ユーザの知識レベルに応じた統一的テキスト伸縮手法

荻野 紫穂* 渡辺 日出雄*
shiho@jp.ibm.com hiwat@jp.ibm.com

1 始めに

近年、様々なデバイスおよび様々な通信環境から、多様なプロフィールを持ったユーザによる情報へのアクセスが行われるようになってきている。このような状況で、情報提供技術には、種々の観点からユーザに最適な形式の情報を提供することが求められている。この分野において、自然言語処理技術が果たすことができる役割は大きなものがある。特に、要約・言い換えの技術分野が直接的に貢献できるものであり、この分野では、様々な方式選択がなされている [6]。

要約の分野では、[2] を基盤として、各種手法が加味されてきた [8][4]。情報抽出においては、主に [7] などに見られるように、特定の固有名詞・日時などの表現や、自然言語データセットからの正解部分抽出などを中心に研究が行われている。これらの出力に関するユーザ毎のカスタマイズについては、[9] は大きな概念を示し、[5] は、既知情報と未知情報の差異を使用して、出力をカスタマイズすると共に文章の結束性を高める方法を提案している。また、[3] は、表現置き換えをいくつかの評価基準を定め、それぞれの評価基準を変えて適用するものと定義している。

本稿では、単なる要約でもなく、理解を助ける言い換えでもなく、ユーザの対象分野の習熟レベルに応じたテキストの書き換え手法（今後本論文ではこれをテキスト伸縮と呼ぶことにする）について提案する。たとえば、

「J1 リーグの鹿島アントラーズの相馬が、...。」

という文章を、サッカー分野の知識が豊富なユーザには、

「J1 鹿島の相馬が、...。」

という文に書き換え提示し、一方、サッカー分野に関してほとんど知識のないユーザには、

「日本のプロサッカーの最高次リーグである J1 のクラブチーム鹿島アントラーズに所属する相馬選手が、...。」

という文を提示するということを目指している。これだけだと、習熟レベルに応じてテキスト量を増減させているだけに見えるが、これと要約技術を組み合わせることで、要約と言い換えの効果が混在した出力を得ようとするものである。

*日本 IBM(株) 東京基礎研究所

2 システム構成

本システムは、分野階層と個々の分野に付随する書き換え辞書、ユーザの習熟度管理モジュール、およびそれらのデータを用いて書き換えを行うテキスト伸縮エンジンからなる。書き換え辞書は、内容書き換え辞書、注釈辞書、長さ変換辞書から構成される。図 1 に構成を示す。

2.1 分野階層とユーザの習熟度

本システムは、シソーラスのような分野間の階層構造を仮定する。各分野ノードには、書き換え辞書が付属している。図 2 に分野階層の例を示す。

ユーザの習熟レベルは、各分野ノード毎に保持・管理される。さらに、個々の辞書エントリ毎に設定も可能である。

ある分野ノードで、あるユーザの習熟レベルが指定されていない場合、祖先分野ノードへ遡ってそのユーザの習熟度が指定されているものを探しそれを用いる。分野階層のルートノードにはデフォールトで全てのユーザに中間の習熟レベル（習熟レベル普通）が付いているものとする。

2.2 辞書

システム辞書は、内容置換辞書、注釈辞書、長さ置換辞書を含む。これら 3 つの辞書は相互に関連しているので、メンテナンスは相互を参照の元に行う。辞書はどれも、各エントリにつき、ユーザの習熟レベルとそれに対応する表記要素を一個以上含む。システム辞書は、原則的にシソーラス状の構造を持つものとする。比較

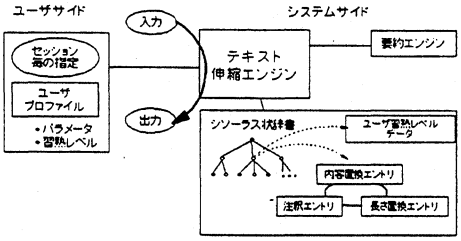

図 1: システム構成図

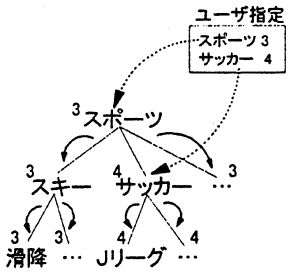

図 2: 習熟レベルの伝播と保護

的単純な実装の場合、辞書はそれぞれの対照分野毎に「サッカー」「経済」などの専門用語辞書を持ち（純粋なシソーラス的に考えて必ずしもレベルが揃っている必要はない）、その専門用語辞書にエントリが含まれることになる。

