日中韓辭典研究所
The CJK Dictionary Institute

深層学習用辞書データベース
DeepLEX

春遍雀來
2020 年 12 月 25 日改訂

1. 深層学習とは何か

「深層学習」と「ニューラルネットワーク」は、人工知能分野における最先端領域として、現在話題となっているトレンドキーワードです。人工ニューラルネットワークによる深層学習は、機械学習の手法をこれまでにない水準に高め、音声認識、サイバーセキュリティ、機械翻訳など、多岐にわたる先進技術で不可欠な役割を果たしています。たとえば自然言語処理（NLP）の分野も、ニューラルネットワークにより著しい発展を遂げており、代表例として Google 翻訳の翻訳品質の大幅な向上があります。

2. 辞書データベースの利点

大規模辞書等の語彙データベースは、自然言語処理における深層学習タスク、たとえばデータセット生成、単語埋め込みの作成、ハイブリッド技術の利用に多大な利点をもたらします。

日中韓辭典研究所（CJKI）は「Lexical Resources for Deep Learning（深層学習用辞書データベース）」、略称 **「DeepLEX 辞書データベース」**と呼ばれる超大規模な語彙データベースの開発を積極的に進めています。**DeepLEX 辞書データベース**は、固有表現抽出（NER）、サイバーセキュリティ、ニューラル機械翻訳（NMT）、音声技術等、広範な分野で深層学習に大きく貢献する能力があります。

2.1 ニューラル機械翻訳（NMT）

ニューラル機械翻訳技術は、従来の機械翻訳と比べ大幅な改善を遂げてきたものの、比較的頻度が低い内容語、特に POI や人名の異表記等の固

The CJK Dictionary Institute, Inc.
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001 JAPAN
Phone: 048-473-3508 Fax: 048-486-5032
Email: jack@cjki.org Web: www.cjk.org

有表現に対しては、性能が劣化するという問題があります。対策として離散確率モデルを用いた辞書等を追加情報としてニューラル機械翻訳システムに統合すれば、精度スコア（BLEU および NIST）の大幅な質的向上を実現できます。詳細に関しては、下記の論文をご覧ください。

Very Large-scale Lexical Resources to Enhance Chinese and Japanese IR and NLP
（中国語および日本語の IR および自然言語処理を向上する超大規模語彙データベース）

2.2 正則化

深層学習のアルゴリズムは、トレーニングデータだけでなく、たとえば固有表現の表記のゆれ（例：「Ichiroo」対「Itirou」）のような未知の入力データも、最適に処理できる必要があり、そのために重要な役割を果たすのが正則化です。正則化はニューラルネットワークの過学習を防ぐ一連の技術から成る手法です。その目的は、前述の固有表現の異表記等の問題領域から、完全に新しいデータが入力された場合に、深層学習モデルの精度を高めることです。その手段として、固有表現の異表記を収録した大規模辞書データベースをベクトルデータの圧縮および各表記に対する有意値の計算に使用すれば、大幅なモデル精度向上につながる可能性があります。

2.3 固有表現抽出 （NER）

従来の固有表現抽出ではルールベース手法を利用しますが、それは特定分野の網羅的な辞書データが存在する場合に性能を発揮します。しかしながら、実際の辞書データは多くの場合に網羅的ではないため最善の性能に至りません。改善策として、人工知能（AI）により表現を自動検出する手法を使用できますが、ラテン文字化された中国語名やアラビア語名の異表記など、データが豊富な領域に関しては、必ずしも十分な再現率と精度が得られません。

高精度を達成するために、最も実用的な解決策は、たとえば CJKI が提供する辞書のように、数千万または数億の項目を含む、包括的かつハー

The CJK Dictionary Institute, Inc.
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001 JAPAN
Phone: 048-473-3508 Fax: 048-486-5032
Email: jack@cjki.org Web: www.cjk.org

