E-026

日英対訳データとニューラルネットワーク機械翻訳を利用した類語表現の抽出

Extraction Japanese Paraphrase from Bilingual Corpora Using Neural Machine Translation

徳原 生輝† 竹内 孔一† 村上 仁一‡ 徳久 雅人†

Seiki Tokuhara Kouichi Takeuchi Murakami Jinichi Masato Tokuhisa

1. はじめに

高度な意味処理を要求される自然言語処理の研究において類義語情報は有用である。しかし、人手で類義語の対応表を作成するのは非常にコストのかかる作業である。そこで、日本語コーパスから類義語情報を自動で獲得することが求められている。類義表現や言い換え表現を獲得する先行研究としては、分散意味論による Skip-gram を拡張した同義語獲得の研究がある[1]。萩原らは同義語であるかどうかの判定において文脈情報が有 用であることを示している[2]。また、Melvin は Zero-Shot Translation が可能なニューラルネットワークを構築したことを発表している。Zero-Shot Translation とは、明示的に学習していない 2 つの言語間での翻訳をすることであり、複数の言語の翻訳システムを 1 つのネットワークで学習させた際に潜在空間 (図 1) において類義表現が近くに分布されることを実験的に示した[3]。上記の特性を利用して翻訳情報から類義語を自動で獲得する手法を提案する。

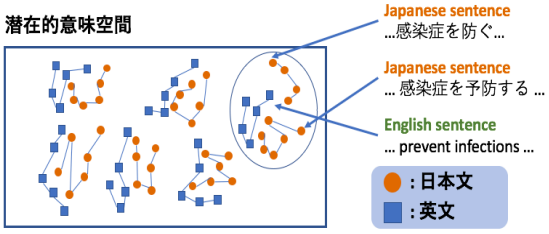

図 1: 潜在的意味空間

2. 類義語抽出アルゴリズム

本章では類義語抽出のアルゴリズムについて記述する。類義語抽出までのステップは大きく分けて以下のようになる。

1. 機械翻訳システムによる翻訳情報の学習
2. 1 で学習した翻訳モデルに類義語候補を入力
3. 類義語候補に対してモデルが生成した翻訳情報を抽出しベクトル化
4. ベクトルのコサイン類似度を比較

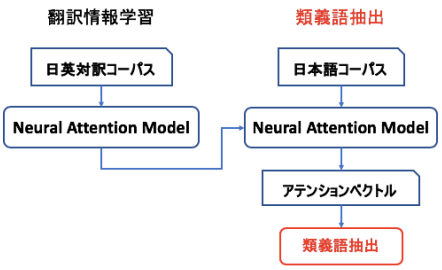

図 2: 類義語抽出の概要図

2.1 ニューラル機械翻訳

類義語抽出の前処理として事前に学習する Neural Attention Model による機械翻訳の概要について述べる。入力文の単語埋め込み表現列には nwjc2vec の分散表現ベクトル[4]を利用する。出力層では one-hot ベクトルで出力するようにする。翻訳モデルには、時系列情報を扱うことの出来る Encoder-Decoder モデルにアテンション機構[5]を適用したものを利用する。また、Encoder、Decoder それぞれの内部層には LSTM を利用する。

2.2 アテンション機構

本研究で使用しているアテンション機構の説明をしていく。入力文の j 番目の単語 X_j を入力した時の内部状態を h_j、出力文の t 番目の単語 y_t を出力する時の内部状態を s_t とおくと図 3 のようになる。

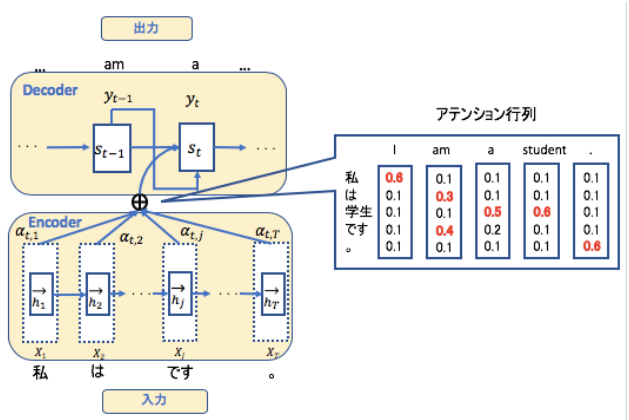

図 3: アテンション機構の概要図

ここで、図 3 の右に示す表はアテンション行列という行列で、式(3)で与えられる。

$$e_{tj} = v^{\mathrm{T}} \tanh(Ws_{t-1} + Uh_j) \quad (2)$$

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T} \exp(e_{tk})} \quad (3)$$

2.3 翻訳情報の抽出

類義語候補それぞれの翻訳情報の抽出方法について述べていく。図 3 の単語 t を出力する場合の式 (3) について考える。Ws_{t-1} と v は一定なので、単語 t を出力するのに単語 j がどれくらい重要視されたかは $U * h_j$ の値に依存する。つまり、似た意味を持つフレーズを翻訳モデルに入力した

場合のそれぞれの$U * h_j$の値は似ていると考えられる。そこで、日本語コーパスを入力した時のそれぞれの$U * h_j$の値をベクトルとして保存しコサイン類似度を計算し値の近い順により有力な類義語候補として選択する。

