簡単なフィルターを用いた二言語シソーラスの自動構築*

影浦峡†, 辻慶太‡, 相澤彰子†
†学術情報センター研究開発部（{kyo,akiko}@rd.nacsis.ac.jp）
‡東京大学大学院教育学研究科（i34188@m-unix.cc.u-tokyo.ac.jp）

Abstract

いくつかの言語処理モジュール（フィルター）を重ね合わせることにより、二言語対訳・対応コーパスから二言語シソーラスを自動構築する手法について報告する。計算機科学分野の対訳データを用いた実験に基づき、手法の有効性を示すと共に今後の改良方針も整理する。

1 はじめに

本発表では、いくつかの言語処理モジュール（フィルター）を重ね合わせることにより、二言語対訳・対応コーパスから二言語シソーラスを自動構築する手法について報告する。組み合わせる手法は、(1) 形態素解析＋名詞句抽出、(2) 統計的単語アラインメント、(3) グラフ理論に基づく最少グラフカットである。いずれもこれまでに良く研究されてきており、これ以上要素技術としての精度を劇的に挙げるのは難しい。逆に、これらを上手く組み合わせて、実際に利用可能なシソーラスが生成するといった課題は応用面から重要である。本報告では、まず手法について説明し、次いで、実際のデータに適用した結果の評価を示す。

2 方法

ここでは、(1) 形態素解析＋名詞句抽出、(2) 統計的単語アラインメント、(3) グラフ理論に基づく最少グラフカットの方法とその流れを簡単に紹介する。

2.1 形態素・語彙解析

この段階では基本的な語彙単位（キーワード候補単位）を抽出する。ここでは、日英の統合的単位の対応をとりやすくするため、最少単位と最大複合単位のみを抽出し、中間構成単位は取らない[1]。

(a) 形態素解析

基本単位を抽出するために、既存の形態素解析ソフトを用いる。今の所、日本語については JUMAN3.5 (黒橋 & 長尾 1998) を、英語については LT_POS/LT_CHUNK (Mikheev 1996) を用いている。

(b1) 最小単位抽出

英語最少単位は、LT_POS により出力された内容語をそのまま取る。日本語については、英語に単位を会わせるために、以下のようなパターンを取る。

```
C_Prefix* (C_Adv|C_Adj|N) C_Suffix*
```

ここでC_は漢字あるいはカタカナのみからなることを示す。

(b2) 最大単位抽出

英語の最大単位は、LT_CHUNK の出力結果にいくつかのアドホックな後処理を加えたものである。特に、タイトル等を処理する場合、LT_CHUNK では名詞を動詞と取り違えがちなので、それを補正する。一方、日本語の最大単位は、以下のパターンで定義する。

```
^C_Adj* (C_Affix|C_Adv|C_Adj|N)+
```

ここで^C は、漢字あるいはカタカナで始まるものを示す。ここでのパターンは JUMAN の切り間違い等を考慮してかなり緩くしてある。これと字種との混合の方が、厳密な言語的名詞句の定義よりも実際上は良い。ここでも、単位を会わせるために、アドホックな処理を加えている。例え

*Automatic Construction of Bilingual Thesauri by Kyo Kageura, Keita Tsuji & Akiko N. Aizawa.

[1]現在のところ、この部分の設計はまだ途上のものである。ここで中間的な単位について取っていないのは、適切な構成素を抽出するにあたっての問題（「オンライン情報検索」では「オンライン検索」と「情報検索」を、「法律情報検索」では「法律情報」と「情報検索」を取らなくてはならないが、この構造付与はさほど簡単ではない。これに対して連接と依存の総組み合わせを取るのには抵抗がある）をまだ解決していないこと、さらに、シンタグマティックな単位の二言語間での対応が難しくなることという問題が残されているからである。

ば、「用」は英語の 'for' に対応することが多く、これは LT_CHUNK では名詞句に含まれないため、日本語でも対応して切ることにしている。

これらの処理の結果、最小単位データと最大単位データができる。これらは、次の統計的アラインメントモジュールでは別々に処理される。

2.2 統計的対訳候補抽出

統計的対訳候補抽出モジュールでは、統計的な対訳対の重み付けと、それに対する（マクロ及びミクロな）事後処理を行う。

(a) 統計的重み付け

統計的対訳対抽出の研究は色々あるが (Dagan & Church 1994; Eijk 1993; Fung 1995; Gale & Church 1991; Kupiec 1993; Melamed 1996; 北村&松本 1997)、繰り返し計算を行わない単純な尺度の中で結果の良かった対数尤度比の検定量を重みとして用いる (辻他 1999)。対数尤度比の定義については Dunning (1993) 等に詳しい。

