パラレルコーパスからの語義の対訳用例の自動獲得

村澤 和弥　　白井 清昭
北陸先端科学技術大学院大学 情報科学研究科
{s1410041,kshirai}@jaist.ac.jp

1 はじめに

本研究は，日本語学習者のための読解支援システムとして，語義曖昧性解消 (Word Sense Disambiguation; WSD) の技術を用いて単語の意味を自動的に判定し，その意味の語釈文と用例をユーザに提示するシステムの構築を目指している．本システムは日本語学習者の辞書引きを支援するものであるが，辞書として和英辞書 EDICT[1] を用いる．EDICT の語釈文は単語もしくは句という単純なものではあるが，語義ごとに日本語の例文とその英訳が用意されている．我々は，日本語学習者が単語の意味を理解する上で，用例を提示することは大きな役割を果たすと考えている．この際，日本語用例だけでなく英語の対訳も提示した方が望ましい．

本システムでは，辞書中の対訳用例を，用例に基づく WSD で利用するとともに，日本語学習者に提示する用例としても用いる．対訳用例の数が多ければ多いほど，WSD の性能の向上が見込めるし，単語の意味の理解を助けるのに適した用例が見つかる可能性も高くなる．本論文では，パラレルコーパスから特定の語義を持つ対訳用例を自動的に獲得し，辞書における語義の用例を拡充する手法について述べる [5].

2 関連研究

語義の用例をコーパスから自動獲得する先行研究について述べる．Fujita らは，岩波国語辞典の例文を利用し，その例文を含む (例文よりも長い) 文をコーパスから自動的に抽出することによって，WSD の訓練データとなる語義付き例文の数を増加させる手法を提案している [1]．単一言語コーパスではなくパラレルコーパスから用例を獲得する研究としては Melo らによるものがある [4]．彼らの手法では，パラレルコーパスおよび語義が対応付けられた多言語辞書を用意し，機械学習された WSD の分類器を用いて原言語，目標言語の両方の単語の語義の曖昧性を解消し，語義の用例を獲得する．この際，2 つの言語の語義を同時にチェックすることで獲得される例文の品質を向上させている．ただし，原言語，目標言語の両方について作成コストの高い語義タグ付きコーパスを必要とするという欠点がある．Kathuria らは，パラレルコーパスから語義の用例を自動獲得し，また獲得された用例に基づく WSD 手法を提案している [2]．本論文はこの手法を拡張・改良するものである．

[1] http://www.csse.monash.edu.au/~jwb/edict.html

S1 { story, talk , conversation, speech, chat }
E_{11}: そんな話は知りたくない。
I don't want to know that kind of story.
E_{12}: その話を私に聞かせないで下さい。
Please let me not hear of that story.
S2 { discussions, argument, negotiation }
E_{21}: 3 時間議論したが，話がまとまらなかった。
After 3 hours of discussion we got nowhere.

図 1: EDICT における「話」の語釈文と例文

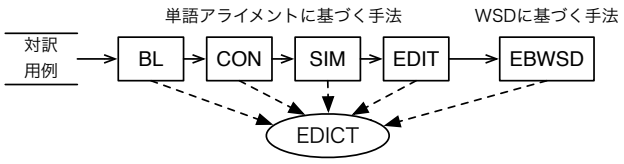


図 2: 提案手法の概要

3 提案手法

本節では，EDICT 中の日本語の語義に対し，その語義を持つ日本語の例文ならびにその翻訳となる英語の文の組 (対訳用例) をパラレルコーパスから獲得する手法について述べる．EDICT における語義の例を図 1 に示す．EDICT では，語釈文は単語または句で表現されている．本研究ではパラレルコーパスとして日英新聞記事対応付けデータ JENAAD[6] を用いた．前処理として，日本語文は MeCab[2]，英語文は Morpha[3] を用いて単語を原形に直した後，GIZA++[4] を用いて単語のアライメントを行った．

提案手法の概要を図 2 に示す．パラレルコーパスにおける対訳用例に対し，Kathuria の手法 (これをベースラインとし，以下 BL と記す)，周辺語のマッチングに基づ

