複数の２言語辞書とコンパラブルコーパスからの多言語辞書の生成

山元　陽祐，　綱川　隆司，　梶　博行
静岡大学大学院情報学研究科

1. はじめに

インターネット技術の発展により、様々な国の人々との交流が容易に行えるようになり、機械翻訳や情報検索においても多言語のものが求められるようになってきた。多言語の機械翻訳や情報検索には、対象言語の対訳語の組から構成される多言語辞書が重要な役割を果たす。多数の言語の対訳語の組においては多義性のある単語の語義が絞られるので、機械翻訳やクロス言語情報検索の精度向上を期待することができる。

最も単純な多言語対訳辞書の生成方法として、２言語対訳辞書をマージする方法が考えられる。例えば、英語-日本語辞書と英語-アラビア語辞書を、共通言語の英語を介してマージすることで、英語-日本語-アラビア語の辞書を生成することができる。しかし、この方法では、媒介となる言語の単語が多義語の場合、誤った対訳語の組(以下、ノイズ)も生成されるため、そのようなノイズを除去する必要がある。

本研究では、(N-1)個の２言語辞書をマージして得られる N 言語の対訳語の組の候補から正しい組を選択するために、各言語のコーパスから得られる文脈情報を用いる方法を提案する。他の言語を介してしかつながらない言語対では、対訳文から構成されるパラレルコーパスの利用は期待できないため、提案方法では、同一分野のそれぞれの言語のコーパスを組み合わせたコンパラブルコーパスを利用する。本稿では、英語を媒介言語とし、日本語と中国語の単言語コーパスを用いて英日中対訳辞書を生成する予備実験について報告する。

2. 関連研究

University of Washington の研究グループは、２言語対訳辞書の集合から翻訳グラフを生成してその構造から対訳関係を推定する方法を提案し、PanDictionary と呼ぶ多言語辞書を構築している(Mausam et al., 2009)。また、ハブ言語を介して対訳語を推定するためコンパラブルコーパスを利用する方法を提案している(Sammer and Soderland, 2007)。本稿で提案する方法は、コンパラブルコーパスを利用するという点でこれに近いが、コンパラブルコーパスの利用方法が異なっている。提案方法では、コンパラブルコーパスからの対訳抽出の代表的な方法である文脈類似度に基づく方法を利用する(Fung and Yee, 1998; Rapp, 1999)。

２言語対訳辞書の集合からの多言語辞書の生成は第３言語を介した対訳辞書の生成の拡張と考えることができる。田中ら(1998)は、２言語対訳辞書をマージする際に、媒介となる単語の数が多いほど正しい対訳語ペアであるというヒューリスティクスに基づく方法を提案した。このヒューリスティクスと品詞や文字の対応に関するヒューリスティクスを組合せた方法も提案されている(張ほか, 2005)。第３言語を介した対訳辞書生成においてコンパラブルコーパスを利用する方法も提案されているが(Kaji et al., 2008)、提案方法とはコンパラブルコーパスの利用方法が異なっている。

3. 提案方法

3.1. 概要

本研究の多言語辞書はさまざまな言語の単語の組の集合で、各組はそれに含まれる単語のどのペアも共通の語義を持つ。提案方法の入力となる２言語対訳辞書も共通の語義を持つ単語の組の集合と考える。

多言語辞書 D は、入力の２言語対訳辞書の和集合からスタートし、以下のように単語の組を逐次追加することによって生成される。

一つ以上の単語 $w_1,\dots,w_K$ を共有する２つの組 $\{w_1,\dots,w_K,\dots,w_I\}$と$\{w_1,\dots,w_K,w'_{K+1},\dots,w'_J\}$が D に含まれ、それらをマージして得られる組$\{w_1,\dots,w_I,w'_{K+1},\dots,w'_J\}$が次の条件を満たすとき$\{w_1,\dots,w_I,w'_{K+1},\dots,w'_J\}$を D の構成要素として追加する。

