コンパラブルコーパスとWebを用いた用語翻訳器

小松原 慶啓　　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

言語横断検索や文書翻訳において、対訳辞書は不可欠な言語資源であるが、科学技術分野では多数の専門用語が用いられるので、辞書のカバレッジを向上させることが課題である。また、単に辞書引きするだけでは対訳辞書に登録されていない語に対して訳語を返すことはできない。対訳辞書に登録されていない語に対して訳語を合成する機能が必要である。

本研究の目的は、上記の課題を解決し、ユーザが入力するタームに対し適切な訳語を出力する用語翻訳器を開発することである。この目的に対し、コンパラブルコーパスから相関値付き対訳辞書を生成し、それを用いた要素合成法によりランク付き訳語候補リストを作成し、さらにWebを用いて訳語を検証する方法を提案する。そして、コンパラブルコーパスとしてJST科学技術文献抄録コーパスを用いた評価実験を行い、提案方法の有効性を示す。

2. アプローチ

対訳辞書のカバレッジを向上するため2言語のコーパスから語の対訳を抽出する手法を導入するが、その前提として、利用するコーパスのタイプを決定しなければならない。パラレルコーパスが利用できれば、語のアラインメント [1] に基づく高精度の対訳抽出方法を利用することができる。しかし、大規模なパラレルコーパスが利用できる分野は限られる。そこで、本研究ではコンパラブルコーパスの利用を前提とする。利用可能な分野が多いという意味では、同じ分野の単言語コーパスを組み合わせた広義のコンパラブルコーパスが望ましい。文脈類似度を利用する方法 [2),3)] などが適用できるが、十分な精度を達成することが困難である。そこで、文書単位のアラインメントがとれる狭義のコンパラブルコーパスを利用することとし、文書対における共起頻度に基づく相関を計算することにより語の対訳を抽出する。これは、対訳文中の共起頻度に基づく相関を計算することによりパラレルコーパスから対訳を抽出する方法 [4)] における“文”を“文書”に置き換えたものであるが、論文抄録のように短い文書の場合には有効であると思われる。

コーパスから抽出される対訳を含めても、対訳辞書のカバレッジを100%にすることは不可能であるので、辞書に登録されていない複合語に対してon the flyで訳語を生成する機能を付加する。複合語の訳語生成には要素合成法 [5),6)] を用いることができる。参照する対訳辞書のカバレッジを高めることにより、要素合成法によって訳語が生成される可能性も高くなるが、要素合成法で得られる多数の訳語候補の中から正しい訳語を選択することが課題となる。このため、コーパスからの抽出時に計算された構成要素の対訳の相関値に基づいて、複合語に対する訳語候補のスコアを計算し、スコアが最大の訳語を選択する。さらに、入力タームと生成された訳語候補の Web 中の共起頻度を利用することにより、誤った訳語候補を除去する [7)]。

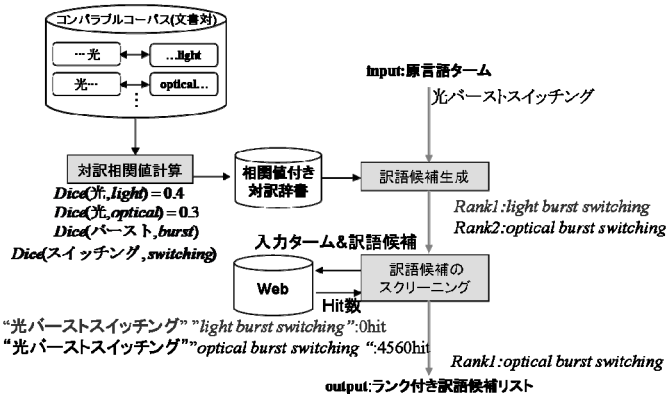

図1　提案方法概要

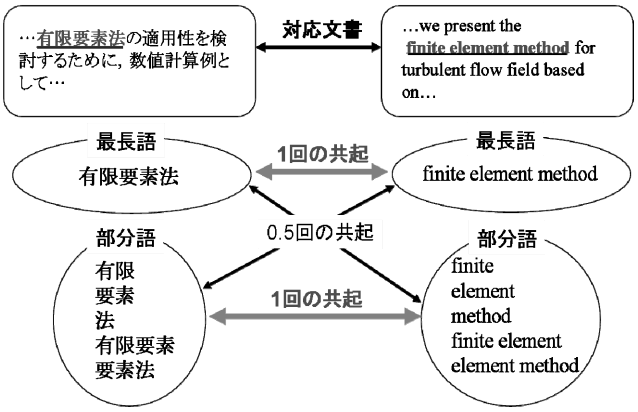

図2　共起回数 g()のカウント

3. 提案方法

提案方法は、図 1 に示すように、(1)対訳相関値の計算、(2)訳語候補の生成、(3)訳語候補のスクリーニングの3ステップから構成される。提案方法は任意の言語対に適用可能であるが、以下においては原言語を日本語、目標言語を英語として記述する。以下、各ステップの詳細を記述する。