[2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[3] http://users.sussex.ac.uk/~johnca/morph.html
[4] http://www.statmt.org/moses/giza/GIZA++.html

く手法 (CON)，単語間類似度に基づく手法 (SIM)，単語間編集距離に基づく手法 (EDIT)，用例に基づく WSD による手法 (EBWSD) の各手法を順に用いて対象語の語義を推定し，語義を推定できた時点でその対訳用例を EDICT に追加する．BL, CON, SIM, EDIT の各手法は単語アライメントの情報を主に利用している．以下，各手法の詳細について述べる．

3.1 単語アライメントに基づく対訳用例獲得

3.1.1 Kathuria の手法

Kathuria らの手法 [2] では，語義の語釈文中の語とパラレルコーパスの英文中の語を照合することで日本語文中の単語の語義を推定する．ここではその概略について述べる．対象語 t の語義 s に対し，日本語文 Ja と英語文 En の組が以下の条件を満たすとき，(Ja, En) を語義 s の用例として獲得する．

(1) 日本語文 Ja が対象語 t を含む．

(2) t^e を t と対応関係にある En 中の語とする．t^e もしくは t^e を含む複合語が，語義 s の語釈文中のいずれかの語もしくは複合語と等しい．

(3) t^e は，t の複数の語義のうち，たかだかひとつの語義の語釈文中の単語とマッチする．

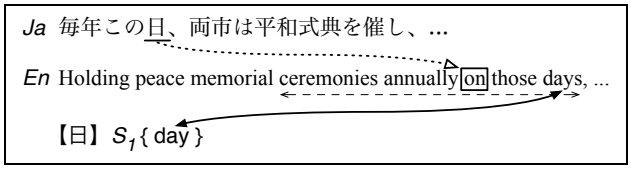
この手法は，パラレルコーパスに存在する全ての語義の用例を獲得するのではなく，信頼性の高い語義の用例のみを獲得するという方針に基づいている．そのため，獲得される対訳用例の数が少ないという問題点がある．以降，より多くの対訳用例を獲得する手法を提案する．

3.1.2 周辺語のマッチングに基づく手法

Kathuria の手法における条件 (2) を以下のように変更する．

(2)' t^e の前後 2 単語の範囲で出現する語もしくは複合語が，語義 s の語釈文中のいずれかの語もしくは複合語と等しい．

以下の例では，対象語「日」と対応関係にある単語 t^e は「on」であるが，その 2 つ後の単語は「day(s)」であり，「日」の語義 S_1 の語釈中の「day」と一致するので，(Ja, En) を語義 S_1 の対訳用例とみなす．



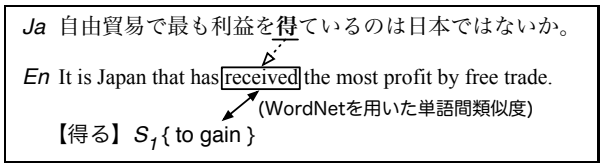
3.1.3 単語間類似度に基づく手法

t^e と語義 s の語釈文中の単語 (w^s とおく) を照合する際，両者が完全に一致しなくても，意味的に類似した単語であれば，具体的には t^e と w^s の類似度が 0.5 以上であれば，(Ja, En) を語義 s の対訳用例として獲得する．単語間の類似度は WordNet を用いて測る．類似度の定義を式 (1) に示す．

$$sim_{WN}(x, y) = \frac{1}{lsp + 1} \quad (1)$$

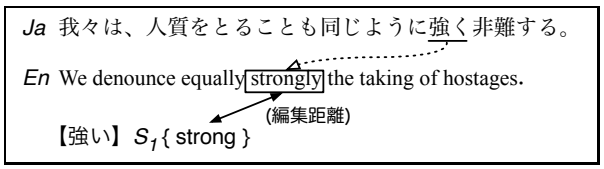
(lsp は単語 x と y の synset を結ぶ最短パスの長さ)

以下の例では，対象語「得る」と対応関係にある単語 t^e は receive であり，「得る」の語義 S_1 の語釈中の gain との類似度が高いので，(Ja, En) を語義 S_1 の対訳用例とみなす．



