対訳文章を用いた専門用語対訳辞書の自動作成
－訳語対応における両立不可能性を考慮した手法について－

石本浩之　長尾眞

京都大学工学部 電気工学第二学科

1　はじめに

対訳文章を用いて専門用語の対訳辞書を自動作成する際の問題として、専門用語の認識、訳語対応の多様性とその解消、幾つもの異なった訳の存在の取り扱い、という問題がある。訳語対応に関するこれまでの研究においては、主として"Ranking Method"という方法が用いられてきた[5, 4, 3]。この方法では、まずどちらか一方の言語において専門用語を正しく認識しておき、それに対応する訳語の候補を相手言語の文章から抽出する。次に対訳辞書や統計情報を用いることによって訳語候補に順序づけを行ない、最も適当と考えられる候補を訳語として決定する。しかし専門用語は新しく作られた語や一般語を組み合わせることによって表わされることが多いなどの様々な特徴を持つ。さらに英語においては前置詞句などの曖昧な語句を含む専門用語などもよく見られる。従って専門分野の文章のように比較的小さい量の文章を対象とする場合、単言語における専門用語の認識には限界がある。一方、専門用語を認識するための情報は相手言語の文章中に含まれることが多く、これを利用することによって専門用語を的確に捉えられる場合がある。そこで我々は、両立不可能性という概念を導入することによって対訳文内部の細かい対応関係を考慮し、専門用語の認識とその訳語の決定を同時に行なう。これらの決定を行なうための情報源としては辞書情報、統計情報、カタカナ表記の訳語推定を用いる。本手法では各対訳文内部における訳語の対応関係を考慮することから、比較的少量のテキストからでも専門用語の訳語対応を高精度に一意に決定できるほか、従来の方法では難しいと考えられる幾つもの異表記の訳語抽出も可能となる。本手法を用いて通信分野の文章について実験を行なったところ、この両立不可能性が有効に働くことを確認した。

2　両立不可能性を考慮した訳語対応の枠組

2.1　両立不可能性

専門用語認識の曖昧性と訳語対応の多様性を同時に表現するために、両立不可能性という関係を導入する。両立不可能性とは、対訳文の内部において考えられる訳語対応の内で、互いに成り立たない訳語対の間に成立する関係である。この関係は基本的には、対訳文中においてある専門用語に対応する訳語はただ一つだけであり、その訳語以外の語には対応しないということを前提にして得られる。さらに、我々は両立不可能性を訳語対応ネットワークとして表現することによって、専門用語認識と訳語対応の曖昧性の解消を効率良く行なう。訳語対応ネットワークは、専門用語の訳語対の候補を節点とし、互いに両立しない訳語対間に枝をはったグラフである。

例えば、例1のように日本語と英語で専門用語がそれぞれ二つずつありこれらの対応関係を明らかにすることを考えてみる。

例 1
…digital network, digital circuit…
…デジタル網, デジタル回路…

ここで[digital =デジタル]、[circuit =回路]という訳語情報だけが得られたとする。このとき「デジタル回路」に対応する訳語は、その構成語における訳語情報から'digital circuit'であることがわかる。一方「デジタル網」の場合には、手がかりとなる訳語情報に[digital =デジタル]しかないために、'digital circuit'と'digital network'のどちらに対応するかわからない。そこで、図1のように可能な全ての訳語対応関係を作り、それらの関係を線で表現する。この図から⟨digital circuit, デジタル回路⟩が成立するとき、⟨digital circuit, デジタル網⟩および、⟨digital network, デジタル回路⟩は成立しないということがわかる。「デジタル回路」と'digital circuit'の対応を正しいと決定すると, 自動的に「デジタル網」は'digital network'に対応することになる（図2）。このような関係を両立不可能性の関係と呼び、図1のように訳語対応の関係を表現したものを訳語対応ネットワークと呼ぶ。

また次の例では、英語の方に前置詞句が存在しているために'closed user group with outgoing access facility'を専門用語として的確に認識することが難しい場合を示している。

