RWCPにおける研究用テキストデータベースの開発

豊浦 潤* 井佐原 均† 荻野 紫穂‡ 桑畑 和佳子§ 高橋 裕信* 徳永 健伸¶
橋田 浩一‖ 橋本 三奈子§ 元吉 文男‖
*RWCP (Real World Computing Partnership) †通信総研
‡日本IBM 東京基礎研究所 §富士通 Internet ソフトウエア部
¶東工大 情報理工学研究科 ‖電総研 知能情報部

1 はじめに

自然言語処理の研究を進める上で、言語データベースが必要なことは言うまでもない。特に研究成果の発表を自由に行なう上で対象となるテキストが、著作権問題をクリアーしていることが重要である。こうした背景の下、新情報処理開発機構（RWCP：Real World Computing Partnership）が取り組んできたテキストデータベース作成の軌跡を以下に紹介する。

2 昨年度までの取り組み

RWCPでは、平成6年度より8年度までRWCデータベースワークショップを設置し、実世界に関するデータの収集と利用を目的として、テキスト・音声・画像・マルチモーダルといったデータベースの作成を行なった。

この期間に作成したテキストデータベースを、表1に示す。

94年度の時点では、対象とするテキストを探すのが、まず難しかった。作成した通産白書によるデータベースは著作権問題はクリアーしていたが、データの規模の点でEDRコーパスに比べ見劣りするものであった。しかし、この間、言語処理学会の努力で、毎日新聞社の新聞記事データが研究目的に使用可能となったため、状況は好転した。95年に作成したRWC-DB-TEXT-95-1は、量的には国内で画期的な規模のものとなった。

上に示したテキストデータベースのうち、形態素解析データはTHiMCOと呼ぶ品詞体系に基づいて作られている[1]。RWC-DB-TEXT-96-2は、岩波国語辞典の語釈文を形態素解析すると同時に、見出し、読みなどの項目をタグ情報で付与し、電子的に取り扱いを容易にしたものである。RWC-DB-TEXT-95-3は、UDCコードと呼ばれる分類コードを新聞記事の主題に応じて付与したもので、これまでにないタイプのデータベースである[2]。RWC-DB-TEXT-96-3は、新聞記事を対象にした情報検索検索のベンチマークで、日本語のベンチマークとしては初の本格的なものである[4]。

また、表1のうち、RWC-DB-TEXT-94-1/2,RWC-DB-TEXT-95-1/3を1枚にまとめたCD-ROMを製作した。そして、使用目的を研究・評価に限定して配布を行なっている（具体的入手方法などは[3]を参照）。

また通信総研などを中心にRWC-DB-TEXT-95-2をもとに単文データ（共起関係データ）の抽出が行なわれている[5]など、一次解析データから二次解析データの作成なども行なわれている。

3 平成9年度の取り組み

平成9年度からは、RWCデータベースワークショップの活動はRWCPの新たな研究領域に発展的に引き継がれ、テキストデータベースの作成のために、新たにWGが組織された。このWGは、平成8年度以前に作成したコーパスのおもに質的拡充、すなわちテキストを解析・解釈して、より多様なタグ情報を作成することを目指すことにした。これは、平成6年度から8年度に作成したデータは形態素解析タグを大量のテキストに付与し、大容量のコーパス構築の目的は実現されたので、今度は、構文情報や意味情報など、より高度なタグの作成を目標とするということである。

こうした条件を満たすコーパスを検討した結果、WGでは、平成9年度は

1. これまで作成した形態素タグの品詞統一
2. 新聞記事に出現する語に対する意味タグの付与

表 1: 平成 6 年度から 8 年度に作成したＲＷＣテキストデータベース

Database 名	内容
RWC-DB-TEXT-94-1	通産省報告書形態素解析データ（人手修正済） （平成 4 ～ 6 年度白書＋付録：延べ 7469 文）
RWC-DB-TEXT-94-2	日本電子工業振興協会報告書形態素解析データ（人手修正済） （自然言語処理の動向に関する調査報告書：3530 文）
RWC-DB-TEXT-95-1	毎日新聞形態素解析差分データ（９１年版～９４年版。全記事） （総形態素数：109,734,585、異なり形態素数：394,845）
RWC-DB-TEXT-95-2	毎日新聞形態素解析差分データ（人手修正済。９４年版。３０００記事） （総形態素数：888,000、異なり形態素数：45,845）
RWC-DB-TEXT-95-3	毎日新聞記事ＵＤＣコード付与データ（９４年版。３００００記事） （総コード数：97095、異なりコード数：14407）
RWC-DB-TEXT-96-1	毎日新聞形態素解析差分データ（９５年版。全記事）
RWC-DB-TEXT-96-2	岩波国語辞典タグ付きデータ （総見出し数：56256）
RWC-DB-TEXT-96-3	毎日新聞情報検索評価用データベース（９４年版。５００００記事） （検索要求：60querrys、正解判定：2 段階）

3. 新聞記事の文の、係り受け解析

の３つを行なうことにした。

１は、平成 6 年度から 8 年度に作成したデータは形態素解析タグに一部見られた品詞体系の不備を除き、品詞付与の方針変更や品詞の追加を行なった。採用した品詞体系は chasen の品詞に準拠しているが若干の相違がある。修正した形態素タグをこれまで解析した新聞記事 5 年分と岩波国語辞典に与え直した。

２は、新聞記事に出現する語に対し国語辞書の語義を意味タグとして与えるもので、辞書には、平成 8 年度に作成した岩波国語辞書のテキストデータベースを用いている。但し、複数の意味タグ付与や、該当する意味タグなしなどの判断も許す。

３は、新聞記事の文の、係り受けを文節単位で解析したものであり、前出の通信総研の作成したデータ [5] を利用することにより効率的に作成される。

２、３については今年度は試作に止めている。本格的なコーパス作成は来年度以降になる予定である。

4　今後の予定

以上、ＲＷＣテキストデータベースの概要を示した。今後は、今年度試作したデータの充実化を図る一方、国語辞典自身への意味タグ付け、本格的な意味処理を目指した GDA(Global Document Annotation) タグ付け、メッセージ理解に利用可能な新聞記事のテンプレート抽出など、新しい種類のタグ情報の作成も検討する予定である。

なお、本文で紹介したデータベースのうち完成しているもので、現在配布している CD-ROM に含まれないデータベースについては、1998 年春以降に CD-ROM 化して配布開始する予定である。

謝辞

データの使用を許可いただいた、毎日新聞社、岩波書店に感謝致します。

参考文献

[1] 井佐原 均, 他, ＲＷＣにおける品詞情報付きテキストデータベースの作成, 言語処理学会第 1 回年次大会, 1995.

[2] 豊浦潤, 他, ＲＷＣにおける分類コード付きテキストデータベースの開発, 信学技報 *NLC96-13*, 1996.

[3] http://www.rwcp.or.jp/wswg/rwcdb/text/.

[4] 木谷強, 他, 日本語情報検索システム評価用テストコレクション, 情処研報 *DBS114-3*, 1997.

[5] 橋本三奈子, 他, コーパスからの単文データの抽出, 言語処理学会第 3 回年次大会, 1997.