未知語の分布表現の翻訳に基づく機械翻訳のドメイン適応

石渡 祥之佑 * 吉永 直樹 †‡ 豊田 正史 † 喜連川 優 †§
* 東京大学大学院 情報理工学系研究科 † 東京大学 生産技術研究所
‡ 情報通信研究機構 § 国立情報学研究所
{ishiwatari, ynaga, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

1 はじめに

現在，一般的に用いられる機械翻訳システムの多くは，対訳コーパスを学習データとする統計的機械翻訳に基づくものである．統計的機械翻訳システムは人手により作成された対訳コーパスから自動的に知識を抽出するため，各言語に固有の文法規則を記述する必要が無く，あらゆる言語対において利用できる点で汎用性に優れている．

統計的機械翻訳の大きな課題として，学習データのドメインとテストデータのドメインが異なる場合に翻訳精度が著しく落ちる [1] 点が挙げられる．これを防ぐためには，利用ドメインに合わせて対訳コーパスを作成するアプローチが有効である．しかし，プロの翻訳者による対訳コーパスの作成は多大なコストを必要とするため，容易に行うことはできない．こうした問題意識から，既存の対訳コーパスで学習されたモデルをテストデータのドメインに適応させる研究が数多くなされている [2] [3] [4].

翻訳対象の文とは異なるドメインのコーパスから学習されたモデルを使う際に問題となるのは，学習データに存在しない未知の語や表現，統語構造の翻訳をいかに行うかという点である [1]．本研究は，これらの中でも特に顕著な問題となる未知語の翻訳に取り組む．未知語，すなわち翻訳モデルに存在しない語に訳語を与えることを考えた場合，最も簡単なアプローチは，ドメインごとに対訳辞書を作成する方法である．しかし，この方法は上述した対訳コーパスと同様，作成に多大なコストがかかるため現実的ではない．これに対し，ドメインに特化しない汎用的な対訳辞書と，対象ドメインの単言語コーパスであれば，比較的容易に利用できる．前者については，たとえば日英であれば EDICT[1] や英辞郎 [2]，その他の言語対でも Open Multilingual Wordnet[3] といった容易に入手できるデータセットが存在する．後者についても対象ドメインの文書を機械翻訳システム利用者から入手するか，ウェブからクロールすることにより容易に獲得可能である．上記の状況をふまえ，本論文で提案する手法は小規模な汎用対訳辞書と，対象ドメインの単言語コーパスのみを用いる．具体的には，単言語コーパスから得られた分布表現の翻訳を汎用対訳辞書により学習することで，未知語に訳語と翻訳確率を与える手法を提案する．

以降，2 節では機械翻訳におけるドメイン適応の関連研究について，3 節では提案手法のベースとなるベクトル表現の翻訳についてそれぞれ説明する．続く 4 節で提案手法を述べ，5 節で提案手法の有用性を評価するための実験を述べる．

2 関連研究

機械翻訳におけるドメイン適応の研究の多くは，対象ドメインの対訳コーパスが十分に得られない状況を想定して行われている．しかし，本研究のように対象ドメインの対訳コーパスが全く存在しない状況を想定している研究は少ない．ドメイン適応に対象ドメインの対訳コーパスを全く必要としない点で，Yamamoto ら [3] の研究は本研究と共通している．彼らは学習データのドメインも，テストデータのドメインも未知である状況を想定し，教師なし学習に基づくドメイン適応の手法を提案した．Yamamoto らの提案手法では，学習データである対訳コーパスをクラスタリングにより分割し，各クラスタをひとつのドメインとみなす．分割されたコーパスを用いて複数の翻訳モデルを学習し，テスト時にはいずれかの翻訳モデルを選択して利用することで，テストデータと関連の強い学習データの影響を受けやすくしている．

また，Mathur ら [2] は個別に作成された複数の対訳コーパスを活用して，翻訳モデルのドメイン適応を行

[1] http://www.edrdg.org/
[2] http://www.eijiro.jp/
[3] http://compling.hss.ntu.edu.sg/omw/

う手法を提案した．彼らは TED talk，ニュース，ソフトウェアのマニュアル等 11 種類の対訳コーパスを元に翻訳モデルを学習し，それらのモデルを線形補間により組み合わせることで近似的にドメイン適応を行う手法を提案した．Mathur らは実験により，対象ドメインの対訳コーパスが全く存在しない場合でも，任意のテストデータに対してある程度ドメイン適応が可能であることを示した．

