制約抽出のための対訳コーパスを用いた半教師ありクロスリンガル用語推定

澤田 悠治 [1]　　小田 悠介 [2,3]
[1] 奈良先端科学技術大学院大学 情報科学領域
[2] LegalForce Research　　[3] 東北大学 データ駆動科学・AI 教育研究センター
sawada.yuya.sr7@is.naist.jp　　yusuke.oda@legalforce.co.jp

概要

機械翻訳では，専門用語や固有名詞に対しては曖昧性の少ない適切な翻訳が求められる．推論時に単語選択を強制する手法が研究されているが，このとき与える語彙は人手によるため，強制する必要のある語彙を自動獲得する手法の開発が望まれる．本研究では，機械翻訳モデルに与える制約を抽出するための用語推定モデルを考え，対訳コーパスの大量のラベルなしデータと少量のシードデータを用いた半教師あり固有表現認識モデルを提案する．ASPEC の日英翻訳データを日英の固有表現認識コーパスとして使った評価実験により，半教師ありモデルによって英語の用語抽出精度が向上すること，単語アライメントを素性に加えることでさらに精度が向上することを確認した．

1　はじめに

機械翻訳の性能は，BLEU などの訳出全体の傾向を測定する指標において向上し続けていることが知られている [1]．一方で，訳抜けや翻訳内容の曖昧さといった，局所的な翻訳の制御に関する問題が依然として残っている．特に，専門用語や固有名詞に対しては曖昧性の少ない適切な翻訳が求められ，特定語彙の翻訳内容を考慮した手法が提案されている [2, 3, 4]．これらの手法では，特定の語彙を目的言語側の訳出時の制約として使用し，制約に含まれる単語を必ず出力するようモデルが制御されている．しかし，制約として使用する語彙は予め人手で作成する必要があり，翻訳言語対ごとの制約の作成は膨大な人手コストを要するため現実的ではない．また，専門用語は日々追加されるため，人手による辞書の使用にはメンテナンスのコストが必要となる．

そこで本研究では，機械翻訳モデルに与える制約を自動的に抽出するための固有表現認識を応用した用語推定モデルを提案する．従来法の固有表現認識モデルは教師あり学習の設定に基づいており，1 万から 10 万文程度の教師データを使用する．しかし，対訳コーパスは各言語で 100 万文程度の規模からなり，既存の固有表現認識コーパスのように人手で用語をアノテーションすることは困難である．本研究では，そのようなアノテーションを持たない対訳コーパスからの用語の抽出を想定し，少量の教師データと固有表現認識を組み合わせた半教師あり学習の枠組みを提案する．具体的には，単言語の固有表現認識モデルとして隠れマルコフモデルに基づく手法を使用し，単語アライメントの結果に基づいてモデルのラベル遷移を制限する機構を導入する．科学技術論文の対訳コーパスである ASPEC[5] の日英翻訳データを用いた実験では，提案手法が両言語で教師あり学習モデルと比較可能な精度で抽出可能であることが分かった．

2　用語推定モデル

本研究で用いる手法の概要を図 1 に示す．本手法は，原言語と目的言語それぞれの固有表現認識モデルと単語アライメントモデルの二つのモジュールで構成されており，以下のステップに従って用語を推定する．

1. 初期パラメータの設定
2. 単言語の半教師あり固有表現認識モデルの学習
3. 固有表現認識モデル同士の単語アライメント
4. 単語アライメントを導入したモデルの再学習

固有表現認識モデル [6, 7] と単語アライメントモデル [8] はどちらも bigram の隠れマルコフモデル（以下 HMM）による推定を行う．半教師あり固有表現認識モデル（ステップ 1,2）と単語アライメント機構の導入（ステップ 3,4）についてそれぞれ説明する．