内容置換辞書

内容置換辞書は、例えば、「形態素解析」について、

表記要素	習熟レベル
形態素解析	5
↑↓	
単語分割のための解析	4
↓↑	
単語への分割と単語の分類の付与	1

のようなエントリを持つ。この場合、ユーザの習熟レベルは1-5の5段階であり、段階数が増えるごとに習熟レベルも高いものとする。各表記要素には、ユーザ習熟レベルとそれに対応した置き換えテキストが含まれる。置き換えテキストは変数を含むパターンであってもよい。

対応置き換えテキストがない習熟レベルについては、そのレベルより下で一番近いレベルの語が代用される。この例の場合、自然言語処理の知識のあるユーザ（レベル5）ならば、何の説明もなく「形態素解析」という用語を使用しても意味が分かるが、知識のないユーザ（レベル1-2など）にとっては、そうした用語は意味をなさないので、そのようなユーザが対象になる場合は、意味の分かる語「単語への分割と単語の分類の付与」などを選択する。

また、表記要素間の矢印は、相互に置換可能かどうかを指し、矢印の方向のある向きにだけ、書き換えが行われる。例えば、

表記要素	習熟レベル
CD	5
↑	
キャッシュディスペンサー	1

のようなエントリがある場合、、「キャッシュディスペンサー」という出現はレベル5のユーザに対して「CD」に書き換えられるが、「CD」という出現は、レベル1のユーザに対しても、「キャッシュディスペンサー」に書き換えられない。このように、この矢印は、書き換え元の表現が多重に書き換え可能な際に、不用意な書き換えを防ぐ役割を果たす。

エントリのキーは、置き換えのソースとなることのできる置き換えテキストである。例えば、「形態素解析」に関するエントリ例のキーは「形態素解析, 単語分割のための解析, 単語への分割と単語の分類の付与」で、「CD」に関するエントリ例のキーは「キャッシュディスペンサー」である。もしキーにマッチする語が与えられた場合は、システムはそのキーから、各表記要素の「どのレベルの表記要素に置き換えられるか」というパスを辿り、指定された習熟レベルに応じた置き換えテキストを探し出す。

表記要素が変数を含むパターンのエントリは、例えば

表記要素	習熟レベル
Xの再現率はN%である	5
↓↑	
正解のうちのN%がXにも含まれる	1

のように記述される。ここでは省略したが、Nが数字にマッチするとか、Xは名詞句とマッチするというような情報はこれらのパターンの一部として記録されている。

注釈辞書

注釈辞書は、例えば、

単語	注釈	習熟レベル
適合率	null	5
適合率	結果に含まれる正解数/全正解数 * 100	4
適合率	正解のうちの何パーセントを実際に出力することができたか	1

のように記述される。nullは、このレベルに関しては、注釈を出さないという指定を示す。例えば、上記の場合、レベル5のユーザに対して注釈は出力されないが、もしレベル5の注釈がnullでなく何も指定がなければ、一番近い下のレベルの注釈である「結果に含まれる正解数/全正解数 * 100」が代用される。nullは、こうしたある種の過置換を防ぐのに使われる。

長さ置換辞書

長さ置換辞書は、長さの微調節に使われる。例えば、「彼からの贈り物」と「彼の贈り物」は、長さが違うだけで、ほぼ同じ意味を指す[1][3]。より長い文にしたい場合は前者を、短くしたければ後者を選ぶとよい。こうしたエントリが、長さ置換の辞書に収められている。長

さ置換辞書は、例えば、

表現要素	習熟レベル
内閣総理大臣 ←→ 総理大臣 ←→ 首相	0
アメリカ合衆国 → アメリカ	0
株式会社エヌ・ティ・ティ・ドコモ ←→ NTT ドコモ ←→ ドコモ	0
XからのN → XのN	0
…	
DF ←→ ディフェンダー	5
DF → 最後列ライン	3

のように記述される。長さ置換辞書において、習熟レベルに関係ない、例えば「彼からの贈り物」を「彼の贈り物」と置き換えるようなものは、習熟レベル0とし、テキスト伸縮の際の指定習熟レベルに関わらず適用できるものとする。

