ウェブを用いた専門用語翻訳支援における多様な情報源からの信頼度情報の提示

外池 昌嗣 † 宇津呂 武仁 †† 影浦 峡 ‡ 佐藤 理史 ‡‡ 阿辺川 武 ‡

† 京都大学 情報学研究科
†† 筑波大学 システム情報工学研究科
‡ 東京大学 教育学研究科
‡‡ 名古屋大学 工学研究科

1. はじめに

翻訳者が文書を翻訳する場面では，既存の対訳辞書に載っていない専門用語が現れたとき，訳を自分で決める必要性が生じる．一般に，翻訳者が専門用語の訳語を探す際には，専門用語対訳集や，既訳の専門文書を収集してその中から、当該用語の訳語を探すことになるが，近年では，ウェブを検索して，当該分野における専門用語の語法を確認するなどの手段も用いられる．本論文では，翻訳者による文書翻訳の作業において，専門用語の訳語を決定する過程を支援することを目的として，専門用語の訳語の候補を生成するとともに，既存の汎用辞書や専門用語辞書，ウェブ上の専門文書等，多様な情報源からの情報を自動的に収集して，各々の訳語候補の確からしさを提示するシステムを設計，実装した．ここで，翻訳者が，通常，当該分野の非専門家であると仮定すると，この翻訳支援の過程においては，当該分野の非専門家である翻訳者が確信を持って専門用語の訳語を決定するために必要となる情報を同定し，これの効果的な提示を実現することが最も重要な課題となる．本論文では，当該分野における訳語既知専門用語を利用して，ウェブ検索エンジンを用いて訳語候補の信頼度を推定したり，あるいは，専門文書中において，訳語と原語の併記を抽出する等，数種類の情報とその信頼性の度合いを提示することにより，専門用語翻訳支援の過程を実現する．

2. 翻訳者の求める情報

例えば，電気工学分野の用語 “suspension insulator” を翻訳する場合を考える．このとき，翻訳者はまず，この用語が既存の対訳辞書に載っているかを調べる．もし，載っていなければ，例えば，suspension と insulator の訳語を辞書で調べ，両者の訳語を合成して訳語の候補を作る．そして，訳語の候補が正しいかを，ウェブを用いて調べる．調べ方としては，まず，訳語候補が実際にウェブの文書で使われている例があるかを調べる．実際に使われているようであれば，次は，訳語候補が現れる文書を見て，確かに電気工学分野の文書で使われているかを調べる．そして，その一方で,「“suspension insulator” の訳語は A である」と明示的に書かれているページを探すなど，多岐にわたる．

このため，本論文では，翻訳者の負担を低減させるため，翻訳者が訳を見つける過程を支援する．具体的には，

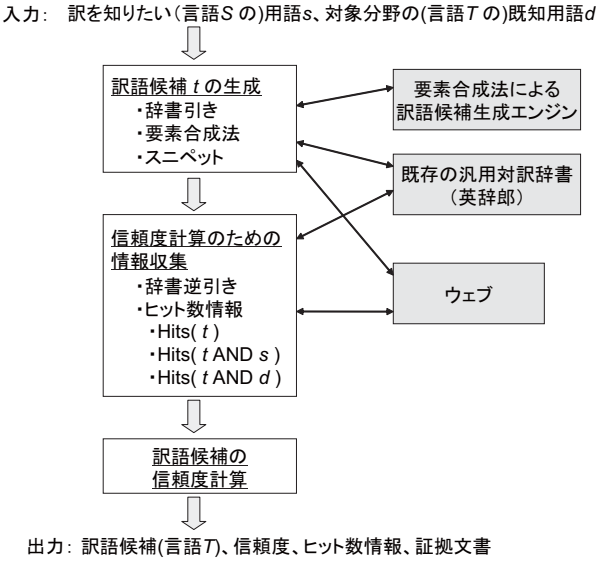

図 1 専門用語翻訳支援の流れ

訳語候補を生成または収集する部分と，訳語候補が正しいかを判断するための情報を収集する部分をシステムが代行し，訳語候補とそれに関する情報を，訳語候補の信頼性の度合いと共にユーザーに提示することを目指す．

