Wikipediaからの対訳用語対の抽出

Extracting Bilingual Lexicons from Wikipedia

岡崎 直観[*1] 劉 瀟[*1] 綱川 隆司[*2] 辻井 潤一[*1*3*4]
Naoaki Okazaki Xiao Liu Takashi Tsunakawa Jun'ichi Tsujii

[*1]東京大学 The University of Tokyo
[*2]静岡大学 Shizuoka University
[*3]英国マンチェスター大学 University of Manchester, UK
[*4]英国国立テキストマイニングセンター National Centre for Text Mining, UK

This paper explores approaches for acquiring bilingual lexicons from Wikipedia, in which encyclopedic knowledge is accumulated by editors all over the world. The first half of this paper reports the usefulness of manually-annotated clues in Wikipedia (inter-language links, redirect pages, and Wiki links) for extracting Japanese-Chinese bilingual lexicons. Comparing the bilingual lexicons obtained by these simple clues with a large-scaled bilingual dictionary, we find that Wikipedia and the existing dictionary have small amount of lexical overlaps. In order to bridge the gap between the existing dictionary and Wikipedia, we build a machine translation system for technical terms. This system obtains translation probability models of phrases and hanzi-kanji letters from the dictionary and language models from all of the Wikipedia articles. We also analyze co-occurrence statistics and distributional similarity of bilingual terms in Wikipedia, and integrate these measures to the translation system.

1. はじめに

対訳辞書は，機械翻訳 [Brown 90] や言語横断検索 [Nie 99] など，複数の言語を扱うアプリケーションにおいて，欠かせない言語資源である．これまで，高い構築コストにも関わらず，様々な言語間において大規模な対訳辞書が作られてきた．しかし，文化・科学の発展に共に新しい概念・実体が生み出されたり，既存の概念・実体に対して別の表現が追加されるなど，言語は絶えず変化している．とりわけ，専門用語は活発に生み出されているため，一般的な対訳辞書ではカバーしきれずに，正しく翻訳できないことがある．そこで，翻訳するドメインに適応した専門用語対訳辞書を個別に用意し，新語を登録・管理することで，翻訳の精度を維持する必要がある．

本研究では，世界中の編集者が共同で構築しているウィキペディアを言語資源として分析し，対訳用語対を自動的に獲得する．ウィキペディアには，言語間リンク，転送ページ，ウィキリンク等，対訳辞書の獲得に有用と思われる情報が付与されており，これらの有用性を検証する．さらに，既存の対訳辞書とウィキペディアの統計情報を統合し，ウィキペディアから対訳用語対を抽出する試みを紹介する．なお，本稿では，今後需要の増加が見込まれる日本語と中国語の対訳用語抽出に焦点を当てる．提案手法の核となるアイディアはこれらの言語に依存しないので，形態素解析や翻訳モデルの構築など，言語に依存する箇所を入れ替えることで，他の言語対にも適用できる．

2. 関連研究

既存の対訳辞書の語彙不足は，古くから問題とされており，文書から対訳項目を（半）自動的に抽出する手法が研究されてきた [Fung 98, Rapp 99, Shao 04, Robitaille 06, Haghighi 08]．これらの研究では，問題設定に応じて並列コーパス，コンパラブル・コーパス，もしくは単言語コーパスが用いられる．2語間の対訳関係の推定には，既存の対訳辞書，綴りの類似性，翻字関係，周辺に存在する語の類似性，コンパラブル・コーパスにおける共起性などの特徴が用いられた．

近年，ウィキペディアを大規模な言語リソースと見なして，固有表現の曖昧性解消 [Bunescu 06]，意味的な関連性の計算 [Strube 06]，シソーラス構築 [中山 06] などのタスクに応用する研究が盛んに行われている．本稿では，ウィキペディアから対訳辞書を自動獲得する手法を紹介する [Erdmann 08, Ramírez 08, 新井 08]．Erdmann ら [Erdmann 08] は，ウィキペディアの言語間リンクで結ばれている記事のタイトルを対訳として抽出し，転送ページ，ウィキリンクのアンカーテキストを用いて対訳語彙項目を拡張する手法を提案している．Ramírez [Ramírez 08] らは，日西英のシソーラスを構築するため，ウィキペディアから言語間リンクを対訳として取り出し，英語の記事の内容やカテゴリ情報をもとに，対訳項目を WordNet シソーラスに関連付ける手法を提案した．新井ら [新井 08] は，ウィキペディアの言語間リンクの接続パターン（方向性など），言語間リンクの数，既存の辞書との比較を報告している．これらの先行研究は，言語間リンク，転送ページ，カテゴリなど，記事同士を編集者が直接的に結びつけた情報に基づいている．第 3. 節で調査するように，これらの情報はウィキペディアのごく一部であり，ウィキペディアのリソースの利用効率という面において，課題が残る．