3.1 コンパラブルコーパスからの相関値付き対訳辞書の生成

日本語と英語のターム(単純語及び複合語)の間の相関を対応文書中の共起頻度に基づいて計算し、日本語の各タームについて、相関値の高い上位 l 個の英語ターム及び相関値を出力する。

日本語ターム jt と英語ターム et の相関は次式で定義される。

$$Dice(jt,et)=\frac{2\cdot g(jt,et)}{f(jt)+f(et)} \quad [1]$$

ここに、$f(it)$, $f(et)$は jt、et が 1 回以上出現した文書数、$g(it,et)$は jt と et が共に 1 回以上出現した対応文書数である。

相関値計算の対象とする語は、語が出現する文書数(f)が θ 以上で、日本語については名詞列、英語については形容詞、名詞からなる単語列である。対応文書数をカウントする際、jw、ew が独立した名詞句(最長語)であるか、ある名詞句の部分(部分語)であるかを区別する。図 2 に例示するように、最長語どうしあるいは部分語どうしの組は 1 回、最長語と部分語の組は 0.5 回とカウントする。その理由は、抄録対で共起した語であっても、部分語として出現した語が最長語として出現した語と対訳関係であることは少ないからである。

3.2 相関値付き対訳辞書を用いた要素合成法

複合語に対して構成要素の訳語から訳語候補を合成する要素合成法を利用する。ステップ1で生成した相関値付き対訳辞書(表 1 に例示する)を用いることにより、訳語のスコアを計算し、ランク付きの訳語候補リストを作成する。

日本語タームに対する英語訳語候補のスコアの定義は以下のとおりである。日本語ターム jt に対する英語訳語候補 et のスコア $S(jt,et)$は、要素合成法によって計算されるスコア $C(jt,et)$と相関値付き対訳辞書にエントリーされている jt と et の相関値 $Dice(jt,et)$の重み付き和とする。

$$S(jt,et)=\lambda\, C(jt,et)+(1-\lambda)Dice(jt,et) \qquad (0\le\lambda\le 1)\ [2]$$

ここに、λは重みを決定するパラメータである。

スコア $C(jt,et)$は以下のとおりである。

・jt, et がそれぞれ 1 単語 jw, ew からなるとき jw と ew の相関値とする。

$$C(jw,ew)=Dice(jw,cw) \quad [3]$$

・jt が p 個の単語 jw_1, …, jw_p の列、et が q 個の単語 ew_1, …,ew_q の列であるととき、jt, et をそれぞれ2つの構成要素に分割し,前方構成要素間のスコアと、後方構成要素間のスコアの平均値とする。jt, et の分割のしかたは一般に複数考えられるが、スコアが最大となる分割を採用する。

$$C(jt,et)=\max_{1\le i\le p,1\le j\le q} avg(S(jw_1\ldots jw_i,ew_1\ldots ew_j),S(jw_{i+1}\ldots jw_p,ew_{j+1}\ldots ew_q)) \quad [4]$$

なお、2つのスコアの平均は、次式ように調和平均を採用する。

$$\mathrm{avg}(x,\ y)=\frac{2xy}{x+y}\ [5]$$

訳語候補の合成とスコアの計算は、図 3 に示すように動的計画法によって行う。、可能な合成訳の全てについてスコアを計算していては、入力タームの構成単語数や、辞書の訳語数によっては、膨大な処理が必要となるためからである。r 語から構成される日本語タームに対しては $r\times r$ の三角行列が生成する。最初に、対角線上のセルに、相関値付き対訳辞書が示す訳語と相関値を記憶する。次に、対角線に近いセルから順に、セルに対応する日本語単語列に対する訳語候補とそのスコアを計算し、スコア順に第 m 位まで記憶する。最終的に入力タームに対応するセル(最右上のセル)に記憶された訳語をスコア順に第 n 位まで出力する。

表 1　相関値付き対訳辞書の例(l=2)

