2T-03

トライ構造を用いた多言語対訳辞書の効率的圧縮手法

住友 徹, 森田 和宏, 泓田 正雄, 青江 順一
徳島大学工学部知能情報工学科

1．まえがき

近年インターネットの普及とともに，種々の言語で記述された文書を閲覧・利用することが可能になった．しかしながら，利用者は全ての言語を理解できないので，理解可能な言語（母国語と呼ぶ）で文書の内容を知る手段が必要となる．これに対して，多言語翻訳システム，クロス言語検索システム[1][2]などの研究が活発に行われている．数種の言語を対象とし，各言語への新語の追加がない場合には，各言語にそれぞれの対訳情報を格納する手法で問題とならないが，言語の種類が多くなり，しかも新語により対訳情報が頻繁に更新される場合は，対訳辞書の効率的な検索・管理が重要な課題となる．本論文では，言語表記を辞書検索部から容易に生成できるダブル配列法を利用して，適当な辞書を使用することで対応している．しかし，検索の前にその要求に対してどの辞書を使用するかを決定しなければならないため，処理が複雑になる．また，2 言語間での対訳辞書であっても英日と日英のように二つの辞書を用意しなければならないため，記憶量を無視することはできない．そこで，ダブル配列[3]を用いてこれら複数の対訳辞書を統合する手法を提案する．

2．多言語対訳辞書

多言語対訳辞書は検索キーの取り扱う言語の数を n とし，言語を $L_1,L_2,...,L_n$ で表す．

ここで言語 L_1 の単語を x_1 とする時，対応する意味を持つそれぞれの言語における単語を $x_2,...,x_n$ で表す．この時，言語 L_1 の単語 x_1 に対する対訳辞書は，キーx_1から検索できるレコードを n-1 のフィールドに分割して，それぞれのフィールドに対応する言語の対訳情報を格納することになる．従って必要となるフィールド数は，全体では，$n(n$-1)必要となる．ある言語から別の言語の対訳情報を検索する場合，検索速度は問題無いが，事前に検索キーの言語を同定しておく必要がある．登録においては，新語が登録された時 n 個のキーの追加とフィールドの追加が必要になる．複数のフィールドに同一の単語情報が登録されることになる．

提案手法では，複数の対訳辞書を一つのトライ上に構築する．ある単語情報 x_i がどの言語に属するかという情報 L_i は，登録された単語情報 x_i に対応するレコードに L_i を登録し，レコードを取得することで単語情報 x_i の言語 L_i を特定できるようにする．これによって事前に言語の同定をおこなう必要がなくなる．また，レコード部のフィールドに単語情報をそのまま登録するのではなく，トライの葉ノードの番号を登録する．単語情報そのものの取得は，登録された葉ノードの番号から根ノードに向けてパスをたどりアークの重みを後方から調べていくことで単語情報を取得する．これによって，同一の単語情報を複数のフィールドにわたって登録する必要はなくなり，またレコードに登録される単語情報が葉ノードの番号だけになることから，レコードの容量は大幅に削減される．

葉ノードから根ノードへの走査はトライを実現するデータ構造によって，その効率が大きく変わってくる．配列やリストでは親ノードの情報を持っていないため，親ノードの検索に時間をとられることになる．かといって親ノードの情報を持たせた場合は，記憶量に問題が生じる．これらの問題を解決するために提案手法では親ノードを容易に特定でき，かつコンパクトなダブル配列を用いる．

3．ダブル配列による多言語対訳辞書

レコードに葉ノードを登録することでレコードの容量の圧縮を行う．しかし，トライに新規にキーを追加する際に，ダブル配列の性質上，ノードの移動が生じ得る．これによって起こる葉ノードの変更は大きな問題となる．葉ノードの変更が生じると全レコードから変更が起こった葉ノードを検索し，検索された全情報を更新しなければならない．

これを回避するために，レコードに葉ノードの情報を，直接持たせるのではなく，別に葉ノード番号の一覧を登録したテーブルを定義する．そして，そのテーブルに登録された葉ノードへの参照番号をレコードに登録する．これによって葉ノードの変更が起きた時に，中間テーブルに登録された葉ノードに

An Efficient Compression Method for Multilingual Dictionaries Using a Trie Structure
Toru Sumitomo, Kazuhiro Morita,
Masao Fuketa and Jun-ichi Aoe
Department of Information Science and Intelligent Systems, University of Tokushima

変更を加えるだけですむ.

そして，この中間テーブルは，葉ノードとレコードの1対1の対応を考えれば，レコード内に対応する葉ノードの番号を保持することで代用できる(図1参照).