(条件) $w_i(i=K+1,\dots,I)$と$w'_j(j=K+1,\dots,J)$の任意のペアが(i)(ii)のいずれかを満たす。

(i) w_iとw'_jが対訳であることを D が示している。すなわち $\{w_i,w'_j\}\in D$。

(ii) コーパスから得られるw_iの文脈$C(w_i)$とw'_jの文脈$C(w'_j)$の類似度$sim(C(w_i),C(w'_j))$が一定の閾値θ以上である。すなわち、$sim(C(w_i),C(w'_j))\geq\theta$。

(ii)のうち、文脈の抽出と表現については 3.2 節、文脈類似度の計算については 3.3 節で記述する。

図１に多言語辞書生成のプロセスを例示する。英語(EN)、ドイツ語(DE)、日本語(JP)、中国語(CN)の４言語を対象とし、EN-DE, EN-JP, EN-CN の３つの２言語辞書が入力である。これらの２言語対訳辞書に含まれる２言語の単語の組の集合が多言語辞書 D の初期値である。

EN-DE と EN-JP の２つ組をマージし、DE-JP のコンパラブルコーパスを用いることによって EN-DE-JP の３つ組が生成される。同様に EN-JP と EN-CN の２つ組をマージし、JP-CN のコンパラブルコーパスを用いることによって EN-JP-CN の３つ組が生成される。さらに、EN-DE-JP と EN-JP-CN の３つ組をマージし、DE-CN のコンパラブルコーパスを用いることによって EN-DE-JP-CN の４つ組が生成される。EN-DE-JP-CN の４つ組は、EN-DE-JP の３つ組と EN-CN の２つ組をマージし、DE-CN および JP-CN のコンパラブルコーパスを用いることによって生成することもできる。

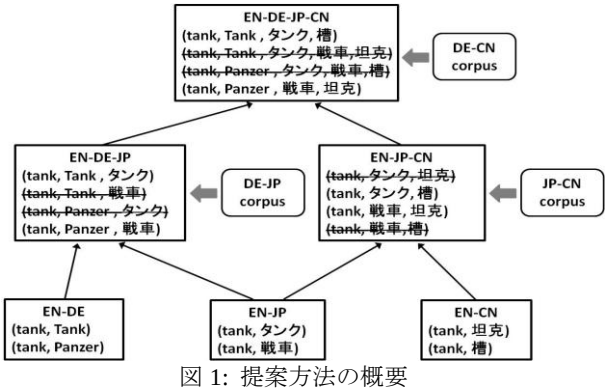


図 1: 提案方法の概要

3.2 文脈の抽出と表現

提案方法では、対訳語候補の単語のうち、マージされた組が共有していた単語以外の各単語、すなわち$w_i(i = K+1, \ldots, I)$と$w'_j(j = K+1, \ldots, J)$の文脈を重み付き関連語集合で表現する。ここで、ある単語 w の関連語とは、wとの共起に基づく相関が高い単語で、重み付き関連語集合$C(w)$とは、w との相関値を重みとしてもつ関連語の集合である。すなわち、$C(w) = \{w/\alpha\}$。ここに、αは単語 w と関連語 w の相関値である。例えば、「戦車」という日本語単語の重み付き関連語集合は以下のように記述される。

C(戦車) = {兵士/1492.7,…,戦闘/3082.3,…,爆弾/619.6,…}

この例では、「戦車」と関連語「兵士」の相関値は 1492.7、「戦車」と関連語「戦闘」の相関値は 3082.3、「戦車」と関連語「爆弾」の相関値は 619.6 である。

関連語の重みとなる相関指標として、様々なものが考えられる。本研究では、いくつかの相関指標をそれぞれ単独で用いる場合を比較するとともに、二つの相関指標を組み合わせて使用する方法を考案した。

(1) 単一の相関指標を使用する場合

単一の相関指標を用いる w の重み付き関連語集合$C(w)$は、コーパス中で w と一回以上共起した単語のうち、wとの相関値が上位 M%の単語を構成要素とし、各単語に wとの相関値を重みとして付与したものである。

