自然言語処理 114－5
(1996. 7. 18)

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

豊浦 潤 徳永 健伸† 井佐原 均‡ 岡 隆一
RWCP つくば研究センタ情報統合研究室
†東京工業大学 大学院情報理工学研究科
†通信総合研究所関西先端研究センター知的機能研究室

概要：RWCP(Real World Computing Partnership) データベースワークショップでは、約3万件の新聞記事に、国際十進分類法のＵＤＣコードを付与したテキストＤＢを開発した。このデータベースは、研究利用に対してに無償で公開される。この種の日本語のテキストデータベースは、これまでに存在しなかった。
このデータベースは、テキスト分類や情報抽出などの、自然言語処理システムの共通ベンチマークとして利用可能である。

Development of RWC Text Database Tagged with Classification Code

Jun Toyoura, Takenobu Tokunaga†, Hitoshi Isahara‡, Ryuichi Oka
Tsukuba Research Center RWCP
Tokyo Institute of Technology†
Communications Research Laboratory‡

The Real World Computing Database Working Group has built a Text Database tagged with UDC (Universal Decimal Classification) Code for about 30,000 newspaper articles. This database is available to public and free for research use. There is no such kind of text database in Japanese.
It can be use for common benchmark to evaluate various natural language processing systems, for example “text categorization”, “information extraction” and so on.

1 はじめに

RWCP*では、リアルワールド・コンピューティングの主要テーマとして、画像・言語・音声などのリアルデータを情報統合する研究を掲げている。こうした研究を遂行するためには、多種多様かつ大量のリアルデータが必要不可欠である。

一方、開発したシステムを公正に評価するためには、他の情報処理システムと共通のデータベースを用いて、それらの性能を比較する必要がある。そのためには、共通利用可能なベンチマーク・データベースが必要である。こうした必要性から、米国では 1992 年に LDC**が設立され、音声・テキストデータを CD-ROM 化し、頒布を開始している。しかし、日本ではこの種のデーターベースの整備が遅れているため、データ共有化の必要性が叫ばれている [1]。

こうした背景から、RWCP では 1994 年に、RWCP Database Workshop を組織し、画像・言語・音声・マルチモーダル・移動ロボット（平成 8 年度より）のデータベースの開発に着手した。上述の要請から、RWC データベースは、具体的には、

- 統計的に十分大きいこと
- データの性質が良好であること
- 研究目的に共通利用可能であること
- 安価に入手できること

などの条件を念頭に作成している。

以下では、特に評価用のテキストデータベース（以下、単にコーパスと呼ぶ）について述べる。まず、RWCP で開発を進行・完了したコーパスを紹介し、次に、今回作成した、分類コード付きコーパスについて説明し、最後に、今後の方針を述べる。

*Real World Computer Partnership(和称:新情報処理開発機構)

**Linguistic Data Consortium

2 RWC テキストコーパス

2.1 コーパスの種類

コーパスを作成する前に、コーパスの利用目的、即ちテキストに、どのような情報を付加するかを決めておく必要がある。RWC のコーパスでは、以下のように、大まかに 2 種類の利用を想定した、情報付加を行なうことにした。

1. 自然言語処理評価
   (a) 形態素解析情報
   (b) 構文解析情報
   (c) 意味解析情報
2. 情報検索評価
   (a) 分類情報
   (b) テンプレート情報

1 に関しては、ここでは詳しく説明しないが、自然言語処理用としては妥当な選択であろう。2 に関しては、情報検索の結果の直接評価である適合率・再現率を求めるベンチマーク [4] は既に存在するので、テキスト分類、情報抽出などへの利用を想定した情報を作成することにした。

また、1,2 は基本的には独立しているが、元にするテキストは極力同じものを使う方向で作成することにした。

2.2 コーパスに使うテキスト

コーパスのベースとなるテキストには以下の条件が必要である。

1. 著作権問題がクリアーされている
2. 安価で、容易に入手できる

3. 電子化されている

4. データ量がある程度大きい

5. 文章の品位が水準を満たしている（文法や用語の誤りがないというレベル）

従来は、1〜5を満たすようなテキストデータは皆無であったが、多くの人々の努力と好意で、現在は、1〜5をほぼ満たすような新聞記事データが研究目的に利用可能になっているので[5]、この新聞記事を中心にコーパスを作成することにした。