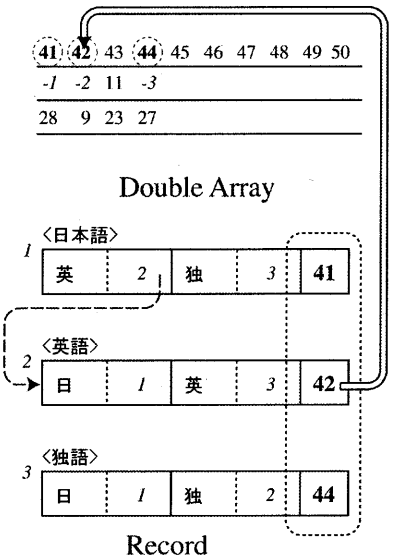


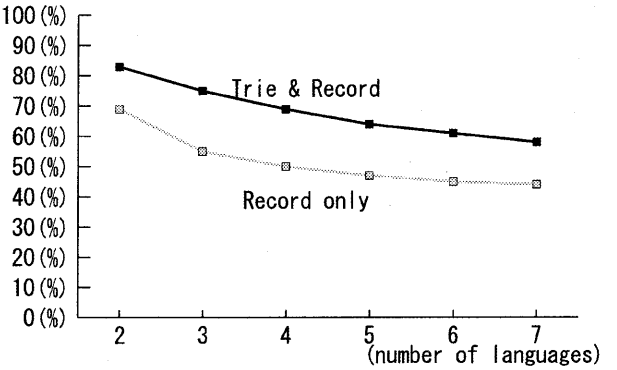
図1　レコードに葉ノードを登録

4．実験と評価

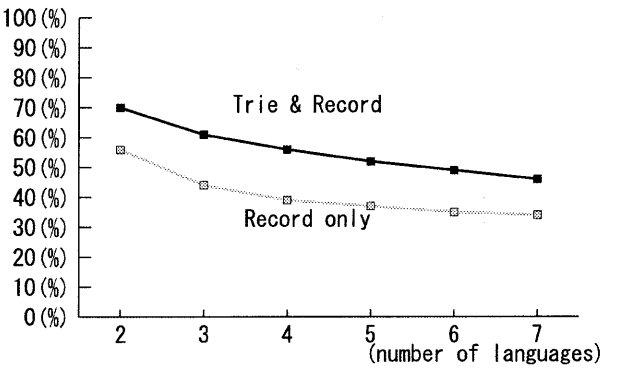
実験に用いたデータは1言語につき2万件の文字列をランダムに生成したものであり，平均語長10，13の文字列長の違う2つのデータを用意して，それぞれにつき従来法と提案手法での記憶量の比較を行った．実験に使用したマシンのCPUはPentiumIII 500MHz，メモリーは256MB，実験用プログラムの開発言語はC++である．またレコード番号は4バイトのデータとして扱い，辞書は1次記憶上に構築した．実験結果の従来法と提案手法の記憶量の比を百分率であらわしたグラフを次に示す(図2)．グラフには全体のデータの比と，レコードだけの比を示した．なお，上が平均語長10のデータ，下が平均語長13のデータのものである．

図2より，言語数や平均語長が大きいほど，提案手法は有効であると言える．

また従来法でキーワードを検索，対訳情報を取得するまでの時間は1.03ミリ秒．提案手法では1.63ミリ秒であった．レコード情報取得時にトライを逆方向に辿る必要があるため，従来法より若干速度は劣るが，十分高速であることが確認できた．



(a)平均語長10



(b)平均語長13

図2　実験結果

5．まとめ

複数の辞書を持つ多言語対訳辞書を，ダブル配列を利用して一つに統合化し，管理を容易にし記憶量を減少する手法を提案した．実験の結果，記憶量に関しては従来法より有効であり，検索の時間計算量においても十分に実用的であることが確認できた．

また今回は1次記憶上で実験を行ったが，2次記憶上で本手法を用いる場合にレコードの大きさが言語数によって固定される事は，データの読み出しや書き込みに有利であると考えられる．今後は2次記憶上で多言語対訳辞書を構築し，本手法の評価を行っていく．

参考文献

[1] 市山俊治,野村直之：多言語間機械翻訳辞書の開発手法，情報処理学会研究報告，NL73-14(Jun.1989).

[2] 藤井敦,石川徹也:技術文書を対象とした言語横断情報検索のための複合語翻訳，情報処理学会論文誌,Vol.41,1038--1045,2000

[3] 青江順一:キー検索技法－トライ法とその応用，情報処理学会論文誌,Vol.34,1244--251,1993