コーパス間の類似度によるコーパス分類と専門分野別辞書構築

3R－2

鳥原 信一

日本アイ・ビー・エム株式会社 東京基礎研究所

1. はじめに

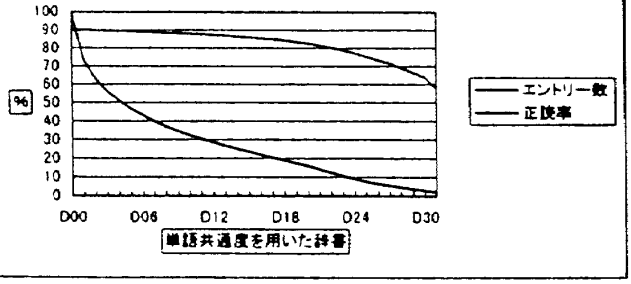
日本語テキスト音声合成システムにおいても、種々の規模のシステムで、そのシステム・リソースの中で最大の精度を得ることが求められている。我々は形態素解析用辞書をスケーラブルにすることを試みる。すなわち、最小で基本語彙辞書を使用しておき、実行時には音声合成するテキストの分野推定を行い、専門分野別辞書を動的に組み込み・切り換えを行うシステムである。本論文では、最初に、基本語彙の決定方法について述べる。次に、専門分野別辞書の構築に当たり、あらかじめ分野が分かっているコーパスをテンプレートとして、新たに利用可能なコーパスとの類似度を相互相関、相互情報量によって判定し、コーパスの分類・収集を行う方法について述べる。最後に、分野ごとに分類されたコーパスからその専門分野に属すると思われる単語の抽出法について述べることにする。

2. 基本語彙辞書

専門分野別辞書の構築に先立って、基本語彙セットを決定する必要がある。まず、我々が利用可能なコーパス群(31 個のコーパス)を我々の標準辞書で形態素解析を行った。標準辞書(D00)に加え、ある単語がいくつのコーパスに共通に現れるかによって、D01 から D31 の辞書を作成した。なお、31 個の辞書を作成する際に、標準辞書にあるエントリーと照合して作成したので、コーパスには現れたが、形態素解析によって解釈された未知語はここには含まれない。図 1. にこれらの 32 個の辞書を用いた正読率(読みのみ)とエントリー数の 関係を示す。また、表 1. に正読率の変化表を示す。

表 1. から分かるように、標準辞書 D00 の約 30,000 エントリーは使用されることがきわめて少ないであろうことは容易に想像できる。正読率が D07 と D08 の間で大きく変化しているので「D07」を基本語彙辞書とすることにする。

図1 エントリー数と正読率



辞書	エントリー数（含む未知語）	正読率	変化率
D00	121,429	90.31	0.05
D01	89,042 (242,577)	90.26	0.27
D06	52,235 (69,669)	89.13	0.23
D07	48,366 (62,958)	88.90	0.46
D08	44,991 (57,421)	88.43	

表 1 正読率の変化表

3.1 コーパス間の類似度

専門分野別辞書を構築するに当たっても、コーパスを利用して行う。あらがじめ分かっている分野のコーパスをテンプレートとして、新たに利用可能となったコーパスを、つぎのような尺度でコーパス間の類似度をもとにコーパスを分類・収集することにする。

$$\text{相互相関} = \sum_{i=1}^{m}\sum_{j=1}^{n}(P(a_i)-P(b_j))^2$$

$P(a_i)$と $P(b_j)$は同一の単語のそれぞれのコーパスにおける確率である。値が小さい方が類似度が高い[1]。

$$\text{相互情報量} = -\sum_{i=1}^{m}\sum_{j=1}^{n}P(a_i,b_j)\{\log(P(a_i))+\log(P(b_j))-\log(P(a_i,b_j))\}$$

$P(a_i)$と $P(b_j)$は同一の単語のそれぞれのコーパスにおける確率である。コーパス A に現れた単語は、コーパス B に必ず現れるとするので、$P(a_i,b_j)=1$ とする。 値が大きい方が類似度が高い[2][3]。

3.2 分野のテンプレートと類似度

「CD-毎日新聞(データ集)」[4]には、どの新聞記事面なのか内容コードがつぎのように付いている(表 2. 参照)。

01	1面	05	社説	12	総合	16	科学
02	2面	07	国際	13	家庭	18	芸能
03	3面	08	経済	14	文化	35	スポーツ
04	解説	10	特集	15	読書	41	社会

表 2 新聞記事内容コード

Sorting corpus by their similarity and building special field dictionaries.