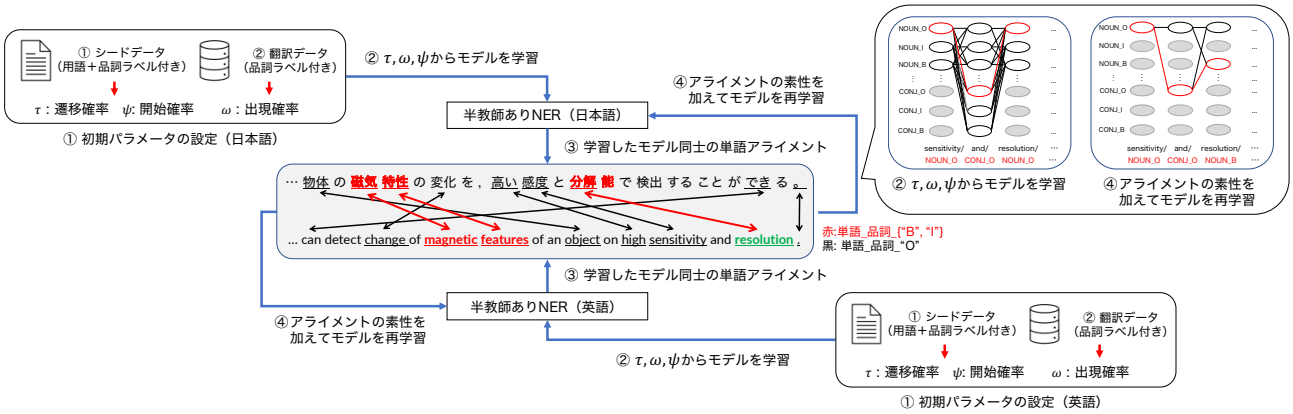

図 1　提案手法の概要図

2.1　半教師あり固有表現認識

半教師あり固有表現認識モデルとして用いる bigram HMM では，語数 N からなる単語系列 $w_{1:N}$ とラベル系列 $z_{1:N}$ に対して，HMM の遷移確率 τ と出現確率 ω，開始確率 ψ をそれぞれ以下のように定義する．

$$\begin{aligned} z_i|z_{i-1}=t,\tau^{(z)} &\sim \mathrm{Mult}(\tau^{(z)}) \\ w_i|z_i=t,\omega^{(z)} &\sim \mathrm{Mult}(\omega^{(z)}) \\ z_1=t,\psi^{(z)} &\sim \mathrm{Mult}(\psi^{(z)}) \end{aligned} \tag{1}$$

ここで，z_i と w_i は各時点 i（$1 \le i \le N$）のラベルと単語を表し，Mult(·) は多項分布を表す．本研究では，HMM の潜在変数 z_i を品詞と用語ラベル（BIO 方式）のタプルとし，用語と品詞がラベルづけされたシードデータと品詞のみがラベルづけされた対訳コーパスから品詞と用語ラベルを同時に推定する．用語ラベルは数千文程度の規模であれば比較的容易に作成可能であり，品詞ラベルは Spacy[1] や Sudachi[9] といった既存ツールから高精度に予測可能である．各パラメータにはデータの特徴を反映させるために，事前分布をそれぞれ個別のアルゴリズムで設定した．まず，各単語がどの品詞となりうるかは用語かどうかを推定する手がかりになるため，品詞と用語ラベルのタプルの遷移確率と開始確率をシードデータ上の数え上げを元に設定する[2]．出現確率については，対訳コーパス中の全ての語彙に対して事前確率を求めるため，各単語でなりうる品詞と用語ラベルのタプルに対して一様に確率を付与する．モデルのパラメータは EM アルゴリズムで推定し，事後分布による推論は Viterbi アルゴリズムを使用する．