3. ウィキペディアと既存の辞書の比較

3.1 言語間リンクに基づく対訳辞書

まず，ウィキペディアの執筆者によって明示的に付与された情報のみを利用し，日中対訳用語対を獲得する予備実験を行った．実験には，2008 年 10 月 19 日付の日本語版ウィキペディ

連絡先: 岡崎直観 <okazaki at is i u-tokyo ac jp>
東京都文京区本郷 7-3-1
東京大学大学院情報学環・学際情報学府
(03) 5841-4120

番号	辞書抽出処理	対訳対の数
1	言語間リンク	113,697
2	1 + 英語表現削除	105,738
3	2 + 数値表現削除	94,395
4	3 + 参照削除	94,184
5	4 + クリーニング	87,536
6	5 + 転送ページ拡張	88,018
7	6 + ウィキリンク拡張	135,495

表 1: ウィキペディアの日中言語間のリンクに基づく対訳辞書

ア[*1]，2008 年 11 月 17 日付の中国語版ウィキペディア[*2]を用いた．表 1 に，日中言語間のリンク（interlanguage link），転送ページ（redirect page），ウィキリンクなどの情報に基づいて，得られた対訳用語対の数を示した．なお，以降の説明では，日本語の語 j と中国語の語 c のペアを $\langle j, c\rangle$ と略記する．

双方向，もしくは片方向の言語間リンクで結ばれている日本語と中国語の記事を抽出し，そのタイトル同士を対訳用語対と見なすと，113,679 件の用語対が獲得できた（表 1 の 1.）．ここから，以下の条件を満たす不適切なペアを削除すると，94,184 件の用語対が残った．

- 2. 片側もしくは両側の語が英数字のみで構成される
  例: ⟨“ページランク”, “PageRank”⟩
- 3. 片側もしくは両側の語が時間表現のみで構成される
  例: ⟨“2008 年”, “2008 年”⟩
- 4. 片側もしくは両方の語に参照表現（#）がある
  例: ⟨“池袋駅#東京地下鉄”, “新線池袋站”⟩

さらに，4. に対して，⟨“Category:社会保障”, “Category:社會保障”⟩における “Category:” のような名前空間を表す接頭辞，⟨“ラブストーリー (韓国映画)”, “假如愛有天意”⟩ に見られるような括弧表現を削除し，用語対の異なり数をカウントしたところ，87,536 件の用語対が得られた（5.）．この用語対の中から 400 件をランダムにサンプリングし，対訳辞書項目としの適切さを人手で評価したところ，83.3%の精度であった．

Erdmann ら [Erdmann 08] は，転送ページと，他の記事からのウィキリンクを用いて，対訳語彙項目を拡充する手法を提案している．転送ページによる語彙拡張では，例えば w_r という記事を w_j に転送する指示があり[*3]，5. の用語対の中に $\langle w_j, w_c\rangle$ というペアが存在するとき，$\langle w_r, w_c\rangle$ を対訳対に追加する．他の記事からのウィキリンク（他の項目へのリンク，もしくは内部リンクとも呼ばれる）を用いる語彙拡張では，ある記事 w_o がアンカーテキスト w_a で記事 w_j へのリンクを貼っていて[*4]，6. の用語対の中に $\langle w_j, w_c\rangle$ というペアが存在するとき，$\langle w_a, w_c\rangle$ を対訳対に追加する．これらの語彙拡張手法は語の曖昧性を増加させるため，間違った対訳ペアを追加してしまう恐れがあるが，約 9.4 万個の対訳対を約 13.5 万個に増加させることができた．

