

第3言語を介した対訳辞書の自動生成

玉村 真一[†], Dashtseren Erdenebat[‡], 梶 博行[‡]

[†] 静岡大学大学院情報学研究所, [‡] 静岡大学情報学部情報科学科

あらまし 本稿では、一方の言語が同一である2つの対訳辞書から新しい言語対の対訳辞書を生成する新しい方法を提案する。提案方法の特徴は、第1言語と第2言語それぞれの単一言語コーパスを用いることにより、媒介となる第3言語の語の多義性によって生み出される誤った訳語を除去することである。提案方法は、それぞれの言語のコーパスから関連の強い語のペアを抽出し、第1言語の関連語ペアと第2言語の関連語ペアのアラインメントをとることにより、第1言語の1つの語の各関連語がそれぞれ第2言語のどの訳語を支持するかを決定する。そして、一定の比率以上の関連語に支持される訳語を選定する。誤った訳語を支持する関連語は非常に少ないので、誤った訳語は除去されることになる。日英と中英の対訳辞書から日中対訳辞書を生成する予備実験を行い、提案方法のフィジビリティを確認した。

キーワード 対訳辞書, リソースの乏しい言語対, 単一言語コーパス, 機械翻訳, クロス言語情報検索

Automatic Construction of a Bilingual Dictionary via a Third Language

Shin'ichi Tamamura, Dashtseren Erdenebat, and Hiroyuki Kaji

Department of Computer Science, Shizuoka University

Abstract This paper proposes a method for constructing a bilingual dictionary for a new language-pair from two bilingual dictionaries that share one of their languages. The main feature of the proposed method is that it uses monolingual corpora of the first and second languages to eliminate spurious translations caused by the ambiguity of intermediary third-language words. The method extracts pairs of associated words from each of the corpora, aligns them across languages, and, for a first-language word, determines which of its associated words supports which of its translations in the second language. Then, it selects translations supported by more than a certain proportion of the associated words and, as a result, eliminates spurious translations, which are likely to be supported by few associated words. A preliminary experiment, in which a Japanese-Chinese dictionary is constructed from Japanese-English and Chinese-English dictionaries, demonstrates the feasibility of the proposed method.

Key words bilingual dictionary, resource-scarce language pair, monolingual corpus, machine translation, cross-language information retrieval

1. はじめに

機械翻訳やクロス言語情報検索のニーズがさまざまな言語対に拡がってきている。しかし、大規模な対訳辞書が利用できる言語対は限られている。このため、新しい言語対の対訳辞書を低コストで作成することが課題となっている。

大規模な対訳辞書が存在しない言語対であっても、それぞれの言語と第3の言語（たいていは英語）の間に大規模な対訳辞書が存在するケースが多い。そこで、第1言語-第3言語の対訳辞書と第2言語-第3言語の対訳辞書から第1言語-第2言語の対訳辞書を合成することが考えられる。第1言語の語と第2言語の語が対訳関係にあるとき、それらに共通の第3言語の訳語が存在することを利用した方法である。しかし、媒介となる第3言語の訳語が多義性をもつ場合、誤った対訳語ペアが生み出されるという問題がある。

この問題に対し、田中ら（1998）は、媒介となる第3言語の訳語の数が多いほど正しい対訳語ペアである可能性が高いという仮説をたて、その有効性を実証した。しかし、正しい訳語、誤った訳語ともに媒介となる訳語が1つという語も多く、この考え方だけで十分な精度を達成することは困難である。田中らの方法を基本としながら、意味カテゴリが一致するペアを優先する方法（Bond, et al., 2001）、あるいは品詞や構成文字の関連性を考慮する方法（Zhang, et al., 2007）も提案されているが、十分な解決には至っていない。

本稿では、第1言語と第2言語のコーパスを利用して誤った訳語を除去する方法を提案する。ここで、両言語のコーパスは同じ分野のコーパスであればよく、パラレル/コンバラブルコーパスである必要はない。対象とする言語対では、通常、大規模なパラレル/コンバラブルコーパスの存在が期待できないことを考慮した結果である。また、2.2で述べるように、提案方法

