個人適応による英日翻訳での訳語候補の順位付け

青木 優　　山本 和英
長岡技術科学大学 電気系
E-mail:{aoki, ykaz}@nlp.nagaokaut.ac.jp

1. はじめに

近年、情報化社会の発達により我々はあらゆる情報を容易に入手することが可能となった。ユーザには複数の選択肢が用意され、個人の興味や知識から必要な情報を判断しなければならないこともある。このような場面では、個人の興味や知識などを学習する個人適応システムが有効であり、その手間を軽減することができる。

英日翻訳における訳語選択を考えた場合、多義性の問題があり、複数個の訳語候補が提示されることが多い。文脈や経験から人手で訳語選択する作業には手間がかかる。本稿では、英日翻訳を対象とした、ユーザプロファイルを用いた訳語候補の順位付け手法を提案する。本稿におけるユーザプロファイルとは、ユーザの専門分野情報や、よく使用する単語などを蓄積したデータである。このようなユーザプロファイルを用いることによりシステムはユーザの情報を学習し、英日翻訳での訳語選択の際、ユーザに最も適した訳語を提示することができる。

2. 関連研究

個人適応の研究としては、野美山ら[1]が個人適応型情報検索システムの作成を行っている。これはあらかじめユーザの興味のある文書集合に含まれるキーワードからユーザ観点の文書ベクトルを作成し、入力された文書記事のベクトルと比較する。二つのベクトルの類似度が高い程、ユーザの興味に合った情報であり、ランキングの上位にくる。また、長谷川ら[2]は電子メールの個人適応ランキングを行っている。メールの送受信履歴から、差出人、対象者、タイプ、期限などの情報を学習し、メールの重要度を計算することで電子メールをランキングする。これらの手法では、興味のある文書や電子メールの送受信履歴などからユーザの個人性を学習しているが、個人性の更新は行っていない。本手法では、訳語候補を順位付けして提示し、さらに選択された訳語をフィードバックさせることで、ユーザプロファイルの更新を行っている。

3. 提案手法

本稿では、訳語候補の中から個人に適した訳語選択をするために必要なデータの総称をユーザプロファイルと定義する。具体的には、ユーザの専門分野、よく使用する単語、過去に選ばれた訳語履歴、選ばれた訳語と共起する単語の蓄積データをユーザプロファイルと呼ぶことにする。

手法では、このユーザプロファイルを利用して訳語候補のスコアを求める。このスコアとは訳語候補の尤もらしさを表す指標である。このスコアが高いほどユーザに適した訳語候補であることを示す。このスコアにより訳語候補を順位付けし、ユーザに提示する。