2.2　単語アライメントを考慮した HMM

対訳文では，一方の言語で固有表現の意味に曖昧性があっても，片方の言語では意味ごとに異なる固有表現として明記される場合がある [10, 11]．例えば “本（Ben）” は中国語で稀に外国人の名前を翻訳する際に使用されるが，英語では人名である場合は “Ben”，それ以外の意味では別の単語が用いられる．本研究はこのような対訳文の用語の記述の違いを利用し，単語アライメントを素性として使用する．また，対訳文のアライメントでは，同じ品詞や同じ用語ラベルであるかはアライメントの手がかりになるため，単語系列と予測したラベル系列を組み合わせた（単語，品詞，用語ラベル）の三つ組のタプルからなる系列を入力としてモデルを学習する．具体的には，語数 N からなる原言語側のタプルの系列 $z^{src}_{1:N}$ と語数 M からなる目的言語側のタプルの系列 $z^{tgt}_{1:M}$ から対応関係にあるペア（$z^{src}_i(1 \le i \le N)$, $z^{tgt}_j(1 \le j \le M)$）を予測し，以下の式よりアライメント行列 $\gamma_i(t)$ を作成する．

$$\forall i,t\gamma_i(t) = \begin{cases} 1 & \text{if } \gamma_i(t) = z^{src}_i \\ 1 & \text{if } \gamma_i(t) = z^{tgt}_j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

ここで，$\gamma_i(t) = 1$ は $\gamma_i(t)$ が単語ペアの各単語で予測された用語ラベルであることを表す．単語アライメントモデルは教師なしの単語アライメントツールである GIZA++[3] を使用し，HMM の各パラメータは IBM Model 1 の lexical translation probability を初期化

1）https://spacy.io/
2）O ラベルから I ラベルへの遷移や z_0 で I ラベルとなるような無効なパターンはシードデータで出現しないため，これらのパターンが起きる確率は 0 になる．
3）https://github.com/moses-smt/giza-pp

	英語				日本語			
	学習用	シード	開発用	テスト用	学習用	シード	開発用	テスト用
文数	1.0M	1,790	1,784	1,812	1.0M	1,790	1,784	1,812
単語数	25.9M	44.9k	44.6k	45.0k	29.6M	51.1k	51.1k	51.0k
用語数	-	5,115	5,780	5,886	-	5,205	5,850	5,886
平均単語数	25.91	25.09	25.02	24.86	29.63	28.54	28.62	28.09
平均用語出現数	-	2.86	3.24	3.25	-	2.91	3.28	3.3
用語の平均単語数	-	2.54	2.46	2.51	-	2.66	2.62	2.61
用語を含む文の比率	-	91.7%	92.6%	95.4%	-	91.7%	92.6%	95.4%

表 1　WAT2021 Restricted Translation Task から作成した固有表現認識データセット

Algorithm 1 Alignment-aware EM

Input: $w_{1:N}, \tau, \omega, \psi, I$

Output: τ, ω, ψ

1: **for** 1 to I **do**
2: Compute forward probabilities: $\forall i, t\ \alpha_i(t)$
3: Compute backward probabilities: $\forall i, t\ \beta_i(t)$
4: Compute binary alignments: $\forall i, t\ \gamma_i(t)$
5: Compute masked forward-backward probabilities:
6: $\hat{\alpha}_i(t) = \alpha_i(t) \odot \gamma_i(t)$
7: $\hat{\beta}_i(t) = \beta_i(t) \odot \gamma_i(t)$
8: Compute posteriors:
9: $p(z_t = i | w_{1:N}, \tau, \omega, \psi) \propto \hat{\alpha}_i(t)\hat{\beta}_i(t)$
10: $p(z_t = i, z_{t+1} = j | w_{1:N}, \tau, \omega, \psi)$
11: $\propto \hat{\alpha}_i(t) p(z_{t+1} = j | z_t = i)$
12: $\times \hat{\beta}_j(t+1) p(w_{t+1} | z_{t+1} = j)$
13: **end for**
14: **return** τ, ω, ψ

