中間言語とのDice係数ベクトルを用いた対訳抽出

李 寧*[1]　小川 泰弘[1,2]　大野 誠寛[1,2]　中村 誠[3]　外山 勝彦[1,2]

[1] 名古屋大学 大学院情報科学研究科　[2] 同 情報基盤センター　[3] 同 大学院法学研究科

{lining, yasuhiro}@kl.i.is.nagoya-u.ac.jp

1 はじめに

グローバル化に伴い，世界中の人々が交流する機会が増えている．国際交流の際に最も使われる言語は英語であるが，英語が公用語でない人口は依然として世界の多数を占めている．したがって，情報交換のため，英語以外の言語の間での情報共有も必要である．そのためには，対訳辞書が必要になる．

対訳辞書を作成するコストを削減するため，パラレルコーパスから対訳語を自動抽出する方法が研究されている．しかし，言語のペアと文書の分野によっては，パラレルコーパスは量が少ない場合，あるいは存在しない場合がある．

一方，対象とする2言語間のパラレルコーパスは存在しないが，それぞれの言語と英語との間のパラレルコーパスは存在する場合がある．特に近年，経済・社会のグローバル化に伴い，多くの国で特許や法令の英訳が公開されている．特定分野の文書の英訳には，元が異なる言語であっても，同じ専門用語が使用される可能性が高い．そのため，2言語間で対応しない文書であっても，それぞれの英訳とのパラレルコーパスから，英語を中間言語として対訳を抽出する手法が考えられる．

そこで本稿では，対訳資源が少ない言語間の情報共有を支援するため，原言語と中間言語のパラレルコーパス，および目標言語と中間言語のパラレルコーパスから対訳を抽出する手法を提案する．提案手法では，原言語の単語と目標言語の単語をベクトルで表現し，その間の類似度が高いペアを対訳として抽出する．その際，ベクトルの要素として，各言語と中間言語からなるパラレルコーパス上で計算したDice係数を用いる．

2 関連研究

対訳の自動抽出は，原言語の単語集合W_s，目標言語の単語集合W_tから，単語対集合$\{(w_s, w_t) \mid w_s \in W_s, w_t \in W_t, w_s$と$w_t$が対訳$\}$を抽出することである．以下，$(w_s, w_t)$を構成する可能性のある$w_s$や$w_t$を単語候補と呼ぶ．

本稿の提案手法と同じく，中間言語を介して対訳語を抽出する方法として，田中ら[1]と張ら[2]らは中間言語との対訳辞書を用いる手法を提案した．田中ら[1]は，w_sの中間言語への訳語集合$W_p(w_s) = \{w_p \mid w_p$は$w_s$の訳語$\}$および，$w_t$の中間言語への訳語集合$W_p(w_t) = \{w_p \mid w_p$は$w_t$の訳語$\}$を辞書引きで網羅し，$W_p(w_s)$と$W_p(w_t)$が2つ以上の単語を共有すれば$w_s$と$w_t$を対訳とした．張ら[2]は，対訳語を抽出する際，中間言語の訳語のほか，品詞対応関係，漢字対応関係などの情報を用いてスコアリングを行った．本研究の提案手法は，対訳辞書ではなく，原言語と中間言語との間，および目標言語と中間言語との間の2つのパラレルコーパスを利用する．

提案手法と同じく，中間言語とのパラレルコーパスを利用した研究として，Tsunakawa[3]らは，英語を中間言語として，GIZA++[1]で原言語と英語，および目標言語と英語のそれぞれのパラレルコーパスをアライメントし，原言語–英語，英語–目標言語の翻訳確率から原言語–目標言語の翻訳確率を計算した．本研究は，翻訳確率ではなく，単語候補をベクトルで表し，その間の類似度で対訳を抽出する．