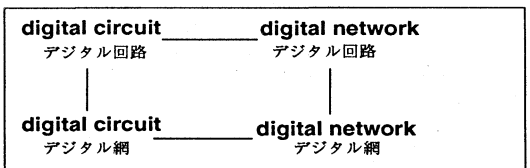

図 1: 訳語対応ネットワーク1

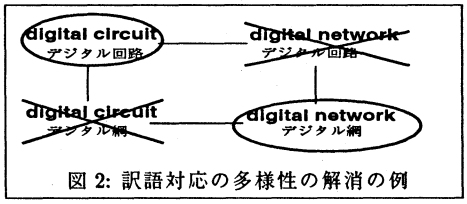

図 2: 訳語対応の多様性の解消の例

例 2
···closed user group with outgoing access facility ···
···出接可閉域ユーザグループファシリティ···

そこで、日本語からは「出接可閉域ユーザグループファシリティ」が専門用語の候補として抽出され、英語からは'closed user group with outgoing access facility'、'closed user group'、'outgoing access facility'の三つが候補として抽出されたとする。これらの訳語対応の関係を訳語対応ネットワークで表わすと図 3となる。ここで「出接可閉域ユーザグループファシリティ」が'closed user group with outgoing access facility'に対応することを決定すると同時に'closed user group'と'outgoing access facility'は間違って認識された候補であったことが分かることになる（図 4）。ただし、「オプショナルユーザファシリティ」と「ユーザファシリティ」のようにどちらも専門用語でありながら同時に部分-全体という関係にある専門用語や、「オプショナルユーザ」と「ユーザファシリティ」のように互いに部分文字列を共有する専門用語もある。この種の専門用語については、先に示した専門用語認識の曖昧性との区別を特にはしていない。よって、両方の言語において、二つの専門用語候補が互いの部分単語列を共有している場合、もしくは二つの専門用語候補が部分-全体の関係にある場合にはそれらの間には両立不可能性は成立しないことにする。

さらに、対訳辞書の構築の際には、ある用語が複数の訳語を持つ可能性も考慮に入れる必要がある。次の例 3では「データ転送」が'data transmission'と訳されているが、例 4では'data transfer'と訳されている。

例 3
···related to each direction of data transmission ···
···データ転送の各方向における···

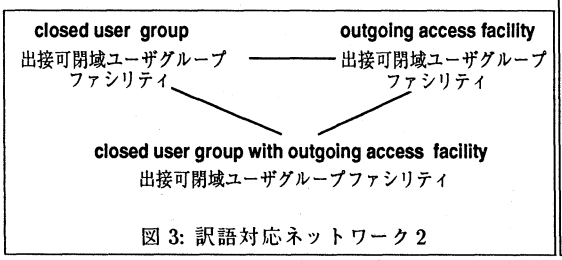

図 3: 訳語対応ネットワーク 2

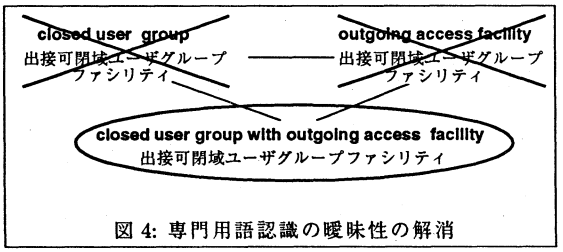

図 4: 専門用語認識の曖昧性の解消

例 4
···in the data transfer state of a logical channel ···
···論理チャネルがデータ転送状態にあるときに、···

これらの関係を訳語対応ネットワークで表現すると図 5となる。データ転送 data transfer の訳語対と両立不可能な関係にある訳語対は、それが含まれている対訳文中の訳語対だけであることがわかる。すなわち、これとデータ転送 data transmission の間には両立不可能性は成立していないため、両者が抽出可能であることが分かる。このように複数の異表記の訳語を持つ場合にでも有効である。

2.2 専門用語の定義

訳語対応における両立不可能性を考慮することによって、曖昧な語句を含む専門用語も扱うことができる。そこで本研究では扱う専門用語を名詞句に限定し、取り扱う文章全体を通じて 2 回以上現れる一番長い名詞句を専門用語候補とする。ただし、日本語では形態素解析、英語では語尾解析を行なう。まず専門用語の候補となる例を例 5に示しておく。

