4E-4

重ね合せ符号を利用する大規模辞書検索システムについて

玉越 靖司
松下電器産業(株)東京研究所

1．はじめに

我々は自然言語処理システム用の大規模な単語辞書と，それに対して辞書引きを行うシステムの開発を進めている。このシステムの目的は，各々の自然言語処理システムで個別に行われている辞書や辞書引きシステムの開発に係わるオーバーヘッドを吸収することである。従って，このシステムにおいて対象とする辞書はマスター辞書の性格をもった汎用大規模辞書である。現在，２０万語規模の日本語単語辞書を念頭に置いている。本稿では，重ね合せ符号を利用する辞書引き手法を提案する。

2．理想とする辞書引き

現在，自然言語処理システムで行われている辞書引きとして最も一般的な方法は転置ファイルを用いる方法である。この方法は大規模な辞書に対しても高速に検索が可能である。転置ファイル方式にハッシングを導入して，我々が目指す規模の辞書を非常に高速に検索する方式もある。しかし，こうした転置ファイルを用いる方法では，特に辞書引きのようにマルチキーの検索を行いたい場合，必要な記憶領域量や辞書のメンテナンスのコストが大きい。

一方，例えばPrologのような言語で記述している自然言語処理システムにおいては，述語で辞書を記述し，検索を言語の単一化機能に委ねている例もある。こうした方法は辞書の規模が小さな場合には有効だが，我々が目指す大規模辞書の検索方式としては適当でない。

以上を考え合わせると，我々が理想とすべき辞書引きシステムの特長は次のようなものである。

(1) 高速な検索が可能である
(2) 大規模辞書の検索が可能である
(3) マルチキーの検索が可能である
(4) 記憶領域量を抑えることが可能である
(5) メンテナンスが容易である

さらに，種々の自然言語処理システムで用いることを考慮すると，それぞれの環境的制約に対応できるように，これらの条件のトレードオフによる設計が容易に可能であることが望ましい。

3．重ね合せ符号を利用するデータ検索手法

重ね合せ符号を利用するデータ検索の手法（以下ＳＣW方式と呼ぶ）[1]について述べる。ＳＣW方式は，本質的にはパンチカードによるハンドソーティングの手法と同じである。これまで，電話番号案内[2]や文献情報検索[3,4,5]，ホーン節の単一化可能項検索[6,7,8,9]などに応用されてきた。

3.1 データの蓄積

マスターファイルＦ中のレコードRiが持つ＃Ri個のキーワード１つ１つに対し，ハッシュ関数を用いてｂｃｗ(Binary Code Word)と呼ぶ長さｂのビット列を作る。このハッシュ関数はｂｃｗのｂビットのうち，ｗビットに'１'を立て，残りは'０'とする。ｗを重みと呼ぶ。＃Ri本のｂｃｗをビット毎に論理和をとった長さｂのビット列ｓ(Ri)をRiのｓｃｗ(Superimposed Code Word)と呼ぶ。こうして作成された＃Ｆ本のｓｃｗを，ｂ本の長さ＃Ｆのビット列としてファイルＳＣＷに格納する。ここで，＃ＦはＦ中のレコードの個数である。

このようにして作られるＳＣWは，Ｆの転置ファイルに比べてはるかに小さく，メンテナンスも容易である。しかも，ｂを決めることにより，ＳＣWのサイズを自由に設計することができる。

3.2 検索

質問Ｑが持つ＃Ｑ個のキーワードについて，蓄積の場合と同じハッシュ関数を用いてｂｃｗを作成し，同様の処理によりマスクｓ(Q)を作る。ｓ(Q)とＳＣＷを照合して，ドロップと呼ばれる検索結果

$$D(s(Q))=\{Ri \mid s(Ri) \wedge s(Q)=s(Q)\}$$

を求める。ｓ(Q)で'１'が立っている位置に対応するＳＣWの長さ＃Ｆの列に対して，ビット毎の論理積をとることによってＤが簡単に求められる。この論理積の結果である長さ＃Ｆのビット列において'１'が立っているビットｉに対応するRiがＤに含まれる。ＳＣＷ方式では，＃Ｆのサイズにほとんど依存せず，高速にＤの検索が可能である。

3.3 フォルスドロップ

Ｑに対する正しい検索結果のレコード集合をＡ(Ｑ)とすると，Ｄ(ｓ(Ｑ))は次の性質を持っている。

$$D(s(Q))=A(Q)\cup Df(s(Q))$$
$$(A(Q)\cap Df(s(Q))=\phi)$$

ここで，Ｄｆは誤って検索されたレコードの集合のことであり，フォルスドロップと呼ばれる。ＳＣＷ方式ではＤからＤｆを除去するオーバーヘッドが必要である。この処理時間コストのオーダーは＃Ｄであり，Ｄの検索に比べてずっと大きい。ここで＃ＤはＤ中のレコードの個数である。しかし，ｂやｗ，ハッシュ関数の設計により，Ｄｆを十分に小さくすることが可能である。

A Looking up System in a Large Scale Dictionary via Superimposed Code Words
Yasushi TAMAKOSHI
Tokyo Research Laboratory, Matsushita Electric Industrial Co., Ltd.

