自動的に作成された洪日辞書の評価

VARGA István　　横山 晶一　　橋本 力

山形大学大学院理工学研究科
dyn36150@dip.yz.yamagata-u.ac.jp, {yokoyama, ch}@yz.yamagata-u.ac.jp

概要

本研究では我々が以前開発した自動的作成方法による洪日辞書の評価を行った。内的評価では、「適合率」と「見出し語の頻度に基づく再現率」が関連研究の方法で作成された辞書の結果を上回った。外的評価では、ウエブ上の評価システムを作成した。この評価システムは、オンラインの特性を生かして検索欄などを容易に追加できる。また、このような自動的に作成された辞書に基づくコミュニティベースの辞書システムについても検討する。

1. はじめに

現在は電子化辞書が自然言語処理の様々な分野、また言語教育用 CAI ソフトウエアの不可欠なツールになっている。機械翻訳においては新たな言語ペアの翻訳システム開発のために、コンピュータ用辞書を早急に作成する必要がある。しかし、多くの言語ペアの場合、人材や時間的なコストが高く、電子化辞書の自動的作成方法が必要になる。特に、使用頻度が低い言語ペアの場合は信用度が高い一般的な参照辞書がないため、辞書作りとその評価が更に問題になってくる。

ハンガリー語と日本語の場合は洪日辞書に必要な信頼できる辞書、また膨大な二言語コーパスや他のデジタル化された資源がほとんど存在しない。我々はすでに、頻度が低い言語ペアの一例としてハンガリー語と日本語を選び、中間言語として英語の辞書やオントロジーを利用し洪日辞書を作成した(Varga & Yokoyama, 2007)。

本論文ではまず、これまでに提案された辞書の自動的な作成方法の問題点について述べ、次に本研究における辞書の作成方法を説明する。作成した洪日辞書の評価についても述べる。また、ウェブ上のコミュニティベースの辞書システムについても述べる。

2. 現在の方法と問題点

一般的な辞書は「記述的」と「規範的」という二つの作成方法に分けられる。自動的辞書作成にもこの二つの方法が適用できる。膨大な二言語コーパスから見出し語ペアを抽出する方法は記述的方法だと考えられる。それに対して中間言語の辞書で見出し語ペアの適応性を判断する方法は規範的である。

二言語コーパスを利用する方法は 80 年代から研究され、様々な精度の高い提案がある(Brown 等、1998；Kay & Röscheisen、1993； Brown、1997)。

中間言語に基づく方法は最初に Tanaka によって提案された(Tanaka & Umemura、1994)。Tanaka の方法は中間言語として英語への”harmonized dictionary”と“inverse consultation”を利用し、結果として仏日辞書を作成した。これ以来様々な方法が提案されたが、それらの多くが対象となっている言語の特徴（漢字表記など）を利用するため、他の言語ペアの場合は適応できない(Shirai & Yamamoto、2001； Paik & Bond & Shirai、2001)。2005 年にどの言語ペアにも適応できる方法として情報抽出に利用されている IDF で計算したスコアで判断する方法が提案された(Sjöbergh、2005)。

現在まで提案された方法の問題点は以下の 2 点にまとめられる。

a.　適応性

使用頻度が低い言語ペアには膨大な二言語コーパスがないため、記述的方法は適用できない。また、規範的方法では対象言語の特徴を用いる方法が別の言語ペアでは扱えない。

b.　見出し語の訳の文字列的制限

これまでの中間言語に基づく方法は中間言語への二言語辞書にある訳の”lexical overlap”(共通の文字列をマッチする方法)で辞書を生成している。しかし、二言語辞書の意味的情報が限られているため、このように生成された辞書の再現率と適合率が低い。

自動的作成方法のためには、ある見出し語の意味の訳だけではなく、意味の徹底的な説明が必要である。二言語辞書からこの情報が得られないため、同じ見出し語の訳が辞書によって異なる。その結果で”lexical overlap”を行うと、再現できない翻訳ペアが多い。