上記 2 つの研究は，いずれも対訳コーパスから得られる翻訳モデルを調整することで，一部の学習データが翻訳に及ぼす悪影響を減らすアプローチと考えることができる．これに対し，本研究の提案手法は未知語に訳を与えるものであり，学習データに存在しない情報を追加するアプローチと考えることができる．この点において，Wu ら [4] の研究は本研究と共通している．しかし，Wu らが対象ドメインの対訳辞書を用いている一方，本研究ではドメイン特化しない汎用的な対訳辞書を用いる点では異なっており，本研究の提案手法の方がより汎用性が高い．

3 単語のベクトル表現の翻訳

異なるドメインの文書間では，出現する語彙の分布に大きな差異が存在する．そのため，学習に用いた対訳コーパスとは異なるドメインに属する文は，翻訳できない未知語を含む可能性が高くなる．以降本節では，単語のベクトル表現を活用することにより，未知語に対しても訳語を与える方法を述べる．

単語のベクトル表現 (以降，単語ベクトル) は，単語の意味を計算処理可能な形式で表現する手法であり，一つの単語の意味を一つのベクトルで表現する．これらのベクトルは「ある単語の意味は，その単語と共に出現している単語群によって特徴づけられる」という分布仮説 [5] [6] に基づき，単言語コーパスから教師なし学習により作成される．得られたベクトル間の類似度を測ると，似た文脈語の分布 (すなわち，似た意味) を持つ単語のベクトル同士は類似度が高くなることが知られている．

単語ベクトルを用いることにより，単語間の意味的な類似性の計算を，ベクトル間の類似度計算 (コサイン類似度など) に帰着させることが可能となる．しかし，一般的な単語ベクトルには異なる言語の単語間の類似性判定に用いることが難しく，言語を横断するアプリケーションに利用しにくいという問題点が存在する．この問題に対し，Mikolov ら [7] は単語ベクトルを線形変換により他の言語のベクトルに翻訳する手法を提案し，言語をまたいで単語の意味的な類似性を比較することを可能とした．翻訳に用いる線形変換 (以降，翻訳行列)$\boldsymbol{W}$ は，訳語ペアの集合 $\{\boldsymbol{x_i}, \boldsymbol{z_i}\}_{i=1}^{n}$ を用いて，以下の最適化問題を解くことによって得られる．

$$\boldsymbol{W}^{\star} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \sum_{i=1}^{n} \|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{z}_i\|^2 \qquad (1)$$

単語ベクトルは単言語コーパスから学習されるため，未知語のベクトルを作成することも可能である．本研究ではこの未知語ベクトルを翻訳することにより，適切な訳語を推定する手法を提案する．ただし，ベクトル翻訳には上述した Mikolov ら [7] の手法ではなく，これを拡張した著者ら [8] の手法を用いる．これは，著者らの手法が Mikolov らの手法を上回るベクトル翻訳精度を達成したためである．また，単語ベクトルの一種である分布表現と呼ばれるベクトル表現手法が翻訳に適しているとの実験結果に従い，本論文でも分布表現を用いて未知語の翻訳を行う．

4 提案手法

機械翻訳において，学習データと異なるドメインに属する文は，未知語，すなわちフレーズテーブルに存在しない語を含む可能性が高くなる．提案手法は前節で述べた分布表現の翻訳により，未知語へ適切な訳語を与えることを可能とする．以降，未知語を含む文の翻訳の流れを述べる．

(1) まず，機械翻訳システムに未知語 x を含む文が入力された場合，事前に原言語の単言語コーパスから学習した x の分布表現 $\boldsymbol{x}$ に翻訳行列 $\boldsymbol{W}$ を乗じることで，目的言語のベクトル $\boldsymbol{W}\boldsymbol{x}$ へと翻訳する．(2) 次に，$\boldsymbol{W}\boldsymbol{x}$ と全ての目的言語の分布表現との比較を行い，コサイン類似度が高いもの上位 10 語を x の訳語候補とする．この際，それぞれの訳語候補への翻訳確率は，$\boldsymbol{x}$ と上位 10 語のベクトルとのコサイン類似度の和が 1 になるように正規化することにより与える．以降，得られた訳語候補及びその翻訳確率はバックオフモデル [9] として扱う．(3) 最後に，事前に学習した翻訳モデルと言語モデル，及び上記で作成したバックオフモデルをデコーダに入力し，翻訳文候補の生成，選択を行う．

上記ステップ (2) において，訳語候補の上位に正解の類義語や対義語，上位語，下位語といった誤訳が出力される場合がある．この現象は，分布表現がコーパス内における語の共起に基づいて作られることに起因する．しかし，ここで出力された誤った訳語候補は，ステップ (3) で言語モデルによりある程度補正されることが期待できる．

表 1: 実験に使用したコーパスの規模

コーパス	日本語	英語
京都関連記事 (対訳)	30MB (440k文)	31MB (440k文)
料理レシピ (対訳)	13MB (150k文)	11MB (150k文)
Wikipedia(単言語)	4.4GB	16GB