3. 評価実験

本章では類義語抽出の評価実験の設定及び実験結果について述べていく。

3.1 翻訳モデルの学習設定

翻訳モデルを学習するにあたって定めた学習の設定について述べていく。1 バッチは 100 事例で学習し、学習回数は 100 回に設定する。また、ロス関数には交差エントロピー誤差関数を利用する。対訳データには辞書から自動抽出した日英対訳データ[6] (約 11 万文)を利用する。

3.2 類義語抽出の実験設定

入力には鳥バンクの「意味類型パターン辞書」に使用されているに使日本語フレーズ約 6 万事例を用いる。出力する類義語候補はコサイン類似度上位 10 件に絞る。比較対象には nwjc2vec の分散表現ベクトルの平均値を利用する。また、鳥バンク内で同一の英語フレーズと対応関係にある日本語フレーズ群を正解データとする。accuracy は各入力に対して出力される類義語候補 10 フレーズのうちに正解データと一致するものがいくつあるかで算出する。

3.3 実験結果

類義語抽出の実験結果は表 1 のようになった。

手法	Accuracy
提案手法	8.9%
nwjc2vec	14.1%

表 1: 類義語抽出精度

類義語抽出の精度は提案手法が nwjc2vec より 5.2%下回る結果となった。

4. 考察

類義語抽出の結果を抽出精度と出力結果それぞれの観点から考察していく。

4.1 抽出精度における考察

類義語抽出の精度の観点から考察をしていく。nwjc2vec の方が accuracy の値が高くなったのは分散表現ベクトルが単語の共起頻度からベクトルを生成しているのに対して提案手法では翻訳情報からベクトルを生成するので翻訳モデルの精度によって類義語抽出の精度も決まるからだと考えられる。改善案としては、学習パラメータのチューニング、翻訳モデルの改善、対訳コーパスの変更などによる翻訳精度の向上が挙げられる。また、入力に使用している「意味類型パターン辞書」に登録されている日本語フレーズに 1 単語のものも多く登録されているので、nwjc2vec の分散表現ベクトルの平均値が実質 1 単語の分散表現ベクトルの値の比較になっているで、ノイズが入らない分 nwjc2vec の方が精度が高くなったと考えられる。

4.2 出力結果における考察

類義語抽出の出力結果の観点から考察をしていく。

手法	出力結果
nwjc2vec	この仕事を仕上げ, この仕事を完成する
提案手法	この仕事を仕上げ, この仕事をし, この仕事を片付け

表 2: 「この仕事を仕上げる」に対する出力結果

表 2 は「この仕事を仕上げる」を入力した場合の類義語抽出の結果の中で正解したもののリストであるが、提案手法では「この仕事を片付け」というフレーズが出力されている nwjc2vec の方では出力されていないので「仕上げる」に対する「片付け」という対応関係が学習出来ていると考えられる。これは翻訳情報から日本語の対応関係が取得出来たことを意味している。

5. おわりに

今回は日英機械翻訳モデルの翻訳情報をベクトル化して類義語抽出をする手法を提案した。結果として分散表現ベクトルの平均値に比べて精度は劣ることになったが、機械翻訳のパラメータのチューニング、モデルの変更、学習コーパスの変更などで翻訳精度の向上を実施すれば類義語抽出の精度向上にも繋がると考えられる。

謝辞

本研究を進めるに当たり、指導してくださいました竹内孔一先生、議論に参加してくださいました竹内研究室の諸氏に心より感謝いたします。鳥取大学の村上仁一先生、徳久雅人先生には、鳥バンクの利用の許諾を頂きました。心より感謝いたします。

使用したツール

- 鳥バンク, http://unicorn.ike.tottori-u.ac.jp/toribank/

参考文献

[1] 城光英彰, 松田源立, 山口和紀, 文脈限定 skip-gram による同義語獲得に関する研究, 言語処理学会第 22 回年次大会, (2016)
[2] Hagiwara Masato, Yasuhiro Ogawa, and Katsuhiko Toyama, Selection of effective contextual information for automatic synonym acquisition, *In Coling/ACL*, pp. 353-350, (2006)
[3] Melvin Johnson, Mike Schuster, QuocV.Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Vi´egas, Martin Wattenberg, Greg Corrado, Macduff Hubhes, Jeffrey Dean, Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, *CoRR*, (2016).
[4] 浅原正幸, 岡照晃, nwjc2vec: 「国語研日本語ウェブコーパス」に基づく単語の分散表現データ, 言語処理学会第 23 回年次大会, (2017)
[5] Dzmitry Bahdanau, KyungHyum Cho, and Yoshua Bengio, Neural Machine Translation By Jointly Learning To Align And Translate, *In Proceedings of ICLR*, (2015)
[6] 村上仁一, 藤波進, 日本語と英語の対訳文対の収集と著作権の考察, 第一回コーパス日本語学ワークショップ, pp. 119-130, (2012)

† 岡山大学大学院自然科学研究科　Okayama University, Graduate School of Natural Science and Technology
‡ 鳥取大学　Tottori University