3. 専門用語翻訳支援システムの設計

本論文で提案する専門用語翻訳支援システムの構成を説明する．以下では，言語 S の用語を，言語 T に翻訳する場合を考える．専門用語翻訳支援システムの構成を図 1 に示す．まず，入力として，訳を知りたい (言語 S の) 用語 s と，その用語が属する分野の (言語 T の) 既知用語 d が与えられることを仮定する．d は，s の訳語候補 t が対象分野の用語としてふさわしいかを評価するのに利用する．ある 1 つの用語の翻訳をしたいときは，その用語の属する分野の言語 T の用語を用意すればよい．一方，ある技術文書全体を翻訳したいときは，技術文書に現れる訳語既知の用語を利用すればよい [1]．出力としては，訳語候補と共に，信頼度，ヒット数情報，訳語候補が実際に現れる文書等を返す．詳細は，4 章で説明する．

システムは，大きく 3 つのステップからなる．最初のステップでは，訳語候補を生成する．次のステップでは，生成された訳語候補が正しいかを評価するための情報を収集する．そして，最後のステップでは，収集した情報に基づいて，訳語候補の信頼度を計算する．以下では，そ

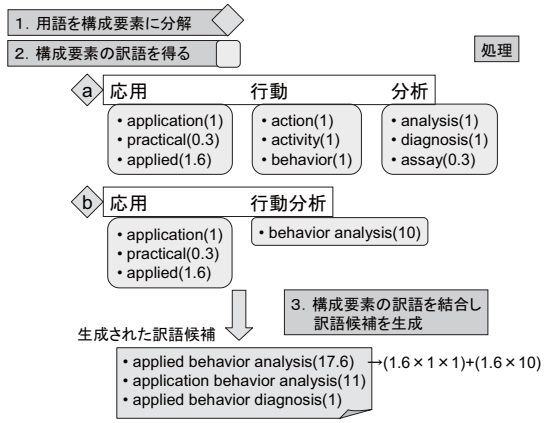

図 2 要素合成法による訳語候補生成

れぞれのステップについて説明する．

3.1 訳語候補の生成

このステップでは，３つの方法で訳を知りたい用語の訳語候補を生成または収集する．

3.1.1 辞 書 引 き

既存の汎用対訳辞書☆に訳を知りたい用語が載っているか調べ，載っていれば，その訳語を訳語候補とする．

3.1.2 要素合成法

要素合成法とは，訳を知りたい用語を構成する単語・形態素の訳語を既存の対訳辞書から求め，これらを結合することにより訳語候補を生成する [2] 方法である．

要素合成法による専門用語翻訳の例として，日本語の専門用語「応用行動分析」の訳語推定の様子を図 2 に示す．

3.1.3 スニペットからの訳語候補収集

要素合成法では，言語 S と言語 T で構成要素の対応が取れないような用語を翻訳することができない．そこで，正しい訳語が出現する可能性のあるウェブページを収集し，そこから訳語候補を抽出することを考える．正しい訳語が現れるウェブページとしては，次の２つの条件が考えられる．

条件 1 翻訳対象の用語を含む

条件 2 翻訳対象の用語の構成要素の訳語を含む

本論文では，ウェブページを収集するのではなく，サーチエンジンの検索結果のスニペットを利用するアプローチを取る [3]．

条件 1 を利用する場合，翻訳対象の用語をクエリーとして，上位 50 件のスニペットを収集する．次に，スニペットから訳語候補を抽出する．抽出するのは，名詞，形容詞，接尾辞，カタカナ語の並びの最長の形態素列である．ただし，これらをすべて抽出したのでは，翻訳対象の用語と無関係な複合語が数多く訳語候補となってしまうので，以下の２つの条件で絞り込みを行う．１つ目の条件は，翻訳対象の用語の構成要素の訳語を１つ以上含むことである．２つ目の条件は，訳語候補が翻訳対象の用語と以下の例のようなパターンにマッチする形式で出現することである．

- 懸垂碍子 (suspension insulator)
- suspension insulator (懸垂碍子)

なお，訳語候補を収集する際には，訳語候補がスニペットに出現した頻度も記録しておく．

一方，条件 2 を利用する場合，スニペットを収集するためのクエリーは，翻訳対象の用語の構成要素の訳語と対象分野の言語 T の用語とする．こちらの場合も，同様に，翻訳対象の用語の構成要素の訳語を 1 つ以上含む複合語を抽出する．

3.2 訳語候補に関する情報収集

このステップでは，収集または生成された訳語候補がどの程度信頼できるのかを評価するための情報収集を行う．ここで収集する情報は，訳語候補を辞書引きして得られる情報と，訳語候補をサーチエンジンで検索して得られるヒット数に関する情報の２種類に大別される．