(b) 事後処理フィルター

事後処理フィルターはマクロ・フィルターとミクロ・フィルターとからなる。マクロ・フィルターは、(i) 正しい対訳は、統計値 X_S 以上の値を取る、(ii) 一つの要素に対して正しい対訳は最大 X_C 個しか存在しないという二つの制約からなる。

ミクロ・フィルターは、一つのコーパス単位（事前にアラインされた、統計量を取る基本となる単位：以下の実験ではタイトルペア）においては、一つの用語に対して対応する用語は一つであるという制約を利用する[2]。コーパス単位毎に、その内部で最も統計スコアの高い対から選び取り、その日本語・英語を含む対を削除することを繰り返すことで、候補が絞られる。

これらの作業は、最少単位データと最大単位データに対して別々に適用し、その結果できたデータを統合する[3]。

2.3 クラスタ生成

こうしてできた対訳候補対データに対して、グラフ理論のアルゴリズムを適用する。説明のために、表 1 のデータが、これまでの処理で生成されたものとしよう。詳細な説明は Aizawa & Kageura (to appear) を参照のこと。

[2]論文全文などではこれは当てはまらないが、対訳が入手しやすい抄録やタイトルでは概ね当てはまると考えられる。

[3]最小単位データと最大単位データとを別々に扱う点も、現在の方針は多分に便宜的なものである。

Japanese keywords	English keywords	frequency
キーワード	information retrieval	1
キーワード	keyword	39
テキスト検索	information retrieval	1
テキスト検索	text retrieval	6
テキスト検索	text search	3
検索指示語	keyword	1
広域情報検索	information retrieval	1
情報検索	information gathering	4
情報検索	information retreival	1
情報検索	information retrieval	320
情報検索	information search	5
情報収集	information gathering	6
情報収集	information retrieval	1
文献検索	bibliographic search	1
文献検索	document retrieval	11
文書検索	document retrieval	19
文書検索	text retrieval	1

表 1. クラスタ生成への入力データ例

まず、ノードが日本語か英語のキーワードを、リンクが対訳関係を示すようなかたちの、初期キーワードグラフを生成する（図 1）。リンクの強さには、現在のところ、対訳のコーパス頻度を用いている。クラスタ生成は、このグラフから誤りリンクを削除するかたちで行う[4]。この問題は、グラフ理論における最初カット問題として定義できる。

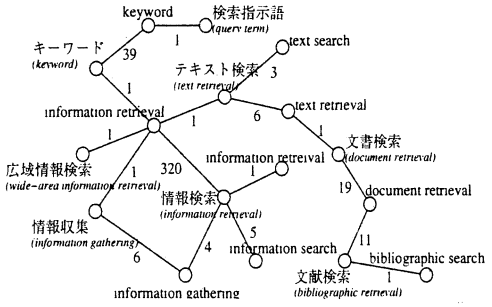

図 1. 初期キーワードグラフ

一定の値を目安に、リンクを削除したときに独立したクラスタを構成するようなリンクを検出し削除する（このとき削除するリンクは一つであるとは限らない）。現在のアルゴリズムでは、まず、リンクの強度が N_α 以上であるか、削除すると一つのキーワードが孤立するか、言語間で綴りが同一かであるものを、削除できないリンクとしてマークする。次に、全てのクラスタにはコアキーワードが含まれなくてはならないという制約を設定するために、頻度 N_β 以上のキーワード

[4]従って、これは、統計的クラスタリング (Deerwester et. al 1990; Grefenstette 1994; Schütze & Pedersen 1997) とは大きく異なる。

をコアキーワードとしてマークする。最後に、合計強度が N_ϵ 以下のリンクを検出し削除する。図2は、$N_\alpha = N_\beta = N\epsilon = 3$ でこのアルゴリズムを適用した例である。実線は削除不可リンクを、黒丸はコアキーワードを、A、B、C、Dは生成されたクラスタを示す。