3.2 既存の辞書との比較

さて，こうして得られた対訳用語対は，既存の日中対訳辞書と比較して，どのような性質を持っているのだろうか？ 我々は，クロスランゲージ社の大規模な日中対訳辞書と，5. 及び 7. で得られた対訳用語対の，完全一致による重なり具合を調べた（表 2）．クロスランゲージ社の辞書には，618,400 件の日本語語彙，623,098 件の中国語語彙，及び 1,069,887 件の対訳語彙対が収録されている．5. で得られた辞書には，83,532 件の日本語語彙，85,111 件の中国語語彙，87,536 件の対訳用語対が収録されているが，これら 2 つの辞書の両方に現れる対訳用語対は，8,656 件しかなかった．言い換えると，クロスランゲージ社の用語対の 0.8%がウィキペディアの辞書（5.）でカバーされ，ウィキペディアの辞書（5.）の 9.9%がクロスランゲージ社の辞書でカバーされていることになる．このように双方の辞書の重なりが非常に小さいのは，双方の言語資源がカバーしている語彙の違いに起因すると思われる．

対訳辞書	日本語	中国語	日中語彙対
既存の対訳辞書	618,400	623,098	1,069,887
多言語ウィキペディア (5.)	83,532	85,111	87,536
両辞書の語彙の積集合	14,927	12,782	8,656
既存の対訳辞書	618,400	623,098	1,069,887
多言語ウィキペディア (7.)	100,823	94,441	133,659
両辞書の語彙の積集合	15,761	13,203	9,009
既存の対訳辞書	618,400	623,098	1,069,887
単言語ウィキペディア	901,072	334,967	N/A
両辞書の語彙の積集合	62,360	26,550	N/A

表 2: クロスランゲージの対訳辞書とウィキペディアの比較

さらに，日本語と中国語のウィキペディアの記事タイトルを単言語の語彙とみなすと，901,072 件の日本語の語，334,967 件の中国語の語が得られる．7. で得られた辞書と比較すると，日本語のウィキペディア記事タイトルの 11.2%，中国語のウィキペディア記事タイトルの 28.2%が，7. の辞書を用いて相手方の言語に翻訳できることが分かる．また，クロスランゲージ社の対訳辞書と比較すると，日本語ウィキペディア記事タイトルの 6.9%，中国語ウィキペディア記事タイトルの 15.2%が，それぞれ相手方の言語に翻訳できる．

これらの予備調査から，ウィキペディアの記事タイトルや言語間リンクで得られる対訳用語対と，既存の大規模対訳辞書との間には，語彙的な重なりが小さいことが分かった．特に，語彙対レベルで重なりを測ると，既存の辞書とウィキペディアの溝は大きく，既存の辞書とウィキペディアの両方の情報を，活用する方法が求められている．

4. 提案手法

4.1 日中用語翻訳システム

そこで，本研究では，既存の対訳辞書から得られる翻訳の知識と，ウィキペディアの統計情報の両方を用い，対訳用語対の抽出を試みる．既存の対訳辞書に対して完全一致で用語の翻訳を行うと，語彙の些細な違いにより訳せないことがある．例えば，⟨“浮動小数点加算”, “浮点加”⟩，⟨“加減算”, “加減”⟩ が対訳辞書に登録されているとき，“浮動小数点加減算” を中国語に訳すのは，コンピュータにとって自明ではない．このとき，⟨“浮動小数点”, “浮点”⟩，⟨“加算”, “加”⟩ という翻訳知識を獲得できていれば，“浮動小数点加減算” は “浮点加減” に訳されると推論できる．

このような柔軟性を既存の対訳辞書に持たせるため，対訳辞書をコーパスと見なして，統計的機械翻訳モデルを構築した．本システムは，通常の句ベース統計的機械翻訳に加え，日本語

*1 http://download.wikimedia.org/jawiki/20081019/jawiki-20081019-pages-articles.xml.bz2
*2 http://download.wikimedia.org/zhwiki/20081117/zhwiki-20081117-pages-articles.xml.bz2
*3 記事 w_r に `#REDIRECT [[`w_j`]]` という Wiki 記述が存在.
*4 記事 w_o に `[[`w_a`|`w_j`]]` という Wiki 記述が存在.