3.2.1 訳語候補による対訳辞書逆引き情報

得られた訳語候補で対訳辞書を引いたとき，訳語候補は辞書に載っているが，その訳語として，翻訳対象の用語は載っていない場合を考える．このような場合，その訳語候補は誤りである可能性が高いと判断できる．例えば，コンピュータ分野の用語 “cooperative agent”(正解訳語は「協調エージェント」) の訳語候補として,「共同代理店」が得られたとき，これを辞書引きすると “participating agency” という用語のみが得られる．この場合，訳語候補「共同代理店」は正しくない可能性が高いと判断できる．同様に，コンピュータ分野の用語 “communication statement”(正解訳語は「通信命令」の訳語候補として,「通信メッセージ」が得られたとき，これを辞書引きすると “communication message” という用語のみが得られる．この場合，訳語候補「通信メッセージ」は正しくない可能性が高いと判断できる．

3.2.2 訳語候補のヒット数に関する情報

訳語候補の単独ヒット数

訳語候補が正しいとき，その訳語候補は，ウェブ上の文書で使用されているはずである．これを確かめるため，訳語候補をサーチエンジン☆☆で検索して得られるヒット数を調べる．ここで，ヒット数がゼロの訳語候補は，正しくないと判断できる．この情報は，要素合成法によって生成された訳語候補が用語として正しいかを検証するときに有効である．

訳語候補と翻訳対象用語のアンド検索ヒット数

翻訳対象用語 s と訳語候補 t が共に現れる文書が存在するとき，翻訳対象用語 s の訳語として，訳語候補 t が利用されていることが期待できる．例えば，論文のタイ

☆ 本論文では，既存の汎用対訳辞書として，英辞郎 (http://www.eijiro.jp/) を用いる

☆☆ 本論文ではサーチエンジンとして Yahoo! Japan (http://www.yahoo.co.jp) を用いる．

トルやアブストラクトに s が現れ，それらの英訳の中に t が現れる場合などが考えられる．このような文書が多数存在する場合，t は正しい可能性が高いと判断できる．

訳語候補と翻訳対象用語のアンド検索で得られるスニペットに出現する頻度

訳語候補と翻訳対象用語のアンド検索ヒット数を調べるだけでは，訳語候補と翻訳対象用語が遠く離れて出現している場合もカウントしてしまう．そこで，訳語候補と翻訳対象用語が近接しているウェブページの数を調べる．実際には，訳語候補と翻訳対象用語のアンド検索で得られる上位 50 件のスニペットに対して，訳語候補と翻訳対象用語が共起しているスニペット数を調べる．

訳語候補と対象分野の用語のアンド検索ヒット数

分野 D の用語 s の訳語候補を t とすると，t が出現する分野 D の文書が数多くウェブ上に存在すれば，t は正しいと考えられる．ここで，対象分野 D の言語 T の用語 d が与えられたとき，ウェブ上に存在する文書のうち，d を含む文書は，対象分野 D の文書であると仮定する．この仮定に基づくと，訳語候補 t が出現する対象分野 D のウェブページ数を数えるためには，サーチエンジンを利用して t と d のアンド検索のヒット数を求めればよい．

3.3 訳語候補の信頼度計算

このステップでは，3.2 節で収集した情報をもとに，訳語候補がどの程度信頼できるのかを評価する．信頼度は，信頼度の高い順に，★3 個，★2 個，★1 個，記号なし，×の 5 段階で付与する．翻訳対象用語を s，訳語候補を t，対象分野の言語 T の用語を d，$Hits(x)$ を x のヒット数とし，訳語候補の信頼度を決めるルールを以下に示す．

- ★★★
  - 要素合成法により s から t が生成
    AND $Hits(s\ AND\ t) \geq 100$
    AND $Hits(d\ AND\ t) \geq 20$
  - 要素合成法により s から t が生成
    AND $Hits(s\ AND\ t) \geq 10$
    AND $Hits(d\ AND\ t) \geq 20$
    AND s が 2 構成要素以上で，s の訳語として対訳辞書に t が存在
- ★★
  - 要素合成法により s から t が生成
    AND $Hits(s\ AND\ t) \geq 10$
- ★
  - スニペット中に括弧を用いた s と t の併記 (“s (t)” または “t (s)”) が存在
  - s が 2 構成要素以上で，s の訳語として対訳辞書に t が存在
  - 要素合成法により s から t が生成 AND $Hits(s\ AND\ t) \geq 1$
- ×
  - t は対訳辞書に存在するが，その訳語として s が載っていない
  - $Hits(t) = 0$

上記のルールは，十数例のサンプルを分析することにより，試行的に作成した．数百から数千個のサンプルに対して学習の技術を適用し，信頼度判定のルールを学習することが次の課題となる．