相関値として、各単語の出現頻度とウィンドウ共起頻度に基づいて計算される以下の(a)~(d)の相関指標の値を用いる。

(a) 対数尤度比(LLR)
(b) 相互情報量(MI)
(c) カイ二乗(χ^2)
(d) discounted 対数オッズ比(LOR)

(2) 二つの相関指標を組み合わせて使用する場合

二つの相関指標を組み合わせて用いる w の重み付き関連語集合$C(w)$は、コーパス中で w と一回以上共起した単語のうち、wとの相関指標 1 の値が上位M_1%以内、wとの相関指標 2 の値が上位M_2%以内である単語を構成要素とし、各単語に x との相関指標 2 の値を重みとして付与したものである。予備実験の結果に基づいて、相関指標 1 として LLR、相関指標 2 として MI もしくは LOR を用いる場合を考える。二つの相関指標を用いる場合も考慮する理由は、単一の相関指標により抽出される関連語の種類に偏りが存在するためである。例えば、MI もしくは LOR を用いる場合は固有名詞（低頻度語）が多く含まれ、LLR を用いる場合は、一般的な単語（高頻度語）が多く含まれる傾向がある。しかし、固有名詞は普通名詞に比べて対訳辞書に含まれているものが少ないため、3.3 節で述べる文脈類似度の計算には寄与しない場合が多い。また、一般的な単語は多くの単語の重み付き関連語集合に含まれ、対訳でない単語対の文脈類似度を大きくしてしまうという問題がある。

そこで、異なる特徴を持つ二つの相関指標を組み合わせて使用する方法では、まず、関連語候補となる単語を、LLR のような高頻度語が大きい値を持つ相関指標により抽出することで、低頻度語のフィルタリングを行う。次に、MI や LOR のような低頻度語が大きい値を持つ相関指標によって、高頻度語のフィルタリングを行う。これにより、文脈類似度の信頼性向上が期待できる。

3.3 文脈類似度の計算

提案方法では、文脈類似度として、重み付き関連語集合の対応率を利用する。重み付き関連語集合対応率とは、図 2 の例のように、相手言語の関連語集合の語と対訳であるような関連語の割合を表す。「関連語集合の対応率が高い⇔文脈の類似度が高い」と言え、関連語集合の対応率が高い単語対は適切な対訳関係である可能性が高いとみなすことができる。

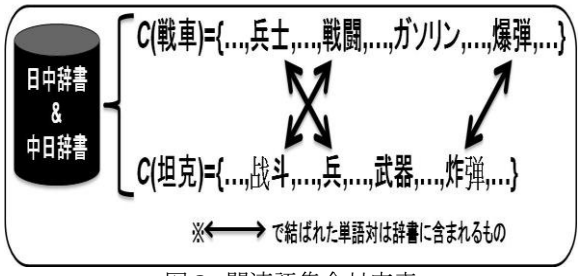


図 2: 関連語集合対応率

言語 1 の単語wと言語 2 の単語w'の関連語集合対応率$sim(C(w), C(w'))$は、$C(w) = \{w_p/\alpha_p\}, C(w') = \{w'_q/\alpha'_q\}$であるとき、次式で表される。

$$sim(C(w), C(w')) = \frac{1}{2}\left\{\frac{\sum_{p\in P}\alpha_p}{\sum_p \alpha_p} + \frac{\sum_{q\in Q}\alpha'_q}{\sum_q \alpha'_q}\right\}$$

ここに、$P = \{p | \exists w'_q \in C(w'), (w_p, w'_q) \in D_{12}\}$

$Q = \{q | \exists w_p \in C(w), (w'_q, w_p) \in D_{21}\}$

D_{12}は言語 1 から言語 2 への対訳辞書、D_{21}は言語 2 から言語 1 への対訳辞書を表す。本方法は、ある言語を媒介として多言語対訳辞書を生成する方法であるので、D_{12}やD_{21}として、既存の対訳辞書を利用できない場合がある。そこで、対訳辞書をマージして得られる対訳語の組の候補の集合で代替する。例えば、図 1 の例では、既存の DE-JP 辞書が存在しないため、DE-JP 間の対応率を計算するために、EN を媒介として得た DE-JP の対訳語候補の集合を用いる。