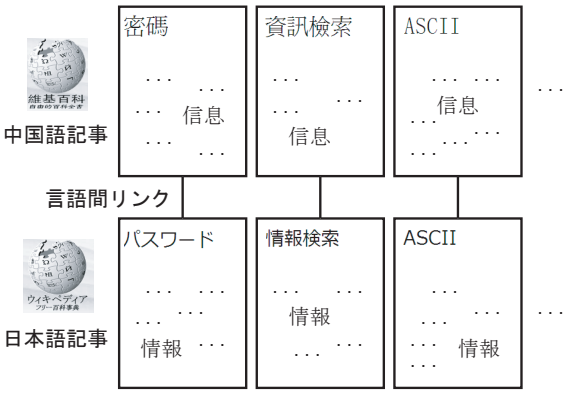

図 1: 言語間リンクで結ばれた対訳用語

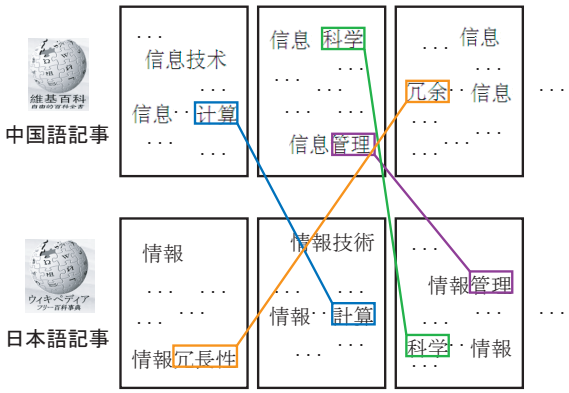

図 2: 2 言語間の単語分布の類似性

と中国語の漢字の翻訳知識を取り込み，日中用語の翻訳精度を向上させている [Tsunakawa 09]．例えば，中国語の「浮」という漢字は日本語でも同じ字「浮」に訳されやすいが，中国語の「站」という漢字は日本語では別の漢字「駅」に訳されやすい．このような日中の漢字の対応関係を知っていると，用語の翻訳の大きな手がかりになる．本システムは日中漢字の対訳関係（漢字翻訳確率）を対訳辞書から学習し，翻訳知識として利用している．翻訳システムの構築には，句ベース統計翻訳のツールキットである Moses[*5] を用いた．言語モデルには，ウィキペディアのすべての記事のテキストから抽出した 5-gram を用いた．日本語の形態素解析には JUMAN[*6] を，中国語の形態素解析には Multilingual Morphological Analyzer (MMA) [Kruengkrai 09] を用いた．

4.2 ウィキペディアからの翻訳知識獲得

既存の対訳用語辞書に加え，ウィキペディアからも翻訳に有用と思われる手がかりを抽出する．言語間リンクで結ばれたウィキペディアの記事は，同一の事柄を別の言語で説明したものであるから，それらの記事中に含まれる語も類似すると考えられる．逆に考えると，中国語の語 c と日本語の語 j があるとき，これらの語を含む記事が多くの言語間リンクで結ばれていれば，c と j は翻訳関係にある可能性が高くなる．例えば，図 1 では，〈“情報”, “信息”〉という対訳ペアが，「パスワード」「情報検索」「ASCII」について書かれた記事に出現している．

中国語の語 c と日本語の語 j の結びつきの強さを，中国語の語 c からウィキペディアの言語間リンクを辿って，日本語の語 j に到達する条件付き確率で測る．

$$P_{\mathrm{co}}(j|c) = \frac{\mathrm{freq}(j,c)}{\mathrm{freq}(c)} \tag{1}$$

ここで，$\mathrm{freq}(j,c)$ は語 j と c を含むウィキペディアの記事を結ぶ言語間リンクの数，$\mathrm{freq}(c)$ は語 c を含むウィキペディアの記事数である．

さらに，単語の分布仮説 [Harris 54] を 2 言語間に拡張し，中国語の語 c と日本語の語 j の類似度を測定する．図 2 は，「情報」という語の周辺に「技術」「計算」「科学」を意味する語が，日本語と中国語の記事に共通して出現している例を示している．今回は，中国語の語 c の前後 5 語を文脈語とし，それぞれの語を既存の辞書で日本語に翻訳してから周辺の単語分布 $R_c(w)$ を作り，日本語の語 j の前後 5 語を文脈語とした単語分布 $Q_j(w)$ と比較する．類似度尺度には，α-skew ダイバージェンス [Lin 98] を採用し，$\alpha = 0.99$ とした．

$$\mathrm{Sim}(j,c) = -\mathrm{D}_{\mathrm{skew}}\left(Q_j(w)||R_c(w)\right), \tag{2}$$

$$\mathrm{D}_{\mathrm{skew}}(Q||R) = \mathrm{KL}\left(R||\alpha Q + (1-\alpha)R\right), \tag{3}$$

$$\mathrm{KL}(Q||R) = \sum_w Q(w)\log\frac{Q(w)}{R(w)} \tag{4}$$

ここで，$\mathrm{Sim}(j,c)$ は j と c の類似度で，値域は $(-\infty, 0]$ である．なお，与えられた語に対してウィキペディア全体から文脈語を効率よく収集するため，転置インデックスシステムである Tokyo Dystopia[*7] を用いた．