4. インターフェース

専門用語翻訳支援システムの情報提示例を図 3 及び図 4 に示す．訳語候補は，生成方法（対訳辞書，スニペット，要素合成法）別に整理して提示される☆．それぞれの訳語候補に対しては，信頼度 (Grade)，辞書に存在するか (Dic)，訳語候補の辞書引き結果 (Back Translation)，要素合成法の辞書スコア (Dic Score)，訳語候補のヒット数 (Hits)，収集したスニペットに訳語候補が現れる回数 (SF)，訳語候補と翻訳対象の用語のアンド検索で得られたスニペットに訳語候補と翻訳対象の用語が共起する回数 (And_SF)，訳語候補と翻訳対象用語のアンド検索ヒット数 (And_Hits(“noise analysis”, Can))，訳語候補と対象分野の言語 T の用語のアンド検索ヒット数 (And_Hits(“電圧”, Can))，要素合成法の順位 (TR)，要素合成法でどのような要素が合成されたか (Alignment) の情報☆☆を提示する．また，スニペットから収集した訳語候補に関しては，実際に訳語候補が現れたスニペットを提示する．

図 3 は，電気工学分野の用語 “noise analysis” の翻訳支援の様子を示している．この “noise analysis” という用語の訳語は，電気工学分野では「雑音解析」であるが，交通工学分野では「騒音解析」となる．図 3 の場合，訳語候補「雑音解析」に★が 3 個ついている．一方，図 4 は，交通工学分野の “noise analysis” の翻訳支援の様子を示している．この場合，訳語候補「雑音解析」の★の数は 2 個に減少している．これは，交通工学分野の用語「公害」と「雑音解析」のアンド検索のヒット数が 20 未満であったためである．

5. むすび

本論文では，当該分野における訳語既知専門用語を利用して，ウェブ検索エンジンを用いて訳語候補の信頼度を推定したり，あるいは，専門文書中において，訳語と原語の併記を抽出する等，数種類の情報とその信頼性の度合いを提示することにより，専門用語翻訳支援の過程を実現した．

今後の課題としては，実際に人間にシステムを利用してもらい，作業がどの程度効率的になったかを調べることと，機械学習により，信頼度を決めるルールを学習す

☆ 翻訳対象の用語をクエリーとしたときのスニペットを利用する方法は，訳語候補の絞り込み方法別に表示される．図 3，図 4 では省略．また，翻訳対象の用語の構成要素の訳語と対象分野の用語をクエリーとしたときのスニペットを利用する方法の結果は，オプションで指定すれば表示される．

