様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年３月３１日現在

機関番号：１３８０１
研究種目：基盤研究（B）
研究期間：2010～2012
課題番号：２２３０００３２
研究課題名（和文）　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

研究課題名（英文）　Automatic Construction of a Sense-disambiguated Multilingual Dictionary

研究代表者
梶　博行（KAJI HIROYUKI）
静岡大学・情報学部・教授
研究者番号：20402232

研究成果の概要（和文）：多言語の機械翻訳や情報検索を支える多言語対訳辞書を既存の２言語対訳辞書から自動生成する新しい方法を開発した。共通の訳語を介して結ばれる２つの言語の語が対訳であるかどうかを、それぞれの言語のテキストデータから抽出される文脈の類似度に基づいて判定する。英日、英中の２言語対訳辞書と日本語、中国語それぞれの新聞記事データから英日中の３言語対訳辞書を生成する実験において適合率 83%、*F* 値 73%を達成し、提案方法の有効性を実証した。

研究成果の概要（英文）：A novel method was developed for producing a multilingual dictionary, which is indispensable for multilingual machine translation and information retrieval, from a set of existing bilingual dictionaries. Whether two words in different languages having one or more translations in common are genuinely translations of one another is judged based on the similarity between their contexts extracted from text corpora. Its effectiveness was demonstrated through an experiment on producing an English-Japanese-Chinese trilingual dictionary from English-Japanese and English-Chinese bilingual dictionaries and Japanese and Chinese newspaper corpora, in which 83% precision and 73% *F*-score were attained.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	5,500,000	1,650,000	7,150,000
２０１１年度	4,300,000	1,290,000	5,590,000
２０１２年度	4,100,000	1,230,000	5,330,000
総　計	13,900,000	4,170,000	18,070,000

研究分野：総合領域
科研費の分科・細目：情報学、メディア情報学・データベース
キーワード：コーパス・多言語処理

１．研究開始当初の背景

経済、社会のグローバル化とともに、さまざまな言語の間の機械翻訳システムや言語横断情報検索システムのニーズが高まっている。これらのシステムには対象とする言語の間の対訳辞書が必要であるが、大規模な対訳辞書が存在しない言語対も多い。そこで、既存の２言語対訳辞書を結合し、多言語対訳辞書を構築することが望まれる。

多言語対訳辞書とは、対象とする３つ以上の言語の語の組であって、どの２つの語をとっても互いに対訳であるような組をエントリーとする辞書である。多言語対訳辞書を構築することにより多言語機械翻訳システム

や多言語情報検索システムを効率よく実現することが期待される。また、語の多義性の構造は言語間で異なるので、多言語対訳辞書のエントリーは個々の言語の語からみると特定の語義を表している。したがって、多言語対訳辞書を語義の曖昧性解消に必要な語義目録として利用することも考えられる。

２．研究の目的

いくつかの言語対の対訳辞書から多言語対訳辞書を自動生成する手法を確立すること、特に、対訳辞書を結合して得られる対訳語の組の候補から正しい組を選択するために疎なコンパラブルコーパスを利用する方法を開発することが本研究の目的である。

２言語対訳辞書の集合から多言語対訳辞書を生成するには、共通の訳語をもつ語をつなぐことが基本となる。しかし、共通の訳語が多義語であると、対訳語の組だけでなく対訳でない語の組すなわちノイズが含まれる。例えば、英日の対訳語の組（tank，戦車）と英中の対訳語の組（tank，坦克）を結合した（tank，戦車，坦克）は対訳語の組であるが、英日の対訳語の組（tank，戦車）と英中の対訳語の組(tank，槽)を結合した（tank，戦車，槽）はノイズである。このような対訳語の組とノイズを分別するために２言語のコーパス（テキストデータ）を利用する方法を確立する。

２言語のコーパスは両言語テキストの対応の程度によってパラレルコーパスとコンパラブルコーパスに分けることができる。パラレルコーパスは対訳の文から構成されるコーパスである。コンパラブルコーパスは、狭義にはほぼ同じ内容がそれぞれの言語で書かれた文書の組の集合であり、広義には同じ分野／ジャンルのそれぞれの言語の単言語コーパスを組にしたものが含まれる。本研究では、広義のコンパラブルコーパス、言い換えると疎なコンパラブルコーパスに適用可能な方法を開発する。２言語対訳辞書が利用できない言語対ではパラレルコーパスや狭義のコンパラブルコーパスも利用できないことが多いからである。これに対し、疎なコンパラブルコーパスは多くの言語対に対して利用可能である。

