WWW上での言語データ収集環境

渕　武志
ＮＴＴ情報通信研究所
fuchi@isl.ntt.co.jp

1　はじめに

辞書やコーパスなどの言語データは、自然言語処理の分野で極めて重要な役割を果たすため、大量、高品質、安価な言語データのニーズは大きい。しかし、高品質な言語データの作成は、今の所、人手に頼る以外に方法がない。そのため、長い時間と大きなコストがかかり、使用制限を課さずに安価で提供されることはまれである。また、特に辞書の内容を充実させるためには、その辞書を実際に使用するユーザからのフィードバックが不可欠である。しかし、そうしたフィードバックの管理や辞書の更新などを人手で行えば、また大きなコストがかかる。このコストを下げるために、言語データの管理のためのシステムがいくつか報告されている[1][2][3][4]が、これらのシステムは閉じたユーザを対象にしていた。そこで本稿では、安価に言語データを構築するために、収集、管理、更新をWWW上で不特定多数のユーザが行うシステムを提案する。

2　システムの要件

不特定多数の情報提供者からの情報を集めるシステムの場合、次の点に留意しなければならない。

・情報提供手段の容易性
・情報内容の平易性
・情報提供行動の動機付け
・情報内容の品質保持
・妨害対策

以下で、これらの点について述べる。

2.1 情報提供手段の容易性

情報提供者にとって、容易に情報を提供できる手段があることが重要である。最近のWWWの普及を考えると、WWW上に情報提供手段を構築することによって、この要件を満たすことができる。

2.2　情報内容の平易性

不特定多数の情報提供者を想定する場合、提供を期待する情報の内容は平易なものにするべきである。例えば、自然言語の例文などは、その言語のネイティブスピーカならば容易に提供可能である。一方、辞書などの場合、一般には詳細な品詞情報などは提供が難しい。そのような場合には、「"する"を付けて動詞として使う」などのように判断が容易な判定方法を示す工夫が必要である。

内容の平易性は、情報の正誤を判断する場合にも重要である。なぜなら、容易に正誤を判断できない情報について、正誤の情報が提供されると期待することはできないからである。

2.3　情報提供行動の動機付け

情報提供手段を用意するだけでは、情報の提供を期待することはできない。不特定

多数の人々に情報提供のための動機を持ってもらうためには、情報提供者への見返りが必要である。見返りは提供される情報自身で十分な場合もある。いずれにしろ、提供される情報を有用な形にして、還元することが不可欠である。

また、これら提供された情報を利用するアプリケーションを供給することが望ましい。アプリケーションが利用されれば、情報の不備を発見する機会も増え、フィードバックの頻度も増すと考えられる。

2.4 情報内容の品質保持

情報提供者が不特定多数の場合、提供される情報には低品質のものも含まれると考えなければならない。従って、低品質の情報をフィルタリングする方法が必要である。このフィルタリングを安価に行うために、不特定多数の閲覧者に情報の正誤判断の情報を提供してもらうことが考えられる。ただし、この方法だとフィルタリング自身の品質も低いと考えるべきであり、判断情報を多数決で用いる等の工夫が必要である。

2.5 妨害対策

本稿で提案するようなシステムの場合、妨害対策が特に重要である。不特定多数の全てが善意であると仮定することはできない。故意に誤った情報を提供することは容易である。従って、誤った情報の影響をいつでも無効にできる仕組みが必要である。

具体的な妨害対策として、誤った情報と判断される情報を数多く提供している提供者を特定し、その人物が提供した全ての情報を無効にすることが考えられる。ただし、誤った正誤判断を故意に提供することで、善意の提供者の情報を無効にさせる形の妨害も想定されるため、正誤判断の情報についても同様の注意が必要である。具体的には、多くの人が正しいと判断した情報について、誤っていると判断を下している判断者について、同様にその判断者からの情報を無効にすれば良いと思われる。

以上の方法は、次の前提がなりたつ場合のみ有効である。

・大多数が正しい情報を提供する。
・大多数が正しい判断を下す。
・簡単に他人に成り済ますことができない。

最後の前提を成り立たせるために、パスワードと電子メイルを組み合わせた方法が有効であると思われる。まず、情報提供者には、情報の提供を開始する前に電子メイルアドレスを入力してもらう。次に、過去に情報提供をしたことがある人の場合には、以前に決めたパスワードを入力してもらう。初めて情報提供をする人の場合には、適当なパスワードを入力してもらい、次回以降にそのパスワードを用いる。システムはその後、その電子メイルアドレス宛に確認の電子メイルを発送し、そのメイルに対する返答を送り返してもらうことによって、確かにその電子メイルアドレスが本人のものであるかを確認する。有効な電子メイルアドレスを大量に取得することは困難[1]であるから、この方法によって一個人が架空の多数に成り済ますことは難しくなる。また、すでに存在する他人の電子メイルアドレスを勝手に用いても、確認のメイルが送られるため、流用は露見することになる。

