翻訳精度に基づく単語クラス自動推定手法

安田 圭志、高井 公一、服部 元、イラクレウス パニコス、石川 彰夫、松本 一則、菅谷 史昭
株式会社 KDDI 総合研究所
{ke-yasuda, ko-takai, ge-hattori, pa-heracleous, ao-ishikawa, matsu, fsugaya}@kddi-research.jp

1 はじめに

近年における音声言語処理の技術の発展に伴い、音声翻訳システムは旅行者にとって現実的なものとなりつつある。その中で観光者が訪れるスポット、ランドマーク、飲食店、宿泊施設などの固有名詞数多く存在し、そのカバレージが、翻訳システムの精度に影響を及ぼすことが分かっている。

本論文では、このような問題を解決するため、クラス言語モデルに基づく機械翻訳システムに対する、クラス自動推定手法について提案する。

提案手法では、固有名詞を含む対訳コーパスを用意し、複数存在する固有名詞クラスの中で最も翻訳性能が高くなるクラスを翻訳自動評価により決定して学習データを作成する。次に得られたクラスと固有名詞を含む文から、畳み込みニューラルネットワーク(CNN)により、単語クラス推定モデルを学習する。生成した CNN モデルは固有名詞を含む文を入力とし固有名詞クラスを推定することができる。このように提案手法では、人手による固有名詞クラスのアノテーションをすることなしに、単語クラス推定モデルの学習を可能とする手法である。

2 関連研究

単語クラスの利用は、音声認識の分野で、データスパースネスの問題を解決するために用いられてきた。この考えは、単語クラス付き対訳辞書を用いる手法[1]などにより、統計的機械翻訳(SMT)などのコーパスパース機械翻訳にも取り入れられている。通常、このような対訳辞書へのクラス付与は、人手により行われている。飲食店や商品名といった固有名詞が日々新出する中で、固有名詞の辞書整備のコストが、実用において大きな課題となっている。

関連研究としては、翻訳対の辞書情報を獲得する方法として、対訳辞書を扱った手法[2]や、翻字を使った手法[3-5]など、いくつかの方法が提案されている。本論文では、対訳表現は上記手法などで抽出した後の、単語クラス自動付与に関する研究である。

単語のクラス推定については、固有表現抽出(NER)や機械翻訳における研究で、様々な方法が提案されている [6,7]。しかしながら、その多くは人手によりアノテーションされた学習データを用いて、教師あり学習を行なう方法である。

図1に示す通り、提案方法においても、教師あり学習を用いるが、学習データを自動的に構築するために、人手によるアノテーション作業が不要である。

3 提案手法

図2に示す通り、提案手法は、データ構築処理部と固有名詞クラスタ推定モデルの学習処理部からなる。ココでは, これらの処理について説明する。

3.1 データ構築

図2に、データ構築部の処理の流れを示す。ここでは、固有名詞を含む対訳コーパスを用意しておき、以下の手順にて学習データを構築する。

1. 対訳コーパスから両言語に固有名詞が1つづつ含まれている対訳文対を抽出する。
2. 抽出した文から固有名詞クラスの一つを選択し、固有名詞を翻訳システムの辞書に登録する。
3. 1の原言語文を上記の2の辞書を用いた機械翻訳システムで翻訳する
4. 2で登録した固有名詞を辞書から取り除く
5. 上記の2から4をすべての固有名詞クラスに対して行う。

上記の手順で得られた固有名詞クラスごとの翻訳結果を、参照訳(上記1の目的言語文)を用いて自動評価し、次式により、最適な固有名詞クラス($\hat{c}$)を得る。

$$\hat{c} = \mathrm{argmax}_{c \in C} S_{RIBES}(T_{REF}, T_{MT}^{c}) \quad (1)$$

ここで、C, T_{REF}, T_{MT}^{c}はそれぞれ、固有名詞クラスの集合、対訳コーパス中の目的言語文、固有名詞クラスcとして登録した辞書を用いた翻訳システムによる翻訳結果である。また、S_{RIBES}はT_{REF}とT_{MT}^{c}の間の RIBES スコア[8]で、次式により計算される。

$$S_{RIBES}(T_{REF}, T_{MT}^{c}) = R_{cor}(T_{REF}, T_{MT}^{c}) \times (l_{com}/l_{MT})^{\alpha} \quad (2)$$