提案手法と同じく，ベクトルで単語候補の意味を表し，ベクトル間の類似度により対訳語を抽出する手法として，Fungら[4]は，単語候補wと小規模対訳辞書にある単語の共起からTF-IDF値を計算し，TF-IDF値をベクトルの要素として単語候補のベクトルを作成した．単語の意味を表すベクトルの要素は，TF-IDF値のほか，Haghighiら[5]が提案したMCCAにおける単語の素性や，Ivanら[6]が提案したBiLDAモデルにおける単語の潜在的なトピックへの所属確率などが利用されている．これらのベクトルは原言語と目標言語のコンパラブルコーパスを使用して作成されている．本稿の提案手法は，原言

[1] https://code.google.com/p/giza-pp/

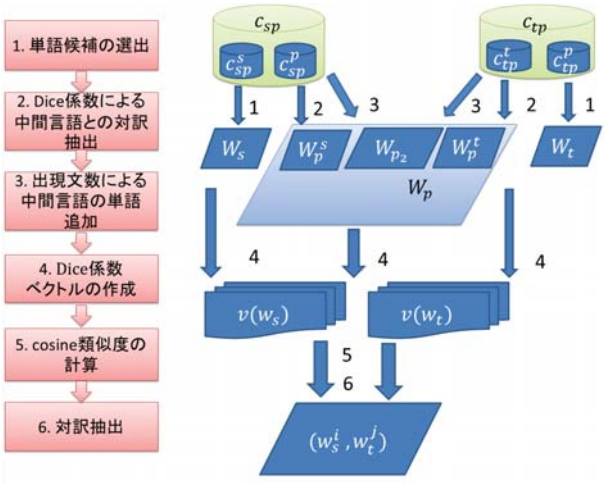

図 1: 提案手法の概要

語と目標言語との間のコーパスではなく，それぞれと中間言語との間のコーパスを用いる．

3 提案手法

提案手法では，単語候補と中間言語の単語の共起程度を示す Dice 係数のベクトルによって単語候補の意味を表し，その間の類似度が高いペアを対訳として抽出する．提案手法の手順を以下に示す．また，その概要を図 1 に示す．

入力: 原言語と中間言語のパラレルコーパス $C_{sp} = (C^s_{sp}, C^p_{sp})$，目標言語と中間言語のパラレルコーパス $C_{tp} = (C^t_{tp}, C^p_{tp})$．ここで，$C^s_{sp}$ と C^p_{sp} は C_{sp} を構成する原言語のコーパスと中間言語のコーパスである．C^t_{tp} と C^p_{tp} も同様である．

出力: 原言語，目標言語の単語対 $\{(w_s, w_t) \,|\, w_s$ と w_tは対訳 $\}$.

手順:

1. 単語候補の選出

   式 (1), (2) により原言語の単語候補集合 W_s と，目標言語の単語候補集合 W_t を求める．ここで，n と m は単語がコーパス内に出現する文数の閾値であり，$f(w, C)$ は単語 w のコーパス C 中における出現文数である．

$$W_s = \{w_s \,|\, f(w_s, C^s_{sp}) \geq n\} \quad (1)$$

$$W_t = \{w_t \,|\, f(w_t, C^t_{tp}) \geq m\} \quad (2)$$

2. Dice 係数による中間言語との対訳抽出

   単語候補の意味を表す中間言語の単語を獲得するため，式 (3) により原言語の単語候補 w_s に対する中間言語の訳語集合 W^s_p ，式 (4) により目標言語の単語候補 w_t に対する中間言語の訳語集合 W^t_p を求める．$d(w_s, w_p)$ は w_s と w_p の Dice 係数で，式 (5) により計算する．$d(w_p, w_t)$ も同様に計算する．$f(w_s, w_p, C_{sp})$ は C_{sp} において，w_s と w_t の共起する文数を表す．中間言語の単語候補集合 W_{p_1} を $W_{p_1} := W^s_p \cap W^t_p$ とする．

$$W^s_p = \{\arg\max_{w_p \in C^p_{sp}} d(w_s, w_p) \,|\, w_s \in W_s\} \quad (3)$$

$$W^t_p = \{\arg\max_{w_p \in C^p_{tp}} d(w_t, w_p) \,|\, w_t \in W_t\} \quad (4)$$

$$d(w_s, w_p) = \frac{2f(w_s, w_p, C_{sp})}{f(w_s, C^s_{sp}) + f(w_p, C^p_{sp})} \quad (5)$$