3. 本研究の辞書作成方法の特徴

二言語コーパスを利用する自動作成方法の精度は高いが、日本語とハンガリー語の場合はこの方法が必要とする膨大なコーパスがまだ開発されていない。ここで新たな中間言語に基づく方法を開発した。これまでの方法と同様に第1ステップでは見出し語ペア候補を二言語辞書によって生成した。例えば、「購入」の英訳として”purchase(別の意味：増力)”と”buy”が抽出され

た。”buy”と”purchase”に対しハンガリー語-英語辞書からハンガリー語の訳は合計 10 件が抽出された。第2ステップでは、これまでの方法と異なり、英語のオントロジー(WordNet)を利用することによって第1ステップで生成された見出し語ペアの適切さを”semantic overlap”で判断する。候補の見出し語ペアのそれぞれの英訳はオントロジーから抽出した情報(意味分類・同意語・反意語・上位関係とそれらの組み合わせ)で拡張し、比較する。例えば、「購入」と”vétel”の共通点は”purchase”であり、オントロジーから抽出した意味分類によって”purchase”が同じ意味になっていると判断でき、正しい見出し語ペアとなる。また、「購入」と”üzlet”では、同意語・反意語・上位関係のスコアを計算して正しい見出し語ペアと判断された。その他の 8 件の見出し語ペアは本研究の方法では不適切な翻訳ペアとして判断された。「購入」には Tanaka の方法では「vétel」としてしか翻訳されず、Sjöbergh の方法では誤った意味しか生成されなかった。

作成した洪日辞書は、ハンガリー語の見出し語が 44664 語、日本語の見出し語が 48973 語、翻訳ペアが 187761 件である。

4. 辞書評価

自然言語の多くの分野と同じように、辞書生成の場合も「再現率」と「適合率」で評価する。しかし、一般的に利用されている、基準になる評価方法が存在しないため、各作成方法の評価のやり方は様々である。評価結果が解釈しにくく、それぞれの方法の評価結果を比較できない。

現在までの評価方法にもいくつかの問題点がある。適合率を計算するためには、多様な評価方法や多くのサンプルがないと辞書の短所と長所を判断できない。

もう一点は再現率の計算の仕方にある。辞書生成の最も複雑な問題は見出し語の曖昧性解消である。頻度が低い見出し語は、曖昧性がほとんどない。それに対し頻度が高い見出し語の曖昧性はより高く、現在の方法では訳の生成に失敗することが多い。この相違は、すべての見出し語を同じ重さで評価しているためで、再現率の結果には反映されない。本論文で評価する、本研究と異なる方法で作成した辞書のすべてが、「カソード」、「感応作用」、「鼓室」のような頻度が低いと判断できる見出し語を翻訳できたが、いくつかの辞書には「食べる」、「辞書」、「人」のような頻度が高い見出し語の訳がなかった。

本研究で行った評価は以下の通りである。

1. 再現率評価では頻度辞書を利用する加重平均で計算した見出し語のすべてを用いる評価を行う。

2. 「1対1」適合率評価では 1 対 1 のエントリーのサンプルを内的に評価する。

3. 「1対多」適合率評価では同じ見出し語ごとにグループ化したサンプルを内的に評価する。

4. ウェブ上の適合率評価のためにウェブ上の評価システムを作成し数人のユーザーに外的評価をさせる。

本研究の方法を現在まで提案された方法と比較するために、更に二つの辞書を作成した。Tanaka の方法と Sjöbergh の方法に適用させるために資源として本研究の方法と同じ洪英辞書と日英辞書を利用した。Tanaka の方法では見出し語ペア 105632 件の洪日辞書を生成した。Sjöbergh の方法では、スコアの閾値が 0.9 で精度が高いと発表されているが、その閾値の場合は生成した辞書にエントリーが 25218 件しかなかった。この数では明らかに再現率の低下につながる。本研究の辞書と比較しうるサイズにするため、閾値を 0.283 に下げる必要があった。こうして作成した辞書は見出し語が 187610 件になった。