に使用した[4]．

単語アライメントモデルから作成したアライメント行列を用いて，EM アルゴリズムにアライメントの素性を導入する．アライメントの素性を考慮に入れた EM アルゴリズムを Algorithm1 に示す．ここで，品詞の集合を T とすると，t は品詞の集合 T と用語ラベルの集合の直積の要素 ($t \in T \times \{B, I, O\}$) である．2.1 節と同様に，シードデータから取得した HMM の初期パラメータ τ, ω, ψ に対して，語数 N で構成された文 $w_{1:N}$ の前向き確率 $\alpha_i(t)$ と後向き確率 $\beta_i(t)$ を計算する（2–3 行目）．そして，それぞれの確率に対してアライメント行列 $\gamma_i(t)$ とのアダマール積をとり，アライメントからなり得ない用語ラベルに対しての出現確率をゼロとした確率の近似値 $\hat{\alpha}_i(t)$，$\hat{\beta}_i(t)$ を計算する（6–7 行目）．これらの近似値を元に事後確率を計算し，全ての文に対して実行した事後確率の平均を次回のパラメータとして更新する．

4）　GIZA++ではタプルの系列をそのまま扱えないため，タプルの各要素をアンダースコア（“_”）で連結した文字列を 1 単語とした単語系列を入力として学習を行なった．

3　評価実験

3.1　データセット

本研究では，実際に対訳コーパスを使用した制約抽出の評価を行うため，WAT2021 Restricted Translation Task[1] の評価データを使用した．本タスクでは，ASPEC の用語や固有名詞に対する翻訳性能の評価に焦点を当て，翻訳時の制約として用いる用語が各文でまとめられている．そこで，ASPEC の用語リストが作成されている dev，devtest，test ファイルから固有表現認識データセットを作成し，それぞれシード，開発用，テスト用として使用する．作成したデータセットの基本情報を表 1 にまとめる．学習用データとして用いるラベルなしデータセットは ASPEC の train-1 ファイルのみを使用し，train-1, dev, devtest, test ファイルそれぞれに対して品詞情報を付与した．日本語の単語分割と品詞タグ付けは Sudachi[5]，英語の単語分割と品詞タグ付けは Moses tokenizer[6] と Spacy をそれぞれ使用した．

3.2　評価方法

日英の全ての用語に対する始点と終点の一致について，適合率・再現率及びそれらの調和平均（F 値）で評価する．シードのみを使用して学習した教師あり HMM をベースラインとし，用語ラベルなしデータ（train-1）を用いて学習した単言語モデル（アライメントなし）および両言語のアライメントを素性

5）　単語分割にはモード A を使用した．
6）　https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

	英語			日本語		
モデル	Precision	Recall	F1	Precision	Recall	F1
ベースライン	54.8	40.8	46.8	**64.9**	54.4	**59.2**
アライメントなし	55.1	46.6	50.5	53.9	52.1	53.0
+ アライメント (再学習 1 回目)	57.9	49.3	53.3	54.0	53.8	53.9
+ アライメント (再学習 2 回目)	**59.0**	**50.0**	**54.1**	55.0	**56.0**	55.5

表 2　WAT2021 Restricted Translation Task による実験結果

入力文	3He と 3H の高運動領域における構造を調べるために， 270MeV での標記反応のテンソル及びベクトル偏極分解能を 全角度領域にわたって測定した。
正解	高運動領域, テンソル, ベクトル, 偏極分解能
ベースライン	3He, 3H, 高運動領域, 270MeV, ベクトル偏極分解能, 全角度領域
+ アライメント (再学習 2 回目)	3He, 3H, 高運動領域, 270MeV, 200MeV, 標記反応のテンソル及びベクトル偏極分解能, 全角度領域

表 3　提案手法による出力例

に加えたモデルと比較する．また，初期の学習段階では固有表現認識モデルの誤りからアライメントにも誤りが生じると考えられるため，固有表現認識モデルとアライメントの学習を交互に 2 回繰り返し，再学習した結果についても示す．ベースライン，提案手法のどちらも開発データで最も高い F 値を示した 4 エポック目のモデルを採用した．