例 5 次の文字列が文章中に現れたとする。
···網で提供されないオプショナルファシリティは、···
···認識しているファシリティの現在値を···
··· *DTE* はオプショナルファシリティを指定せず、···
このとき、以下の専門用語が抽出される。

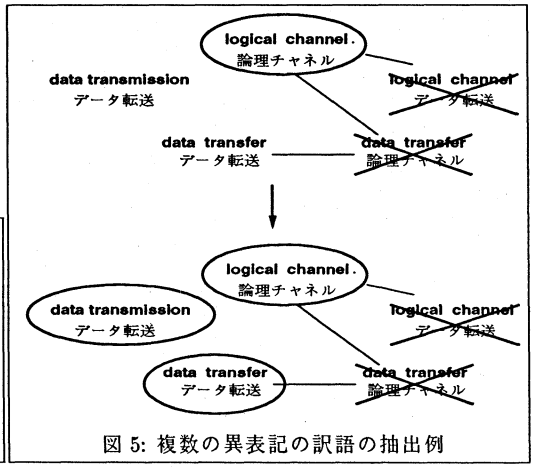

図 5: 複数の異表記の訳語の抽出例

オプショナルファシリティ
ファシリティ

本研究で扱う専門用語の候補の定義は以下のようになる。

$$\left.\begin{array}{ll}\alpha_1 = a_1 a_2 \cdots a_p & a_i \in S, p \geq 1 \\ \alpha_2 = c_1 c_2 \cdots c_q & c_i \in S, q \geq 1 \\ \beta_1 = d_1 d_2 \cdots d_m & d_i \in S, m \geq 1 \\ \beta_2 = e_1 e_2 \cdots e_n & e_i \in S, n \geq 1 \\ \gamma = b_1 b_2 \cdots b_t & b_i \in S, t \geq 1 \end{array}\right\} \quad (1)$$

という単語列があったとする。ここで次のような連接

$$\left.\begin{array}{l}\alpha = \alpha_1 \cdot \gamma \cdot \alpha_2 \\ \beta = \beta_1 \cdot \gamma \cdot \beta_2 \end{array}\right\} \quad (2)$$

が文章中に現れ、さらに

$$\left.\begin{array}{lll} a_p \neq d_m & \text{または} & np(a_{p'} \cdots a_p \cdot \gamma) = false \\ c_1 \neq e_1 & \text{または} & np(\gamma \cdot c_1 \cdots c_{q'}) = false \end{array}\right\} \quad (3)$$

$$np(\gamma) = true \quad (4)$$

の三つの全ての条件を満たすとき、γ は専門用語である。

$np(X) = true$ は X が名詞句であることを表わす。与えられた単語列が名詞句かどうかを判定する際には、各単語の統語範疇が何であるかという知識のみを用いることとし、ある特定の品詞列のみを受理する文法規則を用いる。英語の場合には'integrated service digital network' のように名詞、未定義語、形容詞、分詞の連続を扱うための文法規則や、'closed user group with incoming access' のように前置詞句を含む語を扱うための規則を用いた。一方、日本語の場合には普通名詞、サ変名詞、未定義語の連続を扱うための規則と「データの転送」のように名詞接続助詞を含む語を扱うための規則を用いた。

3 訳語対応ネットワーク上での訳語対決定

3.1 訳語対の候補に与える評価値

訳語対応ネットワーク上で最も評価値の高い専門用語の訳語対から順番に決定を行なう。訳語対の各候補に対して語彙情報に基づく評価値 H_D と統計情報に基づく評価値 H_{Sta}、H_{Occ} の三つの評価値を計算しておき、訳語決定の際にはこの三つ全ての評価値を利用する。