ドコードされた辞書を統合することです。実際、モスクワ大学で実施された研究は、従来の機械学習技術（CRF）がニューラルネットワークモデルに勝る高精度を達成した理由について、辞書機能を使う唯一の技術であることを成功要因として示唆しています。このことは、ニューラルネットワークの時代においても、辞書データベースを基盤とする従来の技術が担う役割が大きいことを実証するものです。すなわち、固有表現辞書は、深層学習技術により取って代わられてはいないことが明示されています。

2.4 サイバーセキュリティ

大規模固有表現辞書は、サイバーセキュリティにおいても、人名、場所の名前、組織名などの通常の固有表現抽出だけでなく、「セキュリティエンティティ」と称されるタイプのサイバーセキュリティ分野特有の固有表現–たとえばハッカーの名前、ハッカーグループ、ソフトウェア製品、ウィルス、電子装置など–の抽出にも対応できるため、主要な役割を果たすことが期待されます。

しかし、サイバーセキュリティ分野の固有表現抽出モデルは、機械学習アルゴリズムに過度に依存し、セキュリティ関連固有表現を無視する傾向により精度が劣化する課題があります。そのソリューションとして、汎用の固有表現辞書を用いたCRF（条件付き確率場）による従来型の固有表現抽出技術、そしてセキュリティ分野特有の固有表現抽出用に精密に調整されたセキュリティ関連固有表現辞書の双方が、サイバーセキュリティに大きく貢献することが見込まれます。

2.5 事前学習モデル

言葉同士の関連性の確立は、多様な問題の解決手段として多くの可能性を秘めています。**BERT**（Bidirectional Encoder Representations from Transformersの略称）、**ELMo** 、**Word2vec**など、そのためのモデルを構築する技術が、近年次々に登場しています。こういった「事前学習モデル」と呼ばれるモデルを構築するために、DeepLEXのデータベー

The CJK Dictionary Institute, Inc.
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001 JAPAN
Phone: 048-473-3508 Fax: 048-486-5032
Email: jack@cjki.org Web: www.cjk.org

スを使用して、注釈付きコーパスなどほかのデータベースと組み合わせることは、複雑な技術ではありますが、特にアラビア語など形態論的に複雑な言語において、有効な結果につながる可能性があります。

2.6 大規模な頻出語彙辞典

日中韓辞典研究所の**DeepLEX 辞書データベース**にあるような頻出語彙辞典を使用すれば、同一の人名の複数の異表記（たとえば１５０を超える「Mohamed」の異表記）や日本語で広く使用される複数の表層格など、表記のゆれから生じる複雑性を大幅に抑制できます。

3. DeepLEX Resources

日中韓辞典研究所の**DeepLEX 辞書データベース**は、固有表現抽出や音声技術等の自然言語処理アプリケーションへの対応に特化して設計された、何千万もの日中韓各語の固有表現を収録しています。何千万もの日中韓およびアラビア語の固有表現からなるそのデータベースの内、一部を以下に紹介します。

1. 中国人名異表記データベース（CNV）

CNV は、中国人の基本人名 160 万項目と主なローマ字異表記を合わせた約 1,000 万項目を収録するデータベースで、標準中国語と四つの方言に対応します。

2. 日本語異表記データベース（JOD）

JOD は同一語の表記の揺れを識別することで情報検索と機械翻訳に貢献します。例えば、[neko] （cat）には、猫、ねこ、ネコ；[kakiarawasu]（write out, publish）には、書き著す、書著す、書き著わす、書著わす等の表記がありますが、この揺れを認識することで精度をあげることができます。

3. 日本人名異表記データベース（JNV）

JNVは日本人の基本人名（姓・名）55万項目とローマ字異表記を合わせた約350万項目を収録するデータベースで、ローマ字異表記は、標準的なローマ

The CJK Dictionary Institute, Inc.
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001 JAPAN
Phone: 048-473-3508 Fax: 048-486-5032
Email: jack@cjki.org Web: www.cjk.org