2.3 これまでに作成したコーパス

以上の検討を経て、これまでに作成したコーパスを、表1に示す。1994年には、未だ新聞記事の利用が可能であるかどうかが未確定だったので、量的には小さいが白書類を用いたコーパスを作成した[2]。1995年から新聞記事を用いたコーパス作成を開始している。1995年までに作成したコーパスは、CD-ROM化して実費配布する*。

3 分類情報

3.1 分類系の選定

分類は高度な知的作業である。特に対象がテキストである場合は分類方法も用途に応じて変わってくる。そのため、テキスト分類を行なっても、何が正しく、何が間違っているのかを決定するのは非常に難しい。そこで、RWCコーパスでは、テキストに出来るだけ普遍性／分解能の高い分類コードを複数個付与することにする。ユーザが目的に応じてこの分類情報を利用できる。

テキストの分類方法としてまず思い浮かぶのは図書館分類である。図書を探しやすいように書架に配置するという目的の書架分類としては、デューイの十進分類：DDC†が有名である。DDCは、分類体系と図書の物理的配置の対応が直観的に分かりやすいため、現在でも多くの図書館がDDCの改良版を採用している。しかし、DDCは書架分類であるため1冊の図書には1つしか分類コードが割り当てないので、複数のテーマを含むような図書の分類が困難である。また現行のDDCはXXX.XXXのように6桁だが、これでは表現力が不十分で新規の概念に対応できないなどの問題点も従来から指摘されている。

また、汎用的な分類系ということでシソーラスのコードも候補に挙げられる。シソーラスとしては、分類語彙表や角川類語辞典が研究ベースで良く用いられているが、これらは 10^3 ～ 10^4 の分類コードしかないので、表現能力の点ではDDCより更に劣ることになる。その他、JICSTシソーラス、日経シソーラスなどは特定分野には強力だが汎用的な分類という観点からは、やはり表現能力が十分ではない。

その他シソーラスではないがEDR概念辞書も候補として考えたが、固有名詞のインデックスが極端に少ないなど新聞記事に対する分類コードとしては、やはり不適格と考えた。また、新聞社でも独自の記事分類を行なっているが、分類の種類が 10^3 程度と少ない。

一方、DDCの問題点を解決するために、コロン分類などのファセット分類が提案されている。これらの一般社会での認知度は必ずしも高くないが、ファセット分類は多次元尺度の分類で、各ファセットの組合せで分類コードを表現するため表現力は高い。そこで、ファセット分類の中で汎用性が高い分類系を探すことにした。その結果、総合的観点からUDCが現時点では最良であると判断し、これを今回採用することにした。

*本報告執筆時には未完成であるが、発表時には、配布可能となっている予定

†Dewey Decimal Classification

3.2 UDC

UDC は、日本では「国際十進分類法」と呼ばれ、FID の管理下にある国際標準規格である。日本では (社) 情報科学技術協会が管理を受託されている。UDC は、系統的階層分類法の一種で、分類コードは標数と呼ばれる。標数は、主標数と補助標数に大別される。主標数は約 60000 項目から成り、DDC に準じた数字で表す階層的な分類コードである。補助標数には、

- 言語，国語の補助標数
- 資料の形式の補助標数
- 場所の補助標数
- 時の補助標数
- 観点の補助標数
- 材料，人の補助標数

の 6 種類が有り、全部で約 4000 項目有る。UDC は、これらの標数を何種かの連結記号で連結して概念分類を行なうものである。今回は 1994 年に発表された「日本語中間版 第 3 版」を用いることにした。本稿は UDC の解説が目的でないので、詳細は UDC 本体 [3] や適当な解説書に譲る。

3.3 新聞記事の選択

今回は、約 30000 件の新聞記事に対して、UDC 標数を付与することにした。具体的には 1994 年度の毎日新聞を用いた。記事を長さの観点から見てみると

短い記事： 死亡記事、人事記事など、文章と言うより羅列に近い記事

長い記事： 選挙の結果、国立大学の試験要項など、やはり羅列に近い記事

傾向が見られた。すなわち、タグ付けの対象としては、短過ぎる記事も、長過ぎる記事も好ましくないと考えられる。よって記事の文字数分布を見て、長さ 400 字～999 字の全 30207 記事について UDC 標数を付与することにした。

3.4 UDC 標数付与作業

実際の UDC 標数付け作業は、図書館関係者でタグ付け作業に精通した者 10 名に依頼し、約 3ヵ月で完了した。可能な限り精確に最下位の標数を付与することを期したが、工数の関係で標数間の連結は大幅に省略した。例えば、主要標数と補助標数も必ずしも連結しなくても良いこととした。