4.1.再現率評価

意味のある、適切な見出し語がどの程度訳されているかを計算するためには、頻度辞書が必要である。本研究で対象となる言語には電子化した頻度辞書がないため、EDR コーパス(Isahara, 2007)を利用した日本語の頻度辞書を作成した。

EDR コーパスは約 20 万の新聞記事から取った、注釈つきのコーパスで、12 万以上の見出し語が 12 品詞によって分類されている。見出し語の平均頻度は 39.6 であった。見出し語の 51.12%の頻度が 1 であったため、これらの単語はエラーと判断して、利用しなかった。

EDR コーパスは日常言語と同じとは断言できないが、このコーパスにある単語の頻度と日常に使われている単語の頻度には相関があると考えられる。

表 1: 再現率評価の結果

辞書作成方法	再現率 (頻度含む)	普通再現率 (頻度無視)
本研究の方法	51.68%	31.61%
Sjöbergh の方法	37.03%	28.50%
Tanaka の方法	30.76%	19.52%
翻訳候補	51.68%	31.61%
日英辞書(一般)	73.23%	54.11%

表 1 のように本研究の方法の再現率は 51.68%で Sjöbergh の方法の 37.03%と Tanaka の方法の 30.76%よりも高い。その上、本研究方法では、本当のペアではないが、候補となったものもカウントされている可能性があるので、それが再現率を高めていると考えられる。

しかし、一般辞書との再現率結果と比較すると、中間言語を利用するために、多くの見出し語への接続がなくなるということも分かった。

頻度を無視した再現率の場合も本研究の結果が最も高いが、曖昧性による困難さが見られる。Sjöbergh の方法の頻度で計算した再現率と普通再現率の差が小さいのは、この方法が頻度が高い見出し語の翻訳に失敗することを示している。

4.2.「1対1」適合率評価

「1対1」適合率評価ではでたらめに見出し語ペア 1000 件のサンプルを選択し、「正しい(○)」・「曖昧(△)」・「誤っている(×)」の基準で手動で評価を行った。この評価方法でも本研究の 79.0%が Tanaka の方法の 62.5%と Sjöbergh の方法の 54.0%を上回った(表 2)。

表 2:「1対1」適合率評価の結果

辞書生成方法	「1対1」適合率		
	○	△	×
本研究の方法	79.0%	6.3%	14.7%
Sjöbergh 方法	54.0%	9.9%	36.1%
Tanaka の方法	62.5%	7.9%	29.6%

4.3.「1対多」適合率評価

「1対多」適合率評価ではでたらめに日本語の見出し語 1000 語のサンプルを選択した。見出し語のすべての訳が適合であれば「正しい(○)」とし、誤りがあるが、多くが正しい場合は「近い(△)」と判断した。誤った訳が 3 つ以上で「誤っている(×)」、訳が無ければ「無い(-)」とした。

この評価方法でも本研究の 72.5%が Sjöbergh の方法の 60.4%と Tanaka の方法の 46.8%を上回った(表 3)。表 3 に示すように Tanaka の方法は実際に翻訳できる見出し語の場合には適合率が高いが、翻訳できない見出し語が非常に多い。

表 3:「1対多」適合率評価の結果

辞書生成方法	「1対多」適合率			
	○	△	×	-
本研究の方法	72.5%	12.9%	14.6%	0%
Sjöbergh の方法	60.4%	13.3%	15.0%	11.3%
Tanaka の方法	46.8%	5.3%	7.3%	40.6%

F 値は頻度で計算した再現率と「1対1」適合率で計算した。本研究では、62.50%であり Sjöbergh の方法の 43.93%と Tanaka の方法の 41.22%を上回っている。

4.4. ウェブ上の評価

本研究の方法で作成した辞書のホームページ上に(http://mj-nlp.homeip.net/mjszotar/)外的評価のためにウェブ上で利用できる評価システムを作成した。この評価システムの評価ページをアクセスすると、サンプル 50 件がはランダムに表示される。両言語の話者がこのように適用した見出し語ペアをすべてマウスで「正しい(○)」、「分からない(△)」、「誤っている(×)」の基準で評価できる。