３．研究の方法

(1) 多言語対訳辞書生成のフレームワークの提案

第３言語を介した対訳辞書生成方法を一般化することにより３言語以上の多言語対訳辞書生成のフレームワークを提案した。対象言語のうちのいくつかの言語対について２言語対訳辞書が利用でき、それらの対訳辞書をつなぐことによりすべての対象言語がつながることを前提とする。また、他の言語を介してしかつながらない言語対についてはコンパラブルコーパスが利用可能であるとする。

【提案方法】

２言語対訳辞書を対訳語の２つ組の集合と考え、すべての入力対訳辞書のエントリーの和集合を多言語対訳辞書 D の初期値とする。そして、以下の手続きを繰り返すことによりエントリーを逐次追加する。

１つ以上の語 $w_1, \cdots, w_K$ を共有する２つのエントリー（$w_1, \cdots, w_K, \cdots, w_I$）と（$w_1, \cdots, w_K, w'_{K+1}, \ldots, w'_J$）が D に含まれ、それらのエントリーの間で共有されない語のすべての組（w_i, w'_j）（i=K+1, …, I；j=K+1, …, J）が次のいずれかを満たすとき、２つのエントリーをマージした（$w_1, \ldots, w_I, w'_{K+1}, .., w'_J$）を D のエントリーとして追加する。

- w_i と w'_j が対訳であることを入力対訳辞書が示している。すなわち、（w_i, w'_j）$\in D$。
- w_i と w'_j が対訳であることがコンパラブルコーパスから推定される。すなわち、w_i が出現する文脈 $C(w_i)$ と w'_j が出現する文脈 $C(w'_j)$ の類似度 $Sim(C(w_i), C(w'_j))$ が閾値 θ 以上である。

図１に、英語(EN)、ドイツ語(DE)、日本語(JP)、中国語(CN)を対象言語とし、EN-DE, EN-JP, EN-CN の３つの対訳辞書と DE-JP, JP-CN, DE-CN の３つのコンパラブルコーパスから EN-DE-JP-CN の４言語辞書を生成する場合を例示する。

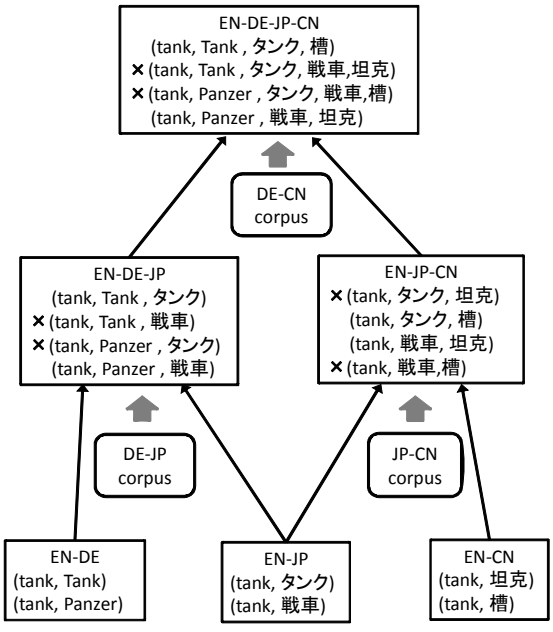

図１　多言語対訳辞書生成方法

(2) 疎なコンパラブルコーパスを用いた対訳／非対訳の判定方法の開発

意味の同じ語は似た文脈で使用されるという分布仮説に基づき、対訳語の組とノイズを分別するため語の出現する文脈の類似度を計算する。この考え方は、疎なコンパラブルコーパスからの対訳抽出に関する多くの研究で採用されている。文脈はそれぞれの言語のコーパスから抽出され、異なる言語の文脈を比較するために種となる対訳辞書を用いるのが一般的である。通常、両言語の文脈をそれぞれ単語ベクトルで表現し、一方を他方の言語に翻訳した上でコサイン係数などの類似度を計算する。しかし、本研究では、いくつかの言語対の対訳辞書をつないで得られるノイズを含む対訳辞書を種辞書として利用せざるを得ないので、通常の方法では文脈類似度の信頼度が低下する。このため以下のような方法を考案した。