では、誤った訳語だけでなく、コーパスの分野であまり使用されない意味の訳語も除去される。したがって、あらゆる分野をカバーする対訳辞書を作成することはできない。このことは提案方法の欠点であるが、自然言語処理では分野へのチューニングが効果的であることを考えると、むしろ望ましい特徴といえる。提案方法は任意の言語対に適用可能であるが、本稿では英語を介して日本語－中国語の対訳辞書を生成する場合について述べる。

2. 提案方法

2.1 課題

図1は、日英対訳辞書と中英対訳辞書を結合して得られる日中対訳辞書の一部である。日本語の見出し語「工場」に対して、「工場」、「厂」など正しい訳語のほか、「作品」、「栽植」、「植物」、「种植」など誤った訳語（以下、ノイズという）が含まれている。これらは、媒介となる「works」や「plant」の多義性に起因している。このようなノイズを含む対訳辞書からノイズを除去することが本研究の課題である。

2.2 基本アイデア

Kaji (2005) は、対象分野の2つの言語のコーパスを用いて対訳辞書を分野に適応させる方法を提案した。それぞれの言語のコーパスから相互に関連の強い語のペアを抽出し、対訳辞書を介して関連語ペアのアラインメントをとることによって、見出し語ごとに各関連語が支持する訳語を決定し、支持する関連語の比率が高い訳語を選定する方法である。この方法における対訳辞書としてノイズを含む対訳辞書を用いれば、ノイズを除去することができる。見出し語（例：「工場」）の関連語とノイズ（例：「栽植」）の関連語の間に対訳関係が成立する可能性は小さく、ノイズを支持する関連語の比率が低くなるからである。なお、この方法の本来の目的から当然のことであるが、ノイズだけでなく、使用するコーパスの分野であまり使われない意味の訳語も除去される。

ここで、上記方法で工夫された点について説明しておく。対訳辞書を介した関連語ペアのアラインメントが基本であるが、それには次の問題点がある。

- (a) アラインメントの失敗： 両言語のコーパスは内容的に必ずしも対応していないので、また対訳辞

日本語	英語	中国語
工場	works	厂, 词, 作品
	plant	安插, 事业, 栽, 栽植, 栽种, 植, 植物, 植物体, 植株, 种, 种上, 种植, 蔚
	workshop	车间, 工场, 工作室, 作坊
	mill	磨坊, 磨矿, 磨
	factory	厂, 厂家, 工厂, 制造厂

日英：EDR, 中英：LDC

図1 日英・英中対訳辞書の結合

書のカバレッジに限界があるので、アラインメントがとれない関連語ペアが多数生じる。

- (b) アラインメントの曖昧性： 対訳辞書を介するだけでは偶然対応が見つかる関連語ペアもあり、複数の関連語ペアとアラインメントが成立する関連語ペアも多い。

(a)や(b)の場合にも関連語が支持する訳語を決定することができるように、「関連語と訳語の関連度」を反復計算する方法となっている。具体的には2.3の(3)に示すが、互に関連する関連語は同じ訳語を支持するという仮説に基づいて、関連語と訳語の関連度を他の関連語と当該訳語の関連度を用いて定義することにより、(a)に対処している。また、関連語ペアのアラインメントと比較して、関連語トリプレットのアラインメントがより信頼できることに着目して、(b)を解決している。すなわち、関連語と訳語の関連度の定義において、見出し語および当該関連語とのトリプレットがアラインメントに成功する関連語が、トリプレットがアラインメントに失敗する関連語より、大きく寄与するようにしている。

2.3 提案方法

図2に示すように、提案方法は以下のステップからなる。日中対訳辞書を生成する場合について述べるが、以下の記述において日本語と中国語を入れ替えると中日対訳辞書を生成する方法になる。