ウェブ上の評価システムを 2007 年 11 月 27 日から公開し、2008 年 1 月 28 日までに 14 名のユーザーが見出し語ペア 208 件を評価した。結果は適合率が 59.13%しかなかった(表 4)。しかし、この結果を容認できない理由は 2 つある。

a. データが統計的に信頼できない

ウェブ上で評価のために提供された見出し語ペアはランダムに示されるが、評価者が 50 件のすべてではなく、自由選択で評価を行った。14 名の評価者は評価件数が 1 件から 50 件まで様々であったため、この結果は統計的ではない。また、評価された 208 件のサンプルも統計的に不充分だと考えられる。

b. 評価者の信頼性が低い

13 名の評価者のうち数人の評価が疑わしかった。例えば、明確に成功している「変える-kicserél」、「年越し-szilveszter」、「立腹-düh」などのような見出し語ペアが「誤っている」とされていた。

表 4: 本辞書のウェブ上で行った適合率の評価結果

外的適合率(ウェブ上)			合計
○	△	×	
59.13% (123 件)	32.69% (68 件)	8.17% (17 件)	100% (208 件)

5. コミュニティベースの辞書システム

現在はコンピュータ用辞書を協働辞書システム化させる傾向が見られる(Bond & Breen、2007)。このようなシステムは一般辞書と比べ、カバー範囲が広く、また容易に修正、追加ができるという利点がある。更に、このような辞書の多くは一般辞書としての利用の上に、フリーソフトウェアで様々な分野で資源として使用できる。そのために、本研究で生成した辞書は修正の目的でフリーソフトウェアとしてのコミュニティベース化を行っている。(http://mj-nlp.homeip.net/mjszotar/)

本システムには洪→日また日→洪の閲覧欄・検索欄・評価欄・追加欄がある。追加、または閲覧や検索結果の画面より外的評価システムと同じように評価によって辞書を修正できる。

コミュニティベースのアプリケーションでは利用者が自由にデータを変形できるため、アプリケーションにノイズがあることが少なくない。利用者が多い場合にはこのノイズも他のユーザーによって自然に修正される可能性が高いが、利用者が多くない場合には他の処理が必要である。本システムの利用者は比較的少ないため、利用者による修正を管理している。まず、利用者によって追加された新たな見出し語ペアは管理者がチェックしなければ辞書に追加されない。また、現段階では削除は不可能である。その代わり、評価を「信頼度」として示している。評価の結果が管理されず辞書に保存され、評価結果が自動的に表示される。

公開から 2008 年 1 月 28 日までにアクセス数は 6770 件、検索欄を含む評価が 405 件、追加が 6 件である。

メールによる利用者の肯定的フィードバックが多かったが、自動的に作成されたため否定的なフィードバックもあった。今後はウェブ上システムに様々なツール、また機能を追加して利用しやすくする予定である。評価者の関与は最も大切で評価者のランキングなどがシステムに追加できる。掲示板、ユーザー登録、漢字情報の表示、または利用者の要請と希望による改良も可能である。

また、システムの利用者が少ない原因の一つとしては、システムが知られていないためだと考えられる。その解決として本システムを他の辞書サーバーに実装する予定である。

6. 問題点

関連研究と比較すると、本研究の最も大きい長所は再現率に見られる。本研究の方法の再現率は元々の見出し語ペア候補の再現率と変わらないが、これまでの方法

の再現率はそれより低い。しかし、本研究の方法は、一般的辞書の再現率には及ばない。

本研究の方法の適合率も関連研究の方法の適合率を上回っている。

本研究の方法の問題点は以下の2点にまとめられる。

a. 再現率に現れる問題点

中間言語の利用、また資源辞書の精度は低い再現率の原因になる。ソース言語から中間言語、またターゲット言語から中間言語につながる見出し語に接続がないものが多い。この原因として、見出し語の訳が慣用句、または説明になっていたり、見出し語がどこかの言語に存在しないと考えられる。このような見出し語は以下のように分類できる。