① 文脈の抽出と表現―重み付き関連語集合―

語 w の出現する文脈を w の関連語の重み付き集合で表現する。関連語の重みとして w とのウィンドウ共起に基づく相関値を用いる。w と共起する語のすべてが w の語義を強く示唆するわけではないので、w と共起する語のうち w との相関値が上位 p %の語を w の関連語として採用する。

語と語の相関指標としてさまざまなものが考えられるので、対数尤度比(LLR)、相互情報量(MI)、対数オッズ比(LOR)、χ^2 スコアの４つを実験的に比較する。また、これらの相関指標は異なる特性をもつので、２つの相関指標を組み合わせる方法も比較する。LLR は高頻度語を過大評価する傾向があるのに対し、MI と LOR は低頻度語を過大評価する傾向がある。そこで、LLR の値が上位 q %に入り、MI または LOR の値が上位 r %に入る語を関連語として採用し、関連語の重みとしてはそれぞれMI、LORの値を用いる。これらのバリエーションをそれぞれ LLR-MI、LLR-LOR と略記する。

② 文脈類似度の計算―重み付き関連語対応率―

ノイズの多い種辞書の影響を抑えるため、文脈を陽には翻訳せずに類似度を計算する。すなわち、文脈類似度を求める２つの語のそれぞれについて「相手の語の関連語の少なくとも１つと対訳関係をもつ関連語の重みの和」と「すべての関連語の重みの和」の比を求め、それらの平均をとる。これを重み付き関連語対応率と呼ぶ。

言語１の単語 w の重み付き関連語集合が $C(w)=\{w_i/\alpha_i \mid i=1, \cdots, M\}$ （α_i が w_i の重み、すなわち w と w_i の相関値）、言語２の単語 w' の重み付き関連語集合が $C(w')=\{w'_j/\alpha'_j \mid j=1, \cdots, N\}$（$\alpha'_j$ が w'_j の重み、すなわち w' と w'_j の相関値）であるとき、重み付き関連語対応率 $Sim(C(w), C(w'))$ は次式で表される。

$$Sim(C(w),C(w')) = \frac{1}{2}\left(\frac{\sum_{i\in I}\alpha_i}{\sum_i \alpha_i} + \frac{\sum_{j\in J}\alpha'_j}{\sum_j \alpha'_j}\right)$$

ここに、$I = \{i \mid \exists w'_j \in C(w'), (w_i, w'_j) \in D_{12}\}$、

$J = \{j \mid \exists w_i \in C(w), (w'_j, w_i) \in D_{21}\}$。

D_{12} は言語１から言語２の対訳辞書、D_{21} は言語２から言語１の対訳辞書である。

(3) 実証実験

(1)(2)の提案方法を用いて、英日および英中の２言語対訳辞書と日本語および中国語のコーパスから英日中３言語対訳辞書を生成する実験を行った。提案方法の中の代替案については実験的に最適案を決定し、またパラメータの値を実験的に決定した。その上で、提案方法の結果と一方の文脈ベクトルを翻訳してコサイン係数を求める方法の結果（ベースライン）を比較した。

実験に使用したデータは次のとおりである。

① 入力データ

- ２言語対訳辞書
  - 英日：EDR 英日／日英対訳辞書
  - 英中：LDC Chinese-English Lexicon
- コンパラブルコーパス
  - 毎日新聞記事（2000 年～2010 年、22.3GB）と LDC Chinese Gigaword の新華社通信記事（2000 年～2010 年、4.24GB）
  - Wikipedia の日本語版（2012 年 8 月 26 日現在、3.1GB）と中国語版（2012 年 9 月 1 日現在、0.7GB）

２つのコーパスを別々に用いた実験を行った。Wikipedia は記事が対応づけられたコンパラブルコーパスであるが、記事の対応情報は利用せず、提案方法をそのまま適用した。

- 文脈類似度計算のための種辞書

EDR 辞書と LDC 辞書を結合したノイズを含む日中辞書。比較のため EDR 日中辞書を用いた実験も行った。

② テストデータ

EDR 辞書と LDC 辞書を結合して得られる英日中辞書のエントリー候補から、毎日新聞記事コーパス中に 2,500 回以上出現する語と新華社通信記事コーパス中に 2,500 回以上出現する語を含む組に限定してランダムに選んだ 3,000 組。日中バイリンガルの学生３人に各組の正誤を判定してもらい、３人の判定結果の多数決により正誤を決定した。3,000 組