見出し語	訳語	相関値
有限	finite	0.6
	finite element	0.3
要素	element	0.4
	component	0.1
法	method	0.5
	result	0.3
有限要素	finite element	0.6
	element method	0.4
要素法	element method	0.5
	finite element method	0.4
有限要素法	finite element	0.5
	finite element method	0.3

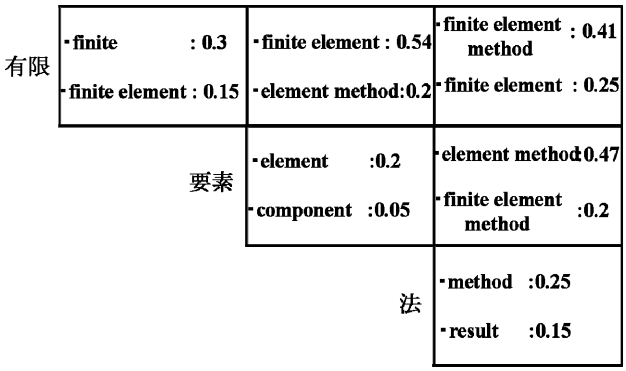

図 3　訳語合成の例(m=2, λ=0.5)

3.3 訳語候補のスクリーニング

要素合成法によって生成される訳語候補には、語として使用されない単語列も含まれる。また、語として使用されるが訳語としては正しくない単語列も含まれる。そこで、入力タームと生成された訳語候補がWeb中に共起するかどうかを判定することにより、訳語候補をスクリーニングする。すなわち、入力タームと訳語候補の組をクエリとした時に検索エンジンから出力されるヒット数が予め定めた閾値 μ 以下の訳語候補をノイズとして除去する。

4.　評価実験

4.1 訳語候補生成実験

4.1.1 実験方法

以下の 3 つの辞書を用いた要素合成法により、テストセットの日本語タームに対し訳語候補を生成した。

(1) 通常の対訳辞書(ベースライン)

人手で作成された対訳を用いる。提案方法と比較するため、ランク付き出力が望ましい。そこで、相関計算に用いた見出し語と訳語の共起文書数(g)を相関値とした。

(2) 相関値付き対訳辞書(提案方法 1)

(3) 相関値付き対訳辞書＋通常の対訳辞書(提案方法 2)

(1)の対訳辞書にエントリーされた訳語について、一定の値を相関値として与え、(2)の対訳辞書とマージした。(1)(2)の両方に含まれる訳語の相関値は、2つの相関値のうち大

きな値を採用した。

4.1.2 使用データ

(1)コンパラブルコーパス

JST 日英論文抄録の情報工学分野 20 年分(107,979 抄録対、60MB)を使用した。日本語抄録は 500～1000 文字程度、英語抄録は 100～300 語程度の短い文書の集合である。論文 ID を介して日英の抄録間の対応は取れているが、文対応は得られない抄録対も多い。

(2)通常の対訳辞書

EDR,EDICT,英辞郎から日本語の名詞とその訳語の組を集めて作成した。

(3)テストセット

デジタル人工知能学事典(人工知能学会,2008)の和英索引より1094ペア、言語処理学事典(言語処理学事典,2010)の和英索引より1661ペアの日英対訳タームを使用した。日本語タームを入力ターム、英語タームをレファレンス訳とした。

4.1.3 パラメータの値

(1)相関値付き対訳辞書の作成

- 語が出現する抄録数の閾値 θ=10
- 各日本語タームに対する英語訳語ターム数の上限 l=20
- 提案方法2における通常の対訳辞書にエントリーされた訳語の相関値:0.1

(2) 要素合成法

- C(*jt*,*et*),Dice(*jt*,*et*)の重みを決定するパラメータλ

デジタル人工知能学辞典の和英索引に含まれる日本語専門用語と英語訳のペアのうち、テストセット以外の1000 ペアを用いて決定した。その結果、ベースラインではλ=0.33、提案方法1ではλ=0.43、提案方法2ではλ=0.40 となった。

- 三角行列の各セルが記憶する訳語の最大数 m=100
- 出力する訳語候補の最大数 n=10

4.1.4 実験結果

要素合成法で出力されたランク付き訳語候補リスト中のレファレンス訳のランクの逆数の平均 MRR (mean reciprocal rank) を算出した。出力された訳語候補リスト中にレファレンス訳が含まれない場合は、そのランクを∞としてMRRを算出した。表1に示すように、3つ方法のうち提案方法 2 が最も良く、提案方法1でもベースラインを大きく上回った。