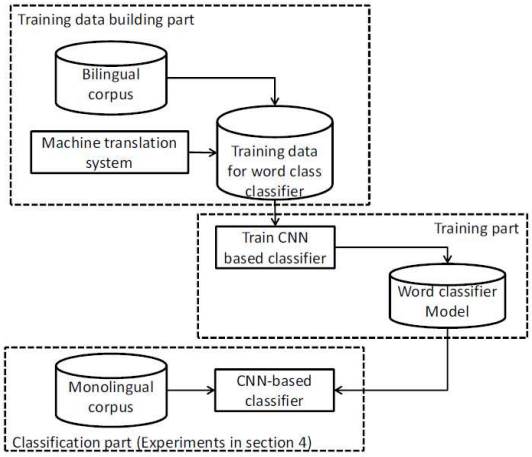

図1 提案手法の処理概要

ここで、l_{MT}, l_{com}, R_{cor}は、それぞれT_{MT}^{c}の総単語数、T_{REF}, T_{MT}^{c}の共通単語数、共通単語間の順位相関係数を示す。またαは$(0 \leq \alpha \leq 1)$ペナルティ値に関するハイパーパラメータである。

最終的には、対訳コーパスの原言語文と、$\hat{c}$の対を学習データとし利用する。

3.2 CNNによる固有名詞クラス推定

本節では、前述の手法により作成した学習データから固有名詞のクラス推定モデルの学習手法について説明する。固有名詞のクラス推定モデルには、CNN を用いる。CNN は画像処理や音声処理[9,10]の分野で高い性能を発揮しているが、近年では、分散表現化[13]した単語を入力することで、テキスト分類[11,12]のような自然言語処理の分野でも応用されている

図3は、固有名詞クラス推定のニューラルネットワークの構成図である。

ここで、$x_i (\in R^k)$は文中のi番目の単語の分散表現で、n 単語からなる文は下記の式で表現できる。

$$x_{1:n} = x_1 \cdots x_i \cdots x_n \tag{3}$$

特徴量として用いる n-gram 長（またはフィルタサイズ）h, j とし、h から j までの特徴量を下記の式で示す

$$c_{h,j,i} = \tanh\left(\mathrm{w}_{h,j} \cdot x_{i:i+h-1} + b_{h,j}\right) \tag{4}$$

ここで、$\mathrm{w}_{h,j}$と$b_{h,j}$はそれぞれフィルタに対する重みとバイアス項である。n-gram 長ごとの式(4)に対し、式(5)の Max pooling 層では $c_{h,j}$の全要素から最も高い値を選択する。

$$c_{h,j} = [c_{h,j,1}, c_{h,j,2}, \cdots c_{h,j,n-h+1}] \tag{5}$$

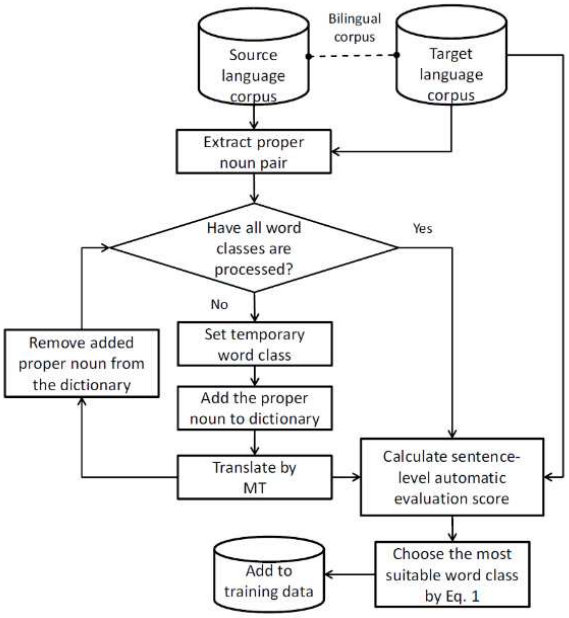

図2 データ構築処理

$$\hat{c}_{h,j} = \max_{i=1:n-h+1} c_{h,j} \tag{6}$$

最後に出力層を結合し、ソフトマックスを用いてクラスの確率$(\hat{y} \in R^{n_c})$を得る。

$$\hat{y} = \frac{\exp(z_q)}{\sum_{p=1}^{n_c} \exp(z_p)}, q = 1, \cdots, n_c \tag{7}$$

n_cはクラスの総数で$z(\in R^{n_c})$は出力層の正規化前の値である。

4 実験

4.1 評価手法

実験では、日本語と英語の対訳コーパスから学習データを構築し、これを用いて CNN によるクラス推定器を学習する。固有名詞推定の評価としては、下記の条件で作成された固有名詞辞書を用いた日英方向 SMT による機械翻訳を行い、翻訳性能を比較した。

