自 然 言 語 処 理 102−11
（1994. 7. 22）

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 —

石本 浩之† 長尾 眞†

† 京都大学 工学部 電気工学第二教室
〒606 京都市左京区吉田本町

あらまし

対訳文章を用いて専門用語の対訳辞書を作成する際の問題として，専門用語の認識，専門用語の複数対応の解消，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．我々は，既存の辞書情報の利用に加え，両立不可能性という概念を導入することによって，これらの問題に対処する．両立不可能性は対訳文内の対応関係を考慮することによって得られる関係であり，単言語での専門用語認識の曖昧性や二言語間での対応の多様性を統一的に扱うことができる．この両立不可能性と既存の辞書から得られる情報によって，専門用語の訳語対を一意に決定できるほか，対訳辞書作成の際に必要となる幾つもの異表記の訳語抽出が可能となる．約 500 文程度の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．

Automatic Construction of a Bilingual Dictionary of Technical Terms from Parallel Texts
: Considering Incompatibility between Correspondences

Hiroyuki Ishimoto† Makoto Nagao†

†Department of Electrical Engineering, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto 606, Japan

Abstract

When constructing a bilingual dictionary of technical terms using parallel texts ,we have some problems to cope with as recognition of technical terms, correspondence ambiguity, possible presence of multiple translation. We propose a method to build a dictionary, using incompatibility between bilingual correspondences of technical terms as well as existing hand-written dictionaries. Incompatibility is the relation extracted by considering correspondences ,which occur in each parallel sentence, and could treat ambiguity of recognition of technical terms and ambiguity of correspondences together as one. Incompatibility works effectively in resolving the ambiguity problems and that multiple translations , which are to be concerned in the construction of a bilingual dictionary, can be extracted.

1 はじめに

文の解析や機械翻訳の際, 未知語の扱いがしばしば問題となり, 文の解析において曖昧性が生じたり, 機械翻訳の際に誤訳を生むことがある. 特に技術用語や専門用語は様々な分野での技術革新が盛んな今日, 日々増加し続けており, 大規模な専門用語辞書を常に作成し続けることが必要となっている. しかし, 大規模な辞書作成には人手やコストが膨大となるため, 専門用語辞書を計算機上で効率的に構築する技術が必要とされている. そのための知識源としては, 対訳文章の利用が考えられる.

現在, 自然言語処理の分野において, 対訳コーパスは豊富な知識源として注目されており, 知識獲得に関して各種の研究がなされている.[4, 1, 3, 6, 7] このとき, 文間の対応, 単語間の対応, 構造間の対応などがしばしば問題となり, 辞書情報, 統語情報, 統計情報を用いることによって対処する方法が多数考案されている. 一方, 専門用語対訳辞書を自動作成する際にはこれらの問題を含めて以下のような難しさがある.

1. 専門用語の認識
2. 訳語対応の多様性とその解消
3. 幾つもの異なった訳の存在の取り扱い
4. 専門用語対の評価

これらの問題に対して, 山本[9] は英語の名詞連続と未知語を専門用語として抽出し, これらに対応する日本語の訳語を辞書情報と頻度情報を用いて抽出している. また, 熊野[10] は, 日本語から合成語や未知語を専門用語として抽出し, 文対応が行なわれたテキストを基に英語の訳語候補を取り出す. さらに, 各訳語候補に対して辞書情報, 統計情報を基に評価を行なうことによって訳語の選出を行なっている. しかし, これらの手法では, どちらか一方の言語において専門用語抽出の精度がかなり高くなければならない. また, 複数の異なった訳の存在の取り扱いが考慮されていないために, ある専門用語が幾つかの異表記の訳語を持つ場合, その抽出は難しい. 一方, Kupiec[5] はあらかじめ各言語から名詞句を抽出しておいて, 統計的手法を用いて名詞句間の対応づけを行なっている. しかし, この方法を用いて専門用語の訳語対の抽出を行なう場合, 専門用語を的確に認識する技術が必要である. また, ある専門分野における対訳文章で, 統計的に意味を持つ量のものを入手することは現在のところは困難である.