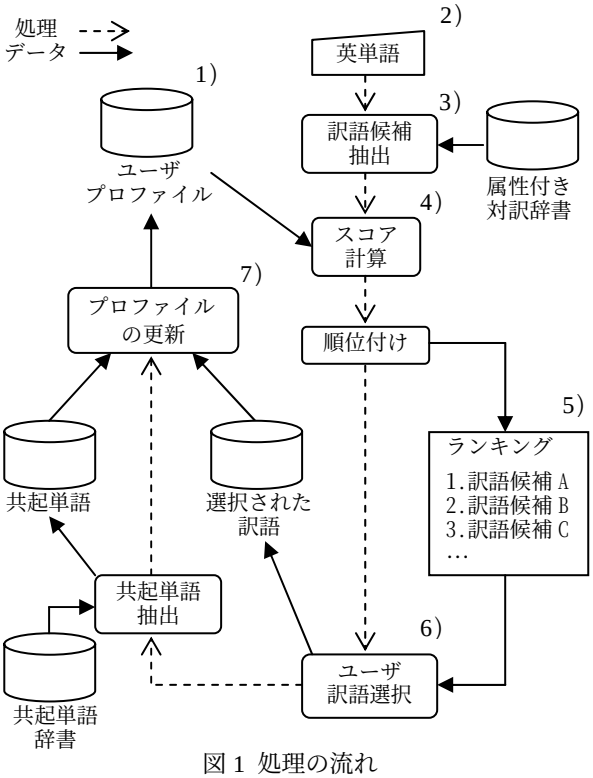

図 1 処理の流れ

1) 個人の特徴が現れやすい文書を入力し、ユーザプロファイルを作成。
2) ユーザが訳語の知りたい英単語を入力。
3) 対訳辞書から訳語候補をリストアップ。
4) プロファイルを参照し、訳語候補のスコアを計算。
5) 訳語候補を順位付けし出力。
6) ユーザが尤もらしい訳語を選択。
7) 選択された訳語等のプロファイルを更新。
8) 2)～7)の操作を繰り返す。

2)～7)の操作を繰り返すことによりユーザプロファイルが更新され、訳語候補のスコアに反映される。その結果、訳語候補のランキングは個人適応されていく。

3.1 辞書の作成

3.1.1 属性付き対訳辞書の作成

英日翻訳を行うためには対訳辞書が必要となるが、通常の英日対訳辞書ではプロファイルを作成するためには情報が乏しい。そこで、英日の言語対の他に、分野情報を属性として付与した。属性付き対訳辞書は例 1 のような英語、日本語、属性値の三つ組とした。

例 1. [computer：計算機：電気・電子]

対訳辞書の作成にはクロスランゲージ専門語辞書を利用した。クロスランゲージ専門語辞書は見出し語を入力

すると、英訳(または和訳)だけではなく、訳語の分野情報も同時に出力される。これを利用して検索結果の Web ページから英訳、和訳、属性値の三つ組を収集し属性付き対訳辞書を作成した。属性値は、クロスランゲージ専門語辞書の分野（基本語、電気・電子、コンピュータ、機械工学、化学用語、医療医学など）21 種類とした。属性を付与することで訳語選択の際にユーザが選択した訳語の属性値がプロファイルに蓄積され、以後の訳語候補のスコア算出の際に反映される。

3.1.2 共起単語辞書の作成

共起単語辞書はコーパス内の一文中で共起する二語の共起情報を用いて作成した。コーパスには毎日新聞 2000 年版(2)を用いた。

3.2 ユーザプロファイルの作成

本稿で用いるユーザプロファイルは大きく二つに分けられる。一つは個人の特徴の現れやすい文書を用いてあらかじめ作成するもので、多用単語プロファイルと分野情報プロファイルである。もう一つはユーザーがシステムを使う度に更新していくもので、訳語履歴プロファイルと共起単語プロファイルである。

3.2.1 多用単語プロファイル

同じ単語の訳語であっても、ユーザの使用頻度の高い訳語が多く用いられると考える。そこでユーザが多用する単語をあらかじめプロファイルに蓄積する。そのために Blog などの個人の特徴の現れやすい文書を用いてプロファイルの作成を行う。

3.2.2 分野情報プロファイル

3.2.1 節で得られた単語の頻度情報から、対訳辞書を用いて単語を属性情報に変換し、その属性値の頻度を分野情報プロファイルとして蓄積する。全ての訳語には分野を示す属性がついているため、その属性値の情報を用いてユーザの分野情報を取得できる。