条件1　人手により単語クラスを付与
条件2　ランダムにクラスを決定
条件3　表1に示す事前分布に基づき、ランダムにクラスを決定
条件4　提案手法によりクラスを付与

なお、翻訳性能の評価には RIBES[8]を用い、条件4においては、10 分割交差検定を行っている。

表1 データセットの固有名詞クラスの詳細

Category	% in the data set
Accommodation	10.33
Attraction	4.9
Building	8.44
Country name	12.15
Foreign First Name	5.57
Foreign Last Name	3.73
Food	7.84
Japanese First Name	4.35
Japanese Last Name	4.17
Land Mark	11.73
Organization	9.29
Shop	4.75
Souvenir	6.54
Others	6.2
Total	100

4.2 実験条件

実験には、Basic Travel Expression Corpus（BTEC）[14]の日英部を使用した。既存の対訳辞書を用いて、固有名詞を1つだけ含む 5,471 文対を BTEC から抽出した。表1は学習データとなる固有名詞クラスの詳細とコーパス内における出現割合である。固有名詞クラスは旅行ドメインで利用する 14 クラスを用いた。データセットの中で出現割合が最も多いクラスは国名で、最も少ないクラスは外国人の姓であった。

CNN によるクラス推定モデルの学習のため、Word2Vec[13]で単語分散表現を事前に学習した。分散表現は Wikipedia のコーパスから学習した。また、CNN 学習時においては、分散表現部分は固定し、チューニングは行っていない。表2と3に、実験に用いたコーパスの詳細と、CNN によるクラス推定で用いたハイパーパラメータ等の設定を示す。

4.3 実験結果

図4は、各種辞書を用いた SMT による翻訳結果に対する自動評価結果である。縦軸は、自動評価で用いた RIBES のスコアである。全ての条件において、基盤となる SMT と、対訳辞書のエントリー同じであり、クラスの付与方法のみが異なる。条件2、3はランダムにクラスを決定するため、10 回の試行の平均値と、標準偏差をエラーバーで表わしている。一方で、条件4は 10 分割交差検定で行ってはいるが、実験は 1 度だけとした。

表2 実験で使用したコーパスの統計情報

Corpus type	# of words	Lexicon size
BTEC（Japanese side）	83,942	9,266
Wikipedia corpus	10,363,151	116,556

表3 CNN パラメータ

Parameters	Setting
Maximum length of input sentence	150 words
Mini batch size	64
Dimension of word-embedding vector (k)	100
Filter window size (n-gram length)	3 to 5-gram
Number of filters for each window size	128
Drop out rate for fully connected layer	0.5
Optimizer	Adam optimizer
# of output units	14

図4で示す通り、提案手法によりクラス付与を行なった辞書は、二つランダム条件よりも高い訳質となっており、さらに人手によるアノテーションをも上回っている。

人手による固有名詞クラスアノテーションよりも 精度が上回る原因について以下に考察する。人手によるアノテーション時、単語単体からクラスを決定しているが、提案手法では、固有名詞を含む文全体からクラスを推定している。このように提案手法で用いるコンテキストの情報が、多義性ある単語などに対いて有利に働いていると考えられる。

5 まとめと今後の課題

本論文では、クラスベースの翻訳システムへの利用を目的とし、固有名詞クラスを自動的に付与する方法を提案した。本手法では、対訳コーパスと翻訳システムとを用いて、学習データを構築し、次に構築したデータから CNN による固有名詞クラス推定モデルを学習した。

評価実験では、このモデルから自動的にテストセットの固有名詞を推定し、最後に推定したクラスを SMT に登録し、得られた翻訳結果を自動評価した。実験結果によると、提案手法によるクラス推定により、人手でアノテーションしたものと同等以上の翻訳性能が得られることが確認できた。

これらの結果から、提案手法は、人手による学習データの作成なしに、クラスベース翻訳システムの固有名詞辞書拡張を可能にしたと言える。

今後の検討課題として、コーパスサイズを増やしてクラス推定の精度を改善する予定である。また、本論文では BTEC コーパスに限定した実験となったが、今後フィールドデータを用いた効果検証を行っていきたい。