比較的小規模な対訳テキストからでも専門用語対訳辞書の作成を行うためには, できるだけ既存の辞書情報だけを用いることによって, 専門用語認識の曖昧性を含めた二言語間での対応の多様性を解消する必要がある. これを実現するためには, 各対訳文内部の訳語の対応関係を考慮する必要がでてくる. そこで, 我々は既存の辞書情報に加えて両立不可能性という概念を導入し, 各対訳文内部の対応関係を考慮することによって専門用語の対訳辞書を作成する方法を提案する. 両立不可能性とは, 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関係であり, 二つの専門用語対の候補が互いに両立するかどうかという情報をもつ. これを用いることによって, 専門用語認識の曖昧性や専門用語対応の多様性を統一的に扱うことができる.

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}

一般の和英辞書における「網」のエントリーには 'net' とあり,'network' はない. よって「デジタル網」にとっては {digital network, digital circuit} の両者どちらに対しても同じくらいの類似度をもつことになり, これだけではどちらに対応するかを決定することは難しい. 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" にとって "digital circuit" に対する類似度の方が高いので, そ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. このような関係を両立不可能性の関係と呼んでいる.

我々はこの両立不可能性という関係を積極的に用いることによって, 専門用語の訳語対の抽出を行なった. 両立不可能性という関係で表わされた曖昧性や多様性の問題は, 各訳語対に対して計算された評価値を基に解消され,

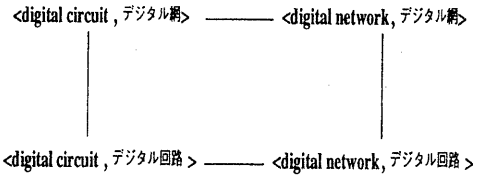

図 1: 専門用語対間の両立不可能性の例

専門用語の訳語対を一意に決定できる．また，専門用語が異表記の訳語を幾つか持つ場合でも，それらの間に両立不可能性が成立していなければ抽出できることになる．一方，専門用語の訳語対の評価値を計算する際のヒューリスティクスには辞書情報，シソーラス，発音に基づくカタカナ-ローマ字変換を用いた．これを以下に詳述する．

2 両立不可能性を考慮した訳語対応の枠組

両立不可能性を利用することによって，専門用語の訳語決定を行なう際，問題の展開方法によっては計算量が膨大となる．そこで，我々は，両立不可能性を図1のような訳語対応ネットワークとして表現する．訳語対応ネットワークは，専門用語対を節点とし，互いに両立しない専門用語対間に枝を張ったグラフである．この訳語対応ネットワークの構築方法の詳細については4節に述べることとし，以下に訳語対応ネットワークを用いる理由を示す．

- 問題を展開する順序が既定されない．
- 専門用語対の候補間にある関係を容易に表現できる．
- 専門用語抽出の曖昧さと対応の曖昧さを同時に表現できる．

図2に本手法全体の枠組を示す．まず，単言語ごとに，文章全体を通しての単語列の重複を考慮しながら，専門用語候補の抽出を行なう．この段階では，考えられるあらゆる専門用語の候補が抽出される．次に，文対応済みの対訳テキストと各言語について抽出された専門用語候補を用いて，専門用語対の候補を限定する．ただし，この段階では専門用語対の候補に曖昧性が残っていてもよい．専門用語の定義と専門用語対の候補の抽出方法については，3節で述べる．次に，各対訳文対における 任意の二つの専門用語対の候補が「互いに両立するかどうか」を調べて，その両立不可能性を訳語対応ネットワークとして表現する．最後に，構築された訳語対応ネットワークの各節点 (専門用語対) に対して評価を行ない，評価最大のもののうち互いに矛盾しない専門用語対から順次確定していく．確定された節点と両立不可能な節点は順次消していくことによって，正しい訳語対応が一意に得られる．

3 専門用語と専門用語対

3.1 専門用語の定義

次に，専門用語の特徴と本研究での定義を述べる．専門用語の特徴として次の四点が挙げられる．

- 名詞句が多い
- 専門用語は複合語からなる場合が多く，個々に見ると専門用語ではないような基本的な単語によって形成される場合が多い．

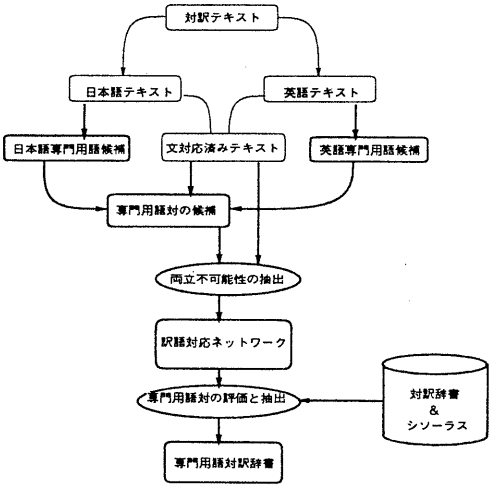

図 2: 全体の枠組

- 単言語では辞書登録されているが，訳語として対訳辞書に登録されていない場合がある．特に日本語では，この現象がカタカナ語によく現れる．
- 前置詞句を含む場合がある