- (1) 日英・中英対訳辞書の結合

日英対訳辞書と中英対訳辞書から、共通の英訳語をもつ日本語の語と中国語の語のペアをすべて求めることにより、日中対訳辞書を生成する。この辞書にはノイズが含まれる。

- (2) 関連語ペアの抽出

日本語コーパス、中国語コーパスそれぞれから関連

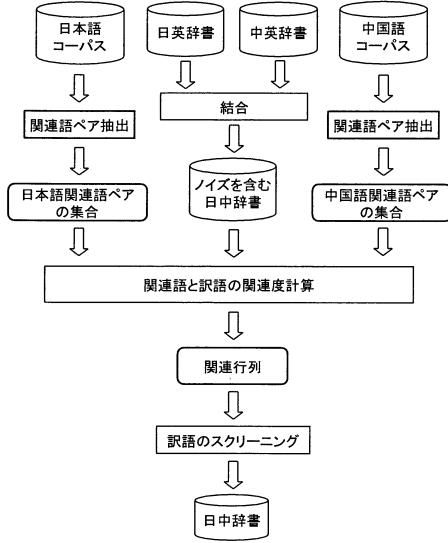


図2 提案方法の概要

語ペアを抽出する¹。関連語ペアとは、予め定めた閾値以上の相互情報量をもつ語のペアである。二つの語 t と t' の相互情報量 $MI(t, t')$ は次式で定義される (Church and Hanks, 1990)。

$$MI(t, t') = \log_2 \frac{\Pr(t, t')}{\Pr(t) \cdot \Pr(t')} \quad [1]$$

ここに、 $\Pr(t)$ は t の生起確率、 $\Pr(t, t')$ は t と t' の共起確率である。

語と語の共起確率を求めるため、語と語の共起頻度をカウントする。本方法の目的にはトピック的な関連語が有効と思われるので、前後の文にまたがるような比較的大きなウィンドウを設定し、ウィンドウ内に共起する頻度をカウントする。関連語ペアとして内容語のペアを対象とするが、両言語のウィンドウの範囲が釣りあうよう、ウィンドウサイズは機能語を除き内容語の数で指定する (3.の実験ではウィンドウサイズを ± 25 内容語とした)。

(3) 関連語と訳語の関連度の計算

日本語の各見出し語に対して、(2)で抽出された関連語と(1)で得られた訳語の間の関連度を計算する。関連語と訳語の関連度は[2]~[4]式で再帰的に定義され、反

¹ 実験システムを作成するにあたって、(独) 情報通信研究機構から中国語形態素解析プログラムの提供を受けた (Nakagawa and Uchimoto, 2007)。

復計算される。式中の添え字 n は反復計算のサイクルを示す。なお、初期値は[5]式で与えられる。

- 見出し語 x の第 i 関連語 x'_i と第 j 訳語 y_j の関連度:

$$C_n(x'_i, y_j) = MI(x'_i, x) \cdot \frac{C_n(x'_i, y_j) + \alpha \cdot C_n(x'_i, y_j)}{\max_k [C_n(x'_i, y_k) + \alpha \cdot C_n(x'_i, y_k)]} \quad [2]$$

ここに、 $MI(x'_i, x)$ は x'_i と x の相互情報量、 α は関連度の第1因子に対する関連度の第2因子の相対的な重みを表すパラメータである。

- 関連度の第1因子:

$$C_n(x'_i, y_j) = \sum_{\substack{x'' \in A(x) \\ \wedge x'' \in A(x'_i)}} C_{n-1}(x'', y_j) \quad [3]$$

ここに、 $A(x)$ は x の関連語の集合である。

- 関連度の第2因子:

$$C_n(x'_i, y_j) = \sum_{\substack{x'' \in A(x) \\ \wedge x'' \in A(x'_i)}} \max_{\substack{y' \text{ s.t.} \\ (x'_i, y') \in D \\ \wedge y' \in A(y_j) \\ \wedge \exists y'' [(x'', y'') \in D \\ \wedge y'' \in A(y_j) \\ \wedge y'' \in A(y')]}} MI(y_j, y') \cdot C_{n-1}(x'', y_j) \quad [4]$$

