Vol. 23 No. 2 情報処理学会論文誌 Mar. 1982

ショートノート

医薬品情報のための文章のコード化規則と復元アルゴリズム†

岡田 正彦††

医薬品情報のデータベースを作成するための基礎的研究として，日本語で記述された薬の説明文をコード化し，また復元する方法を開発した．
コード化の規則は，そのルールをよく理解し，しかも医薬品自体にも知識のあるものが変換作業を行うという前提で作製した．実際のコード化は174語から成る辞書と11の規則を利用するだけで可能である．
復元規則も簡単で，28種類の項目から成っているだけである．

1. まえがき

医薬品情報のコンピュータサービスシステムの必要性が高まってきており，国内でも，すでに厚生省の承認医薬品ファイルを初めとする多数の専用ファイルが作られている．しかし，これらのファイルの多くは事務処理向けで，製品名，用量，薬価などの簡単な項目しか登録しておらず，医療用情報としては不十分である．ところが，医薬品データに関して医師や薬剤師が必要とする十分な情報をカバーするためには，必然的に文章でしか表現できない項目も扱わざるを得なくなってくる．

そこで，われわれは医薬品添付文書中の「作用」の項に限定して，文章をコード化する規則と，それを元の文章に復元する方法を検討した．コード化に関しては，特に次の三つの立場から規則を定めた．

1）文章コード化の作業はあらかじめ定めた規則を十分習熟し，なおかつ薬剤についての専門知識のあるもの（薬剤師など）が行うこと．

2）コード化されたデータに対して論理的な検索が可能であること．

3）文章としてデータを出力する必要がある場合にも，短時間で復元可能なアルゴリズムが存在すること．

なお，医薬品データベースでは，通常，内容の変更はあまり行われない．新薬の発売時や薬剤再評価の実施時などにはデータの追加や更新が行われるが，頻度は少ない．したがって，データ入力は特定のスタッフが担当するだけでよく，手作業によるコード化で実用上十分である．

本文では，方法の詳細と応用例について述べる．

2. コード化規則

以下，医薬品添付文書に記載されている元の文章を原文，それをコード化した文章を入力文，復元アルゴリズムによって翻訳された文章を出力文と呼ぶことにする．

入力文は，**表1**に示す7種類の辞書と対照しながら作製するが，各辞書の内容は数値や記号を極力さけ，英単語を基本にした．**表2**と**表3**には，前置詞と接続詞の辞書を具体的に示す．三つの表から明らかなように，各辞書に登録した単語は，総計174語ときわめて少ない．ただし，名詞だけは，内容がひじょうに多岐にわたるため，そのすべてを辞書に収録することは実用的でない．そこで，名詞に関しては，頻繁に使用されしかも重要なキーワードとなりうるものだけを辞書中に定義した．

なお，各辞書に登録した単語は，文献[1]に収録された200品目の医薬品の説明文を調べ，そこに使われた単語375個（固有名詞，物質名などを除く）から同義語を除き174語にまとめたものである．

表1 辞書の種類とサイズ
Table 1 Size of each dictionary.

名称	語数
前置詞	13語
接続詞	11語
形容詞	19語
副詞	16語
自動詞	9語
他動詞	55語
名詞	51語

† Free Text Encoding and Decoding Algorithm for Drug Information by MASAHIKO OKADA(Department of Neurophysiology, Brain Research Institute, Niigata University).
†† 新潟大学脳研究所神経生理学部門

表 2 前置詞の辞書
Table 2 Dictionary of prepositions.

入 力	出 力
VIA	オカイシテ
AT	ノ
AFTER	ノアトノ
BETWEEN	ノアイダノ
BY	ニヨッテ
FOR	トシテ
FROM	カラ
IN	ノナカノ
INTO	ノナカヘ
TO	ニ
WHEN	ノトキ
WITH	トトモニ
OF	ノ

表 3 接続詞の辞書
Table 3 Dictionary of conjunctions.

入 力	出 力
BECAUSE	ナゼナラバ
NOT	コトハナイ
OR	アルイワ
OF	ノ
' (apostrophe)	フ
; (semicolon)	, (句点)
: (colon)	. ソノケッカ
, (comma)	, (句点)
. (period)	. (終止符)
?	ト カンガエラレテイル.
=	スナワチ

コード化の規則の主な部分は，次の通りである（全体では 11 項目から成っているが詳細は省略）.

