されないまま年月が過ぎていることが多く、近年創成された新しい概念が含まれていないことがあります。従来のツールに新しい言葉を正しく認識させるには、辞書に新語を追加し常にメンテナンスする義務が発生し、これには技術的な知識も必要となる場合があります。「基本的な学習」機能を謳うツールもありますが、それには予めアノテーションされたコーパスが必要となり、これにもメンテナンス作業が不可欠です。

IRIS NLP Japanese によるテキスト解析

日本語の非構造化データを IRIS NLP Japanese がどのように処理するかを以下に説明します。

例　文：
マララさんは全ての子どもに教育の機会を与えるよう世界の指導者に呼び掛けた。

この文から IRIS NLP は4個のエンティティを検出します：
　マララさん　全ての子ども　教育の機会　世界の指導者

「エンティティ」の検出には、文中の「エンティティ以外」の部分を見つけるよう設計された、コンパクトな言語モデルが使用されます。この例では、「は」「に」のような助詞や動詞などがその「エンティティ以外」の部分です。実際の「エンティティ」検出には、曖昧さの残る部分を専用の言語ルールが解決し、「エンティティ以外」の部分を

特定します。この手法では、「エンティティ」に存在しうる全ての単語を予め知っておく必要はなく、簡潔でドメインに依存しない言語モデルを維持するため、高速な処理を実現できるのです。

また、IRIS NLP Japanese は独自の「Entity Vector」アルゴリズムを利用し、この文の「パス」を検出します：

マララさん 教育の機会 世界の指導者 すべての子ども

「パス」内の4個の「エンティティ」の順序は、文中の順とは異なります。「Entity Vector」アルゴリズムは、日本語の構文と位置情報だけを基に、文中全てのエンティティをそれぞれの関連性順に並べ替え、エンティティ同士の繋がりの強さを示します。

IRIS NLP の「エンティティ」 vs. 単語

従来のオープンソース日本語形態素解析ツールは、大規模な辞書を利用して文を最小の単語単位に分割しようとします。

例文：
敗血症は腎盂腎炎から至ったケースという。

ツールAによる分かち書き：
敗血症／は／腎盂／腎／炎／から／至っ／た／ケース／と／いう／。

ツールBによる分かち書き：
敗血／症／は／腎盂／腎炎／から／至／っ／た／ケース／と／い／う／。

従来のツールは：

- 辞書に言葉が存在すれば、疾患名を正しく認識します。ツールAでは「敗血症」の3文字を一単語と認識されましたが、ツールBでは「敗血」と「症」を名詞と接尾詞の2語とみなされています。
- 辞書に言葉が存在しない場合、疾患名を一つの意味のある単語として認識しません。ツールAでは「腎盂腎炎」という言葉は「腎盂」「腎」「炎」（2個の名詞と1個の接尾詞）と、計3語と認識されてしまいました。ツールBは、「腎盂」「腎炎」の2個の名詞として処理されました。

「敗血症」を名詞「敗血」と接尾詞「症」の2語に分割することは、決して間違いではありませんが、「敗血」という言葉を見ただけでは、予めその言葉を知らない限りそれが疾患名の一部だとは一目で理解できません。また、「腎盂腎炎」と「腎炎」は全く異なる症状ですが、2語もしくは3語に分割されてしまうと、著者が「腎盂腎炎」として書いた言葉が、不明瞭になってしまいます。

IRIS NLP による解析結果：
敗血症 は **腎盂腎炎** から 至った **ケース** という。

- 「敗血症」「腎盂腎炎」などが辞書に登録されているわけではありませんが、IRIS NLP はこれらの疾患用語を文の文法構造だけを基に抽出しました。ドメイン知識の辞書登録には保守作業も必要ですし、本質的に完璧に用語を網羅することはできません。IRIS NLP では、事前に辞書を作成する必要もなく、よってその保守作業も不要なため、より堅牢性の高いテクノロジーと言えるでしょう。

例文：
安倍晋三首相は、安全保障関連法の成立を最大の成果と強調した。

ツールAによる分かち書き：
安倍／晋／三／首相／は／、／安全／保障／関連／法／の／成立／を／最大／の／成果／と／強調／し／た／。

ツールBによる分かち書き：
安倍／晋三／首相／は／、／安全／保障／関連／法／の／成立／を／最大／の／成果／と／強調／し／た／。

- この例でも、従来のツールは接頭詞、接尾詞などのレベルまで精細に文を分割しています。ここまで細かく分かち書きされてしまうと、個々の単語の意味だけでは話題そのものが何なのか、理解できなくなってしまいます。

IRIS NLP による解析結果：
安倍晋三首相 は、 **安全保障関連法の成立** を **最大の成果** と 強調した。

ここでは、個々の単語に比べ、単語クラスターである「エンティティ」の有用性が高い点を例証します。

安倍晋三首相
安全保障関連法の成立
最大の成果

これらのフレーズは「エンティティ」もしくは「エンティティクラスター」と呼ばれます。個々の単語よりもある程度意味が限定されるため、正確な情報を得やすく、話題そのものがわかりやすくなります。

この意味のある「エンティティ」の抽出に加え、個々のエンティティの文脈情報を導き出すことにより、IRIS NLP はさらにパワフルな情報を提供します。特に、「パス」はエンティティ同士の関連性を示し、IRIS NLP の「Proximity」（近接）メトリクスの基礎となるので重要です。「パス」や「Proximity」から、「安倍首相が安全保障関連法に関し何かしたらしい」こと、「敗血症は腎盂腎炎と何やら関係があるらしい」ことなどを、事前の知識なく推測することができます。

従来のツールには、漢字の読みを推定するなど、別の機能も備わっているものもありますが、大量の非構造化テキストに対しオープンエンド型の発見ができるという点は、IRIS NLP の明白な利点です。

IRIS NLP Japanese により識別される論理テキスト単位

日本語は本質的に西洋の言語と異なる部分があるため、IRIS NLP Japanese によって識別される要素も若干他言語の IRIS NLP とは異なります。

文 – 西洋言語と同様、「文」とは文の終わりを示す文字（日本語の場合は「。」）で区切られたテキストを一単位とするものです。

エンティティ – 元来 IRIS NLP の西洋言語対応のために考案された「Concept（概念）」と「Relation（関係）」は、日本語ではその境界があまり明白でありません。このため、日本語解析の際、IRIS NLP Japanese では文の意味のある部分を「エンティティ」と表現します。これは全言語共通の技術レベルでは「Concept」として識別されるものです。

「Entity Vector」に基づいたパス – 西洋言語での「パス」は、一般に1個以上の CRC（Concept-Relation-Concept）シーケンスに基づいており、その結果、パス内の Concept と Relation は原文と同じ順序に並びます。日本語の「パス」は、弊社独自の「Entity Vector」アルゴリズムに基づいており、その結果多くの場合、パス内の「エンティティ」の順序は原文での順とは異なったものになります。この順序は、エンティティ同士の関連性と繋がりの強さを示しています。

お客様事例:

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。

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（メトフォルミンアレルギー）」 vs. 「metformin （メトフォルミン）」）、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 によって検出される「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てい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/