一般に，専門用語には上に述べたような特徴があるものの，その定義は非常に難しい．文を構文解析すれば，範囲の限定は比較的容易にはなるが，構文解析に誤りがないことを前提とすることはできない．以上のことから，本研究で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．専門用語の候補となる例を例2例3に示す．

例 2 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．

オプショナルファシリティ

ファシリティ

例 3 また次の文字列が文章中に現れたとする．

… one or more closed user groups and …

… the closed user group with outgoing access facility …

… Closed user group with outgoing access selection. …

このときには，以下の専門用語が抽出される．

closed user group

closed user group with outgoing access

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．ただし，日本語では形態素解析，英語では語尾解析を行ない，名詞句解析は正規文法の範囲内で行なう．

定義 1 (専門用語) 各言語における単語 S の上の列で,

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

という単語列があったとする. ここで次のような連接

$$\alpha = \alpha_1 \cdot \gamma \cdot \alpha_2$$

$$\beta = \beta_1 \cdot \gamma \cdot \beta_2$$

が文章中に現れ, さらに

$$a_p \neq d_m \quad \text{または} \quad a_p \doteq d_m = space$$

$$c_1 \neq e_1 \quad \text{または} \quad c_1 = e_1 = space$$

$$np(\gamma) = true$$

の全ての条件を満たすとき, γ は専門用語であるとする. ただし, $np(X) = true$ は X が名詞句であることを表わす. 名詞句解析の文法規則は現在のところ以下のように規定しており, この文法で NPもしくは名詞句 として受理される単語列をすべて名詞句とした.

英語

$NP \rightarrow noun$	$NP \rightarrow undefined$
$NP \rightarrow adj \cdot NP$	$NP \rightarrow noun \cdot NP$
$NP \rightarrow undefined \cdot NP$	$NP \rightarrow adv \cdot NP'$
$NP' \rightarrow adv \cdot NP'$	$NP' \rightarrow adj \cdot NP$
$NP \rightarrow noun \cdot PP$	$PP \rightarrow prep \cdot NP$
$NP \rightarrow undefined \cdot PP$	

日本語

名詞句 → 名詞	名詞句 → 未定義語
名詞句 → 名詞・名詞句	名詞句 → 未定義語・名詞句
名詞句 → 名詞・名詞修飾句	名詞修飾句 → 名詞接続助詞・名詞句

3.2 専門用語対の候補

定義1に従って抽出された専門用語候補と文対応済みの対訳テキストを用いて, 専門用語対の候補を抽出する. 専門用語対の候補となるためには, 専門用語候補の対応が2回以上現れなければならないとした.

図3の場合には, まず以下の専門用語が候補として抽出される.

専門用語候補 γ	α_1	α_2	β_1	β_2
user facility	*space*	*space*	*space*	and
facility	user	*space*	and	*space*
ユーザファシリティ	*space*	*space*	*space*	と
ファシリティ	ユーザ	*space*	と	*space*

このとき, ⟨user facility, ユーザファシリティ⟩, 及び ⟨user facility, ファシリティ⟩, ⟨facility, ユーザファシリティ⟩, ⟨facility, ファシリティ⟩ が専門用語対の候補となる.

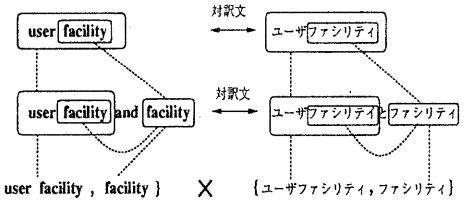

図 3: 専門用語対の候補となる例

4 訳語対応ネットワーク

抽出された専門用語対の候補において互いに両立しない関係を訳語対応ネットワーク上に表現しておく. これによって, 専門用語対の候補の中から正しい専門用語対を一意的に, かつ効率的に選択できる. これは, あらかじめ抽出された専門用語対の候補を再度対訳テキストに参照させることによって行なわれる. まず, 訳語対応ネットワークを以下に定義しておく.

