様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月　 ３ 日現在

機関番号：１４６０３
研究種目：研究活動スタート支援
研究期間：2020～2021
課題番号：２０Ｋ２３３２５

研究課題名（和文）知識グラフを統合したニューラル機械翻訳

研究課題名（英文）Neural Machine Translation Integrated with Knowledge Graph

研究代表者
渡辺　太郎（Watanabe, Taro）

奈良先端科学技術大学院大学・先端科学技術研究科・教授

研究者番号：９０３９５０３８

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：深層学習により大幅な性能向上を果たしたニューラル機械翻訳は、モデルの学習のために、大量のデータを必要とする。ところが、データを増やすだけでは、固有表現や、誕生日などの属性、所属先など他のオブジェクトとの関連性など、日々更新される知識を翻訳するのは難しい。本研究では、物事の属性および関連性を記述し、かつ、不完全ながらも多言語化された知識グラフを統合した機械翻訳を実現することで、問題が解決できるかを解明する。
本研究では、単語単位ではなく、サブワード単位に学習された知識グラフのベクトル表現を統合した機械翻訳モデルを実現した。機械翻訳実験の結果、人手評価で、固有表現が正しく翻訳されることを示した。

研究成果の学術的意義や社会的意義
本研究における知識グラフと機械翻訳を統合したモデルにより、知識を反映した機械翻訳を実現した。本手法により、人名や地名等の固有表現をより正しく翻訳できることを示している。今後は、知識グラフを更新することで機械翻訳モデルの再学習を全く必要としない機械翻訳モデルを実現することにより、各ドメインへと容易に適用可能、かつ、カスタマイズ可能なシステムの実現を目指す。

研究成果の概要（英文）：Neural machine translation demands huge data when training the translation model, although its performance has been drastically improved by deep learning. However, simply increasing training data does not assure that the trained model can fluently translate named entities, properties, e.g., date of birth, or relations with other objects, e.g., affiliations, since such knowledge will be updated almost every day. This work investigates a method to solve the issue by integrating multilingual a knowledge graph into machine translation, which is knowledge representation denoting attributes and relations of objects with partially multilingual annotation. In this research, we proposed a machine translation model which integrates representations from knowledge graph that is trained by subword unit, not word-wise unit. Experimental results on machine translation tasks showed that named entities are translated correctly after our manual investigations.

研究分野：自然言語処理

キーワード：機械翻訳　知識グラフ

１．研究開始当初の背景

深層学習を応用したニューラル機械翻訳は、統計的手法に基づく機械翻訳と比較して大幅な性能向上を果たしたが、まだ解けていない問題が数多くある。特に、原言語と目的言語の文が対応付けられた対訳データが少ない場合や、対象とするドメインのデータがない場合、また、低頻度語や長文を入力とする場合、極端に翻訳の性能が悪くなる。これは、ニューラル機械翻訳の汎化性能がまだ弱く、現在の技術では、非常に大量の対訳データを集めてモデルを学習する必要がある。

現在、対訳データをウェブなどから自動的に収集し、対応付誤りなどのノイズを含んだデータを除去する手法や、単言語のデータを大量に集め既存の機械翻訳システムにより翻訳することでデータを追加するなどの手法が提案されている。網羅的にデータを増やすだけでは、個人名や会社名などの固有表現や、誕生日などの属性、所属先など他のオブジェクトとの関連性など、意味を保って翻訳するのは難しい。また、日々更新される知識を新たに追加するのは難しく、従来法ではデータの追加およびモデルの再学習を要する。さらに、DARPA XAI プロジェクトが問題提起したように、ニューラルネットワークで学習されたモデルは人間が理解できるような表現ではなく、例えば誤った翻訳が生成されたとしても、その原因を追求するのは容易ではない。このように、データ量に依存せずとも高精度な機械翻訳が可能なのか、また、説明可能な翻訳が生成できるのか、が問われている。

２．研究の目的

大量のデータを必要とする課題およびモデルが説明可能ではないという課題を解決するため、本研究では、知識グラフを統合したニューラル機械翻訳を実現する。知識グラフは、WordNet など一般的な名詞の同義語や、Wikipedia を元にした YAGO (https://datahub.io/collections/yago)や DBPedia (https://wiki.dbpedia.org)など固有の物事の属性および関連性を記述したものであり、BabelNet (https://babelnet.org)のように不完全ながらも多言語化されている。例えば、図１のように「ギネス氏」を翻訳する場合、知識グラフ上で「ケノービ」を介して「スター・ウォーズ」と関連付けられることから、``Alec Guinness"と翻訳され、ビール醸造会社ギネスの創業者``Arthur Guinness"として翻訳されない。逆に、知識グラフをたどることで``Alec Guinness"という訳語が選択された理由が分かる。