表 1 には、MRR の値のほか、レファレンス訳が 10 位までに入ったテストターム数を、(a)レファレンス訳が対訳辞書に登録されていたものと(b) レファレンス訳が要素合成法で合成されたものの内訳とともに示した。この結果から以下のことが読み取れる。

(a)コーパスから生成した相関値付き対訳辞書は通常の対訳辞書と比べて専門用語のカバレッジが高い。

(b)相関値付き対訳辞書を用いた要素合成法は通常の対訳辞書を用いた要素合成法より多くの訳語を合成することができる。

また、デジタル人工知能学事典、言語処理学事典の 2 種類のテストセットについて、MRR での結果に明らかな差異が見られた。テストセットの比較から、言語処理学事典の入力タームは、人工知能学事典と比べ、コーパス出現数が低い「国名」や、「言語名」が多いことが解った。今回の実験では、相関値付き対訳辞書生成に、小規模な論文抄録コーパスを用いているため、このような分野におけるタームの特徴が精度に大きく影響したものと考えられる。

表 2　訳語候補生成実験結果

人工知能学事典	ベースライン	提案方法1	提案方法2
MRR	0.22	0.40	0.44
TOP10に正解を含むテストターム数(1094語中)	306語	517語	558語
辞書に登録されている訳語候補	148語	350語	384語
要素合成法により生成された訳語候補	158語	167語	174語
言語処理学事典	**ベースライン**	**提案方法1**	**提案方法2**
MRR	0.20	0.31	0.35
TOP10に正解を含むテストターム数(1661語中)	450語	602語	690語
辞書に登録されている訳語候補	296語	408語	491語
要素合成法により生成された訳語候補	154語	194語	199語

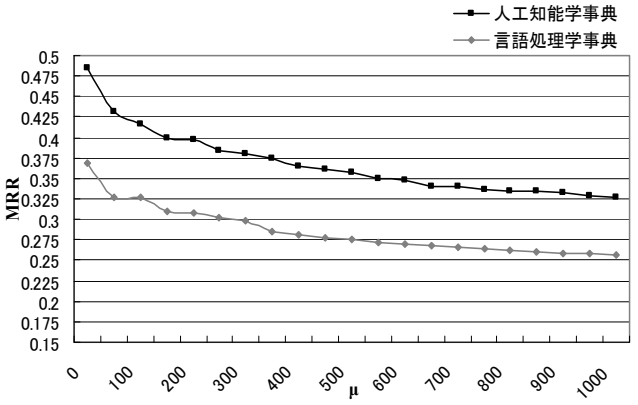

図 4　スクリーニング実験結果

4.2 訳語候補のスクリーニング実験

4.2.1 実験方法

訳語候補生成実験で最も MRR の良かった提案方法2にステップ3を追加する実験を行った。ステップ2の要素合成法で出力する訳語候補の最大数 n を 50 とした。訳語候補の各々を入力タームとともにクエリとして検索エンジン「Yahoo!Japan」に入力した。

4.2.2 実験結果

ヒット数の閾値(μ)を 50 刻みで変動させ、MRR を計算した結果を図 4 に示す。MRR は μ=0(頻度 0 の訳語候補を除去)で最大となり、以降は減少する傾向にあった。スクリーニング前と比較し、人工知能学事典、言語処理学事典の両テストセットとも 0.03 程度の MRR の向上が見られた。除去できるノイズは、Web 上で殆ど出現しないタームが殆どであった為、μ の変化が大きく結果に影響することはなかった。

5. 今後の課題

5.1 訳語候補生成率の向上

人工知能学事典のテストターム中、提案方法 2 で上位 10 位以内に正解訳語を生成できなかったタームについて、その原因を分類し、その対策とともに表 3 にまとめた。

正解訳語の生成に失敗する最大の原因は、入力ターム及びその構成要素のコーパス中の出現頻度が小さいため、相関値付き対訳辞書に登録されず、また通常の対訳辞書にも登録されていないことである。これに対しては、一般にはより大きなコーパスを用いて相関値付き対訳辞書を生成する以外に解決策はない。ただし、対訳辞書に登録されていなくても解決できる語もある。アルファベット列を含む語は、アルファベット列に対する訳語は同じアルファベット列であるとして訳語を合成することが可能である。