定義 2 (訳語対応ネットワーク) 訳語対応ネットワークは無向グラフ $G = [V, E]$ である. V は専門用語対を表わす節点の集合であり, $E = \{(u, v) \mid u, v \in V\}$ は両立不可能性を表わす無向枝の集合である. 専門用語対 $u \in V$ の評価値は $h(u)$ で与えられ, その頻度は $freq(u)$ で与えられる. また, 専門用語対 u, v 間の両立不可能な関係 $(u, v) \in E$ の頻度は $freq((u, v))$ で与えられる.

専門用語対の評価値の計算方法については 5.1で述べることとし, 両立不可能性という関係はどういう場合に生じるかについて以下に述べる. 専門用語対の候補のうち, 両立不可能な関係にあるのは, 以下の三つに場合である. ただし, 図中の $[E1, \cdots, E7]$ と $[J1, \cdots, J8]$ は対訳文であり, その各要素は単語 (形態素) である. また, $Te1, Te2, Te3, Tj1, Tj2$ は専門用語の候補であり, $\langle Te1, Tj1\rangle, \langle Te1, Tj2\rangle, \langle Te2, Tj1\rangle, \langle Te2, Tj2\rangle, \langle Te3, Tj1\rangle$ は専門用語対の候補である.

Type1 一対訳文中で一つの専門用語は一つの専門用語にのみ対応する.

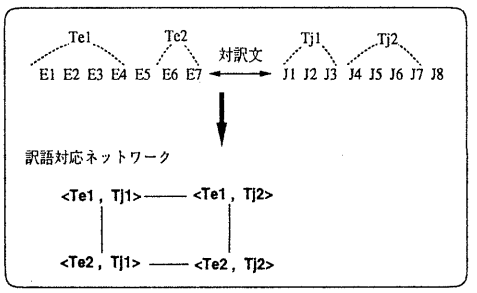

図 4: Type1

Type2 専門用語の候補抽出の際に曖昧性がある場合，そのうちの一つだけが正しい．例えば図5の場合には単語列 $[E1,\cdots,E7]$ に専門用語の候補 $Te1, Te2, Te3$ が混在しているが，そのうちの一つだけが $Tj1$ に対応できる．

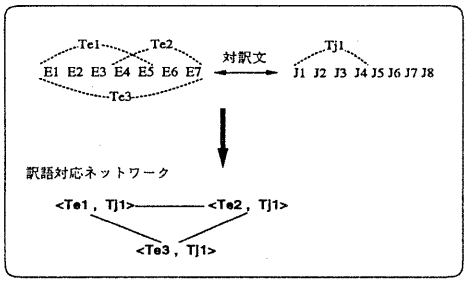

図 5: Type2

Type3 両言語において 専門用語の構成語が専門用語である場合には，大きな専門用語の対とその構成語の対の間には両立不可能性は成立しないことにする．例えば，図6の場合，$Te1$ は $Te2$ に含まれ，$Tj1$ は $Tj2$ に含まれているが，$\langle Te1, Tj1\rangle$ と $\langle Te2, Tj2\rangle$ とは両立できる．ただし，$Te1$ と $Tj2$ が成立する場合には $Te2$ と $Tj1$ は成立しない．

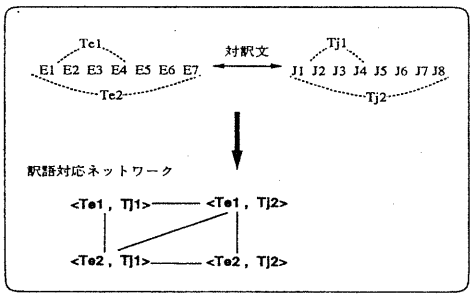

図 6: Type3

5 専門用語対の評価と抽出

訳語対応ネットワークにおける各専門用語対の評価を行ない，この評価値と両立不可能性を用いて専門用語の訳語対の集合を求める．そこで，まず各専門用語対の評価値の算出方法を示し，次に専門用語対の候補の中から専門用語対を選出していく過程について述べる．