ここに、 D は対訳辞書に含まれる対訳語ペア全体の集合である。

- 関連度の初期値:

$$C_0(x'_i, y_j) = MI(x'_i, x) \quad [5]$$

[2]式からわかるように、関連度 C は「関連語と見出し語の相互情報量」と「2つの因子 C' と C'' の重みつき和を (最大値が1になるように) 正規化した値」の積である。第1因子 C' は、アラインメントとは無関係に定義されており、2.2 で述べた問題点(a)を解決する。第2因子 C'' は、関連語トリプレットのアラインメントに成功する関連語のみから定義されており、2.2 で述べた問題点(b)を解決する。

反復計算の収束の様子などの詳細は (Kaji and Morimoto, 2005) を参照されたい。

(4) 訳語のスクリーニング

(3)で計算された関連語-訳語関連行列を2値化する。すなわち各行 (関連語) において最大値をもつ要素を1、それ以外の要素を0にする。そして、列 (訳

語)ごとに、1の要素の比率を算出する。この比率は、各関連語が1つの訳語を支持すると考えたときに各訳語を支持する関連語の比率を表すので、支持度と呼ぶ。最後に、支持度が予め定めた閾値以上となる訳語を選定する。

関連語－訳語関連行列と支持度 σ の例を図3に示す。 σ の閾値を0.1とすれば、この例では、日本語見出し語「工場」の中国語訳語として「工厂」と「厂」が選定される。

3. 予備実験

3.1 実験方法

提案方法により EDR 日英対訳辞書と LDC Chinese-English Translation Lexicon から日中対訳辞書と中日対訳辞書を生成する予備実験を行った。使用したコーパスは次のとおりである。

- ・日本語コーパス：毎日新聞記事(2000年1月～2005年12月、632Mバイト)
- ・中国語コーパス：新華社通信記事(1991年1月～1995年12月、300Mバイト)

日本語、中国語とも、コーパス中の出現頻度が800以上の名詞に限定して関連語ペアを抽出した。相互情報量の閾値は0と低めに設定したが、一つの語に対する関連語を相互情報量が上位の300語に制限した。

次に、抽出された関連語ペアの集合に基づいて関連語－訳語関連行列を計算した。すなわち、出現頻度が800以上の見出し語について、出現頻度が800以上の訳語候補を対象として関連行列を計算した。ただし、どの訳語候補の出現文脈も見出し語の出現文脈と接点をもたないため、関連行列が初期値のまま終わる見出し語があり、それらは除外した。

このようにして関連語－訳語関連行列が計算された見出し語(日本語3695語、中国語1462語)を対象として、以下の訳語スクリーニング実験を行った。

(i) 実験 A (単方向スクリーニング)

2.3で述べた方法により、日中対訳辞書と中日対訳辞書をスクリーニングした。いずれの方向でも、支持度 σ の閾値は0.1とした。

(ii) 実験 B1 (双方向スクリーニング1)

日中対訳辞書と中日対訳辞書の両方をスクリーニングし、得られた2つの対訳辞書に共通に含まれる対訳

工場	工厂	厂	植物	栽植	...
機械	7.36	6.53	3.11	2.60	...
製造	6.97	7.15	2.89	2.73	...
稼働	5.40	4.83	2.63	3.01	...
飼料	1.67	0.97	6.32	5.70	...
⋮	⋮	⋮	⋮	⋮	⋮
σ	0.553	0.407	0.022	0.018	...