3. 出現文数による中間言語の単語追加

   より多くの中間言語の単語との Dice 係数により単語候補の意味を表すため，手順 2 で獲得した中間言語の単語のほか，コーパス内における出現文数の高い単語を追加する．具体的には，式 (6) により W_{p_2} を求め，$W_p := W_{p_1} \cup W_{p_2}$ とする．

$$W_{p_2} = \{w_p \,|\, n < f(w_p, C^p_{sp}) < k \text{ かつ } m < f(w_p, C^p_{tp}) < l\} \quad (6)$$

4. Dice 係数ベクトルの作成

   以上で求めた W_p の要素 w_p を用いて，Dice 係数を計算し，それらを要素とする Dice 係数ベクトルを式 (7)，(8) により求める．

$$\boldsymbol{v}(w_s) = (d(w_s, w^1_p), \cdots, d(w_s, w^{|W_p|}_p)) \quad (7)$$

$$\boldsymbol{v}(w_t) = (d(w_t, w^1_p), \cdots, d(w_t, w^{|W_p|}_p)) \quad (8)$$

5. cosine 類似度の計算

   すべての $\boldsymbol{v}(w_s)$ と $\boldsymbol{v}(w_t)$ 間の cosine 類似度 $sim(\boldsymbol{v}(w_s), \boldsymbol{v}(w_t))$ を式 (9) により求める．

$$sim(\boldsymbol{v}(w_s), \boldsymbol{v}(w_t)) = \frac{|\boldsymbol{v}(w_s) \cdot \boldsymbol{v}(w_t)|}{|\boldsymbol{v}(w_s)||\boldsymbol{v}(w_t)|} \quad (9)$$

6. 対訳抽出

   Dice 係数ベクトルが互いに一番類似する単語候補を対訳として抽出する．具体的には，式 (10), (11) を満たすすべての対訳ペア (w^i_s, w^j_t) を抽出する．

$$w^i_s = \arg\max_{w_s \in W_s} sim(\boldsymbol{v}(w_s), \boldsymbol{v}(w^j_t)) \quad (10)$$

$$w^j_t = \arg\max_{w_t \in W_t} sim(\boldsymbol{v}(w^i_s), \boldsymbol{v}(w_t)) \quad (11)$$

4 実験

提案手法の有効性を確認するため，ベースライン手法を設定し，それと提案手法を比較した．

4.1 ベースライン手法

ベースラインでは，中間言語の訳語が同じである w_s と w_t を対訳ペアとする．すなわち，提案手法で作成した W_s，W_t，W_{p_1} を利用し，すべての $w_p \in W_{p_1}$ に対し，w_p との Dice 係数が最も高い原言語の単語 w_s，および w_p との Dice 係数が最も高い目標言語の単語 w_t を対訳ペアとする．具体的には，式 (12) により求める．

$$\{(w'_s, w'_t) \mid w'_s = \arg\max_{w_s \in W_s} d(w_s, w_p), \quad w'_t = \arg\max_{w_t \in W_t} d(w_t, w_p), w_p \in W_{p_1}\} \tag{12}$$

4.2 実験設定

法令の分野において，中国語と日本語のパラレルコーパスの量は少ないが，中国と日本の法令の英訳が公開されたことにより，法令の中英，日英パラレルコーパスは大量に存在する．今回の実験では，提案手法を用いて，法令の中英と日英のパラレルコーパスからの中国語と日本語の対訳を抽出する．実験で使用した中英パラレルコーパスは中国大陸法令およびその英訳 23,405 文であり，日英パラレルコーパスは，日本語法令およびその翻訳 193 本，計 90,263 文である．

