大規模単言語コーパスと基本語対訳辞書を用いた専門用語の訳語獲得

石川 裕貴　　中澤 敏明　　黒橋 禎夫
京都大学大学院 情報学研究科
{ishikawa, nakazawa, kuro}@nlp.kuee.kyoto-u.ac.jp

1　はじめに

機械翻訳などにおいて、専門用語の訳語情報は非常に重要である。しかし、専門用語は日々新たに生成されるため、その対訳を人手で整備するのは困難である。本論文では、原言語および目的言語の大規模単言語コーパスと基本語対訳辞書を用いて、専門用語の訳語を自動的に獲得する手法を提案する。

単言語コーパスと対訳辞書を用いた専門用語の訳語獲得を行う既存研究において、外池ら [3] は専門用語を構成する単語の類似度を利用し、Fung と Yee [1] は専門用語が現れる文脈の類似度を利用している。これら既存の手法では、専門用語を構成する単語の辞書中の訳語を組み合わせて訳語候補を生成しているが、本手法では目的言語の単言語コーパスから訳語候補となりうる単語列を抽出するため、構成的でない訳語も獲得できる。また、文脈の類似度および専門用語を構成する単語の類似度を組み合わせて訳語を選択する。

提案手法の全体像を図 1 に示す。提案手法ではまず訳語候補生成部で、入力された専門用語から訳語候補を生成する。次に構成語類似度計算部、共起語類似度計算部でそれぞれの訳語候補と入力された専門用語との類似度を計算する。最後に、構成語類似度のスコアと共起語類似度のスコア双方を考慮して訳語選択を行う。なお、手法の説明においては入力を日本語、出力を英語とする。

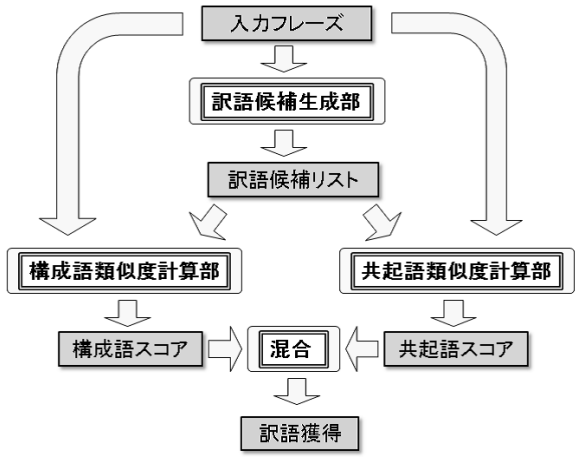

図 1: 提案手法の全体像

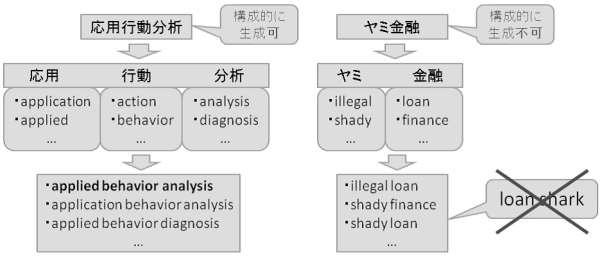

図 2: 構成的に生成不可能な専門用語

2　訳語候補の生成

2.1　構成的に生成不可能な専門用語

専門用語の訳語の獲得については様々な手法が提案されているが、訳語候補を生成する際には、構成要素の辞書訳を組み合わせて生成するのが一般的である [1, 3]。例えば、「応用行動分析」の訳語候補は、図 2 の左側のように、「応用」、「行動」、「分析」という構成要素それぞれの辞書訳を組み合わせて生成される。この場合は、“applied behavior analysis” という正しい訳が訳語候補に含まれるが、「ヤミ金融」のように構成要素「ヤミ」、「金融」の訳語の組合せでは正解の”loan shark” が生成できず、訳語候補に正解が含まれない専門用語もある。

「応用行動分析」のように、構成要素の辞書訳を組み合わせることによって正解が生成できる専門用語を**構成的に生成可能**と定義し、「ヤミ金融」のように構成要素の辞書訳を組み合わせても正解が生成できない専門用語を**構成的に生成不可能**と定義する。