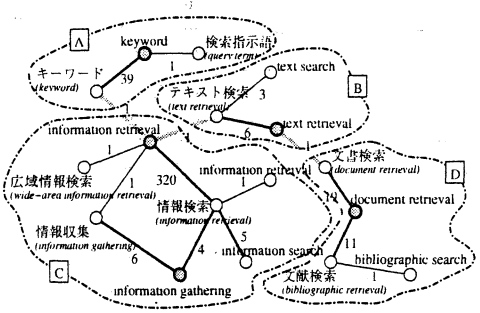

図 2. グラフ理論によるクラスタ生成

3 実験と評価

計算機科学の学会発表論文 25534 表題対からなるコーパスを用いて手法の評価実験を行った。形態素・語彙解析を行った結果、最小単位データでは、日本語が延べ 178091 語・異なり 14938 語、英語が延べ 154554 語・異なり 12634 語、最大単位データでは、日本語延べ 89742 語・異なり 38813 語、英語延べ 80018 語・異なり 41693 語となった。

対訳候補抽出は、X_S も X_C も共に 10 に設定して実験を進めた[5]。結果として抽出された対訳候補対は 28905 対、日本語が 24855 異なり語、英語が 23430 異なり語であった。このうち、20071 対は頻度 1、3196 対は頻度 2 であった。また、8242 が最少単位対、20663 が最大単位対であった。1 日本語あたり平均 1.16 英語、1 英語あたり平均 1.23 日本語が対応していた。

この対訳候補対から作られた初期キーワードグラフは 19527 の独立したサブグラフから構成された。最大のサブグラフには 2701 対（全体の 9.3%）が含まれていた。クラスター生成処理は、$N_\alpha = 4$、$N\epsilon = 10$、$N_\beta = 1$ という設定で行った。これは、いくつかのパラメータの組み合わせの中からクラスタの大きさ等をもとに経験的に決めた[6]。893 対訳対の対訳関係が削除され、20357 のクラスタが生成された。最大クラスタの構成対は 64 となった。表2に、クラスタの大きさ（構成対訳対の数）とクラスタ数を示す。

size	no. of clusters	size	no. of clusters
1	16693	5-9	322
2	2354	10-19	52
3	504	20-64	22
4	410		

表 2. クラスタの大きさと数

結果の評価は、生成されたクラスタ内部の一貫性とリンク削除の精度の二点から行った。クラスタの内部一貫性を評価するために、一ペアのみからなる「小」クラスタ 1000 サンプルを「正解」(c)、「ほぼ正解」(m)、「誤り」(w) に、また 2～9 ペアからなる「中」クラスタ 200 サンプルとそれ以上の「大」クラスタ 50 サンプルについては「正解」(c)、「ほぼ正解」(m)、「混合」(h)、「誤り」(w) に分けて評価した。結果を表3に示す。「ほぼ正解」や「混合」を含めて概ね 9 割近いクラスタが意味的なまとまりを持ったものであることがわかる[7]。「混合」や「誤り」の原因として、統合的単位のズレと統計的重み付けの失敗とが挙げられる。

	c	m	w	
小	711	173	116	
	c	m	h	w
中	56	85	17	42
大	7	17	23	3

表 3. クラスタ内部の一貫性評価

削除の精度については、削除されたリンクのうち 400 サンプルについて、「正しく不適切なリンクを削除した」(c)、「誤って正しいリンクを削除した」(w)、「関連語リンクを削除した」(p) で評価した。表4は、その結果である[8]。

c	p	w	計
264 (66.0%)	77 (19.3%)	59 (14.8%)	400

表 4. リンク削除の妥当性

正しいのに削除されてしまったリンクは全体の 15%程度である。実は、より詳しく見ると、削除されたもののうち、パラメータ N_α が指定するリンクの強度が強いものに、誤って削除されたリンクが偏って多く含まれていることがわかった。

[5] X_S が 10 以下のもので最小単位データ 100 標本のうち正解対訳は 3、最大単位データ 100 標本では 5 であった。

[6] パラメータの影響については Aizawa & Kageura (to appear) を参照。

[7] 「中」クラスタの精度が低いのは、2 ペアからなるクラスタにおいて一要素不適切なものがあったものは全て「誤り」としたからである。3 ペア以上だと、一要素だけ不適切でも必ずしもクラスタとして「誤り」とはならない。