手順 1 の前処理として，中国語文を NLPIR[2] で分割し，その中の名詞，動詞，形容詞を抽出した．日本語文は MeCab[3] で分割し，大品詞が名詞，動詞，形容詞である内容語を抽出した．パラレルコーパスの英語文は，Standford Parser[4] でトークン化し，小文字化とレンマタイズの処理を行った．手順 1 の閾値 m と n は 3 に設定した．その結果，中国語の単語候補集合 W_s の要素数は 4,202 語，日本語の単語候補集合 W_t の要素数は 5,769 語となった．手順 2 で作成した W_{p_1} の要素数は 1,014 語となった．

以上の設定で予備実験を行い，手順 3 の k を 940 とし，l を 2,600 と定めた．このとき，W_p の要素数は 3,061 語となった．

[2]http://www.nlpir.org/
[3]https://code.google.com/p/mecab/
[4]http://nlp.stanford.edu/software/lex-parser.shtml

表 1: ベースラインと提案手法の精度比較

手法	抽出数	正解数	精度
提案手法	1,630	749	46.0%
提案手法 (上位 1,046 語)	1,046	573	54.8%
ベースライン	1,046	486	46.5%

表 2: 抽出された対訳 (一部) とその正誤判定

中国語	ベースライン訳	提案手法訳
场所	敷地 (×)	場所 (○)
设计	意匠 (×)	設計 (○)
临床	臨床 (○)	修練 (×)
超过	超え (○)	超える (○)
应急	対策 (×)	応急 (○)
退税	還付 (×)	既 (×)
配置	装備 (×)	漁ろ (×)
撤回	取り下げ (○)	取り下げる (○)

4.3 実験結果

ベースライン手法と提案手法による抽出数と精度を表 1 に示す．提案手法の抽出精度は 46.0%で，ベースライン手法の精度と同程度であるが，提案手法の正解抽出数は 749 語で，ベースライン手法の 486 語より多い．そこで，提案手法の抽出結果から cosine 類似度順の上位 1,046 語を抽出し，評価した結果，精度は 54.8%であり，ベースライン手法の 46.5%より高かった．

本稿の目的は，対訳資源が少ない言語間でより多くの対訳を抽出することである．提案手法は，ベースライン手法より多くの正解対訳を抽出している．また，精度は高くないが，類似度が高い対訳ペアでは，ある程度の精度を保っている．ゆえに，提案手法は有効である．

4.4 Dice 係数ベクトルの有効性

ベースライン手法と提案手法との間で精度に差が生じた原因を調べるため，ベースライン方法と提案手法の抽出結果から一部を選出し，比較した．表 2 に抽出された結果と正誤判定を示す．

ベースライン手法では，中間言語の単語 w_p との Dice 係数が最も高い $w_s \in W_s$ と $w_t \in W_t$ を対訳とした．しかし，Dice 係数による抽出結果は，必ずしも正しい訳語ではない．たとえば，表 2 において，ベースライン手法では，英単語 “site” との Dice 係数が最も高い中国語は “场所” であり，これは正しい訳語である．しかし，“site” との Dice 係数が最も高い日本語の単語は「敷地」であり，これは誤った訳語である．この場合，ベースライン手法は誤って “场所” と「敷地」を対訳として抽出した．

表 3: “场所”,「敷地」,「場所」の Dice 係数ベクトル

単語	site	place	premise	center	…
场所	0.32	0.31	0.24	0.24	…
敷地	0.37	0	0	0	…
場所	0.12	0.47	0.04	0.02	…

表 4: “临床”,「臨床」,「修練」の Dice 係数ベクトル

単語	clinical	nurse	hospital	…
临床	0.97	0	0	…
臨床	0.95	0.18	0.10	…
修練	0.77	0.13	0.09	…