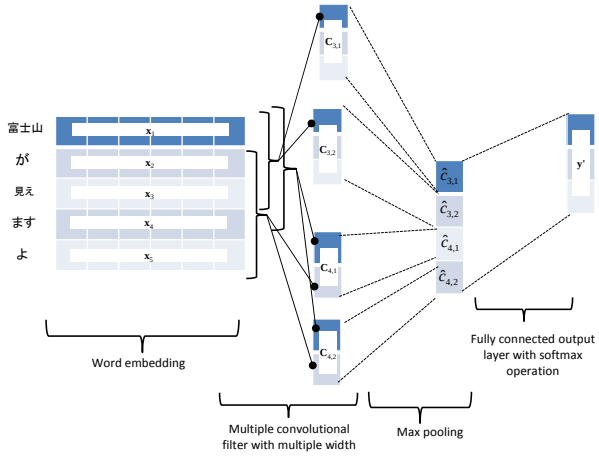

図3 固有名詞クラス推定の CNN の構成

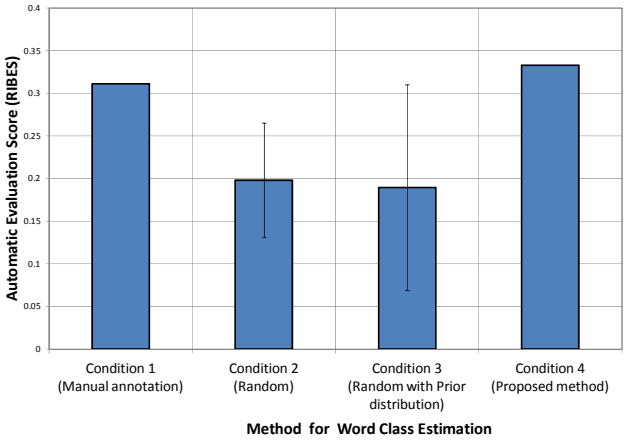

図4 RIBES による翻訳自動評価結果

謝辞

本研究は、総務省「グローバルコミュニケーション計画の推進 -多言語音声翻訳技術の研究開発及び社会実証- Ⅰ.多言語音声翻訳技術の研究開発」の一環として実施したものです。

参考文献

[1] Okuma, H. et al. (2008). “Introducing a translation dictionary into phrase-based SMT”. Trans. of IEICE, Inf. & Sys., 91-D, pp.2051–2057.

[2] Tonoike, M. et al. (2005). “Translation Estimation for Technical Terms using Corpus collected from the Web” . Proc. of the PACLING, pp.325–331.

[3] Al-Onaizan, Y. et al. (2002). “Translating named entities using monolingual and bilingual resources”. Proc. ACL, pp.400–408.

[4] Sato, S. et al. (2009). “ Web-Based Transliteration of Person Names”. Proc. of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp.273–278.

[5] Finch, A. et al. (2011). “ Integrating a joint source channel model into a phrase-based transliteration system”. Proc. of NEWS2011, pp. 23–27.

[6] Ma, X. et al. (2016). “End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF ”. Proc. of ACL, pp. 1064–1074.

[7] Yasuda, K.. et al. (2017). “Building a location dependent dictionary for speech translation systems”. Proc of CICLING.

[8] Isozaki, H. et al. (2010). “Automatic evaluation of translation quality for distant language pairs. Proc. of EMNLP,pp. 944–952.

[9] Krizhevsky, A. et al. (2012). “Imagenet classification with deep convolutional neural networks” . Proc. of NIPS, pp. 1097–1105.

[10] Abdel-Hamid, O. et al. (2014). “Convolutional neural networks for speech recognition”. IEEE/ACM, pp. 1533–1545.

[11] Kim, Y. (2014). “Convolutional neural networks for sentence classification” . Proc. of EMNLP, pp. 1746–1751.

[12] Kalchbrenner, N. et al.(2014). “Convolutional neural networks for modeling sentences” . Proc. of COLING, pp. 655–665.

[13] Mikolov, T. et al.(2013). “Distributed representations of words and phrases and their compositionality”. Proc of NIPS, pp. 3111–3119.

[14] Kikui, G. et al. (2003). “Creating corpora for speech-to-speech translation”. Proc. of EUROSPEECH, pp. 381–382.