- 慣用句
- 活用している単語
- 頻度が低い見出し語
- ある文化のみで利用される見出し語(料理名、生活に関する単語:「寿司」、「畳」)
- ある言語のみで利用される見出し語(反義概念を含んだ単語:「上下」、「男女」;接頭語付で用いられる単語:「kiröpít=飛ばされる」、「elfogyaszt=食べ切る」)
- 様々な品詞(助動詞、助詞、連体詞など)

b. 適合率に現れる問題点

適合率に関する問題は2種類ある。最も大きい原因は見出し語の曖昧性である。もう一つは、単語の意味範囲による差だと考えられる。名詞、形容詞と形容動詞が比較的容易に訳されるが、意味範囲の広い動詞の正確な訳が認識しにくい。

資源辞書の精度も翻訳に失敗する原因になることがある。同じ中間言語として利用されている英語の単語がそれぞれの辞書に異なる意味でしか保存されていない。例えば「loaf=パン;ぶらつく」の連結で「ぶらぶら=vekni(パン)」のような誤っている見出し語ペアが生成された。また、見出し語には無用な意味も保存され、それが有意義な意味を隠してしまう場合も少なくなかった。

外的評価の不適切さの原因として、評価方法の説明がユーザーまで届かなかったこと、または評価者が少なかったことがあげられる。前節で述べた改良によって外的評価システムも改善されると考えられる。

7. 終わりに

本論文では先行研究で生成した洪日辞書に対し複数の内的評価と外的評価を行った。内的評価の結果より本研究の方法は関連研究の方法の再現率と適合率に上回ったということが分かった。また、オントロジーに基づく方法は辞書のみに基づく方法より効果があるが、辞書を手作業で修正する必要があるということも分かった。

外的評価の結果はデータが統計的に信頼できなかったため容認できなかった。

辞書を修正するために現在のウェブ上のシステムを改良し、コミュニティベースのシステムとして実用させる予定である。

謝辞

辞書システムを詳細に検討いただくとともに、システムの改良に対し種々助言いただいた Jim Breen 氏(Monash 大学、オーストラリア)と MTA-Sztaki 研究所(ブダペスト、ハンガリー)に感謝します。

参考文献

Bond, F., Breen, J.W. (2007): "Semi-Automatic Refinement of the JMdict/EDICT Japanese-English Dictionary", NLP2007, Shiga, Japan.

Breen, J.W. (1995): "Building an Electric Japanese-English Dictionary", Japanese Studies Association of Australia Conference, Brisbane, Queensland, Australia.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., Roossin, P. (1998): "A Statistical Approach to Language Translation", In COLING-88, 1, 71-76.

Brown, R.D. (1997): "Automated Dictionary Extraction for Knowledge-Free Example-Based Translation", In Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation, Santa Fe, 111-118.

Isahara, H. (2007). "EDR 電子化辞書の現状", NICT-EDR symposium, 1-14.

Kay, M., Röscheisen, M. (1993): "Text-Translation Alignment", Computational Linguistics, 19(1), 121-142.

Paik, K., Bond, F., Shirai, S. (2001): "Using Multiple Pivots to align Korean and Japanese Lexical Resources", In NLPRS-2001, 63-70, Tokyo, Japan.

Sjöbergh, J. (2005): "Creating a free Japanese-English lexicon", In Proceedings of PACLING, 296-300, Tokyo, Japan.

Shirai, S., Yamamoto, K. (2001): "Linking English words in two bilingual dictionaries to generate another pair dictionary", In ICCPOL-2001, 174-179, Seoul, Korea.

Tanaka, K., Umemura, K. (1994): "Construction of a bilingual dictionary intermediated by a third language", In Proceedings of COLING-94, 297-303, Kyoto, Japan.

Varga, I., Yokoyama, S. (2007): "Japanese-Hungarian Dictionary Generation using Ontology Resources", Proceedings of MT Summit XI, 483-490, Copenhagen, Denmark.