もう一つの問題点として，中間言語の単語の曖昧性によって，中間言語訳の抽出結果が正解だとしても，原言語の単語と目標言語の単語が対訳でない可能性がある．例えば，表 2 には，中国語の単語 “设计” と日本語の単語「意匠」が誤って対訳として抽出された例が示されている．“设计” と「意匠」の英訳は，同じく “design” であるが，中国語の “设计” は，様々な分野で使用され，日本語の「設計」の意味も含む．日本語の「意匠」は美術，工業作品などの分野のみで使われるため，中国語の単語 “设计” より意味が狭い．

そのような問題点に対し，提案手法の Dice 係数ベクトルは，単語候補の意味をより正確に表示することができる．Dice 係数ベクトルの表現力を示すため，“场所”,「敷地」,「場所」，それぞれの Dice 係数ベクトルの一部を表 3 に示す．“site” との Dice 係数が最も高い原言語の単語と目標言語の単語は，それぞれ “场所” と「敷地」であるが，Dice 係数ベクトルで単語候補の意味を表示することによって，中国語の単語 “场所” と類似する意味を持つ訳語として，日本語の単語「場所」を抽出できている．

4.5 誤り分析

ベースライン手法では，“临床” の正しい訳語「臨床」が抽出されたが，提案手法では，誤った訳語「修練」が抽出された (表 2)．“临床”,「臨床」,「修練」の Dice 係数ベクトルの一部を表 4 に示す．

この誤った抽出の原因は 2 つあると考えられる．1 つ目の原因は，「臨床」と「修練」が日本語コーパスでよく共起するため，両単語の Dice 係数ベクトルが類似しているという点である．もう 1 つの原因は，“hospital”，“nurse” など，「臨床」との Dice 係数が高い英単語が，中英パラレルコーパスで中国語の “临床” と共起しない点である．コーパスを調べた結果，「臨床」の多くは，大学設置基準という法令に含まれていた．一方，“临床” は中华人民共和国药品管理法にしか出現しない．両法令が対象とする法領域が異なるため，“临床” と「臨床」の Dice 係数ベクトルは類似していなかった．その結果，両単語候補は対訳として抽出できなかった．

以上の考察から，両コーパスの分野の差も提案手法の性能に影響する．

5 おわりに

本稿は，Dice 係数ベクトルを用いて，原言語と中間言語との間および目標言語と中間言語との間のパラレルコーパスから対訳ペアを抽出する手法を提案した．また実験により，本稿の提案手法の有効性を確認した．今後はコーパスの分野の差を考慮し，提案手法を改良する予定である．

参考文献

[1] 田中 久美子, 梅村 恭司, 岩崎 英哉: 第三言語を介した対訳辞書の作成. 情報処理学会論文誌, Vol.39, No.6, pp.1915-1924, 1998.

[2] 張 玉潔, 馬 青, 井佐原 均: 英語を介した日中対訳辞書の自動構築. 自然言語処理, Vol.12, No.2, pp. 63-85, 2005.

[3] T. Tsunakawa, N. Okazaki, J. Tsujii: Building Bilingual Lexicons Using Lexical Translation Probabilities via Pivot Languages. *LREC-2008*, pp.1664-1667, 2008.

[4] P. Fung and L. Y. Yee: An IR Approach for Translating New Words from Nonparallel, Comparable Texts. *COLING-1998*, Vol.1, pp.414-420, 1998.

[5] A. Haghighi, P. Liang, T. Berg-kirkpatrick, D. Klein: Learning Bilingual Lexicons from Monolingual Corpora. *ACL-2008*, pp.771-779, 2008.

[6] I. Vulić, W. De Smet, M.F. Moens: Identifying Word Translations from Comparable Corpora Using Latent Topic Models. *ACL-2011*, pp.479-484, 2011.