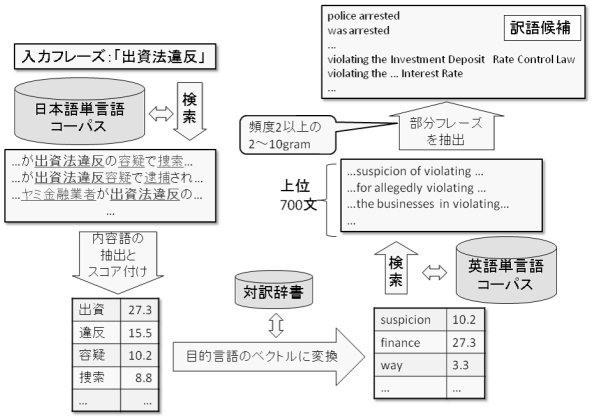

図 3: 訳語候補の生成

本研究では、構成的に生成不可能な専門用語であっても正しい訳語を獲得するために、構成要素の辞書訳を組み合わせて訳語候補を生成するのではなく、目的言語のコーパスから訳語候補を収集する。

2.2 訳語候補の生成

訳語候補生成の流れを図 3 に示す。訳語候補生成部では、入力の専門用語が出現する文集合を bag-of-words とみなし、そこに含まれる単語の対訳が多く出現する文を目的言語のコーパスから選べば正しい訳語が含まれている可能性が高い、という仮説に基づいて訳語候補を生成する。訳語候補生成部では、TFIDF をベースとした手法を用いる。

まず、入力の専門用語を含む文を原言語コーパスから検索する。検索結果に含まれる内容語 ST_j の頻度 STF_{ij} を計数し、原言語コーパス中の内容語の文書頻度 SDF_j を使ってそれぞれの内容語にスコアを与える。本研究では、1 文書を 1 文として扱うので、文書頻度 SDF_j はある内容語が出現する文の数となる。スコア $SourceScore_{ij}$ は次のように計算される。ここで、SN は原言語コーパス中に含まれる全文数とする。

$$SourceScore_{ij} = STF_{ij} \log \frac{SN}{SDF_j} \qquad (1)$$

次に、対訳辞書で訳語を調べ、それぞれの訳語にスコア $SourceScore_{ij}$ を与えることにより、目的言語の特徴ベクトルを作成する。作成した特徴ベクトルと、目的言語のコーパスから作成した転置インデックスとの内積をとることにより、目的言語のコーパスに含まれるそれぞれの文にスコアをつけ、スコアが上位の文を抽出する。目的言語の転置インデックスには、ある文 TD_k に内容語 TT_l が TTF_{kl} 回出現するとしたときに、以下の式で計算される $TargetIndexScore_{kl}$ を

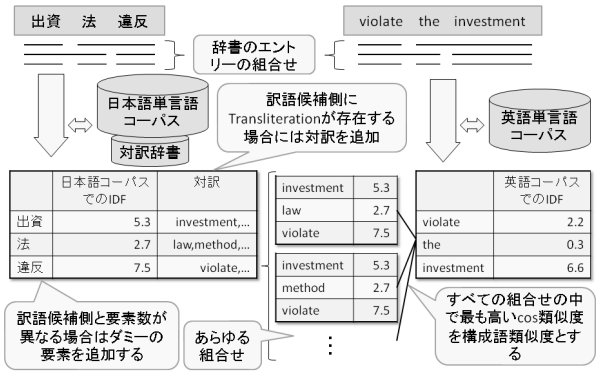

図 4: 構成語類似度の計算

与える。ここで、TN は単言語コーパス中に含まれる全文数とし、内容語 TT_l の文書頻度を TDF_l とする。

$$TargetIndexScore_{kl} = TTF_{kl} \log \frac{TN}{TDF_l} \qquad (2)$$

最後に、抽出された文に含まれるあらゆる部分単語列を収集し、付属語で終わっているものと、冠詞以外の付属語から始まるものを取り除いて訳語候補とする。

3 類似度計算と訳語選択

構成語類似度 [3] と共起語類似度 [1, 2] の 2 つの類似度を用いて訳語選択を行う。構成語類似度は、専門用語に含まれる構成要素どうしの類似度であり、共起語類似度は専門用語と共起する内容語どうしの類似度である。それぞれの訳語候補について上記 2 種類の類似度を考慮したスコアを求め、もっともスコアが高いものを訳語として選択する。