図 1に実際に付与した例を示した。最下行が今回作成したデータで、記事番号と UDC 標数のリストの形式となっている。1 番目の標数:(46) は、テキストの形式を表す補助標数で、テキストが新聞であることを示している。2 番目の標数:(093.3) もテキストの形態を表す補助標数で、このテキストが歴史的資料として日記を含んでいることを示している。3 番目の標数:929.731(44) は、主標数が国王を、補助標数がフランスを示している。

最終的に付与された UDC 標数は、延べ 97095 これらのうち相異なるものは 14407 であった。1 記事平均では、3.2 標数が付与されたことになる。

3.5 UDC 標数付コーパスの利用方法

本コーパスの利用方法について触れよう。UDC の解説を見れば明らかであるが、UDC 標数の種別は正規表現のマッチングをとることで簡単に識別できる。例えば、perl の場合なら

```
/\"\d+\"/
```

で時間の補助標数を

```
/^\d{3}/
```

で主標数の上位 3 桁をとることが簡単にできる。

これらの標数を用いた検索が可能である。但し、検索の結果の精度は保証するものではない。古い話しになるが、Cranfield の実験でも UDC の検索精度は 75%程度に止まっている。検索の評価を行なう場合は、正解としてではなく、最初の段階でのフィルタリング、あるいは参照として利用することを推奨する。

また、場所の補助標数を参照することにより場所別の分類、時間の補助標数を参照することにより時間別の分類など、多次元尺度による記事の分類も容易に行なえるであろう。

4 おわりに

以上、UDC 標数を付与したコーパスの開発について述べた。表 2 に、これまで述べたコーパスのスペックをまとめた。この種のコーパス作成は初めてだったため、様々な反省点もある。例えば、上述の 929.731(44) が具体的にはルイ１６世を指しているといった、スコープノートの類が必要ではないかとの意見もあった。この情報を加えれば、このコーパスは MessageUnderstanding の評価にも利用できると考えられるが、工数の関係で今回は付与を断念した。また、付与した UDC 標数は、今回のコーパスでは単なる羅列になっているが、記事の主題に相当する標数を差別化するなど、標数間で何らかの順位つけを行なった方が良かった。

そもそも人間の主観が介在するようなこの種の作業は、手間とコストがかかる一方、作業者の個性に依存が避けられず、コーパス自体の客観的精度を評価するのが困難である。こうした困難に対する有効な対策は具体的に思いつかないが、今回のコーパスを試金石として、当面はユーザの意見を待ちたいと思う。

今後は、このコーパスの改良を目指す一方、2(b) の形のコーパス作成も検討を進めたい。

謝辞

このコーパスを作成する機会を与えて下さった、RWCP 島田所長、分類コードの選定に当たり助言・指導頂いた、図書館情報大学 石川先生、UDC 標数付与の監修を頂いた、元 UDC 実行委員会副委員長 市川先生、新聞記事の利用を快諾頂いた、毎日新聞社メディア事業局 蓑田部長、川見氏に感謝致します。

参考文献

[1] 板橋 秀一, 言語データ共有化 (LRSI) について, *LRSI* シンポジウム, 1994.

[2] 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, pp.181-184, 1995.

[3] (社) 情報科学技術協会, 国際十進分類法 *ISBN4-88951-029-X*, 丸善, 1994.

[4] 芥子 育雄, 他, 情報検索システム評価用ベンチマークＶｅｒ．１．０（ＢＭＩＲ－Ｊ１）について, 情処研資 *DBS106-19*,pp.139-146, 1995.

[5] http://cactus.aist-nara.ac.jp/lab/resource/cdrom/Mainichi/MS.html

表 1: RWC で開発中／完了したコーパス

データセット名	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。通商白書平成４～６年版等）
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、９４－計－４）
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）
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）
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）

＼ＩＤ＼００００００１０
＼Ｔ１＼［余録］変化
＼Ｔ２＼　フランスのルイ十六世はずぼらだったが、日記だけはきちんとつけた。その日、狩りで一匹の獲物もなかったので、日記にただ一言「何もなし」…

00000010 (046) (093.3) 929.731(44)

図 1: タグ付けした記事の例

表 2: RWC-DB-TEXT-95-3 のスペック

総記事数	30207
延べ標数	97095
異なり標数	14407
うち補助標数 (言語)	87
うち補助標数 (資料)	167
うち補助標数 (場所)	6223
うち補助標数 (時)	162
うち補助標数 (観点)	404
うち補助標数 (材料)	964