辞書は、システム外の外部辞書（インターネットリソースなど）を使用してもよい。そうした外部辞書を使用した場合などで、習熟レベル付与がないエントリは、習熟レベルが中程度と見なす。

習熟レベルの競合

各分野ノードや辞書間には、ユーザ指定（指定のない場合はデフォルト）によるオーダーを想定する。習熟レベルおよび置き換えテキストの競合が起こった場合、システムはまず、このオーダーに従って選択すべきエントリを決める。オーダーを使用しても競合がある場合は、習熟レベルの高いほうのエントリが選択される[1]。

書き換えの適用条件

これらの書き換え知識は常に適用可能とは限らない。たとえば、サッカーの分野で「鹿島」を「J1リーグの鹿島アントラーズ」に置き換える場合、

(1) J1の鹿島
(2) 鹿島アントラーズ
(3) 鹿島の相馬が...

(1)(2)の例では置き換えをしないが、(3)では行うという様にしなければならない。

このため、書き換え知識を適用する場合、書き換え対象単語の近傍に書き換え後の表現を構成する単語が含まれないことが条件となる。ここで、近傍とはその単語の前後長さn文字の範囲、あるいは、その単語と係り受け関係でn閉包となる単語群、などの指標により決定する。

[1]競合の解決方法に関しては有効性の確認などが充分になされておらず、更に研究の余地がある。

2.3 ユーザの習熟レベルの微調整

ユーザの習熟レベルは、インタラクションにより変化する。ユーザは以下のインタラクションをすることができる。

- 特定の単語の習熟レベルを上げる。これは、より簡潔な記述を求める場合である。
- 特定の単語・フレーズの習熟レベルを下げる。これは、より詳細な記述を求める場合である。

これらのインタラクションにより、分野内の書き換え辞書エントリのユーザの習熟レベルが変化する。よって、個々の分野毎に、辞書エントリーのユーザの習熟レベルの平均を計算し、その分野の習熟レベルをその平均値に一番近いものにする。ただし、個別に指定されたデータに関しては、この再調整においても変更されない。

2.4 表示パラメータ

システムは以下のような表示パラメータを参照して、出力テキストの形式と長さを決定する。これらのパラメータは、セッションやプロファイルの中でユーザが指定することもできる。

- テキスト長
  語数、字数などの指定や、元テキストとの相対長、画面大での調節などにより指定できる。
- 注釈表示のパラメータ
  注釈をどう表示するかに関係する。マウスを置くとポップアップする、欄外注、用語集の表示、用語集へのリンク、丸括弧などによる文章中への埋め込みなどを指定する。

特に注釈表示パラメータで埋め込みが指定されるかどうかは、結果のテキスト長に影響する（埋め込みが指定されれば一般的にテキストは長くなる）。

3 処理手順

3.1 変換ステップ

テキスト伸縮エンジンは、以下のステップで作業を行う。

1. 入力テキストを解析する。
2. テキスト中の表現を、内容置換エントリと注釈エントリとを参照して、知識レベルに応じた表現に書き換える。この際、どの語がどこからどこまでに展開されたかをアノテートする。ただし、注釈辞書に関しては、表示パラメータが埋め込みになっている

場合にだけ展開し、そうでない場合はアノテーションだけを保持する。置き換え対象語の近傍に置き換えテキストの一部がある場合は、置き換えを行わない。

3. 要約エンジンにより各単語に重要度を付与する。次に、重要度により、指定されたサイズに一番近いサイズの出力データを選択する。
4. 上記データを、長さ置換エントリにより指定サイズまで微調整する。
5. テキストの出力形式を作成する。「注釈として展開された」とアノテートされた部分が出力テキストにあったら、注釈表示パラメータに従って表示形式を選択する。注釈の伸縮をしたい場合は注釈に対し同じアルゴリズムを実行する。
6. テキストを出力する。

3.2 ユーザインタラクションのフィードバック

ユーザが上記のステップで出力されたテキストを読んで、分からない語があると、その語をハイライトし、記述レベルを下げる（より分かりやすくやさしい表現にする）よう要求する。