4．ＳＣW方式の辞書引きへの応用

ＳＣW方式を自然言語処理の辞書引きに応用することを提案する。

4.1 対象とする辞書

まず，辞書引きの対象となる辞書の性質を規定する。我々は２０万語規模の日本語単語辞書で「見出し語情報」「品詞情報」「意味情報」を持っているマスター辞書[10]を想定している。

これらの情報は階層構造になっている。すなわち，１つの「見出し語情報」は１つ以上の「品詞情報」を持っている。また，１組の「見出し語情報」「品詞情報」は１つ以上の「意味情報」を持っている。

この辞書引きシステムを用いる自然言語処理システムとして，日本語解析と日本語生成の両方を考慮する。日本語解析システムでは，主に「見出し語情報」と「品詞情報」をキーとする辞書引きが行われる。日本語生成システムでは，主に「意味情報」をキーとする辞書引きが行われる。

4.2 ＳＣW方式の辞書引きの特徴

ＳＣW方式が持つ一般的な性質を以下にまとめる。

(1) ＃Ｆにほとんど依存せず，高速にＤが求まる
(2) 検索時間コストの大部分を占めるＤf除去は，オーダー＃Ｄで抑えられる
(3) ｂを調節することにより，ＳＣWのサイズを決定できる
(4) ｂを大きくすると，ＳＣWは大きくなるが，＃Ｄfは小さくなり検索が高速化される
(5) ｓ(Ｒi)中の'１'の平均個数がｂ／２になるようｗを調節することにより，＃Ｄfが最小となる
(6) マルチキーの検索が容易で，しかも＃Ｑが大きいほど＃Ｄが小さくなり，検索が効率的である
(7) Ｒiに関するメンテナンスはｓ(Ｒi)だけでよく，簡単である

こうした特徴は，本稿で理想とする辞書引きに適合しており，ＳＣW方式が辞書引き方式として有望であることを示している。

しかし，辞書が持つキーワードのｂｃｗを単純に重ね合せるだけでは，十分に＃Ｄfを小さくすることができない。そこでＳＣW方式の拡張として，キーワードの性質により，ｓｃｗ中でのｂｃｗの位置（フィールド）を分ける方法を採る。

4.3 見出し語情報のフィールド

ＳＣW方式では，それぞれの異表記に対応するｂｃｗを作成することにより，容易に異表記の辞書引きを実現することが可能である。また，見出し語の文字列の前方文字から順にキーワード化することにより，前方一致の辞書引きも可能である。

これらの機能を実現するためには，見出し語情報のフィールドにおいて多重のｂｃｗの重ね合せを行う必要がある。従って，見出し語情報のフィールド幅ｂｍは，他のフィールド幅に比べてかなり大きくしなければならない。

4.4 品詞情報のフィールド

品詞情報では，品詞や活用など，キーの種類の数が限られている。また，ｂｃｗの重ね合せの数が決まっている。一般に，重ね合せが少ないとｓｃｗ中の'１'が少なくなり，Ｄfの発生が抑えられる。逆に言うと，ｂに対するｗを相対的に大きくできることによって，Ｄの精度が上がる。

こうした特徴を利用して，品詞情報のフィールド幅ｂｈはｂｍに比べて小さく設定することが可能である。

4.5 意味情報のフィールド

意味情報によるキーは一意の概念番号である。従って，意味情報のフィールドではｂｃｗの重ね合せが行われず，他のレコードとの重複によるＤfがない。このことは，＃Ｆ＝２０万のとき意味情報のフィールド幅ｂｉ＝２１とすれば，Ｄfを全く発生させないことを意味する。

4.6 頻度別ＳＣWによる高速化

辞書引き全体の効率のためには，使用頻度の高い見出し語を特に高速に引く必要がある。そこで，＃Ｄを小さくすることによってオーダー＃Ｄの処理時間コストがかかるＤf除去のオーバーヘッドを小さくする。すなわち，辞書中の単語頻度情報を利用して頻度別ＳＣWを作り，高頻度のＳＣWから順に検索を行う。

4.7 辞書の定量的解析による準最適化

スタティックな辞書に対して定量的な解析を行うことにより，またこの辞書引きシステムを用いる自然言語処理システムの特性に応じて，各々のフィールドで準最適なｂやｗ，ハッシュ関数，頻度別ＳＣWのサイズを設計することが可能である。

5．おわりに

ＳＣW方式は処理時間や記憶領域量，メンテナンスのコストに優れ，環境的制約によるこれらのトレードオフが可能であるということに着目して，ＳＣW方式による辞書引きシステムを提案した。ＳＣW方式ではＤfの除去というオーバーヘッドがあるが，その軽減が十分に可能であるとの見通しを得た。現在，この方法の有効性を確認するために，実験システムを作成している。準最適化の手法とその結果については次の機会に報告したい。

参考文献

[1] Roberts, C.S. : Partial-Match Retrieval via the Method of Superimposed Codes, Proceedings of the IEEE, Vol.67, No.12, 1979.
[2] Gabbe, J.D. et al. : Applications of Superimposed Coding to Partial-Match Retrieval, Proc. COMPSAC 78, 1978.
[3] 有川他：重ね合せ符号と逐字サーチを利用する文献情報検索システムについて(1)，情報処理学会第25回全国大会，1982.
[4] 有川他：重ね合せ符号と逐字サーチを利用する文献情報検索システムについて(2)，情報処理学会第27回全国大会，1983.
[5] 有川他：重ね合せ符号と逐字サーチを利用する文献情報検索システム，情報学シンポジウム講演論文集，1985.
[6] Wise, M.J. et al. : Indexing PROLOG Clauses via Superimposed Code Words and Field Encoded Words, Proc. 2nd Int. Logic Progr. Conf., 1984.
[7] 森田他：スーパーインポーズドコードを用いた構造体の検索方式，情報処理学会第33回全国大会，1986.
[8] 和田他：KBMS PHI:重ね合せ符号を用いた関係演算の処理方式，情報処理学会第34回全国大会，1987.
[9] 森田他：MPPMを用いた知識ベースマシン(3)-構造体のインデックス方式に関する一考察-，情報処理学会第35回全国大会，1987.
[10] 単語辞書（第２版），EDR TR-006，1988.