3.1.4 単語間編集距離に基づく手法

t^e と w^s を照合する際，両者が完全に一致しなくても，両者の編集距離が十分に小さければ，(Ja, En) を語義 s の対訳用例として獲得する．ここでは編集距離の閾値を 2 以下と設定した．以下の例では，対象語「強い」と対応関係にある単語 t^e は strongly であり，「強い」の語義 S_1 の語釈中の strong との編集距離は 2 なので，(Ja, En) を語義 S_1 の対訳用例とみなす．



3.2 WSD に基づく対訳用例獲得

3.1 項で述べた手法は t^e と w^s の同一性や類似性に着目して語義を推定している．しかし，語義を決めることのできない対訳用例の多くは，単語アライメントの段階で t^e(対象語 t と対応関係にある英単語) が特定できていない．ここでは，より多くの対訳用例を獲得するために，用例に基づく語義曖昧性解消の手法を適用して対象語の語義を決定する．

用例データベースは，EDICT に記載されている対訳用例と，3.1 項で述べた手法でパラレルコーパスから獲得した対訳用例の両方を用いる．パラレルコーパスにおける日英の文の組を $TE_i = (S_i^J, S_i^E)$，用例データベース中の対訳用例を $TE_j = (S_j^J, S_j^E)$ とし，両者の類似度 $sim(TE_i, TE_j)$ が最大となる TE_j' を求める．TE_i と TE_j' の類似度が閾値 T_s 以上なら，TE_i における対象語の語義は TE_j' の語義と同じであるとみなし，その語義の用例として EDICT に追加する．

$sim(TE_i, TE_j)$ は，式 (2) のように，日本語文の類似度と英語文の類似度の重み付き和とする．

$$sim(TE_i, TE_j) = \alpha \cdot sim_s(S_i^J, S_j^J) + (1-\alpha) \cdot sim_s(S_i^E, S_j^E) \quad (2)$$

sim_s は日本語文および英文の文間類似度である．以下にその詳細を述べるが，日本語文の類似度は Kathuria らが提案した用例に基づく WSD 手法 [2] に準じており，英語文の類似度は本研究で提案するものである．

文の類似度 sim_s は，式 (3) のように，コロケーションの類似度 col，統語的関係の類似度 syn，対象語の類似度 tar の和を基に算出する．l は言語 (J または E) を表わす．また，σ は実数を 0 から 1 までの値に変換するシグモイド関数である．

$$sim_s(S_i^l, S_j^l) = \sigma\left(col(S_i^l, S_j^l) + syn(S_i^l, S_j^l) + tar(S_i^l, S_j^l) \right) \quad (3)$$

コロケーションの類似度 $col(X, Y)$ は，対象語の直前・直後の文脈の類似度を測る指標である．まず，対象語を含む単語 n-gram (n は 6 以下) を抽出する．2 つの文から抽出した n-gram のうち一致するものがあれば，その最長の n-gram の長さに応じたスコアを与える．コロケーションの類似度として与えるスコアの一覧を表 1 に示す．これらの重みは直観により決めている．

表 1: コロケーションの類似度

n		6	5	4	3	2
$col(X, Y)$	(日本語)	1	0.75	0.5	0	0
	(英語)	1	0.8	0.6	0.4	0.2

統語関係の類似度 $syn(X, Y)$ は，対象語と統語関係にある語の類似度である．文 X と Y から，対象語 t と同じ統語的関係にある語 w_x と w_y を取得する．全ての w_x と w_y の組に対する単語間類似度の和を $syn(X, Y)$ とする．その定義を式 (4) に示す．

$$syn(X, Y) = \sum_{w_x, w_y} sim_w(w_x, w_y) \quad (4)$$
$$\text{但し } rel(t, w_x) \in X \ \ \& \ \ rel(t, w_y) \in Y$$

日本語文の類似度を測る場合，統語的関係は，格関係を表わす助詞 (「が」「を」「に」など)，連体修飾，連用修飾のいずれかとする．対象語を含む統語的関係は CaboCha[5] による文節の係り受け解析結果から同定する．単語間の類似度 sim_w は分類語彙表 [3] を用いて測る．一方，英語の類似度を測る場合，英文を Stanford Parser[6] で依存構造解析した結果[7] から得られる統語的関係を用いる．単語間の類似度 sim_w は式 (1) で測る．