5 実験

5.1 実験設定

まず，機械翻訳におけるドメイン適応の実験を行うため，異なるドメインの日英対訳コーパスを用意した．本研究では，京都翻訳フリータスク [10] で公開されている Wikipedia の京都関連記事の対訳コーパスと，ユーザ投稿型レシピサイトである Cookpad[4] の料理レシピの対訳コーパスを利用した．前者のドメインには日本の歴史上の人物や寺社仏閣に関する語彙が多く，だ／である調で記述されている．一方，後者のドメインには調理器具や食材に関する語彙が多く，です／ます調や砕けた表現が頻出する．このうち，料理レシピの対訳コーパスは学習データ (144,178 文) とテストデータ (10,000 文) に分割した．各データのデータサイズは表 1 に示す．

次に，Moses[5] を用いて，これらのデータから言語モデルと翻訳モデルの学習を行った．目的言語の言語モデルは京都関連記事のデータと料理レシピのデータを用いて行い，2 カ国語間の翻訳モデルは京都関連記事の対訳コーパスのみを用いて行った．

続いて，未知語の翻訳に用いるために日英 2 カ国語の分布表現と，日英間の翻訳行列を著者らの先行研究 [8] の手順に従って用意した．分布表現は Wikipedia のダンプデータ [6] から作成したものと，Cookpad の料理レシピデータから作成したものをそれぞれ用意した．分布表現間の翻訳行列は Open Multilingual Wordnet[7] の一部を学習データとし，教師あり学習により得た．

最後に，4 節で述べた提案手法を用いて，翻訳モデルに対してドメイン適応を行う．翻訳前に料理レシピドメインへの適応を行った場合と，ドメイン適応を行わなかった場合の翻訳品質の差異を BLEU[11] により評価する．

5.2 実験結果

実験結果を表 2 に示す．表 2 の**全データ**は料理レシピコーパスのテストデータ 10,000 件を用いて評価を行ったものであり，**未知語あり**は 10,000 件のうち，未知語を含む文のみを用いて評価を行ったものである．テストデータに含まれる未知語の統計量は表 3 に示すとおりである．

表 2: 各手法の翻訳品質 (BLEU)

手法	全データ 日英	全データ 英日	未知語あり 日英	未知語あり 英日
ベースライン	5.58	3.37	5.36	3.16
提案手法 (general)	6.09	3.50	5.91	3.46
提案案手法 (in-domain)	**7.29**	**3.67**	**7.23**	**3.89**
対訳コーパス	20.88	16.69	20.72	17.01

表 3: テストデータ (10,000 件) が含む未知語の統計量

	日英	英日
未知語 (出現回数)	21,218	4,639
未知語 (語彙数)	3,464	1,613
未知語を含む文	8,742	3,636

表 2 の**ベースライン**，**提案手法 (general)**，**提案手法 (in-domain)** は全て京都関連記事の対訳コーパスを用いて翻訳モデルの学習を行い，料理レシピドメインでテストを行ったものである．**提案手法 (general)** は Wikipedia から学習された分布表現を，**提案手法 (in-domain)** は Cookpad の料理レシピデータから学習された分布表現をそれぞれ用いてベクトル翻訳を行い，未知語に対するバックオフモデルを構築した手法である．**対訳コーパス**は料理レシピの対訳コーパスを用いて翻訳モデルの学習に用いたものである．ただし，本研究では対象ドメインの対訳コーパスは利用できないと仮定しているため，この手法により得られる翻訳品質が本論文における upper bound となる．

表 2 から，提案手法を用いて未知語の翻訳を行うことにより，翻訳品質が向上することがわかる．さらに，ベクトル翻訳に用いる分布表現の作成には，一般的なドメインのコーパスよりも，テストデータと同じドメインのコーパスを利用した方が良いという知見が得られた．これは，ベクトルを構成する単語の共起頻度に関する情報が，対象ドメインと強い相関を持つためであると考えられる．また，**対訳コーパス**と比較すると，他の手法の翻訳品質が非常に低くなっているが，これは京都関連記事ドメインと料理レシピドメインにおける語彙や言葉遣い等が著しく異なり，細かい表現の訳し分けが非常に困難となるからである．

さらに考察を深めるため，英日翻訳における出力例を表 4 に示す．表から，**ベースライン**では翻訳できなかった “rum” や “brandy”，“broccoli”，“preheat” といった未知語が提案手法では翻訳できていることが読み取れる．2 つの提案手法の差異に着目すると，**提案手法**

[4] http://cookpad.com/
[5] http://www.statmt.org/moses/
[6] http://dumps.wikimedia.org/から入手．日本語は 2014/11/04 版，英語は 2014/10/08 版．
[7] http://compling.hss.ntu.edu.sg/omw/