5. 実験

ウィキペディアの言語間リンクで獲得した辞書（5.）を正解とみなし，中国語の用語を日本語に翻訳する実験を行った．この実験では，システムに中国語の用語を与え，出力された日本語訳が正解の辞書と一致するかどうか，正解率と文字ベース BLEU で測定した．正解率は，(正しく翻訳された用語の数)/(翻訳した用語の総数) である．本来，BLEU は単語の n-gram の一致率を測る尺度であるが，今回の実験では一つの用語を構成する単語の数が少ないので，文字単位の n-gram（bi-gram まで）で一致率を測った．

実験に用いたシステムは，4.1 節で説明した日中用語翻訳システムと，4.2 節で説明したウィキペディアの統計情報を使う．入力した中国語の語は，日中用語翻訳システムを経て日本語に翻訳される．ウィキペディアの統計情報と組み合わせるときは，翻訳システムから出力される 10-best 翻訳に対し，言語リンクにおける共起確率，単語の類似度などでリランキングを行った．リランキングでは，日本語訳 j の出力順位，中国語の語 c に対する日本語訳 j の翻訳確率推定値 $P_{\mathrm{SMT}}(j|c)$，言語リンクにおける共起確率 $P_{\mathrm{co}}(j|c)$，単語の類似度 $\mathrm{Sim}(j,c)$ を素性とし，最大エントロピー法で素性の重みを学習した．

表 3 に，用語翻訳システムのみを用いた場合と，ウィキペディアの統計情報を追加した場合の，正解率と BLEU スコアを示した．純粋な用語翻訳システムのみの正解率は 22.3%で，ウィキペディアの統計情報を組み込むことで，正解率は 25.1%に増加した．25%程度の正解率しか達成できなかった要因とし

*5 http://www.statmt.org/moses/
*6 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
*7 http://tokyocabinet.sourceforge.net/dystopiadoc/

システム条件	正解率	BLEU
用語翻訳のみ	.223	.375
+Wikipedia 素性	.251	.381

表 3: 言語間リンクで結ばれているタイトルの翻訳実験

て，自動評価の厳密さ*8，未知語，翻字，繁字体と簡字体の差異などが挙げられる．今回の実験において，翻訳元の用語の56.3%において，翻訳システムが学習時に遭遇していない未知語が存在していた．未知語を含む用語を正常に翻訳出来ないとすると，正解率の上限値は 43.7%くらいであり，既知の語で構成されている用語だけを評価すれば，57.4%が正しく翻訳できたことになる．未知語は特に人名や地名に多く見られるが，今回の翻訳システムには翻字（transliteration）に関する処理を一切行っていないため，翻訳対象の人名や地名がそのまま出力され，誤訳と見なされるケースが目立った．

また，字体の違いという中国語ウィキペディア特有の問題もある．中国語のウィキペディアでは，台湾や香港で用いられる繁体字で執筆することが原則となっており，中国大陸やシンガポールなどで用いられる簡字体は，サーバー側の自動変換で表示している．しかし，繁体字と簡字体の差は文字に留まらず，「ビリヤード」が繁体字では「撞球」，簡字体では「台球」と表されるように，用いる語彙が異なるケースも多い．今回の実験では，用語翻訳システムは簡字体の辞書で学習しており，翻訳対象のウィキペディアは繁体字を簡字体に自動変換したものであるから，字体の影響を受けやすい条件であったと言える．

最後に，中国語のウィキペディアの全てのタイトルを日本語に翻訳する実験を行った．得られた日本語訳のうち，ランダムに選んだ 400 個を人手で評価したところ，正解率は 35.5%であった．また，日本語ウィキペディアの既存のタイトルに一致する日本語訳を数えたところ，19,868 件見つかった．言い換えれば，翻訳元の中国語の用語と，提案したシステムで翻訳した日本語の訳を対訳用語対と見なしたとき，19,868 個の対訳用語対が得られたことになる．このうち，7,908 個は言語間リンクで抽出した辞書（7.）に含まれておらず，まだ言語間リンクで結ばれていない記事対を発見することができた．