3.1 構成語類似度の計算

構成語類似度の計算手法を図 4 に示す。まず、入力の専門用語と、訳語候補の双方を構成要素に分割する。この時、構成要素は形態素ではなく辞書のエントリーとしているので、構成要素の組合せは複数存在する。それぞれの構成要素の IDF を計算し、これを要素とする特徴ベクトルを生成する。

次に、各構成要素の辞書訳のあらゆる組合せを作ることにより、目的言語側の特徴ベクトルに変換する。この時、訳語候補の中に、構成要素の Transliteration が存在する場合には、辞書に動的に追加する。

最後に、目的言語側に変換した入力専門用語の特徴ベクトルと、訳語候補の特徴ベクトルとの cos 類似度をとることにより、構成語類似度とする。また、構成

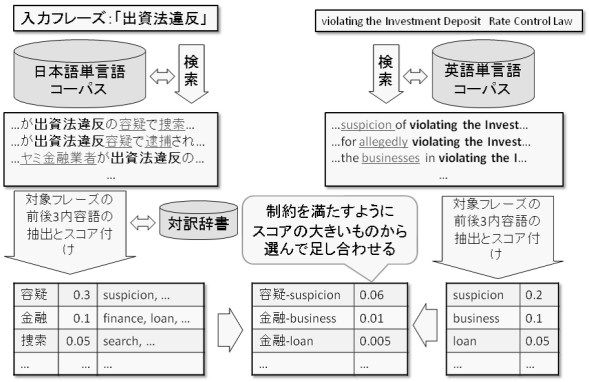

図 5: 共起語類似度の計算

要素数の違いによる類似度への影響を小さくするために、ダミーの要素を追加しておく。もし訳語候補側の構成要素数が入力側の構成要素数よりも少ない場合には訳語候補側の特徴ベクトルに、多い場合には入力側の特徴ベクトルに追加する。ダミーの要素の成分は、構成要素数が多いほうの特徴ベクトルの成分のうち、対訳の対応がない要素の成分の平均とする。

3.2 共起語類似度の計算

共起語類似度の計算手法を図 5 に示す。まず、入力された専門用語と、訳語候補のそれぞれを含む文を、各単言語コーパスから検索する。次に、入力専門用語と訳語候補それぞれの前後 3 内容語の頻度を計数し、IDF を掛け合わせることにより特徴ベクトルを生成し、それぞれのベクトルノルムが 1 となるよう正規化する。

次に、入力側の特徴ベクトルの要素を辞書で調べ、訳語候補側の特徴ベクトルの要素として存在するものの成分どうしを掛け合わせてスコアを計算しておく。入力側、訳語候補側ともに同じ要素は 1 回しか選択できないという制約のもとで、スコアの大きい順に対訳を選択する。最後に、選択された組合せのスコアの総和を計算して共起語類似度とする。

3.3 構成語類似度と共起語類似度の混合

それぞれの訳語候補に対して計算された構成語類似度 $CompScore$ と共起語類似度 $ContScore$ を用いて、以下のスコア $MixScore$ を計算する。

$$MixScore = \lambda ContScore + (1-\lambda) CompScore \quad (3)$$

ここで、λ は共起語類似度を混合する割合を決めるパラメータ $(0 \leq \lambda \leq 1)$ である。$MixScore$ がもっとも高い訳語候補を訳語として選択する。

4 実験と考察

4.1 実験設定

日本語を入力として英語訳語を獲得する実験を行った。単言語コーパスとして、時期とドメインを揃えた新聞コーパス各 100 万文を利用した。また、対訳辞書は研究社の日英、英日辞書を混合して用いた。辞書のエントリー数は日本語側、英語側ともに約 19 万エントリーである。共起語類似度と構成語類似度を混合するパラメータ λ を変えながら精度を測定した。

4.2 評価セットの作成

以下の条件をすべて満たすようなフレーズ対を英日新聞対訳コーパスから収集し、人手で整形して評価セットを作成した。評価セットには、正解を複数もつものも含んでいる。

- 英語フレーズは名詞が 3 語以上連続している
- 日本語フレーズは 3 形態素以上
- 100 万文でのヒット数が英日ともに 10 以上
- 英日ともに対訳辞書に見出し語として含まれていない

条件を満たす 181 フレーズ対を収集し、評価セットとして用いた。181 フレーズ対のうち、構成的に生成可能なフレーズは 73 フレーズ対 (34.0%) であった。