このうち H_D は専門用語の構成語に幾つ訳語が含まれるかを表わし、対訳辞書、シソーラス、およびカタカナ表記の語に対する訳語推定によって求める。ただし、カタカナ表記の語に対する訳語推定とは、対訳辞書に登録されていないカタカナ表記の語に対して、その語の元となった語（英語）を発音や綴りの規則性を利用することによって推定しようとするものである [2, 1]。これは、カタカナ表記の語をまずできるだけ英語の表記に近いような一つのアルファベットの文字列に変換しておき（カタカナ-ローマ字変換）、それと英語との間で文字の並びと発音の共通性を考慮しながら対照を行なう。例えば「ファシリティ」という語を、簡単な規則を用いて、'fasiriti' というローマ字に変換したとする。これと'facility' を比べてみると、'c' と's'、'r' と'l' などを考慮すればほとんど同一視できることがわかる。H_D の計算例を以下に示しておく。

例 6 「マルチリンク送信」と'multilink send' の訳語対に与える H_D は以下のようにして求める。

「マルチリンク」をカタカナ-ローマ字変換によって'maruchirink' に変換する

'maruchirink' と'multilink' では、'r' と'l'、'c' と't' を考慮すると全体として'm,r,c,i,r,i,n,k' の 8 文字が一致しているとみなせる。そこでこれらの間の類似度を

$$\frac{\text{一致した文字数}}{\max\{\text{'maruchirink' の文字数}, \text{'multilink' の文字数}\}}$$

で与えるものとすると、これは 8/11 = 0.73 となる。一方、「送信」と'send' の訳語対が対訳辞書から得られたとすると、H_D は次のようになる。

$$\begin{aligned} & H_D(\text{「マルチリンク送信」}, \text{'multilink send'}) \\ = \ & 0.73 + 1 = 1.73 \end{aligned}$$

一方、H_{Occ} は全テキストを通して対応が何回現れたかを計算したものであり、H_{Sta} は訳語対としての関係の強さを計算したものである。日本語の専門用語 t_j と英語の専門用語 t_e の間の訳語対としての関係の強さは、次の形で与える。

$$H_{Sta} = \frac{(t_j \text{と} t_e \text{の共起頻度})^2}{(t_j \text{の出現回数}) \times (t_e \text{の出現回数})} \quad (5)$$

H_{Sta} は $[0, 1]$ の値をとり、t_j と t_e が常に同じ対訳文中で現れる場合には、$H_{Sta} = 1$ の最大値をとる。

3.2 訳語対の決定順序

訳語対の決定は、訳語対応ネットワーク上に示された訳語対の候補の中で最も適当であると考えられるものから順番に行なう。決定された訳語対と両立不可能な訳語対の候補はもはや成立しないものとして訳語対応ネットワーク上から消去していく。

訳語対の候補の中で最も適当であるというのは H_D が最も高い候補とした。ここで H_D が最も大きい訳語対の候補が複数個存在し、それらの間に両立不可能性が成立していたとする。決定される訳語対の間には両立不可能性が成立することが許されないから、H_D だけでは決定ができないことになる。これは語彙情報の不足が原因であると考え、次に統計情報を利用する。つまり H_D だけでは決定ができない場合には H_{Sta} で比較し H_{Sta} の大きい訳語対の方を決定する。さらに H_{Sta} でも決定できない場合には H_{Occ} を使って決定する。さらに、この三つの評価値を利用してもどちらかを決定できない場合には、

表 1: 実験結果

名	テキスト量 (K-Byte)	抽出数	対応の誤り	正しい	正解率	人手で抽出した数	再現率
X.4	0.95	26	0	26	100.0%(26/ 26)	47	55.3%(26/47)
X.20bis	24.3	64	3	61	95.3%(61/ 64)	87	70.1%(61/87)
X.22	16	242	27	215	88.8%(215/242)	-	-
X.24	23	170	5	165	97.1%(165/170)	-	-
X.25	643	813	67	746	91.8%(746/813)	-	-
X.29	36	418	27	391	93.5%(391/418)	-	-