以上のように、異なる言語の文脈間の類似性を測る際に、一方の文脈を翻訳してコサイン係数などの類似度を計算するのではなく、関連語集合間の対応関係を利用する。その理由は、媒介言語を介して得られたノイズを含む対訳語候補の集合を対訳辞書として使用するためである。

4. 予備評価

4.1 実験の目的

英語を媒介言語として英日中の 3 言語対訳辞書を生成する実験を行った。実験の目的は次の二つである。

(1) 文脈類似度計算方法の比較

異なる言語の文脈間の類似度として、文脈を翻訳せずに関連語集合間の対応率を求める提案方法と、一方の文脈を翻訳し、コサイン係数を求める従来方法の二通りを比較する。

(2) 重み付き関連語集合生成方法の比較

重み付き関連語集合を生成する際に、単一の相関指標を使用する方法と、二つの相関指標を組み合わせて使用する方法を比較する。

4.2 使用データ

(1) コーパス

- 日本語: 毎日新聞 2000 年-2010 年(22.3 GB)
- 中国語: 新華社通信(LDC Chinese Gigaword 第 5 版) 2000 年-2010 年(4.24 GB)

(2-1) 英語を介した日中辞書

- EDR 日英辞書と LDC 中英辞書を反転して得られる英中辞書をマージした辞書(12.5MB, 566842 エントリ)

(2-2) 英語を介した中日辞書

- LDC 中英辞書と EDR 英日辞書をマージした辞書 (5.81MB, 240139 エントリ)

(3) テストデータ

英語を媒介として得られた日中の単語対の中からランダムに抽出した、ある一定の値以上の出現頻度を持つ単語対 1647 個。各日中単語対に対し、三名の独立した評価者による対訳かどうかの正誤判定結果を多数決して得られた正誤タグを付与した。付与したタグの内訳は正解 518 個、不正解 1129 個である。

4.3 実験結果

重み付き関連語集合の要素を選択する際のパラメータ M, M_1, M_2 は、別途用意したトレーニングデータの日中単語対を使用した実験において最も良い結果を示した値を、それぞれの相関指標毎に設定した。また、共起語を抽出する際のウィンドウサイズは $W = 10$ とした。

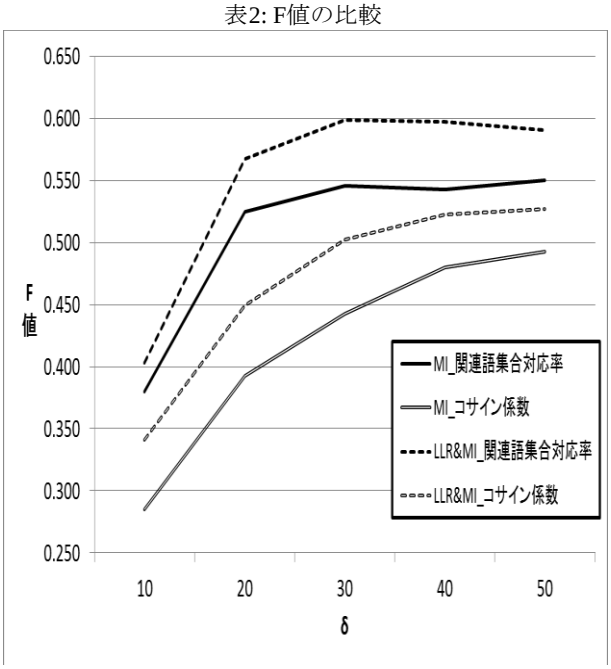
1647 個の単語対それぞれについて、日中単語間の関連語集合対応率を求め、対応率が上位 δ%の単語対を選択した時の精度、再現率、F 値を算出した。