[5] http://taku910.github.io/cabocha/
[6] http://nlp.stanford.edu/software/lex-parser.shtml
[7] Universal dependencies, enhanced の出力を用いる．

対象語の類似度は，語義曖昧性解消の対象となる単語の類似度であり，以下のように定義する．

$$tar(S_i^J, S_j^J) = 0 \quad (5)$$
$$tar(S_i^E, S_j^E) = sim_{WN}(t_i^e, t_j^e) \quad (6)$$

日本語文の類似度を測るとき，対象語は常に同じなので，対象語の類似度は考慮しない (式 (5))．一方，英語の場合，t_i^e，t_j^e を文 S_i^E，S_j^E において日本語の対象語と対応関係にある英単語とし，これらの類似度と定義する (式 (6))．

前述のように，単語アライメントによって t^e が決まらない場合がある．このときは以下の処理を行う．

- EDICT 内の対訳用例については，語義の語釈にマッチする単語を t^e とする．EDICT 内の用例は語義が決まっており，かつ短い例文が多いので，語義の語釈にマッチする単語はたかだか 1 つとする．
- パラレルコーパス中の対訳用例については，EDICT における全ての語義の語釈にマッチする単語を検出し，それらを仮に t^e と定める．全ての t^e の候補について，式 (3) の類似度を算出し，その最大値を英語文間の類似度とする．

4 評価実験

対象単語として表 2 に示す 20 単語を選定した．パラレルコーパス (JENAAD) から対象語を含む文を検索し，対象単語当たりの平均で 1,443 個，合計 28,861 個の対訳用例を得た．

表 2: 対象単語

人，中，出る，言う，時，入る，自分，持つ，強い，情報，開く，認める，日，出す，交渉，得る，対応，前，時間，地方

パラレルコーパスから得られた対訳用例に対し，3 節で述べた各手法で獲得できた対訳用例数を表 3 に示す．「用例数」は 20 個の対象単語の全てについて獲得された用例の総数，「平均」は対象単語 1 語当たりに獲得された用例数を表わす．また，「TOTAL」は各手法を用いて獲得された用例数の累計である．手法 EBWSD における閾値 T_s は 0.6 とした．

20 個の対象単語に対して合計 9407 個の対訳用例を獲得できた．すなわち，パラレルコーパスに含まれる 28,861 個の用例の 33%について，提案手法で語義を推定することができた．ベースラインを除いて，アライメントに基づく手法は獲得される用例数が少ない．その中でも獲得用例数が多いのは WordNet による単語間類似

表 3: 獲得された対訳用例数

	BL	CON	SIM	EDIT	EBWSD	TOTAL
用例数	4318	122	836	145	3986	9407
平均	215.9	6.1	41.8	7.25	199.3	470.35

表 4: 対訳用例の正解率

	BL	CON	SIM	EDIT	WSD1	WSD2
Mi-A	0.836	0.758	0.886	0.987	0.752	0.793
Ma-A	0.831	0.763	0.849	0.833	0.599	0.646

度に基づく手法 (SIM) である．一方，WSD に基づく手法 (EBWSD) は対訳用例の数が多い．これは，単語アライメントによって対象語と対応関係にある英単語が決まらない場合にも語義を推定しているためである．

次に，各手法によって獲得された対訳用例の正解率を調べた．正解率は，語義を正しく決めることができた対訳用例の割合である．それぞれの対象単語について，各手法によって獲得された対訳用例の中から 50 個をランダムに選択し (獲得用例数が 50 以下の場合は全て)，正解率を人手で算出した [8]．20 個の対象単語に対する正解率のマイクロ平均 (Mi-A)，マクロ平均 (Ma-A) を表 4 に示す．手法 EBWSD については，閾値 T_s を 0.6 に設定し，対象語に対応する英単語 t^e が特定されているとき (WSD1) とされていないとき (WSD2) に分けて正解率を算出した．各手法の正解率は 75%以上であり，十分に高いと言える．また，アライメントに基づく手法の方が WSD に基づく手法よりも全体的に正解率が高い．