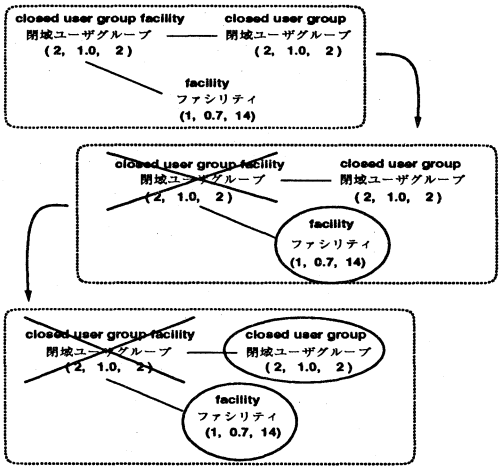

図 6: 訳語決定の例

その評価値をもつ訳語対の候補に対する決定は保留しておき、次に大きい評価値をもつ訳語対の候補に対して抽出を行なう。この段階で決定が保留された訳語対の候補は、他の訳語対が決定されていく過程で、再評価によりどちらか決定できることがある。この例を図 6に示す。ただし、各訳語対の候補には (H_D, H_{Sta}, H_{Occ}) という形で評価値が与えられているものとする。

4 実験と考察

我々は、CCITT (国際電信電話諮問委員会) の国際標準勧告集のうち六つの X 勧告 (X.4、X.20 bis、X.22、X.24、X.25、X.29) の文章を実験対象に選び、その各対訳文章から本研究による手法によって専門用語の訳語抽出を試みた。ただし本手法では、両立不可能性を表現した訳語対応ネットワークから、各訳語対の評価値に基づいて、正しい訳語対を一意に決定することを目的としている。従って辞書情報の欠如その他で評価値が極めて低い訳語対に関しては、その抽出を行なわない。そこで、H_D が 0 の訳語対候補は抽出の対象としなかった。この実験結果を表 1に示す。表中には抽出された専門用語の訳語数、訳語の誤り、正解率を示している。ここでの正解率とは抽出された専門用語の訳語対に対する正しい訳語対の割合である。さらに、実験文章のうち X.4、X.20bis の場合についてのみ、再現率を計算した。ただし、ここでの再現率とは同じ対訳文章を用いて人手で抽出した専門用語対に対する本手法で抽出された正しい専門用語対の割合である。ただし、人手で抽出した数は、本手法で抽出された正しい専門用語対の数に抽出されなかった専門用語対の数を加えたものである。この実験結果から、文章量が比較的に小さいものについても、全般的にかなり高い正解率で専門用語訳語対が一意に決定されることが分かる。すなわち、対訳文の内部における訳語対応の両立不可能性を利用することによって、専門用語の認識の曖昧性と対応の曖昧性の解消が行なえることを示している。しかし高い正解率が得られている一方で再現率が非常に悪い実験結果もある。これらの原因は次の三点である。

- 一回しか対応が現れない
- 語彙情報や統計情報の欠如で評価値が低い
- 専門用語認識の際に用いた文法規則だけでは捉えられないものがある

5 おわりに

訳語対応における両立不可能性を考慮した本手法では、比較的少量の対訳文章からでも正しい訳語対を一意的に高精度で抽出できることを示した。一方であらかじめ設定した文法で捉えられない現象が再現率の低下を招いた。よって単言語でより的確に専門用語を認識するための技術を確立することが今後の課題である。

参考文献

[1] 石本浩之. 二言語対訳テキストを用いた専門用語対訳辞書の自動構築. 修士論文, 京都大学工学部, 1995.

[2] 石本浩之, 長尾眞. 対訳文章を利用した専門用語対訳辞書の自動作成：訳語対応における両立不可能性を考慮した手法について. 情報処理学会研究報告, Vol. 94, No. 31, 1994.

[3] 熊野明, 平川秀樹. 言語情報と統計情報を用いた対訳文書からの機会翻訳辞書作成. 情報処理学会研究報告, No. NL100–12, pp. 89–96, May 1994.

[4] J. Kupiec. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31th Annual Meeting of ACL*, pp. 17–22, June 1993.

[5] F. Smadja. How to compile a bilingual collocational lexicon automatically. In *Proceedings of the AAAI-92 Workshop on Statistically-Based NLP Techniques*, pp. 65–71, 1992.