次に考えられる原因は、入力タームあるいは構成要素の正しい訳語の相関値が低いことである。特に派生語の場合、接辞部分を含まない語に対する訳語との相関値も高くなる傾向があり、正しい訳語との相関値が最大になるとは限らない。これは共起文書数に基づく相関の限界といえる。派生語を含む語に対する訳語候補生成率を向上するには、接尾辞の訳語として「NULL」を追加することが考えられる。

正解訳語の生成に失敗するその他の原因は、形態素解析の結果と対訳辞書との不整合の問題で、特にカタカナの入力タームに対してみられた。複数のカタカナ語の列が1つの形態素とみなされて要素合成法が適用できなくなる場合と、逆に1つのカタカナ語が複数の形態素に分割されるため当該カタカナ語の訳語が利用できない場合がある。これを解決するにはカタカナ語に対する形態素解析の精度を向上させなければならない。

5.2 訳語候補のスクリーニング方法の改良

今回行った訳語候補のスクリーニング実験において、Web と検索エンジンを用いたノイズ除去は、若干の MRR の向上を示したものの、十分な成果が出せなかった。これは、タームの構成単語数でヒット数が大きく変化すること(構成数が少ない語とのヒット数が多くなる)に起因し、故にヒット数 0 の語を除去すること以外に有効なスクリーニングができなかった。このことから、単純な Web のヒット件数では、スクリーニングとして不十分であると解った。

Web と検索エンジンを用いた訳語候補のスクリーニング方法を拡張していく方法としては、hit 数だけでなく、クエリとする入力タームと訳語候補各々の構成単語数を考慮してノイズ判定の尺度を設定すること、あるいは、web のヒット数以外の情報として、単言語コーパスから入力タームと訳語候補の文脈ベクトルを其々抽出し、その類似度を用いることで解決することができる可能性がある。

6. おわりに

コンパラブルコーパスの対応文書中の共起頻度に基づく相関から相関値付き対訳辞書を生成し、要素合成法に組み込んだターム翻訳方法を提案した。コンパラブルコーパスとして日英論文抄録を用いた訳語生成実験では、提案方法の MRR は 0.44 であり、通常の対訳辞書のみを用いた要素合成法の 0.22 を大きく上回った。

表 3 訳語候補生成に失敗する原因の分析

原因	比率	例	改善策
入力タームまたは構成要素の出現頻度が低い	62%	足場掛け:*scaffolding* 枝分かれ構造:*branching structure*	トレーニングコーパス拡大
アルファベットを含む語	3%	アナログLSI:*analog LSI* 単純LR法:*simple LR parsing*	アルファベット入力はそのまま出力する
接尾辞を含む語	16%	遺伝的アルゴリズム :*genetic programming*	接尾辞の訳語として「NULL(空文字)」を追加
カタカナからなるターム	19%	イメージベースモデリング :*Image based modeling*	形態素解析の向上

今後は、訳語候補生成率を向上させるため、アルファベット列や接尾辞の処理を改良するとともに、より大規模なコーパスを用いた評価実験を行うことが必要である。また、訳語候補のスクリーニングにおいて構成単語数に利用することなどを検討する予定である。

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。また、科学技術文献抄録コーパスの研究利用を許可いただいた(独)科学技術振興機構に感謝申し上げます。

参考文献

1) Peter F.,Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer: The mathematics of statistical machine translation: parameter estimation, *Computational Linguistics*, Vol. 19, No. 2, pp.263-311 (1993).

2) P. Fung and L.Y. Yee: An IR approach for translation new words from nonparallel, comparable texts, In *Proc. of COLING-ACL 1998*, pp 414-420 (1998).

3) R. Rapp: Automatic identification of word translations from unrelated English and German corpora, In *Proc. of ACL 1999*, pp. 519-526 (1999).

4) 北村 美穂子, 松本 祐治:対訳コーパスを利用した翻訳規則の自動獲得, 情報処理学会論文誌, vol.37, no.6, pp.1030-1040 (1996).

5) R. Y. Cao and H. Li: Base Noun Translation Using Web Data and the EM Algorithm, In *Proc. of COLING 2002*, pp. 127-133 (2002).

6) 外池 昌嗣, 宇津呂 武仁, 佐藤 理史: 要素合成法を用いた専門用語の訳語候補生成・検証, 言語処理学会第 11 回年次大会発表論文集, pp.13-16 (2005).

7) M. Nagata, T. Saito, and K. Suzuki : *Using the Web as a bilingual dictionary*. In *Proc. of ACL 2001 Workshop on Data-Driven Methods In Machine Translation*(2001).