図3 関連語－訳語関連行列と支持度

語ペアを選定した。共通に含まれるという条件は厳しいので、 σ の閾値は0.025と低く設定した。

(iii) 実験 B2 (双方向スクリーニング2) -

実験 B1と同様に日中対訳辞書と中日対訳辞書の両方をスクリーニングし、得られた2つの対訳辞書の少なくとも一方に含まれる対訳語ペアを選定した。少なくとも一方に含まれるという条件は甘いので、 σ の閾値は0.25と高く設定した。

(iv) 実験 C (再スクリーニング)

単方向スクリーニングされた対訳辞書を用いて、再度、単方向スクリーニングを行った²。 σ の閾値は、サイクル1では0.025、サイクル2では0.1とした。

3.2 実験結果

3.2.1 見出し語あたり訳語数

日中対訳辞書のスクリーニング前後の見出し語あたり訳語数の分布と平均を図4に示す。ここで、スクリーニング前と実験Aについては実験対象の見出し語全体に対する値を示したが、実験B1、B2、Cの値は若干異なる見出し語集合に対する値である。実験B1では、中日対訳辞書のスクリーニング結果との共通部分をとったため訳語数が0になった見出し語を除外している。実験B2では、中日方向のスクリーニングで訳語として選定されたために日中対訳辞書の見出し語に加えられた語を含んでいる。実験Cでは、サイクル2で対象語からはずれた見出し語が若干あり、それらを除外している。

図4からわかるように、どの実験でもスクリーニン

² 再スクリーニングする理由は、ノイズを含む対訳辞書を用いた関連語ペアのアラインメントは誤りが多いと思われるからである。サイクル2では、スクリーニングされた対訳辞書の使用を関連語ペアのアラインメントのために限定し、関連語－訳語関連行列はスクリーニング前の訳語すべてを対象として計算した。

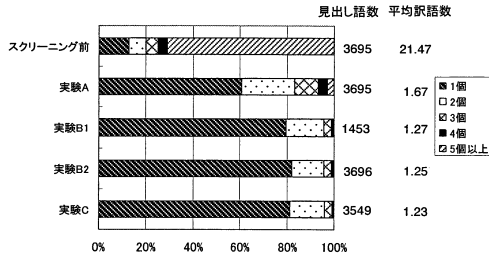


図4 見出し語あたり訳語数の分布

グにより訳語数が大幅に減少し、大部分の見出し語に対する訳語数が1になっている。これは、提案方法が、ノイズだけでなく、コーパスで使用されていない意味の訳語も除去したためである。

3.2.2 訳語スクリーニング結果の例

表1に訳語スクリーニングの結果を例示する。日本語の見出し語「電車」、「新聞」、「ホール」と中国語の見出し語「報」、「大厅」、「论文」について、訳語候補（スクリーニング前の訳語）と各実験によって選定された訳語を示している。また、EDR 日中辞書の訳語を参考として示している。

表1から、提案方法がある程度意図どおりに働いていることがわかる。同時に、改良の余地が大きいことを表している。例えば、「電車」に対する「训练」などのノイズは除去されたが、「新聞」に対するノイズ「论文」は除去されなかった。表2に、訳語の支持度とともに、各訳語を支持する関連語を例示する。「火车」と

「训练」の文脈の差違に比べると、「報」と「论文」の文脈の差違は小さいので、関連語と訳語の関連度の計算で混乱が生じたものと思われる。

また、訳語として正しい場合でも、コーパスで用いられている意味の訳語であるとは限らないことに注意が必要である。表2を見ると、「電車」に対する「火车」、「新聞」に対する「报」などについては、訳語とそれを支持する関連語の関係は妥当と思われるが、「ホール」に対する訳語「漏洞」とそれを支持する関連語「バーディー」、「バット」の関係は妥当でないように思われる。

3.3 結果の検討

(1) パラメータの値

訳語スクリーニングの精度は期待したレベルに達していない。1つの大きな原因として、出現頻度の閾値が大きすぎたことがあげられる。高頻度の見出し語を対象とするのはよいが、訳語や関連語の閾値はもっと小さい値（30程度）にすべきであった。訳語に対する閾値が高すぎたために、選定されるべき訳語が足りられるケースが多くみられた（例えば、「ホール」に対する訳語「穴」や「洞」）。また、関連語に対する閾値が高すぎたため、対応する相手言語の関連語ペアをもつ関連語ペアが少なく、関連語-訳語関連行列の精度が低下したと思われる。他のパラメータを含め、最適値を明らかにすることが今後の課題である。