5.1 専門用語対の評価

専門用語対 $\langle t_e, t_j\rangle$ の評価値は $[0,1]$ の実数であらわされ，次の手順にしたがって計算する．

1. t_j を単語列 (形態素の列) とみなして，その順序を入れ換える．それを t'_j とする
2. t'_j 中の各単語を相手言語に翻訳する (日本語 ⇒ 英語)．それを t_{je} とする
3. t_{je} と t_e を文字列とみなし，パターンマッチングをおこなう．
4. すべての並べ変えとあらゆる翻訳について，マッチングの程度を表わす評価値を計算し，その中で最大のものをとる．

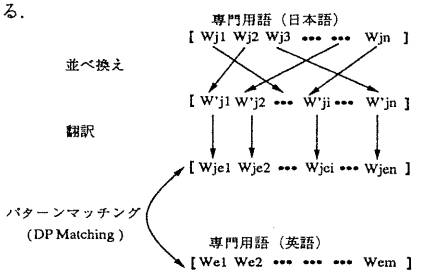

図 7: 専門用語対の評価法のながれ

これらの詳細を以下に述べる．

5.1.1 相手言語への翻訳

ある単語の翻訳を行なうとき，その単語の適切な訳語が必ずしも対訳辞書にあるとは限らない．よって，シソーラスから取り出した同義語も訳語の候補とした．また，日本語のカタカナ語はたとえ未定義語であっても，発音の規則性に基づいたカタカナ - ローマ字変換を行なうことによって，原言語にある程度復元できると期待できる．相手言語への翻訳関数を g とするとこれは次の仕事を行なう．

- 対訳辞書からの訳語とシソーラスからの同義語を訳語群と呼ぶことにすると，この訳語群の中から候補を一つずつ順番に選択する．
- 対訳辞書，シソーラスから候補が得られないが，カタカナ - ローマ字変換が可能な場合にはその変換を行なったものを返す
- 対訳辞書，シソーラス，カタカナ - ローマ字変換のどれからも候補が得られない場合は空列を返す

5.1.2 パターンマッチング

各専門用語は抽出されるときに単語に区切られているにも関わらず，文字列のパターンマッチングで評価値を計算する理由は三つある．

- 表記のゆれに対応する

  例 4 辞書では論理 ⇒ "logic" と翻訳されるが，これと"logical channel"(論理チャネル) 中の "logical" とのマッチングを考えたとき単語単位ではマッチしない．
- カタカナ - ローマ字変換によって完全な英語に変換できないため，ある程度のノイズを吸収する必要がある

  例 5 チャネル ("channel") ⇒ "chaneru" での r と l.

• 専門用語における日本語の形態素数と英語の単語数が必ずしも一致しない.

5.1.3 専門用語対の評価値

専門用語対 $\langle t_e, t_j \rangle$ の評価値 $h(\langle t_e, t_j \rangle)$ の計算方法を以下に定義する. ただし, $0 \leq h(\langle t_e, t_j \rangle) \leq 1$ とする.

まず, t_e,t_j がそれぞれ以下の単語列であったとする.

$$t_e = [w_{e_1} \ldots w_{e_m}]$$
$$t_j = [w_{j_1} \ldots w_{j_n}]$$

つづいて単語列の並べ変えを行なう関数を f, 相手言語に翻訳する関数を g とおいて,

$$\begin{aligned} t_{je} &= g(f(t_j)) \\ &= [w_{je_1} \ldots w_{je_n}] \end{aligned}$$

となったとする. すると 評価値は

$$h(\langle t_e, t_j \rangle) = \max_{f,g} \left\{ dp_{m,n}(t_e, t_{je}) \times \frac{1}{max\{m,n\}} \right\}$$

$$dp_{i,j}(t_e, t_j) = \begin{cases} (\text{if i} = 0 \text{ or j} = 0) \\ 0 \\ (\text{otherwise}) \\ max \begin{cases} dp_{i-1,j}(t_e, t_j) \\ dp_{i,j-1}(t_e, t_j) \\ dp_{i-1,j-1}(t_e, t_j) + sim(w_{e_i}, w_{je_j}) \end{cases} \end{cases}$$

$sim(x,y)$ は原則として $x = y$ の場合は 1 を返し, それ以外は 0 を返す. ただし, x または y がカタカナ - ローマ字変換をした文字列に含まれていて x と y が "b" と"v", "r" と"l" などの場合には $sim(x,y) = 1$ を返す.