システムは、その要求が来ると、その単語のデータのユーザの知識レベルを現在のレベルより一つ下げる。(個々の単語データに関してユーザの知識レベルは特に指定がない場合、それが含まれる分野ノードの知識レベルとなる。)

逆に、ユーザは、冗長に感じる表現にマークをして簡潔な表現にするように要求することができる。この場合、マークされた部分をデータとして含む書き換えエントリを検索し、レベルを一つ上げた記述を提示する。検索結果が複数ある場合は、ユーザに選択してもらう。検索しても見つからない場合は、その旨を表示し、可能であれば、簡略な記述をユーザに記述してもらう。その際、必要なデータは、分野と記述とレベルである。レベルはデフォールトでは現在のレベルの一つ上にする。

3.3 実施例

変形テキストの下線部は、元テキストから変換された部分を示す。

元テキスト（読売新聞サイト [10]2002/01/11/01:29） Ｊ１リーグ鹿島に所属する元日本代表ＤＦの相馬直樹（３０）が、東京Ｖに期限付き移籍することが１０日、明らかになった。１１日にも、正式決定する。相馬は、１９９８年フランス・ワールドカップ（w杯）に左サイドバックとして出場、鹿島の数々のタイトル獲得にも貢献してきた。しかし、左ひざのじん帯断裂で長期離脱した昨季、鹿島がアウグスト（ブラジル）を獲得したことから、復帰後は出場機会が減り、他チームへの期限付き移籍を希望。左サイドが手薄な東京Ｖも獲得を目指していた。

サッカー知識レベル５ Ｊ１ 鹿島に所属する相馬直樹が、東京Ｖに期限付き移籍することが１０日 判った。相馬は、１９９８年 仏・ワールドカップに ＬＳＢ として出場した。しかし、左ひざのじん帯断裂で長期離脱した昨季、鹿島がアウグストを獲得した ため、出場機会が減り、他チームへの移籍を希望。東京Ｖも獲得を目指していた。

サッカー知識レベル１ 日本サッカー最高次リーグである Ｊ１ リーグ の 鹿島アントラーズ に所属する相馬直樹が、東京ベルディ に期限付き移籍することが１０日 判った。相馬は、１９９８年 仏・ワールドカップ大会 に左サイドバック [注１] として出場した。

[注１] 自ゴール前守備４人の陣形で４人中の外側２人。

この場合、注釈の表示パラメータは欄外注となっているが、埋め込みが選択されれば本文はもっと長くなり、ポップアップやリンクが指定されれば本文はもっと短くなる。

4 終わりに

ユーザの習熟レベルとそれに対応する表現とをペアにしたデータ形式を用いて、ユーザのレベルに合わせてテキストを伸縮する方法を提案した。今後は、効率的な置き換え辞書の作成法、エントリの競合などの課題解決に向けて研究を行う予定である。

参考文献

[1] 福島孝博, 江原暉将, 白井克彦,1999,"文短縮化のための文字数圧縮規則", 言語処理学会第５回年大会論文集, pp. 221-224.

[2] Luhn, P. H., 1958, "The Automatic Creation of Literature Abstract," IBM Journal, vol. 2.

[3] Murata, M. and Isahara, H., 2001, "Universal Model for Paraphrasing - Using Transformation Based on a Defined Criteria," Proceedings of Post-conference Workshop of the Sixth Natural Language Processing Pasific Rim Symposium, pp. 47 - 54.

[4] Nagao, K. and Hashida, K., 1998, "Automatic text summarization based on the Global Document Annotation," Proceedings of COLING-ACL '98.

[5] 仲尾由雄,1998, "文書要約装置およびその方法", 特開平10-207891.

[6] 佐藤理史, 2001, "なぜ言い換え／パラフレーズを研究するのか", 第七回言語処理学会年次大会ワークショップ予稿集, pp. 1 - 2.

[7] 関根聡, 江里口善生, 2000, "ＩＲＥＸ－ＮＥの結果と分析", 第六回言語処理学会年次大会ワークショップ予稿集, pp. 25 - 32.

[8] Watanabe, H., "A Method for Abstracting Newspaper Articles by Using Surface Clues," Proc. of 16th International Conference of Computational Linguistics, pp. 974-979.

[9] 山本知, 林謙治, 海老名修, 1993, "情報提供システム", 特開平 5-265677.

[10] http://www.yomiuri.co.jp/