表 4: レシピドメインにおける英日翻訳の出力例．ベースラインにおける未知語，およびその訳語は太字で示す．

入力文	when the chocolate has melted completely , add the brandy or rum to mix in .
ベースライン	チョコレート が 完全 に 溶け て 、 **rum** の **brandy** を 混ぜ た もの で ある 。
提案手法 (general)	チョコレート が 完全 に 溶け て 、 **砂糖** を 入れ て 、 **砂糖** を 混ぜ た もの で ある 。
提案手法 (in-domain)	チョコレート が 完全 に 溶け て 、 **ラム 酒** を 加え 、 混ぜ た もの で ある 。
対訳コーパス	チョコ が 溶け たら 、 **洋酒** を 入れ て 混ぜる 。
参照訳	チョコ が 全て とけ たら **洋酒** を 加え て 混ぜ ます 。

入力文	you can also use broccoli .
ベースライン	**broccoli** も 使 わ れ る こ と が で き る 。
提案手法 (general)	**野菜** も 使 わ れ る こ と が で き る 。
提案手法 (in-domain)	**ブロッコリー** も 使 わ れ る こ と が で き る 。
対訳コーパス	**ブロッコリー** を 使っ て も O K です 。
参照訳	**ブロッコリー** バージョン も 美味 。

入力文	preheat the oven to 200 ° c .
ベースライン	オーブン は 200 度 に **preheat** 。
提案手法 (general)	オーブン は 200 度 に **加熱** 。
提案手法 (in-domain)	オーブン は 200 度 に **予熱** 。
対訳コーパス	オーブン を 200 ℃ に **予熱** し て おき ます 。
参照訳	オーブン を 200 度 に **予熱** し て おく 。

(general) では多くの翻訳が関連語や上位語となっている（“rum” → “砂糖” や，“broccoli” → “野菜” など）一方で，**提案手法 (in-domain)** ではより正解に近い訳語（“rum” → “ラム酒”，“broccoli” → “ブロッコリー” など）を出力できていることがわかる．このことからも，分布表現を学習する際のコーパスはテストデータと同じドメインのものを利用する方が有効であると言える．

6 おわりに

本論文では，統計的機械翻訳の翻訳モデルのドメイン適応を行うため，単語の分布表現を用いた手法を提案した．提案手法は，教師あり学習に基づき分布表現を翻訳することにより訳語を探し，自動的に未知語の翻訳候補と翻訳確率を与える．

実験では京都関連記事ドメインから料理レシピドメインへのドメイン適応を行い，日英翻訳ではベースラインから BLEU +1.7，英日翻訳では BLEU +0.3 の結果が得られた．

今後の課題としては，(1) 未知語以外の語のベクトル翻訳によるドメイン適応，(2) 分布表現の翻訳から句ベクトルの翻訳への拡張，(3) 2 節で述べた関連研究との詳細な比較が挙げられる．

謝辞

本研究の一部は JSPS 科研費 25280111 の助成を受けたものです．また，本研究で実験に用いた料理レシピ対訳コーパスは，クックパッド株式会社の原島純氏よりご提供頂きました．心より感謝致します．

参考文献

[1] Marta R Costa-Jussà. Domain adaptation strategies in statistical machine translation: a brief overview. *The Knowledge Engineering Review*, Vol. 30, No. 05, pp. 514–520, 2015.

[2] Prashant Mathur, Fondazione Bruno Keseler, Sriram Venkatapathy, and Nicola Cancedda. Fast domain adaptation of SMT models without in-domain parallel data. In *Proceedings of COLING*, pp. 1114–1123, 2014.

[3] Hirofumi Yamamoto and Eiichiro Sumita. Bilingual cluster based models for statistical machine translation. *Proceedings of EMNLP-CoNLL 2007*, pp. 514–523, 2007.

[4] Hua Wu, Haifeng Wang, and Chengqing Zong. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of COLING*, pp. 993–1000, 2008.

[5] Zellig S. Harris. Distributional structure. *Word*, Vol. 10, pp. 146–162, 1954.

[6] John R. Firth. A synopsis of linguistic theory. *Studies in Linguistic Analysis*, pp. 1–32, 1957.

[7] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint*, 2013.

[8] Shonosuke Ishiwatari, Nobuhiro Kaji, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. Accurate cross-lingual projection between count-based word vectors by exploiting translatable context pairs. In *Proceedings of CoNLL*, pp. 300–304, 2015.

[9] Philipp Koehn and Barry Haddow. Interpolated backoff for factored translation models. In *Proceedings of AMTA*, 2012.

[10] Graham Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pp. 311–318, 2002.