のうち正（対訳３つ組である）が 1,053 組、誤(対訳３つ組でない)が 1,947 組であった。

③ 評価指標

さまざまなセッティングで提案方法を実行し、3,000 組のテストデータを文脈類似度の降順に並べ、上位 10%, 20%, 30%, ,,, を対訳語の組と判定したときの適合率、再現率、*F* 値を求めた。異なるセッティングの優劣を比較する際は、それぞれの最大 *F*値を比較した。*F* 値は適合率と再現率の調和平均であるが、適合率と再現率の重みを２：１とした。その理由は、コーパスを用いる提案方法では、対訳語の組であっても、コーパス中に用例が含まれなければ棄却されるからである。再現率はコーパスのカバー率という側面をもつ。重要なのは上位にランクされた組の精度である。

４．研究成果

(1) 疎なコンパラブルコーパスを用いた対訳／非対訳の判定方法の有効性の実証

① 文脈の抽出と表現に関し、関連語を抽出するために２つの相関指標（具体的には対数尤度比と相互情報量）の組合せが有効であることを明らかにした。表１は相関指標の比較結果である。

表１ 相関指標の比較

(a) 新聞記事コーパス

相関指標	適合率	*F*値
LLR	0.721	0.631
MI	0.704	0.616
LOR	0.788	0.689
χ^2	0.717	0.622
LLR-MI	**0.833**	**0.729**
LLR-LOR	0.829	0.725

(b) Wikipedia コーパス

相関指標	適合率	*F*値
LLR	0.664	0.646
MI	0.711	0.691
LOR	0.792	0.693
χ^2	0.717	0.632
LLR-MI	**0.796**	**0.696**
LLR-LOR	0.708	0.688

② 文脈類似度の計算に関し、提案方法である重み付き関連語対応率の有効性を実証した。表２は重み付き関連語対応率とコサイン係数(ベースライン)の比較結果である。

表２ 文脈類似度計算方法の比較

(a) 新聞記事コーパス

文脈類似度	相関指標	適合率	*F*値
重み付き関連語対応率	LOR	0.788	0.689
	LLR-MI	**0.833**	**0.729**
コサイン係数	LLR	0.622	0.609
	LLR-MI	0.783	0.691

(b) Wikipedia コーパス

文脈類似度	相関指標	適合率	*F*値
重み付き関連語対応率	LOR	0.792	0.693
	LLR-MI	**0.796**	**0.696**
コサイン係数	LLR	0.708	0.531
	LLR-MI	0.775	0.684

表３から明らかなように、参照する種辞書がノイズを含まない辞書であっても、重み付き関連語集合がコサイン係数より優れている。

表３ 文脈類似度計算方法と種辞書の組合せの比較

文脈類似度	種辞書	適合率	*F*値
重み付き関連語対応率	英日・英中	0.833	0.729
	EDR 日中	**0.842**	**0.743**
コサイン係数	英日・英中	0.783	0.691
	EDR 日中	0.817	0.721

（注）新聞記事コーパス、LLR-MI 使用

(2) 多言語対訳辞書自動生成の実現可能性の実証

(1)の成果を得たことにより、２言語対訳辞書の集合からの多言語対訳辞書自動生成の実現に近づいた。疎なコンパラブルコーパスに適用可能な方法であるので、さまざまな言語を対象にすることができる。

提案方法のようなコーパスを用いる方法で得られる結果は当然のことながらコーパスに依存する。これは長所でもあり短所でもある。コーパスの分野に適応した多言語対訳辞書が生成されるので、機械翻訳などの応用システムは対象分野で最良のパフォーマンスを示すことが期待される。しかし、汎用的な多言語辞書を構築するには、さまざまなコーパスを用いて得られた対訳語の組を累積していくことが必要である。新聞記事コーパスから得られた結果と Wikipedia 記事コーパスから得られた結果をマージすることによりカバー率が向上することを確認しており、これが実際的な方法であるといえる。