図 8に 論理チャネルと logical channel 間の評価値の計算例を示す. ただし, ここでは "チャネル" の訳語が辞書から得られない場合を想定しており, その部分だけカタカナ - ローマ字変換を行なっている.

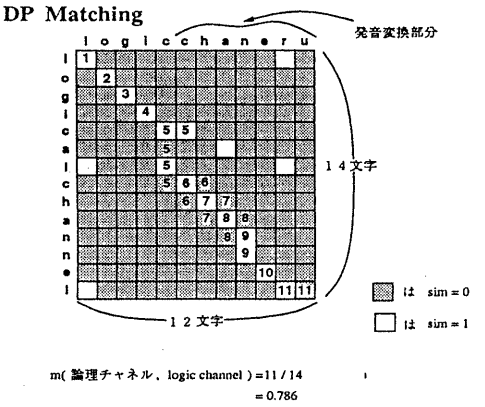

図 8: 専門用語対の評価の例

5.2 訳語対応ネットワークからの専門用語対の抽出

構築された訳語対応ネットワークの各節点 (専門用語対) に対して評価を行ない, 評価最大のもののうちで互いに両立不可能性が成り立たっていない専門用語対から順次確定していく. このとき, 同じ評価値を持つ専門用語対の間に両立不可能性が成立するときには, 頻度を基準に決定を行なう. 評価値と頻度を利用しても両立不可能性が解消できない場合には, その時点での決定は保留しておいて, その次に評価最大となるものに対して計算をした後, 再度評価を行なう. 次に, 確定された節点と両立不可能な節点は順次, 消していくことによって, 正しい訳語対応が一意に得られる. ここで, 図 9に訳語対応ネットワークからいかにして一意に訳語対が決定されるかを示す. ただし, 節点は ⟨専門用語, 専門用語, 評価値, 頻度⟩ とした.

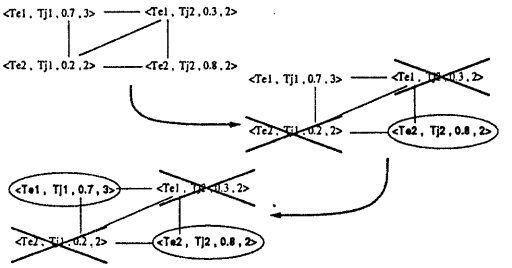

図 9: 訳語対応ネットワークからの訳語対抽出の過程

訳語対応ネットワーク $G = [V, E]$ から専門用語対の抽出を行なうアルゴリズムを以下に示す.

1. 専門用語対の集合 V の要素のうちで, 評価値が最大となる専門用語対の集合 P を求める. ただし, 評価値が最大で互いに両立不可能な関係にあるものについては, 頻度が最大となる専門用語対を選ぶ. ここで, $h(u) < x, u \in P(x$はしきい値) である場合には終了.
2. 集合 P の要素のうちで, 互いに両立不可能な関係にない専門用語対の集合 C を求める.
3. 集合 C の要素数 $|\ C\ | = 0$ の場合には, 集合 V の要素のうちで, 次に評価値が最大となる専門用語の集合を新たに集合 P として (ただし, 評価値が最大で互いに両立不可能な関係にあるものについては, 頻度が最大となる専門用語対を選ぶ.) 2に戻る.
4. 集合 C に含まれる専門用語対を正しい訳語対として抽出する.
5. 訳語対応ネットワークにおいて, 集合 C に含まれるすべての要素及び, 集合 C 中の各要素と両立不可能な関係にあるすべての要素とを集合 V から排除し, これを集合 V' とする. さらに, 排除された専門用語対に接続されていたすべての枝を集合 E から排除し, 集合 E' とする. 制約ネットワーク $G = [V', E']$ を新たに定義し, 1へ戻る.