4.3 実験結果

λ の値と、1 位が正解である割合との関係を図 6 に示す。また、構成的に生成可能な専門用語とそうでないものに分けて精度を計算した結果も示した。

精度が最も高かったのは $\lambda = 0.55$ のときで、63.0% であり、スコアの上位 10 個に正解が含まれる精度がもっとも高かったのは $\lambda = 0.60$ のときで、77.2%であった。また、訳語候補生成部で生成された訳語候補の中に正解フレーズが含まれるものは、181 フレーズ中 171 フレーズ (94.5%) であった。

構成的に生成可能な専門用語とそうでないものとに分けた結果を見ると、前者の最高精度は $\lambda = 0.55$ のときの精度 98.6%であり、後者の最高精度は $\lambda = 0.70$ のときの精度 40.4%であった。

表 1: 訳語獲得例 (出力欄のスコアは ($MixScore$,$ContScore$,$CompScore$))

入力	出力		判定
都市 基盤 整備 公団	Urban Development Corporation	(0.523,0.590,0.441)	○
宇宙 開発 委員会	Space Activities Commission	(0.515,0.345,0.723)	○
国立 循環 器 病 センター	other organs	(0.321,0.391,0.234)	×
国立 循環 器 病 センター	the Japan Organ	(0.293,0.301,0.285)	×
国立 循環 器 病 センター	the Organ	(0.288,0.327,0.241)	×
国立 循環 器 病 センター	the National Cardiovascular Center	(0.286,0.271,0.305)	○
防衛 施設 庁	the Defense Agency	(0.566,0.371,0.804)	×
防衛 施設 庁	the Defense Facilities Administration Agency	(0.537,0.342,0.776)	○

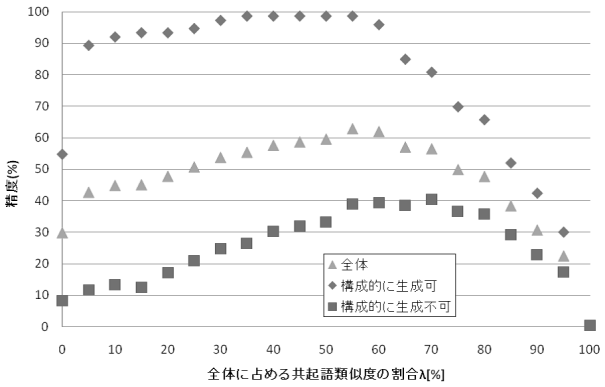

図 6: λ を変えたときの精度の変化

4.4 訳語獲得例と考察

精度が最も高かった $\lambda = 0.55$ のときの訳語獲得の例を表 1 に示す。下線が引いてある構成要素は対訳辞書で対応づいていないものである。

はじめの 2 つは、構成的に生成不可能だが正しい訳語が獲得できた例である。3 つめの例では、正しい訳語に入力フレーズの辞書訳があまり含まれないため、構成語類似度では他の訳語候補と差がついていない。入力フレーズの文脈には「心臓」「移植」などが多く出現するので、“heart” や”transplant” などの単語が文脈に多く出現する訳語が獲得されてしまっている。例えば、2 番目の訳語は”the Japan Organ Transplant Network” の一部であり、”transplant” と強く共起する。4 つめの例では、正しい訳語には入力の構成要素の辞書訳がすべて含まれているが、余分なものも多く含まれてしまっているため、構成語類似度が低くなっている。そのため、 入力フレーズと形態素数が近く、似通った文脈で出現する”the Defence Agency” が獲得されている。改善策としては、エントロピーやコーパス中のヒットカウントを利用して大きなフレーズの一部であるものは訳語候補から除外することなどがあげられる。

5 おわりに

本論文では、構成的に生成できない専門用語であっても訳語候補を生成する手法と、構成語類似度と共起語類似度を併用した訳語選択手法について提案した。また、評価セットを用いて実験を行った結果、本手法の有効性が示せた。今後の課題としては、異なるコーパスや異なる言語対でも本手法が有効かどうか検証することがあげられる。

参考文献

[1] Pascale Fung and Lo Yuen Yee. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics*, pp. 414–420, 1998.

[2] Reinhard Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 519–526, 1999.

[3] 外池昌嗣, 宇津呂武仁, 佐藤理史. ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定. 自然言語処理, Vol. 14, No. 2, pp. 33–68, 20070410.