Shinichi Torihara, IBM Research Laboratory, IBM Japan

これらの新聞記事面のうち、「分野」と言えると思われるものを91年および92年で類似度を計算してみた(表3.参照)。

'91 / '92	国際	経済	家庭	文化	読書	科学	芸術	スポ	社会
国際	6.1	1.2	3.3	2.6	6.8	4.7	5.4	7.5	1.0
経済	2.2	6.1	3.3	4.6	5.8	4.7	6.5	7.4	1.0
家庭	4.6	4.3	6.1	1.2	4.8	3.7	3.4	5.5	2.0
文化	3.7	4.5	2.3	6.1	2.8	1.6	2.2	5.4	3.0
読書	6.7	7.5	4.4	2.2	0.8	1.3	3.1	8.6	5.0
科学	6.5	6.2	2.1	1.4	3.8	1.3	3.6	4.7	4.0
芸術	6.6	7.5	2.4	1.2	4.8	3.7	6.1	8.3	5.0
スポ	3.6	3.6	5.4	3.5	4.8	2.7	6.2	6.1	1.0
社会	2.6	1.4	1.2	3.5	6.8	4.7	5.3	7.1	6.0

表3 新聞記事面類似度(相互相関、相互情報量:ランク順)

これらによると、「読書」、「科学」、「社会」は、このままでは分野とするのは困難である。また、相互相関および相互情報量で類似度がかなり異なるものがあり、さらに研究が必要である。

3.3 新たなコーパスと類似度

あらかじめ分野のテンプレートを用意しておき、新たなコーパスを類似度を用いて、分類できるか検証してみた(表4.参照)。「経済社説」は、「受け皿銀行」についての論説であるが、新しい話題のためか「経済」の分野との類似性を示さなかった。

'92 / 記事	国際	経済	家庭	文化	芸術	スポ
経済速報	1.2	0.0	3.4	2.3	4.5	5.1
スポ速報	0.4	0.2	3.5	1.3	2.1	0.0
経済社説	4.1	3.2	0.5	2.4	1.3	5.0
スポ社説	2.5	1.0	3.3	3.2	4.4	0.1

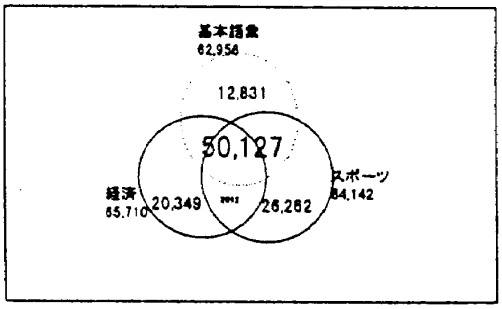
表4 新聞記事面とあらたな記事との類似度

(相互相関、相互情報量:ランク順)

4.分類済みコーパスからの専門分野エントリー抽出

ここでは、分類・収集されたコーパスから専門分野エントリーを抽出する。「経済」と「スポーツ」から基本語彙を抜いて「経済」、「スポーツ」を マッピングしてみた。図2.を参照されたい。

これらを、その他の分野についても繰り返すことになるが基本語彙であまり使用されないエントリーおよび「経済」と「スポーツ」と共通に現れる単語についても再検討が必要である。



経済の単語例
公取委
NAFTA
TNC
ドンクル
資本

スポーツの単語例
UFIS
TOPO
美しだ
ハウエル
打線

図2 基本語彙、経済、スポーツの単語関係

5.おわりに

相互相関および相互情報量によってコーパス間の類似度を計算することにより、コーパスの分類を行った。「読書」、「科学」、「社会」は、その範囲が広く、「分野」として扱うのは現段階では不適当であることが分かった。また、新たな話題のコーパスは、分類不可能であった。コーパス分類のための類似度の精度向上および新しい話題への対応についてさらに研究を進めるつもりである。さらに、分野に分類されたコーパスから専門分野エントリーを抽出したが、この過程で得られた結果を基本語彙エントリーにフィードバックをかけるつもりである。

6.謝辞

「CD-毎日新聞(データ集)」を利用して、大規模なコーパスを対象とした研究ができました。毎日新聞社に感謝致します。

参考文献

[1]前川守,「1000万人のコンピュータ」
3文学編」,岩波書店,1995
[2]梅村恭司,個人のためのデータマイニング,bit,共立版
6-1998
[3]中川聖一,「情報理論の基礎と応用」,近代科学社,1992
[4]毎日新聞社,「CD-毎日新聞(データ集)」
91,92,93,94,95年版