アルゴリズム 3において集合 P に含まれる要素のうち互

閾値	抽出された訳語対の数	対応の誤り	専門用語として不適当	正しい訳語対の数	正解率	人手で抽出した訳語対の数	再現率
0.1	314	42	5	267	85.0 % (267/314)	327	81.7 % (267/327)
0.2	301	36	5	260	86.4 % (260/301)	327	79.5 % (260/327)
0.3	282	24	2	256	90.8 % (256/282)	327	78.3 % (256/327)

表 1: 実験結果

いに両立不可能な関係にある専門用語対に対しては, 再び1に戻ったときに再計算が行なわれる. また, 5で両立不可能性を用いて訳語対応ネットワークから節点と枝を排除できるのは, 次の条件を満たす場合にのみ限定した.

$$freq(u) = freq((u,v))$$

または

$$freq(v) = freq((u,v))$$

ただし, $u, v \in V$ である. これは, u と v は両立可能であるにもかかわらず, 両立不可能性が一度でも成立してしまった為にどちらか一方が排除されてしまう, ということを防ぐためである.

6 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一部分を用いて実験を行なった. これは電気通信の分野についての勧告集であり, 英語で編集されている. この勧告集は主要各国の言語に翻訳されつつあり, 我々は日本語のテキストと英語のテキストを用いた. 実験に用いたテキストは日本語で 458 文で, 英語で 465 文であり, 一文あたりの文字数は平均, 日本語で 57.6文字, 英語で 140.0文字 である. 実験に用いた英語品詞辞書の見出し語数は約 55,000 であり, 日本語の形態素数は約 36,000 である. 専門用語対の評価値の計算には, 見出し語数約 50,000 の和英辞典 (講談社学術文庫の和英辞典) [12] と約 100,000 の英単語を網羅したロジェのシソーラス[13] を用いた. また, 文対応は人手で行なった.

6.1 実験結果

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. よって, 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. 抽出を行なうかどうかの判定はしきい値を基準に行なう. 本研究では, 三種類のしきい値 $(0.1, 0.2, 0.3)$ を設定し, それぞれの場合について実験を行なった. その結果を表 1に示す. この表には各しきい値に対して, 抽出された専門用語の訳語数, 訳語の誤り, 訳語としては間違っていないが専門用語として不適当なものの数, 正解率, および再現率を示している. ここでの正解率とは抽出された専門用語対の数に対する専門用語として適当であるものの割合である. また, 再現率は, 同じテキストを用いて人手で抽出したときの専門用語対の数に対する本手法で抽出された専門用語として適当であるものの割合である. ただし, 専門用語を人手で抽出する際には, 特に名詞句には限定せず, 専門用語として考えられるものはすべて抽出している. この実験結果からしきい値をかなり低く設定しても, かなり高い正解率で専門用語の訳語対が一意に決定されていることがわかる. このことから, 辞書情報の不十分なところを両立不可能性がうまく補っていると言える.

次に本手法で実際に得られた専門用語の訳語対を示す.

例 6 未知語や複合語の例 (カタカナ – ローマ字変換の効果)

インターフェース= interface

スループットクラスネゴシエーション

= throughput class negotiation

例 7 前置詞句を含むの例

出接可閉域ユーザグループファシリティ

=closed user group with outgoing access facility

パケットシーケンスの番号付与

=sequence numbering of packets

例 8 シソーラスの効果

着呼 禁止= Incoming call barred

(対訳辞書からは禁止の訳語として prohibit, ban などはあるが bar はなかった.prohibit の同義語として bar が得られた)

例 9 異表記の訳語が複数個得られた例

facility = {機能 (対訳辞書から), ファシリティ (カタカナ – ローマ字変換から)}

direction of data transmission

= {データの伝送方向, データ伝送方向}

6.2 失敗例と考察

誤った専門用語対を抽出してしまう例としては次のようなものがあった.

- 文字列のパターンマッチングは, カタカナ語などの未知語や, logic と logical などの表記の揺らぎに対しては有効に働いたが, 例 10などの場合のようにかえって悪影響を及ぼす場合もあった.

例 10 パターンマッチングによる悪影響

… 当面当事者間の合意による …

… agreed for a period of time …

が常にペアとなって現れ, 当事者の訳語 "person concerned"

と "period of time" が評価値 0.31 を持つために
当事者 =a period of time が抽出された
これらの問題は, 文字列のパターンマッチングと単語列のマッチングとを組み合わせることによって解決できると考えられる.

- 専門用語抽出の際用いた文法規則では抽出されない
例 11
… facility code not allowed .
という文からは専門用語として facility code not allowed が抽出されず, 結果的に非許容ファシリティ符号 = facility code という対応が抽出された.
文法規則の整備によってある程度改善されることも考えられるが, それによる弊害も考慮する必要がある.
- 正しくない専門用語対を抽出した場合, これと両立不可能な専門用語対が訳語対応ネットワークから排除されるため, さらに間違ったものを選択してしまう場合があった. 訳語対応ネットワークはもともと非連結グラフの集合ではあるから一つの間違いが全体に与える影響は小さいものの, 評価方法を充実させる必要がある.