提案方法と従来方法の二通りにおいて、関連語集合を生成する際に単一の相関指標を用いた場合と複数の相関指標を用いた場合それぞれで最も良い結果を得た指標毎に、その結果を以下の表 1 に示す。また、表 2 に F 値をグラフとして示す。

表1: 実験結果

関連語集合	類似度計算方法	相関指標	δ 10	δ 20	δ 30	δ 40	δ 50	
単一の相関指標により生成	関連語集合対応率	MI	**0.380**	**0.525**	**0.546**	**0.543**	**0.550**	F値
		MI	**0.787**	**0.675**	**0.559**	**0.485**	**0.448**	精度
		MI	**0.250**	**0.429**	**0.533**	**0.617**	**0.713**	再現率
	コサイン係数	MI	0.285	0.392	0.443	0.480	0.492	F値
		MI	0.591	0.505	0.453	0.429	0.401	精度
		MI	0.188	0.321	0.433	0.545	0.637	再現率
二つの相関指標により生成	関連語集合対応率	LLR&MI	**0.403**	**0.567**	**0.599**	**0.597**	**0.591**	F値
		LLR&MI	**0.835**	**0.729**	**0.613**	**0.533**	**0.481**	精度
		LLR&MI	**0.266**	**0.464**	**0.585**	**0.678**	**0.765**	再現率
	コサイン係数	LLR&MI	0.341	0.449	0.502	0.522	0.527	F値
		LLR&MI	0.707	0.578	0.514	0.467	0.429	精度
		LLR&MI	0.225	0.367	0.490	0.593	0.682	再現率

表2: F値の比較



4.4 結果の検討

4.4.1 文脈類似度計算方法

表 2 に示すように、提案方法は、重み付き関連語集合を生成する際に単一の相関指標を用いる場合、二つの相関指標を用いる場合それぞれにおいて、従来方法を上回る F 値を得た。一方の文脈を翻訳して異なる言語の文脈間の類似度を求める従来方法では、単語の多義性によって翻訳後の文脈にノイズが発生してしまうという問題がある。一方、提案方法は文脈を翻訳しないため、比較する関連語集合の内容は変化しない。したがって、集合にノイズが含まれてさえいなければ、用いる対訳辞書の質に左右され難い手法であると言える。

以下の表 3 と表 4 に、文脈類似度の計算に用いる対訳辞書を変更して同様の実験を行った結果を示す。表 3 は既存の日中、中日辞書を用いた場合の提案方法と従来方法それぞれの F 値、表 4 は既存の日中、中日辞書の各見出し語について、その訳語候補のうちの最も出現頻度が高い単語のみを訳語とした、一つの見出し語が一つの訳語しか持たない日中、中日辞書を用いた場合の F 値を示したものである。

表3: 実験結果(既存の辞書を使用した場合)

類似度計算方法	相関指標	δ 10	δ 20	δ 30	δ 40	δ 50
関連語集合対応率	LLR&MI	**0.421**	**0.574**	**0.650**	0.650	**0.664**
コサイン係数	LLR&MI	0.406	0.566	0.642	**0.651**	0.660

表4: 実験結果(一つの見出し語が一つの訳語しか持たない辞書を使用した場合)

類似度計算方法	相関指標	δ 10	δ 20	δ 30	δ 40	δ 50
関連語集合対応率	LLR&MI	**0.406**	**0.565**	**0.645**	0.635	**0.621**
コサイン係数	LLR&MI	0.397	0.553	0.640	**0.637**	0.609

表 3,表 4 から、提案方法では、文脈類似度の計算に既存の辞書やカバー率が低い辞書を用いた場合においても、従来方法を上回る結果を得ることができた。このように、対訳辞書の質に影響されにくいという特徴を持つ提案方法は、利用可能な対訳辞書が多くないマイナーな言語対の対訳関係を考える際に更に有効であると考えられる。

また、重み付き関連語集合の要素を選択する際のパラメータMの最適値は、重み付き関連語集合の対応率に基づいて文脈類似度の計算を行う場合が M=4~5%、コサイン係数に基づいて計算する場合が M=10~15%となった。このことから、提案方法では、より少ない関連語数、すなわち、低コストでの文脈類似度計算が期待できる。