字表記からその他の一般的な表記、混合型表記まで幅広く網羅します。

4. アラビア語全活用形データベース（ArabLEX）

ArabLEXはアラビア語のあらゆる派生、屈折等の活用形を網羅する膨大なデータベースで、収録項目数は6億項目に上ります。品詞、詳細な文法情報、ローマ字表記等も提供できます。固有表現抽出（NER）、音声技術、応答生成等のAIシステム開発に有用で、アラビア語自然言語処理に最適なデータ資源であります。

5. アラブ人名データベース（DAN）

DAN はアラブ人名とそのローマ字異表記を合わせた約650万項目を収録する包括的なデータベースで、母音付きと母音無しのアラビア語表記を共に収録します。

6. 日本語多言語地名 POI データベース（JMP）

JMPは日本の地名とPOI（駅、学校、空港等の地点）を、中国語、日本語、韓国語、ヨーロッパ諸言語、アジア諸言語に翻訳した大規模データベースであります。14言語による多様な分野のPOIを310万項目収録します。

上記のデータベースは、世界最大級の IT 企業により、音声技術、形態素解析、機械翻訳等の自然言語処理および AI（人工知能）アプリケーションにおいて、ならびに自然言語生成、音声技術などの AI においても活用されています。**DeepLEX** の辞書データベースは、AI チップのアーキテクチャへの統合により、埋め込み型機械翻訳、音声合成（TTS）、自動音声認識（ASR）技術にも対応が可能です。

The CJK Dictionary Institute, Inc.
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001 JAPAN
Phone: 048-473-3508 Fax: 048-486-5032
Email: jack@cjki.org Web: www.cjk.org

About Jack Halpern

春遍雀來（ハルペン・ジャック）

春遍雀來は日中韓辭典研究所の取締役社長で、辞書編纂家であります。16年を費やして完成させた『新漢英辞典』を含め、編集長として編纂した漢字学習辞典はその多くが学習用参考書の定番となっています。昭和女子大学で特別研究員を務めた経歴もあります。

ドイツ生まれの春遍雀來はフランス、ブラジル、日本、アメリカ等の6カ国に移り住み、日本には40年以上在住しています。彼は熱心なポリグロットで、18言語を学び、10言語を流暢に話します。特に日本語と中国語の辞書編纂に熱心で、言語学習と辞書編纂に人生を捧げています。

日中韓辭典研究所（CJKI）

日中韓辭典研究所（The CJK Dictionary Institute, Inc.）は日本埼玉県にあり、5,000万項目に上る日中韓各語とアラビア語の大規模辞書データの開発と継続的な拡張を主な業務とする研究所であります。日本語学習辞典の定番である『新漢英字典』とその他多数の辞書の編集長春遍雀來（ハルペン・ジャック）が取締役社長を務めます。

CJKIは高品質な辞書データを提供する世界有数の企業で、高成長の東アジア市場に進出するIT企業を支援します。中国語の各種方言を含む日中韓各語の大規模辞書データ（一般語彙・固有名詞・専門用語）、600万項目からなる包括的なアラビア語全活用形データベース（ArabLEX）、収録項目数が数千万に達する多言語固有名詞辞書等、ソフト開発に有用な多様な包括的データ資源を提供します。

CJKIは日中韓各語のデータ資源を提供する主要な企業であります。高品質かつ包括的な辞書データとコンサルティングを提供することを通じて、日中韓各語とアラビア語の情報処理技術開発に寄与し、富士通、シャープ、ソニー、IBM、Google、Microsoft、Yahoo、Amazon、Baidu等有名なIT企業とソフト開発者に貢献しています。

The CJK Dictionary Institute, Inc.
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001 JAPAN
Phone: 048-473-3508 Fax: 048-486-5032
Email: jack@cjki.org Web: www.cjk.org