一方, 対応としては間違っていないが, 専門用語として不適当なものの例としては, 以下のようなものがあった.

例 12
データ伝送の両方向 = both direction of data transmission

また, 対応がとれない現象は以下の場合に見られた.

- 一回しか対応が現れない
- 辞書情報の欠如で評価値が低い
- 言語によって文法範疇が異なる
例 13 日本語では名詞句であるが英語ではそうではない例
… 転送することによって …
… by transferring a second …

これらの問題を解決するためには, 対訳辞書の整備, 専門用語の確実な抽出などが必要であると考えられる. しかし, 日本語と英語間のように, 表現方法がしばしば異なる言語を対象とする場合には, 単に専門用語抽出の文法規則を整備するだけでは困難であり, 文の内部構造を考慮した対応づけが必要であると考えられる.

7 おわりに

対訳テキストを用いて専門用語の対訳辞書を作成するとき, 専門用語の認識, 専門用語の複数対応の解消, 幾つもの異なった訳の存在の取り扱いが問題となる. 我々は, 専門用語および専門用語対の候補を抽出する時点では一意に絞り込まず, これらの曖昧性を両立不可能性として訳語対応ネットワーク上に表現し, 既存の辞書情報を用いて正しい訳語対を一意に求めることによって, 専門用語対訳辞書を作成するという方法を提案した. この方法では, 従来の方法で問題となる専門用語抽出の精度がそれ程問題とはならず, さらに, 対訳辞書作成の際に必要となる幾つもの異表記の訳語抽出が可能である. 本論文では, 本手法が比較的少量の対訳テキストから専門用語対を抽出する際に有効であることを示したが, 大規模な対訳テキストに対しての応用も考えている. 今後の課題として, 文の構造照合[11] を利用して候補の限定を行なうことや, 統計情報から対訳辞書の拡張を行なうことによって[8] 未知語を含む専門用語に対処することなどを考えている. また, 専門用語の文法学習による抽出精度の向上も考えている.

謝辞

本研究を進めるにあたって有益なコメントを数多くいただいた宇津呂武仁氏 (奈良先端科学技術大学院大学) に感謝します.

参考文献

(1) V. Sadler and R. Vendelmans. Pilot implementation of a bilingual knowledge bank. In *COLING-90*, Vol. 3, pp. 449–451, 1990.

(2) W. Gale and K. Church. A program for aligning sentences in bilingual corpora. In *ACL-91*, pp. 177–184, 1991.

(3) H. Kaji, Y. Kida, and Y. Morimoto. Learning translation templates from bilingual text. In *COLING-92*, pp. 672–678, 1992.

(4) J. Klavans and E. Tzoukermann. The BICORD System: Combining lexical information from bilingual corpora and machine readable dictionaries. In *COLING-90*, Vol. 3, pp. 174–179, 1990.

(5) J. Kupiec. An algorithm for finding noun phrase correspondences in bilingual corpora. In *ACL-93*, pp. 17–22, June 1993.

(6) Y. Matsumoto, H. Ishimoto, and T. Utsuro. Structural matching of bilingual texts. In *ACL-93*, pp. 23–30, June 1993.

(7) T. Utsuro. *Lexical Knowledge Acquisition from Bilingual Corpora*. Doctorial Thesis, Kyoto University, 1993.

(8) T. Utsuro and et al. Bilingual text matching using bilingual dictionary and statistics. In *COLING-94*, 1994. *forthcoming*

(9) 山本由紀雄, 坂本仁. 対訳コーパスを用いた専門用語対訳辞書の作成. 情報処理学会研究報告, No. NL94–12, 1993.

(10) 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.

(11) 石本浩之, 宇津呂武仁, 松本裕治, 長尾眞. 日英対訳文間の構造照合. 情報処理学会研究報告, Vol. 93, No. 41 (93–NL–95), pp. 81–88, May 1993.

(12) 清水護, 成田成寿 (編). 和英辞典. 講談社学術文庫, 1979.

(13) S. R. Roget. *Roget's Thesaurus*. Crowell Co., 1911.