4.4.2 重み付き関連語集合の生成方法

文脈類似度を関連語集合対応率で計算するとして、二つの相関指標を組み合わせて生成した重み付き関連語集合を用いた場合、δ=10 の時に 83.5%の精度を示し、単一の相関指標から生成した重み付き関連語集合を用いた場合の精度 78.7%を上回った。また、文脈類似度をコサイン係数で計算するとしたときも、二つの相関指標を組み合わせて生成した重み付き関連語集合を用いた場合の精度が高くなる傾向が見られた。

表 5 に、LLR と MI を単独で用いた場合と、LLR と MI を組み合わせた場合でそれぞれ抽出される、日本語の単語「石油」との相関値が特に高い関連語の一部を実例として示した。

表 5: 抽出される関連語の実例

"石油"の関連語(日本語)					
LLR単独		MI単独		LLR&MI	
大手	生産	シブネフチ	SPR	OPEC	液化
ロシア	公団	ウルーム	テキサコ	採掘	備蓄
石炭	備蓄	ETBE	キグナス	ガソリン	原油
経済	日本	サウジアラムコ	探鉱	コンビナート	ブレント
精製	企業	ルクオイル	モービル	ストーブ	天然ガス
機構	パイプ	新日鉱	JXHD	IEA	油田

3.2 節で述べた、各相関指標が持つ特徴のとおり、LLR により抽出される関連語の例には、「経済」,「日本」,「企業」などの、「石油」に関する文脈以外にも現れるような一般的な単語が多く、MI により抽出される関連語には、「石油」に関連した会社名や組織名を表す固有名詞が多い事が分かる。一方で、LLR と MI を組み合わせた場合に抽出される関連語は、「OPEC」,「ガソリン」,「コンビナート」など、対訳辞書に含まれていて、かつ、「石油」に関する文脈にしか現れないような単語が多い。

以上のことから、性質の異なる二つの相関指標を組み合わせて用いることで、見出し語を強く特徴付けるような多くの関連語を効果的に抽出できることが分かった。

6. おわりに

ある言語を介して得られる多言語の対訳語候補から正しい対訳語を選択するために、コンパラブルコーパスから抽出した文脈情報を用いる方法を提案した。本稿では、英語を媒介言語として英日中の対訳辞書を生成する実験を行い、その実現可能性を確認した。

予備実験では、文脈類似度の計算方法として考案した関連語集合対応率が対訳辞書の質に影響されにくいことを確認した。また、文脈を表現する重み付き関連語集合を生成する際に、性質の異なる二つの相関指標(対数尤度比と相互情報量)を組み合わせて用いる方法が有効であることを明らかにした。

今後の課題は、関連語集合の生成方法や対応率の計算方法を改良するとともに、他の言語対にも適用・評価することである。

謝辞：本研究は、一部、文部科学省科学研究費補助金 基盤研究(B)「多義性が解消された多言語辞書の自動構築に関する研究」(課題番号 22300032)の支援を受けた。

参考文献

Fung, Pascale and Lo YuanYee. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics*, pp. 414-420.

Kaji, Hiroyuki, Shin'ichi Tamamura, and Dashtseren Erdenebat. 2008. Automatic construction of a Japanese-Chinese dictionary via English. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pp. 699-706.

Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Michael Skinner, and Jeff Bilmes. 2009. Compiling a Massive, Multilingual dictionary via probabilistic inference. In *Proceedings of the 47th Annual Meeting of the ACL*, pp. 262-270.

Rapp, Reinhard. 1999. Automatic identification of word translation from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the ACL*, pp. 519-526.

Sammer, Marcus and Stephen Soderland. 2007. Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons. In *Proceedings of Machine Translation Summit XI*, pp. 399-406.

田中久美子,梅村恭司,岩崎英哉. 1998. 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.

張玉潔,馬青,井佐原均. 2005. 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.