手法 EBWSD において，対訳用例間の閾値 T_s を変化させたときの正解率を調べた．正解率を調べるために，対象単語毎に，単語アライメントによって t^e が決まる場合と決まらない場合のそれぞれにおいて，日本語文間の類似度と英語文間の類似度がともに 0 より大きい対訳用例を 50 個ランダムに選択し (50 個未満のときは全て)，これらの対訳用例の語義が正しく決定されているかを人手で判断した．

結果を図 3 に示す．この実験では式 (2) における α を 4 通りに設定している．$\alpha = 1$ (日本語文の類似度のみ使用)，$\alpha = 0$ (英語文の類似度のみ使用)，$\alpha = 0.5$ および α を最適化したときである．α の最適化とは，対象単語毎にテストデータの正解率が最大となる α を 0.1, 0.3, 0.5, 0.7, 0.9 の中から選択したときの結果であり，日本語文と英語文類似度を組み合わせる手法の正解率の上限を示している．図 3 から，閾値 T_s を上げると正解率が向上する傾向が見られる．また，日本語文と英語文の両方の類似度を考慮したとき ($\alpha = 0.5$) の正解率は，日

[8] 手法 EBWSD のサンプリング手法については後述する．

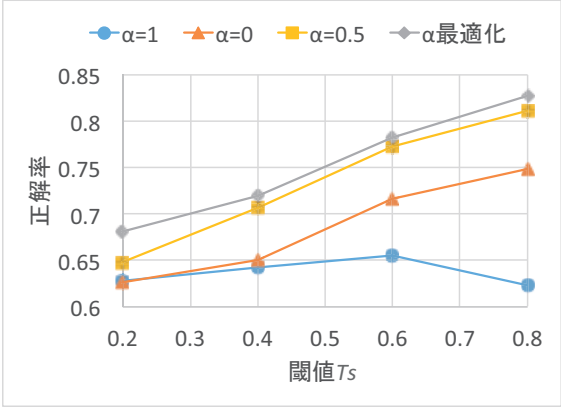


図 3: EBWSD の正解率

本語文のみもしくは英語文のみを用いるときよりも正解率が高い．さらに，$\alpha = 0.5$ の結果は α を最適化した結果とあまり差がない．以上より，WSD の際，日本語文と英語文の両方の類似度を考慮する提案手法の有効性が確認された．

5 おわりに

本論文では，パラレルコーパスから語義の対訳用例を自動獲得する手法について述べた．先行研究と比べてより多くの対訳用例を獲得し，また日本語文と英語文の両方の類似度を考慮することで獲得用例の正解率が向上した．一方，提案手法では，獲得される対訳用例は比較的よく使われる語義のものが多いという問題点もある．パラレルコーパスの量を増やせば多くの語義の用例が得られるが，効率も悪い．使用頻度の低い語義の対訳用例を効率良く獲得する手法を探究することが今後の課題である．

参考文献

[1] Sanae Fujita and Akinori Fujino. Word sense disambiguation by combining labeled data expansion and semi-supervised learning method. In *Proceedings of IJCNLP*, pp. 676–685, 2011.

[2] Pulkit Kathuria, 白井清昭. パラレルコーパスから自動獲得した用例に基づく語義曖昧性解消. 情報処理学会研究報告, Vol. 2012-NL-207, No. 3, pp. 1–8, 2012.

[3] 国立国語研究所（編）. 分類語彙表. 大日本図書, 2004.

[4] Gerard de Melo and Gerhard Weikum. Extracting sense-disambiguated example sentences from parallel corpora. In *Proceedings of the 1st Workshop on Definition Extraction*, pp. 40–46, 2009.

[5] 村澤和弥. 日本語読解支援システムのためのパラレルコーパスからの対訳用例の自動獲得. Master's thesis, 北陸先端科学技術大学院大学, 3 2016.

[6] Masao Utiyama and Hitoshi Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of ACL*, pp. 72–79, 2003.