(2) 双方向スクリーニング

支持度の閾値によって結果が変わってくるので、今

表1 訳語スクリーニング結果の例

	見出し語	訳語候補	実験A	実験B1	実験B2	実験C	EDR日中辞書
日中辞書	電車	带带, 吊吊, 火车, 教养, 练, 列车, 培训, 培养, 培育, 培植, 训, 训练, 斗车	火车	火车	火车	火车	电气列车, 电车
	新聞	报, 报纸, 论文, 论文儿, 纸, 纸头, 纸张, 报界, 报刊, 新闻界	报, 论文	报, 论文	报	报, 论文	新闻, 新消息, 报纸
	ホール	大堂, 大厅, 殿, 霍尔, 堂, 厅, 洞, 洞穴, 洞子, 空穴, 孔, 孔洞, 孔隙, 窟, 窟窿, 窿, 漏洞, 穴, 阱, 堀	大厅, 漏洞	大厅	大厅	大厅, 漏洞	球穴
中日辞書	报	うわさ話, お披露目, ニュース, ベーパー, ベーブル, リポート, ルボ, レボ, レポート, 会籍, 便り, 償い, 石継ぎ, 噂話, 報, 報劣, 報劣金, 報告, 報道, 寄せ文, 届け出, 御報, 情報, 新聞, 流伝, 爆音, 物言, 申しわたし, 申し出, 矢返, 知らせ, 礼金, 言づて, 言伝, 評判, 話, 調書, 謝礼, 返し, 通報, 他41	ニュース	ニュース, 便り, 新聞	ニュース, ベーパー, 報道, 届, 新聞, 調書, 謝礼	ニュース, 報道	日報
	大厅	ホール, 会堂, 会館, 僧堂, 堂, 堂宇, 大広間, 大間, 広座敷, 広敷, 広敷き, 広間, 廊下, 殿, 殿堂, 館	館, 廊下	館, 廊下, ホール	ホール, 会館, 廊下, 殿堂, 館	廊下, 館	大広間, 大間, 広間, 講堂
	论文	ばあ, ぱー, テーゼ, パア, ベイパー, ベーバ, ベーパー, ベーブル, リポート, レポート, 作文, 地紙, 定立, 横紙, 提題, 新紙, 新聞, 新聞紙, 書きもの, 書留, 横紙, 正, 用紙, 紙, 紙製, 西洋紙, 論文, 論考	論文	論文, 新聞	ベーパー, リポート, 作文, 紙, 論文	論文	考, 論, 論文, 論考

表2 支持度と支持する関連語の例

	見出し語	訳語*	支持度 σ	支持する関連語の例**
日中辞書	電車	火车	0.946	快速, 運休, 線路, 痴漢, 路面, 駅員
		培养	0.027	環状, 欲望, 正面, 列車, 支社
		训练	0.027	運転士, 指令, 発, 床
	新聞	报	0.866	新聞社, わこ, 読売, 配達, 活字
		论文	0.134	紙面, PTA, 製紙, 表彰式, 記者
	ホール	大厅	0.829	ソプラノ, 紀尾井, ベートーベン, 3500円, リサイタル, バイオリン
漏洩		0.171	バーディー, パット, ボギー, ウッズ	
中日辞書	報	ニュース	0.849	経済参考報, 専発, 文章, 评论員
		便り	0.040	谈话, 连, 公众, 上周, 干
		新聞	0.030	社论, 评论, 党风, 纪律, 全军
	大厅	館	0.599	展览, 新闻中心, 陪同, 屏幕, 大会堂
		廊下	0.365	营业, 轮椅, 遗体, 象棋, 鲜花
		ホール	0.027	佳节, 典礼, 一片, 演出, 代
	論文	論文	0.896	刊物, 学术, 专家学者, 博士学, 专著
		新聞	0.043	角度, 设想, 规律, 学科, 讨论