---

[1] 実際には、架空の電子メイルアドレスを作ることは容易なので、接続元のマシンのIPアドレスと組み合わせる必要がある。

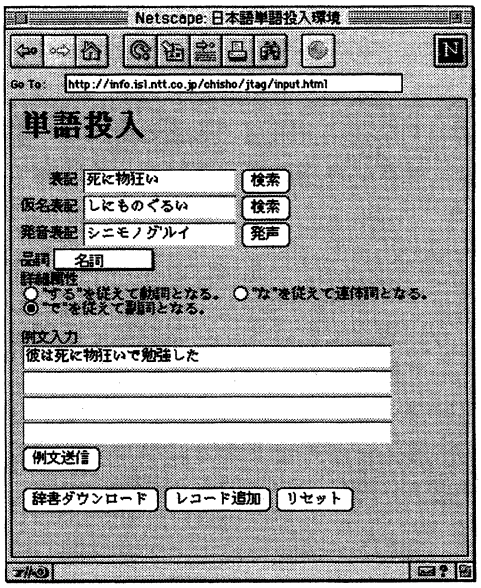

図1 データ投入部の画面イメージ

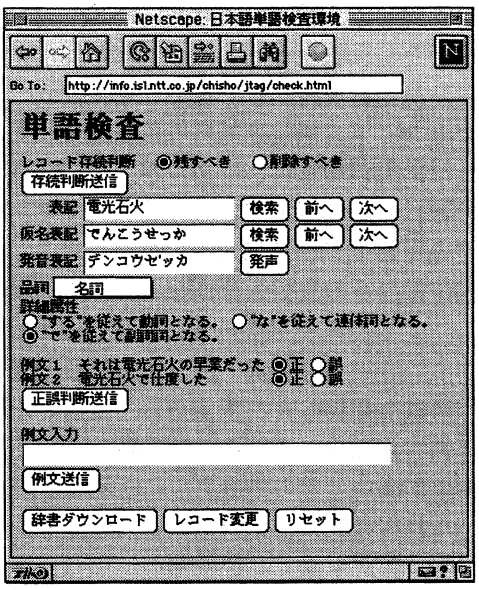

図2 データ選別部の画面イメージ

3 システムの構成

現在開発中の単語辞書の収集整備環境について、以下で述べる。システムは、データ投入部、データ選別部、提供者確認部、提供者選別部から成り、データ投入部とデータ選別部はWWWのページとして作られている。

データ投入部（図1）は、日本語の単語に関するデータを投入するWWWページである。投入するデータは、日本語のネイティブスピーカならば誰でも投入できる程度に設定されている。データが投入されると、提供者のメイルアドレスと共にデータベースに保存される。

データ選別部（図2）は、単語を検索し、データの検査、変更をするWWWページである。検索されたデータに対して、存続か削除かを選別したり、例文について正誤を投入するなどもできる。削除を選択しても、必ずしも実際にデータが削除されるわけではなく、存続判断を行った人の人数がある程度以上に達し、その中で削除の判断をした人の割合が過半数であった場合に初めて削除される。存続か削除かの判定を早めるため、削除が示唆された単語を列挙する方法も提供する。

提供者確認部は、あるユーザが本システムに初めてアクセスしたときに起動し、ユーザの電子メイルアドレスと今後の利用の際に用いるパスワードを入力するように求める。ユーザは、これらの入力をすませた時点で、すぐにシステムへのアクセスを許される。このユーザから情報が提供された場合、ユーザのアクセスが終了した時点で、システムは電子メイルを先の電子メイルアドレス宛に送付する。この電子メイルには、情報提供者の確認のためにメイルに対する返答を送るように要請する文章と、提供された情報の一覧が含まれている。返答が送られてきた時点で、提供者の確認ができた

とみなし[2]、提供された情報を実際にデータベースに反映させる。ある期間以上、返答がなかった場合には、確認が取れなかったとみなして、提供された情報を破棄する。

提供者選別部は、誤った情報を提供する情報提供者を特定する。単語に関するデータの正誤はデータ選別部で判定される。正誤情報自身の真偽の問題は、複数の正誤情報による多数決を用いることによって解決する。特定の提供者がある程度以上の量の誤情報を提供したと判定された場合、その提供者が提供した他の情報も誤っている可能性が高い。その場合、システムは提供された情報のリストをシステムの管理者に提示し、判断を仰ぐ。誤情報の原因は、悪意による他、提供方法等の誤った理解によるものも考えられる。そこで、原因が何であるかは、管理者が判断する。悪意によるものと判断した場合には、その提供者からの情報を全て破棄する。誤解によるものと判断した場合には、適切な理解を促す電子メイルを送付する。誤情報の提供者は少ないと仮定できるならば、管理者の負担はあまり大きくならないと考えられる。

4 今後の予定

現在、日本語形態素解析システムで用いる単語辞書の収集整備環境を開発中である。また、ここで作られる辞書を利用する日本語形態素解析システム[5]も公開する予定である。URL は以下の通り。

http://info.isl.ntt.co.jp/chisho/jtag

[2] 情報を提供した覚えがない旨の返答が送られてくる場合もあるので、確認の返答をするための書式を決めておく必要がある。

現在のところ、本システムは辞書やコーパスなどの言語データを対象にしている。しかし、このシステムは、時刻表や地図のデータなどにも応用できると考えられるため、順次、適用範囲を広げていく予定である。

5 まとめ

WWW上で言語データを収集する環境について提案した。本システムを用いれば使用制限のない大量の言語データを安価に収集することが可能となる。システムは開発中だが、順次、WWW上で公開していく予定である。

参考文献

[1] 小倉健太郎，篠崎直子，森本逞，「言語データベース収集支援システム」，情報処理学会　第 36 回全国大会，1988.

[2] 小倉健太郎，篠崎直子，森本逞，「形態素情報収集支援システム」，情報処理学会第 38 回全国大会，1989.

[3] 安達久博，三輪和弘，中沢正幸，鈴木美穂，「大規模電子化辞書開発における高機能辞書エディタ」，情報処理学会第 41 回全国大会，1990.

[4] 落合尚良，森義和，奥井伸司，「電子化辞書管理のための自然言語インターフェースシステム」，情報処理学会　第 46 回全国大会，1993.

[5] 渕武志，松岡浩司，高木伸一郎，「保守性を考慮した日本語形態素解析システム」，情報処理学会　自然言語処理研究会，Vol.97，No.4，1997.