例えば、ユーザの分野情報で「電気・電子」の出現頻度が高ければ、訳語候補の中で属性値が「電気・電子」の訳語候補はスコアを上げる。

3.2.3 訳語履歴プロファイル

システムが提示した訳語候補の中から、ユーザが選択した訳語を訳語履歴プロファイルに蓄積する。以前に選択された訳語が訳語候補中に存在する場合、スコアを上げる。

3.2.4 共起単語プロファイル

選択された訳語と共起する単語を共起単語辞書より抽出し、その共起単語の頻度をプロファイルに蓄積する。

共起単語は同じような状況下で使われやすい。例えば以前”suspicion”という単語を「容疑」と訳したとする。その後、”arrest”という単語を訳す際に「逮捕する」と「阻止する」でどちらが選択されやすいか考えたとき、以前訳した「容疑」と共起しやすい単語である「逮捕する」の方が選択されやすいと考える。プロファイルで更新される情報が、分野情報、訳語履歴だけでは、以降の訳語候補スコアに反映される情報として不十分である。分野情報は、ユーザがどの分野に属するかという指標の一つであり、つまりグループ適応に近く、個人適応のスコアとしては適切とは言えない。グループ適応とは、ユーザがどのグループに属するかというもので、特定の個人を指すものではない。また、訳語履歴プロファイルは過去に選ばれた訳語の中に訳語候補がなければスコアは変化しないという問題がある。そのため、この共起単語情報を用いることで、ユーザプロファイルを効率的に更新し、より個人適応されたスコアを求めることができる。

3.3 訳語候補スコアの計算

以上のプロファイルを用いて、訳語候補のスコアを決定する。訳語候補スコアは各プロファイルより求めるスコアと初期値の合計である。このスコアに基づき訳語候補を順位付けする。

3.3.1 初期値の設定

プロファイルが集まっていない初期段階ではスコアにあまり差が見られず、順位付けが難しい。また、訳語候補中には使用頻度が低い単語が存在することもある。その例を例 2 に示す。初期値を設定することで、使用頻度の低い訳語候補の順位を下げることができる。

例 2. break：開墾する、船楼端、裂谷

初期値はコーパスから抽出した単語単位での頻度情報を用いて(1)式より計算する。コーパスには毎日新聞 2000 年版(2)を用いた。

$$S_F(w_i) = \frac{f(w_i)}{F} \quad (1)$$

$S_F(w_i)$:訳語候補 w_i の初期値

$f(w_i)$:訳語候補 w_i のコーパス中での出現頻度

F :コーパス中の全単語頻度の合計

3.3.2 訳語候補スコアの計算

まず、各プロファイルより求まるスコアを計算する。

各プロファイルには、単語とその頻度（分野情報プロファイルでは単語の代わりに属性値とその頻度）が蓄積されている。 この情報を用いて(2)式より、各プロファイルのスコアを求める。

$$S_P(n, w_i) = \frac{p(n, w_i)}{P(n)} \quad (2)$$

$S_P(n, w_i)$:プロファイル n のスコア
$p(n, w_i)$:訳語候補 w_i (もしくはその属性値)のプロファイル n での出現頻度
$P(n)$:プロファイル n の全単語頻度の合計
n : {多用単語,分野情報,訳語履歴,共起単語}

以下に分野情報プロファイルのスコアの計算例を示す。

訳語候補

Circuit / 回路/ 電気・電子
Circuit / 回線/ コンピュータ
Circuit / 巡回/ 基本語

分野情報プロファイル

基本語	=48
電気電子	=29
数学	=10
機械工学	=9
コンピュータ	=4

図 2:訳語候補及び分野プロファイルの例

図 2 のような分野プロファイルの場合、下線部の訳語候補「回路」の分野情報プロファイルのスコアは次のように求められる。

$$S_P(\text{分野情報}, \text{回路}) = \frac{29}{48+29+10+9+4} = \frac{29}{100} = 0.29$$

多用単語、訳語履歴、共起単語についても同様にしてスコアを求める。

次に、各プロファイルのスコアと初期値を用いて、(3)式より訳語候補の最終的なスコアを求める。

$$S_{total}(w_i) = S_F(w_i) + \sum_n \left(S_P(n, w_i) \times \lambda(n) \right) \quad (3)$$

$S_{total}(w_i)$:訳語候補 w_i のスコア
$\lambda(n)$:プロファイル n のスコアの重み

この $S_{total}(w_i)$ に基づき訳語候補を順位付けする。$S_{total}(w_i)$ が高い訳語候補がランキング上位に位置し、ユーザに適した訳語である確率が高い。

3.3.3 プロファイルへの重みづけ

本手法で重要となるプロファイルにはそれぞれ特徴があり、各プロファイルによって重要度が異なってくる。そのため、各プロファイルのスコアには重み付けが必要となる。重みの初期値を 1 として、各プロファイルのスコアに重み付けをする。