[8] 実際には、リンク削除の妥当性は、それによって分割されたクラスタの内部一貫性との関係で評価されなくてはならないが、それは今後の課題としたい。

4 改良点と今後の課題

実験評価から、何点か改良すべき点が明らかになった。それらは以下のように整理できる。

(1) 誤りの多くは対訳抽出のレベルで、統合的単位のミスマッチにより引き起こされている。例えば 'patten matching'→ 'パターン' ← 'pattern' などである。この多くは、日本語のカタカナ語が関与している。従って、形態素・複合語処理モジュールで、例えばカタカナと英語の文字列類似度情報などを利用することで統合的単位をより正確に揃えればパフォーマンスは向上すると考えられる。

(2) 統計処理段階で、統計的重み付けから来る誤りもかなり見られた。これは一部には、統合的単位の取り方の誤りあるいは揺れから来るものもあるが、他に、統計処理の前に語尾処理等を行えば避けられるかもしれないものもある。

(3) クラスタ生成段階では、前述したように、グラフ上のコンテクストに関係なく一定のパラメータによりリンク削除を行うと、リンクの強度が強いものの削除で誤りが多く生じる。一方、頻度の高いものからなるクラスタには多数の高頻度ペアが混合して集まる傾向が見られた。ここから、グラフの大きさや中心となるキーワードの頻度に応じて動的にパラメータを変えてやることで、誤削除を押さえつつ削除効率を上げることが可能であると考えられる。他に辞書情報を用いることも頻度の高いものからなるクラスタの処理には有効であろう。

現在我々は特に、形態素・複合語処理モジュール及びクラスタ生成モジュールの改良を行っている。上記の3点についての改良が一通り終わったところで、実際に作られたシソーラスを、言語横断情報検索のために検索式拡張やユーザーナビゲーション、そして専門翻訳用オンライン辞書といった応用の観点から評価して行く予定である。

付言： 本研究は、日本学術情報振興会未来開拓学術推進事業「高度分散情報資源活用のためのユビキタス情報システムに関する研究」の一環として行われた。

参考文献

[1] Aizawa, A. N. and Kageura, K. (to appear) "A graph-based approach to the automatic generation of multilingual keyword clusters." In: Bouligault, D., Jacquemin, C. and l'Homme, M-C. (eds.) *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins.

[2] Dagan, I. and Church, K. (1994) "Termight: Identifying and translating technical terminology." *Proc. of the Fourth ANLP*. p.34–40.

[3] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990) "Indexing by latent semantic analysis." *JASIS*. 41(6), p. 391–407.

[4] Dunning, T. (1993) "Accurate methods for the statistics of surprise and coincidence." *Computational Linguistics*. 19(1), p. 61–74.

[5] Eijk, van der P. (1993) "Automating the acquisition of bilingual terminology." *Proc. of the 6th EACL*. p. 113–119.

[6] Fung, P. (1995) "A pattern matching method for finding noun and proper noun translations from noisy parallel corpora." *Proc. of 33rd ACL*. p. 233–236.

[7] Gale, W. A. and Church, K. W. (1991) "Identifying word correspondences in parallel texts." *Proc. of DARPA Speech and Natural Language Workshop*. p. 152–157.

[8] Grefenstette, G. (1994) *Explorations in Automatic Thesaurus Discovery*. Boston: Kluwer Academic.

[9] 北村美穂子, 松本祐治 (1997) 「対訳コーパスを利用した対訳表現の自動抽出」『情報処理学会論文誌』 38(4), p. 727–736.

[10] Kupiec, J. (1993) "An algorithm for finding noun phrase correspondences in bilingual corpora." *Proc. of 31st ACL*. p.17–22.

[11] 黒橋禎夫, 長尾真 (1998) 『日本語形態素解析システム Juman 3.5 ユーザーズマニュアル』 京都: 京都大学.

[12] Melamed, I. D. (1996) "Automatic construction of clean broad-coverage translation lexicons." *2nd Conference of the Association for Machine Translation in the Americas*. p. 125–134.

[13] Mikheev, A. (1996) "Learning part-of-speech guessing rules from lexicon." *COLING'96*, p. 770–775.

[14] Nagamochi, H. (1993) "Minimum cut in a graph." In: Fujisige, S. (ed.) *Discrete Structure and Algorithms II* (Chapter 4). Tokyo: Kindaikagakusha.

[15] Schütze, H. and Pedersen, J.O. (1997) "A cooccurrence-based thesaurus and two applications to information retrieval." *Information Processing and Management*. 33(3), p. 307–318.

[16] 辻慶太, 影浦峡, 芳鐘冬樹 (1999) 「対訳コーパスからの訳語対抽出」『1999 年度日本図書館情報学会春期研究集会』 p. 25–28.