6. 結論

本稿では，ウィキペディアから対訳辞書を獲得する手法として，記事の編集者が明示的に付与した情報を使うアプローチと，既存の大規模辞書から構築した用語翻訳システムを用いるアプローチを紹介した．ウィキペディアの言語間リンク等で対訳辞書を構築しても規模が小さいこと，既存の辞書を利用してウィキペディアの用語を翻訳しても，未知語の翻訳が課題になることが分かった．今後は，簡体字と繁体字の扱いを改良すると共に，翻字モデルを導入し，固有表現における未知語の取り扱いを改善していきたいと考えている．

謝辞

本研究の一部は，文部科学省科学研究費補助金特別推進研究「高度言語理解のための意味・知識処理の基盤技術に関する研究」および科学技術振興調整費・重要課題解決型研究等の推進「日中・中日言語処理技術の開発研究」の助成を受けています．また，日中対訳辞書を提供して頂いた株式会社クロスランゲージに感謝いたします．

*8 自動評価では，例えばシステムの翻訳が「デカルト座標系」，正解が「直行座標系」のとき，両者は同じ物であっても，システムの訳を正解と判定できない．

参考文献

[Brown 90] Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S.: A statistical approach to machine translation, *Computational Linguistics*, Vol. 16, No. 2, pp. 79–85 (1990)

[Bunescu 06] Bunescu, R. and Marius Pa s.: Using Encyclopedic Knowledge for Named entity Disambiguation, in *Proc. of EACL 2006*, pp. 9–16 (2006)

[Erdmann 08] Erdmann, M., Nakayama, K., Hara, T., and Nishio, S.: Extraction of Bilingual Terminology from a Multilingual Web-based Encyclopedia, 情報処理学会論文誌, Vol. 49, No. 7, pp. 2468–2479 (2008)

[Fung 98] Fung, P. and Yee, L. Y.: An IR approach for translating new words from nonparallel, comparable texts, in *Proc. of Coling 1998*, pp. 414–420 (1998)

[Haghighi 08] Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D.: Learning Bilingual Lexicons from Monolingual Corpora, in *Proc. of ACL-08: HLT*, pp. 771–779 (2008)

[Harris 54] Harris, Z. S.: Distributional Structure, *Word*, Vol. 10, pp. 146–162 (1954)

[Kruengkrai 09] Kruengkrai, C., Uchimoto, K., Kazama, J., Wang, Y., Torisawa, K., and Isahara, H.: An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging, in *Proc. of ACL-IJCNLP 2009*, p. (to appear) (2009)

[Lin 98] Lin, D.: Automatic retrieval and clustering of similar words, in *Proc. of COLING-ACL 1998*, pp. 768–774, Montreal, Quebec, Canada (1998)

[Nie 99] Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R.: Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web, in *Proc. of ACM SIGIR 1999*, pp. 74–81 (1999)

[Ramírez 08] Ramírez, J., Asahara, M., and Matsumoto, Y.: Japanese-Spanish Thesaurus Construction Using English as a Pivot, in *Proc. of IJCNLP 2008*, pp. 473–480 (2008)

[Rapp 99] Rapp, R.: Automatic identification of word translations from unrelated English and German corpora, in *Proc. of ACL 1999*, pp. 519–526 (1999)

[Robitaille 06] Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S., and Utsuro, T.: Compiling French-Japanese Terminologies from the Web, in *Proc. of EACL 2006*, pp. 225–232 (2006)

[Shao 04] Shao, L. and Ng, H. T.: Mining new word translations from comparable corpora, in *Proc. of Coling 2004*, p. 618 (2004)

[Strube 06] Strube, M. and Ponzetto, S. P.: WikiRelate! Computing Semantic Relatedness Using Wikipedia, in *Proc. of AAAI 2006*, pp. 1419–1424 (2006)

[Tsunakawa 09] Tsunakawa, T., Okazaki, N., Liu, X., and Tsujii, J.: A Chinese-Japanese Lexical Machine Translation through a Pivot Language, *ACM Transactions on Asian Language Information Processing (TALIP)*, Vol. 8, No. 2, p. (to appear) (2009)

[新井 08] 新井 嘉章, 福原 知宏, 増田 英孝, 中川 裕志：Wikipedia の言語間リンクに関する分析, 第 22 回人工知能学会全国大会, pp. 2D3–02 (2008)

[中山 06] 中山 浩太郎, 原 隆浩, 西尾 章治郎：Wikipedia マイニングによるシソーラス辞書の構築手法, 情報処理学会論文誌, Vol. 47, No. 10, pp. 2917–2928 (2006)