多用単語スコアはユーザの特徴が最も表れると考え、重みを一番高くする。分野情報はユーザの特徴を表しているが、個人適応というよりはグループ適応に近いため分野情報スコアは初期値よりは高くし、多用単語よりは低くする。訳語履歴スコアもユーザの特徴が現れるが、同じ訳語が何度も選ばれるとは限らない。よって、分野情報と同様の重要度と考える。共起単語情報には毎日新聞コーパスの情報を用いているため、個人適用としての重要度は低い。

これらのことを考慮して各重みは、多用単語スコアの重みは 3、分野情報スコアと訳語履歴スコアの重みは 2、共起単語情報の重みは 1 とした。

さらに、属性付き対訳辞書の 3 分の 2 の属性値を基本語が占めており、割合が圧倒的に多い。そこで訳語の属性値が基本語の場合に限り、分野スコアを低くした。本稿では 20 分の 1 とした。 これより、ランキング上位が基本語に偏ることを防ぐ。

3.4 訳語選択とプロファイルの更新

出力された訳語候補ランキングの中から、ユーザは尤もらしい訳語を選択する。訳語選択により更新されるプロファイルは以下に示す三つである。

・訳語履歴：選択された訳語の頻度を＋1。
・分野情報：選択された訳語の属性値の頻度を＋1。
・共起単語：選択された訳語と共起する単語の頻度を加算。

4. 評価実験

実験方法は以下の通りである。個人の特徴が現れやすい文書として被験者の Blog40 記事を用いた。 被験者は学生 1 名である。

Step.1　被験者の Blog を入力し、多用単語、分野情報のユーザプロファイルを作成する。

Step.2　学習用データとして比較的簡単な英単語 100 語を入力する。被験者は出力された訳語候補の中から尤も訳語として適した一語を選択し、選択された訳語をシステムに学習させる。このとき、選択された訳語の順位を保存しておき、ランキング推移を評価する。

Step.3　評価用データとして属性付き対訳辞書からランダムで選んだ英単語 10 語を入力する。 学習によってユーザ適応した結果、ユーザが選択した訳語がどれだけ上位にきているかを評価する。

4.1 ランキング推移

被験者は英単語 100 語を入力し、出力されたランキングから訳語を選択する。学習の結果、図 3 に示すような結果が得られた。選択順位/候補数が低いほどランキングで上位の訳語候補が選択されたことを示す。

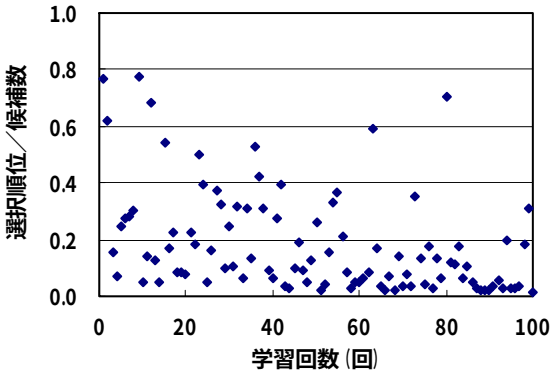

図 3:ランキング推移

図 3 より、学習回数が増加するにつれて、選択された訳語の順位が上位である割合が高くなっていることが読み取れる。

4.2 ランキングの評価

被験者は評価用英単語 10 個を入力し、出力されたランキングから訳語を選択する。その結果、選択された訳語の平均順位は 6 位と、良い結果は得られなかった。

ランキング結果の一例を挙げると、”Table”の訳語が一般的な「食卓」や「テーブル」ではなく、「表」が上位にランキングされ、ユーザは「表」を選択している。プロファイルの分野情報や使用頻度の高い単語のスコアが効いている例である。

次に、和訳すると複数の日本語表現になる単語として teacher (先生、教師、教員、担任)が挙げられるが、このような単語の場合でも、多用単語プロファイルにより、ユーザの使用頻度の高い訳語が上位にくることがわかった。