* $\sigma \geq 0.025$ 以上の訳語のみを示す。 ** 関連度の大きい順

回の実験で結論を出すことはできなかったが、双方向スクリーニング1は少数の確実な訳語に絞るのに適していると思われる。双方向スクリーニング2は訳語候補を多めに選定し、最終判断を人間に委ねる使い方に適していると思われる。

(3) 再スクリーニング

サイクル1と2の結果にあまり違いがなかった。ノイズを含む対訳辞書を用いても、関連語トリプレットのアラインメントの曖昧性がそれほど増大しなかったと推察される。すなわち、関連語-訳語関連行列の反復計算アルゴリズムは、対訳辞書のノイズに対する耐性が十分であるといえる。

(4) 既存の電子化辞書との比較

EDR 日中辞書は、EDR 日英対訳辞書をベースとして(独)情報通信研究機構が開発中の電子化辞書である(Zhang, et al., 2007)。表1からわかるように、そのカバーレージはまだ十分といえない。提案方法は、コーパスに基づいて使用頻度が高い訳語を選定するので、EDR 日中辞書の増補を支援するツールとしても有用と思われる。

最後に、提案方法は、従来の対訳辞書には含まれていない有用な情報を提供することを付け加えておく。支持度 σ は、どの訳語がメジャーな訳語であることを示している。 σ が最大の訳語をデフォルト訳語とすればよい。また、各訳語を支持する関連語を訳語選択に利用することができる。

4. おわりに

第3言語を介して合成された第1言語-第2言語の対訳辞書に含まれる誤った訳語を除去するため、第1言語と第2言語の単一言語コーパスを用いて見出し語の関連語が支持する訳語を決定し、一定の比率以上の関連語に支持される訳語を選定する方法を提案した。そして、日英と中英の対訳辞書から日中対訳辞書を作成する予備実験を行い、フィージビリティを確認した。提案方法は、コーパスの分野に適応した対訳辞書を作成する、どの訳語がメジャーであるかを明らかにする、訳語選択に利用できる関連語を訳語に付与するなど、従来方法にない特徴をもつ。

提案方法はいくつかのパラメータを含むが、それらの最適値を明らかにした上で精度を評価することが今後の課題である。また、当面の目標である日中対訳辞書の作成では、漢字の関連性を利用する手法(Zhang, et al., 2007)を組み合わせたことも興味深い。

謝辞: 本研究の一部は科学技術振興調整費によるプロジェクト「日中・中日言語処理技術の開発研究」および(株)日立製作所との共同研究として実施した。

参考文献

- Bond, Francis, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. 2001. Design and construction of a machine-tractable Japanese-Malay dictionary, In *Proc. Machine Translation Summit VIII*, pp. 53-58.
- Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, Vol. 16, No. 1, pp. 22-29.
- Kaji, Hiroyuki. 2005. Adapting a bilingual dictionary to domains, *IEICE Trans. Inf.&Syst.*, Vol. E88-D, No.2, pp. 302-312.
- Kaji, Hiroyuki and Yasutsugu Morimoto. 2005. Unsupervised word-sense disambiguation using bilingual comparable corpora. *IEICE Trans. Inf.&Syst.*, Vol. E88-D, No.2, pp. 289-301.
- Nakagawa, Tetsuji and Kiyotaka Uchimoto. 2007. A hybrid approach to word segmentation and POS Tagging. In *Proc. 45th Annual Meeting of the ACL*, pp. 217-220.
- 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.
- Zhang, Yujie, Qing Ma, and Hitoshi Isahara. 2007. Building Japanese-Chinese translation dictionary based on EDR Japanese-English bilingual dictionary. In *Proc. Machine Translation Summit XI*, pp. 551-557.