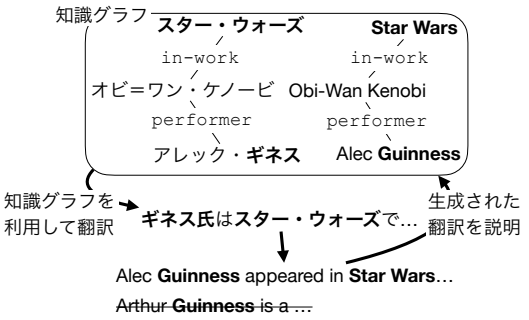

図 1 知識グラフと統合した機械翻訳

入力文には明示的に表現されない意味を知識グラフのリンクをたどって推論することでより高精度な機械翻訳を実現、また、生成された訳文を知識グラフと関連付けることで説明可能、という点で独自性がある。さらに、以下のような研究に発展しうる点で、創造性がある。

- 知識グラフでは、必ずしも全ての物事が多言語にマッピングされているとは限らない。この不完全な知識グラフの言語間の対応関係を自動的に学習する。
- 知識グラフは、必ずしも最新の知識を表現しているとは限らない。多言語テキストから自動的に多言語知識グラフを構築、あるいは知識を追加。
- 日々、新たな知見が得られるが、データの追加やモデルの再学習をせず、知識グラフを更新するだけで、機械翻訳のモデルを自動的に更新する。

３．研究の方法

本研究では、知識グラフをニューラル機械翻訳と統合することにより、大量のデータに依存せずとも高精度な機械翻訳を実現でき、かつ、説明可能な翻訳を生成可能であることを明らかにする。知識グラフがニューラル機械翻訳と統合できるよう、知識グラフで記述された物事の属性および関連性を数値ベクトルへと表現する手法を実装する。実装したシステムを機械翻訳研究で標準的なベンチマークとして使われる WMT18 の英独翻訳タスク(http://www.statmt.org/wmt18)にて評価する。本研究では、学習データが評価時のドメインと一致していなくとも知識グラフの情報を利用して翻訳が可能であるかを検証する。具体的には、機械翻訳モデルの学習には、EU の法律文書および議会の議事録を利用し、知識グラフには DBPedia を利用し、評価にはニュース記事を利用する。この実験により、提案手法が大量のデータに依存せずとも知識グラフの情報を利用することでより高精度な機械翻訳を実現できることを明らかにする。

４．研究成果

本研究では、2 つの手法を提案した。一つは、原言語を固有表現抽出器で固有表現を同定し、そ

の位置を特殊な記号へ変換する。その特殊な記号に対し、予め学習された知識グラフのベクトル表現を用いる手法である(KG-tag NMT)。もう一つは、特殊な記号への変換をせず、単語よりもより粒度の細かい、サブワード単位に学習された知識グラフのベクトル表現を用いる手法である(Subworded-KG NMT)。KG-tag NMT では、固有表現単位にベクトル表現を学習するため、知識グラフの知識を直接反映できるが、原言語の入力文と必ずしも対応する固有表現が存在するとは限らない。Subworded-KG NMT では、固有表現をサブワード単位へ分割し、エンコーダ・デコーダの枠組みで学習する手法を実現した。このため、固有表現を直接反映したベクトル表現ではないが、カバレッジが非常に大きくなり、入力文のほぼ全ての固有表現に対し、知識グラフを反映可能である。また、固有表現のベクトル表現として、サブワード単位の埋め込みベクトルだけではなく、ニューラルネットワークで表現されたエンコーダの層を利用することで、コンテキストを反映したベクトル表現を実現した。

実験結果を表 1 に示す。Transformer [1]に基づくベースラインと比較し、従来法の知識グラフのベクトル表現を埋め込む手法(KG [2])およびマルチタスク学習による手法(Multitask learning [3])では、評価値 BLEU [4]が大幅に下がった。提案法である KG-tag でも同様な傾向が見られるのに対し、Subworded-KG では性能の向上が見られた。これは、知識グラフにおける固有表現を反映することにより、法律文書のドメインで学習されたモデルが全く異なるニュースのドメインを翻訳できることを示しており、本手法の有効性を示している。

		BLEU[%]
Transformer [1]		28.5
KG [2]		25.9
Multitask learning [3]		27.6
KG-tag		25.1
Subworded-KG	Embedding	28.6
	First layer	27.5
	Last layer	28.9

表 1 ドメイン適用実験結果

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. NIPS 2017.
[2] Diego Moussallem, Axel-Cyrille Ngonga Ngomo, Paul Buitelaar, and Mihael Arcan. Utilizing knowledge graphs for neural machine translation augmentation. K-CAP 2019.
[3] Yang Zhao, Lu Xiang, Junnan Zhu, Jiajun Zhang, Yu Zhou, and Chengqing Zong. Knowledge graph enhanced neural machine translation via multitask learning on sub-entity granularity. COLING 2020.
[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. ACL 2002.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 坂井優介，渡辺太郎，藤田篤
２．発表標題 知識グラフ埋め込みを用いたニューラル機械翻訳におけるエンティティ表現の改良
３．学会等名 言語処理学会第27回年次大会
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関