(3) 今後の課題と展望

① 第４、第５の言語への適用

本研究では英日中の３言語対訳辞書の生成実験にとどまったが、提案方法は任意個の言語を対象とすることができる。新たに対象とする言語に必要なものは、既に対象となっている言語の少なくとも一つとの間の対訳辞書、単言語コーパスおよび形態素解析プログラムである。この条件を満たす第４、第５の言語への適用を検討したい。

② 文脈類似度に基づく対訳／非対訳の判

定方法の改良―語の文脈から語義の文脈へ―

文脈類似度を利用する現在の方法には基本的な問題点がある。それは、抽出される文脈はそれぞれの語が表すすべての語義の文脈の総和であり、個々の語義を特徴づける文脈ではないということである。分布仮説は語義ごとにいえることであり、似た語義を持つ語でも異なる語義で用いられるときの文脈が似ているわけではない。このため、多くの語義をもつ語はどの訳語とも文脈類似度が高くならないことがある。特定の語義で用いられる頻度が高い語はそれ以外の語義での訳語との文脈類似度は低くなる。したがって、文脈類似度を利用する方法の性能を高めるには語義ごとに文脈を求めることが必要になる。

１つの方法として、同一の語義を特徴づける関連語どうしの相関が高いことを利用して関連語のクラスタを求めることが考えられる。個々の語がもつ語義の数、そのうちコーパス中に用例が含まれる語義の数とも未知であるので難しい問題であるが、改良の方向を示している。

(4) 関連研究との比較

本研究と同様な多言語辞書の構築をめざした研究として University of Washington の研究グループによる PanDictionary プロジェクトがあげられる。そこでは２つのアプローチが試みられている。１つは、２言語対訳辞書の集合から翻訳グラフを生成し、その構造から多言語の対訳関係を推定する方法である。もう１つはコンパラブルコーパスを利用する方法で、対訳語の組の候補中のハブ言語（英語）の語を含む文と類似のスポーク言語文の中に、候補中のスポーク言語の語が含まれるとき、対訳語の組であると判定する。後者は本研究と近いが、ハブ言語とスポーク言語の類似文の検索を基本としており、本研究より密なコンパラブルコーパスを必要とする方法と思われる。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計９件）

① 山元陽祐, 綱川隆司, 梶博行, “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,” 言語処理学会第 19 回年次大会, 2013 年 3 月 14 日, 名古屋大学.

② Hiroyuki Kaji, Takashi Tsunakawa and Yoshihiro Komatsubara, “Improving Compositional Translation with Comparable Corpora,” Proceedings of the 5th Workshop on Building and Using Comparable Corpora, pp. 134-142, May 26, 2012, Lütfi Kırdar Convention & Exhibition Centre, Istanbul, Turkey.

③ 榊原徹也, 綱川隆司, 梶博行, “コンパラブルコーパスを用いた WordNet の自動翻訳,” 言語処理学会第 18 回年次大会, 2012 年 3 月 15 日, 広島県立大学.

④ 小松原慶啓, 綱川隆司, 梶博行, “コンパラブルコーパスと Web を用いた用語翻訳器,” 言語処理学会第 18 回年次大会, 2012 年 3 月 15 日, 広島県立大学.

⑤ Hiroyuki Kaji, Takashi Tsunakawa and Yoshihiro Komatsubara, “Term Translation Using Comparable Corpora and the Web,” The 11th Japan-China Natural Language Processing Joint Research Promotion Conference, October 29, 2011, Hotel Plaza Miyazaki, Miyazaki, Japan.

⑥ Takashi Tsunakawa and Hiroyuki Kaji, “Word Translation Disambiguation with Using Syntactic Co-occurrence Information and Word Classes,” The 10th Japan-China Natural Language Processing Joint Research Promotion Conference, November 5, 2010, Soochow University, Suzhou, China.

〔その他〕

ホームページ等

http://nlp.cs.inf.shizuoka.ac.jp/

６．研究組織

(1)研究代表者

梶　博行（KAJI HIROYUKI）

静岡大学・情報学部・教授

研究者番号：20402232

(2)研究分担者

許山　秀樹（NOMIYAMA HIDEKI）

静岡大学・情報学部・教授

研究者番号：10257230

綱川　隆司（TSUNAKAWA TAKASHI）

静岡大学・情報学部・助教

研究者番号：30611214

(3)連携研究者

なし