☆☆ 図 3，図 4 では Alignment の欄を省略．

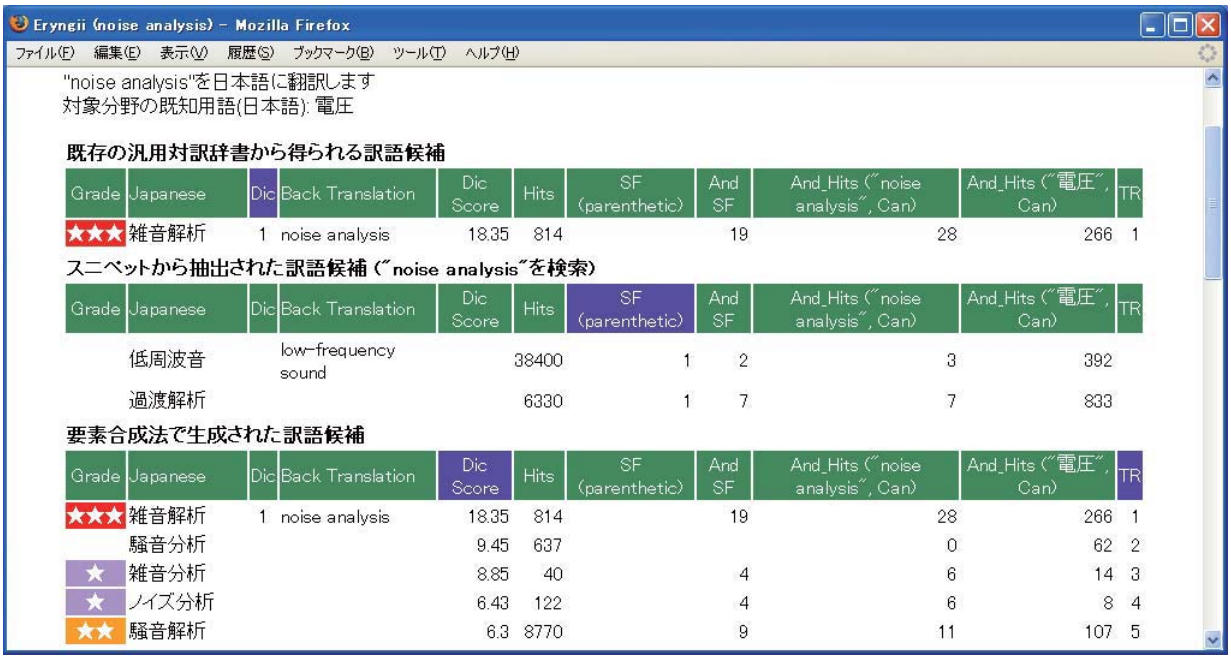

図 3 情報提示例 (翻訳対象用語：“noise analysis”，対象分野：“電気工学”，対象分野の日本語用語：“電圧”)

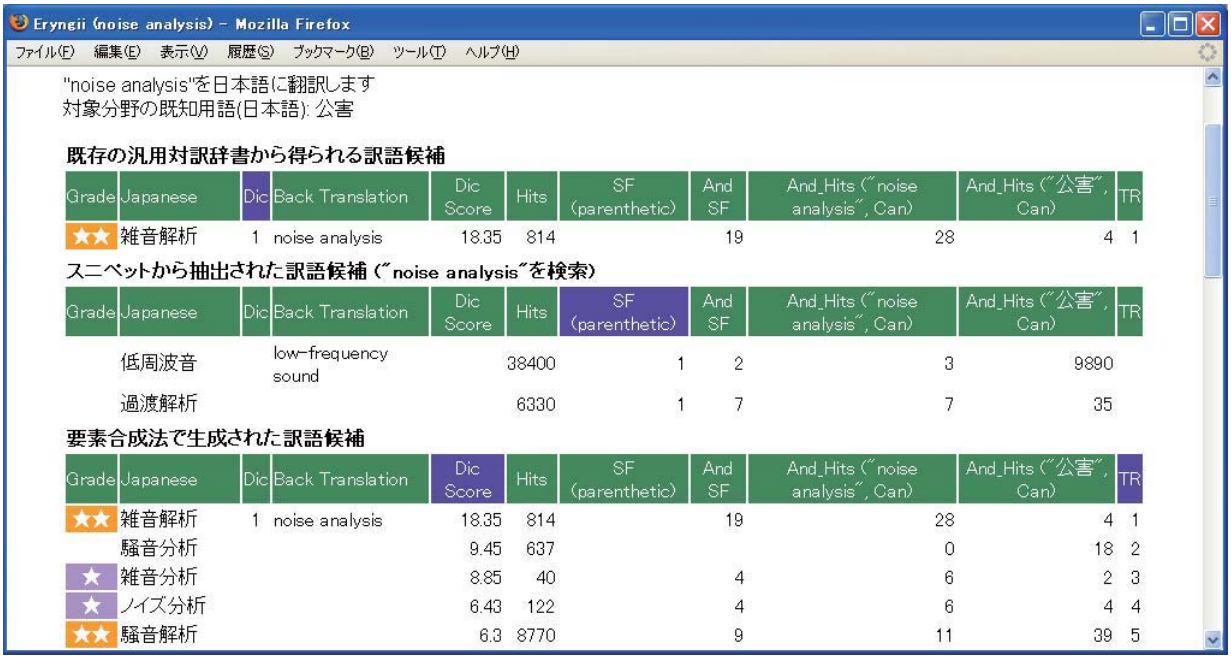

図 4 情報提示例 (翻訳対象用語：“noise analysis”，対象分野：“交通工学”，対象分野の日本語用語：“公害”)

るすることと，椎茸プロジェクト [4] で開発されている翻訳支援ツールにこの専門用語翻訳支援システムを組み込むことが挙げられる．

参 考 文 献

1) 馬場康夫, 外池昌嗣, 宇津呂武仁, 佐藤理史: 対訳辞書とウェブを利用した専門文書中の用語の訳語推定, 言語処理学会第 12 回年次大会論文集, pp. 416–419 (2006).
2) 外池昌嗣, 宇津呂武仁, 佐藤理史: ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定, 自然言語処理, Vol. 14, No. 2 (2007).
3) Huang, F., Zhang, Y. and Vogel, S.: Mining Key Phrase Translations from Web Corpora, *Proc. HLT/EMNLP*, pp. 483–490 (2005).
4) Bey, Y., Boitet, C. and Kageura, K.: The TRANSBey Prototype: an Online Collaborative Wiki-based CAT Environment for Volunteer Translators, *Proc. LREC-2006 LR4Trans-III*, pp. 49–54 (2006).