5. 考察

評価実験では Blog を入力し、初期でのユーザプロファイルを作成したが、実験結果からわかるように学習無しの状態ではランキングの精度が非常に低い。しかし、選択された訳語からユーザプロファイルを更新し、システムに学習させることで、ユーザが選択すると思われる訳語候補がランキング上位に現れる傾向が見られた。英日翻訳における個人適応が有効に働いていると言える。分野情報、訳語履歴プロファイルのみの学習ではランキングにあまり変動が見られなかったが、共起単語プロファイルを加えたことにより学習時のスコアの変化が大きくなり、ランキングに良い影響を与えた。共起単語プロファイルが有効に働いていたと言える。

一方、ランキングの評価では平均順位 6 位と、良い結果は得られなかった。この原因としては評価用の英単語が少なすぎたことが考えられる。複数の被験者がシステムを評価し、検討する必要がある。

また、次のような問題点も挙げられる。

訳語候補の中に一般的に使用頻度が高い訳語候補があると、初期値、多用単語スコアが高くなり、ランキング上位にきてしまう問題がある。例 3 では文書中で頻出する「ない」という単語が”lack”の訳語候補として上位にランキングされている。

例 3. lack：欠ける、不足、欠如、ない

同じ意味の訳語候補でも表記揺れがあったり、似たような意味の訳語が複数出力されることがある。これにより、ランキング表示での訳語候補の数が多くなり、ユーザがランキングから訳語選択を行うことが困難になる問題がある。この例を例 4 に示す。例 4 では、表記的にも意味的にも大差はなく、選択が困難である。このような表記揺れの場合での対応を検討する必要がある。

例 4. break：こわす、こわれる、壊す、壊れる

本手法は文脈などを一切考慮していない。そのため、多義語の場合、訳語候補の分野やユーザが多用する日本語表記である訳語候補がランキングの上位にきてしまう問題がある。この例を例 5 に示す。例 5 では、「岸辺」「堤防」の訳語選択を考えた場合、本手法での個人適応は効果的であるが、「銀行」「岸辺」の訳語選択を考えた場合、必ずしもユーザにとって適切な訳語が選ばれるとは限らない。しかし、このような文脈等の考慮は本手法では対象としていない問題である。

例 5. bank：銀行、岸辺、堤防、頼りにする

6. おわりに

ユーザプロファイルを用いることにより、英日翻訳での訳語候補を順位付けする手法を提案した。本手法では個人を特徴付ける文書より、使用頻度の高い単語や分野など情報を収集し、そのプロファイルを用いて各訳語候補のスコアを計算、ランキングで提示した。評価実験の結果、システムを繰り返し使用することで、ユーザ適応された訳語候補が上位に現れる傾向が見られた。

課題として、ユーザプロファイルの収集方法の検討が挙げられる。 そもそも Blog のような個人の特徴が現れやすい文書を入力するという仮定で行っているが、そのような文書をどうやって入手するかという問題がある。初期状態での効果的なユーザプロファイルを如何に作成するか、また、如何に効率よく学習していくかが今後の検討課題である。

使用した言語資源及びツール

(1) クロスランゲージ専門語辞書
http://reg.crosslanguage.co.jp/webdic/webdic.html

(2) 毎日新聞全記事データベース 2000 年版,毎日新聞社

(3) 形態素解析器「茶筌」Ver.2.3.3,奈良先端科学技術大学院大学 松本研究室
http://chasen.naist.jp/hiki/ChaSen/

参考文献

[1] 野美山浩、紺谷精一、渡辺日出雄、串間和彦、堤秦治郎：個人適応型情報検索システム、情報処理学会研究報告 FI-42-8,pp.49-56,1996.

[2] 長谷川隆明：送受信履歴と情報抽出に基づく電子メールの個人適応ランキング、情報処理学会研究報告 NL-132-3,pp.17-24,1999.

[3] 永井野亮、佐川雄二、杉江昇：知的電子化英和辞書前置詞の曖昧性解消システムの実装、情報処理学会研究報告 NL-147-2,pp.1-6,2002.