3.3　実験結果

実験結果を表 2 に示す．英語では，用語ラベルなしデータを加えて学習すると F 値が 3.7 ポイント，アライメントの素性を加えるとさらに 2.8 ポイント増加した．ASPEC の日英翻訳データは日本語をピボットとして作成されており，英語では表現を簡略化して訳される事例が存在する．例えば，“低侵襲で機能障害が非常に少ない” という文が英語では “simple and postoperative function can be maintained” と訳されており，日本語と英語で文章のニュアンスが異なる．また，“口腔癌患者” が “patients” と訳されるなど，一部の単語が省略されることで用語か一般的な記述かが曖昧になる事例も存在する．このような事例に対して，用語ラベルなしの学習データとアライメントの素性を加えることで，各単語が用語かどうかの曖昧さが改善されたと考えられる．日本語を対象にした設定では，提案手法はベースラインを上回らなかったものの，英語と同様にアライメントの素性を加えることで F 値が 2.5 ポイント増加し，再現率についてはベースラインと比べて 1.6 ポイント上回った．

実際に各モデルから出力された用語の例を表 3 に示す．アライメントの素性を加えて再学習したモデルでは，“200MeV” や “全角度領域” のような制約以外の用語らしい単語列が抽出されている．ASPEC には WAT2021 Restricted Translation Task の制約リストに入っていない用語があり，ベースラインでは制約リストに入った用語のみが抽出されるよう最適化されている傾向が見られる．提案手法は “標記反応のテンソル及びベクトル偏極分解能” のように用語の範囲同定には改善の余地があるものの，用語らしい単語列を抽出する上ではベースラインより好ましい結果が得られたと考えられる．

4　まとめ

機械翻訳に与える語彙制約を自動抽出するための用語推定モデルを提案した．具体的には，対訳コーパスを用いた半教師あり学習による固有表現認識の枠組みと，HMM を用いた固有表現認識モデル上で単語アライメントによりラベル遷移を制限する機構を提案した．ASPEC 日英翻訳データを用いた実験の結果，英語で教師あり学習による HMM を上回る精度を示し，日本語でも正解以外の用語と思われる単語が制約として抽出される傾向が見られた．今後の課題としては，用語の範囲同定の精度改善や既存の制約リストの高品質化が挙げられる．本論文で採用したモデルよりも高性能な手法の適用も考えられ [12][13]，これらによる性能の改善にも取り組む．

参考文献

[1] Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pp. 1–45, 8 2021.

[2] Guanhua Chen, Yun Chen, Yong Wang, and Victor O.K. Li. Lexical-constraint-aware neural machine translation via data augmentation. In *Proceedings of IJCAI 2020: Main track*, pp. 3587–3593, 7 2020.

[3] Matt Post and David Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1314–1324, New Orleans, Louisiana, 6 2018.

[4] Katsuki Chousa and Makoto Morishita. Input augmentation improves constrained beam search for neural machine translation: NTT at WAT 2021. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pp. 53–61, 8 2021.

[5] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2204–2208, 5 2016.

[6] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, 6 2016.

[7] Jason P.C. Chiu and Eric Nichols. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 357–370, 2016.

[8] Stephan Vogel, Hermann Ney, and Christoph Tillmann. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, 8 1996.

[9] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: a Japanese tokenizer for business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 5 2018.

[10] Mengqiu Wang, Wanxiang Che, and Christopher D. Manning. Joint word alignment and bilingual named entity recognition using dual decomposition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1082, 8 2013.

[11] Wanxiang Che, Mengqiu Wang, Christopher D. Manning, and Ting Liu. Named entity recognition with bilingual constraints. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 52–62, 6 2013.

[12] Ke M. Tran, Yonatan Bisk, Ashish Vaswani, Daniel Marcu, and Kevin Knight. Unsupervised neural hidden Markov models. In *Proceedings of the Workshop on Structured Prediction for NLP*, pp. 63–71, 11 2016.

[13] Ying Luo, Hai Zhao, and Junlang Zhan. Named entity recognition only from word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8995–9005, 11 2020.