1. （主語）＋述語＋（目的語）を文の最小単位とし，これを文単位と呼ぶ．ただし，（ ）内の品詞はなくともよい.

2. 文中の単語は，原則として辞書中に定義されたものだけを使用する．ただし，名詞はこの限りでない.

3. 動詞の辞書中にない単語であっても，名詞の辞書中にある単語を使い，「DO＋名詞」の形で，述語として使用できる.

4. （ ）は次のような機能を有する.

1）入力文がいくつかの文単位で構成されている場合，従属文に相当する方の文単位を（ ）で囲む.

2）前置詞のかかる範囲が不明確な場合，それを（ ）で囲んで明らかにする.

3）動詞の受ける範囲が不明確な場合，（ ）で明示する.

5. 入力文の先頭を除く各文単位には主語を付ける.

3. 復元アルゴリズム

以下に述べる復元アルゴリズムは，データベースの利用者が，情報を文章として出力させたい場合に，コード化されているデータを短時間に元の文章に翻訳させるための方法である（全体では 28 項目から成っているが詳細は省略）.

1. 入力文を単語単位に分解し，配列 $\boldsymbol{w}=\{w_1, w_2, \cdots, w_l\}$ にセットする．ただし，l は単語数である.

2. 配列 $\boldsymbol{w}$ 中の単語を各辞書と照合し，その品詞名と，辞書中の出現順位をそれぞれ配列 $\boldsymbol{p}=\{p_1, p_2, \cdots, p_l\}$ と $\boldsymbol{k}=\{k_1, k_2, \cdots, k_l\}$ にセットする.

3. 配列 $\boldsymbol{w}$ 中の各文単位について，nest の深さを調べ（かっこの個数で判定），配列 $\boldsymbol{n}=\{n_1, n_2, \cdots, n_l\}$ にセットする.

4. 配列 $\boldsymbol{n}$ をもとに，nest が最も深い文単位から順に，すべての文単位に対して，いくつかの単語の並び方を変える.

5. 配列 $\boldsymbol{w}$ の各要素に対して，次の挿入，置換，または削除を行う.

5-1. $\xi=1$.

5-2. w_ξ（入力文全体の最初の単語）が名詞でなければ，先頭に「コレハ」を挿入する（主語が省略されている）.

5-3. w_ξ が名詞，$w_{\xi+1}$ も名詞で，かつそれが文単位の先頭である時，「ハ」を間に挿入し，5-7. へ進む.

5-4. w_ξ が名詞，$w_{\xi+1}$ も名詞で，かつそれが文単位の先頭でない時，「ガ」を間に挿入し，5-7. へ進む.

5-5. w_ξ が名詞で，かつ $w_{\xi+1}$ が動詞の時，「オ」を間に挿入し，5-7. へ進む.

5-6. w_ξ が前置詞の時，それより後方の「,」を「ト」と置換して，$k_\xi=-1(k_\xi\in\boldsymbol{k})$ とした後，5-7. へ進む．以下同じ要領で，日本語に翻訳した単語については，配列 $\boldsymbol{k}$ の対応する要素を負数としておく.

5-7. $\xi=\xi+1$.

5-8. $\xi\leqq l$ の時，5-3. へもどる.

6. 配列 $\boldsymbol{w}$ 中の未翻訳の単語（配列 $\boldsymbol{k}$ の対応する要素が非負）を辞書により日本語に置換する．また，辞書中にない単語は，そのままとする．この時点で，配列 $\boldsymbol{w}$ の内容は出力文となる.

4. 応 用 例

医薬品添付文書に実際に現われた例を二つ選び，本法を適用した入力文，出力文を以下に示す.

(a) RECOVER (PAIN , SHUCHOO , ENSHOO) AT (KOOKUU , INKOO).

コレ ハ KOOKUU , INKOO ノ イタミ ト SHUCHOO ト ENSHOO オ カイゼン スル.

(b) DO EFFECT SELECTIVELY TO (ICHOOKAN , TANKAN , NYOOROKEI) : RECOVER (PAIN BY KEIREN) , ; INHIBIT (ZENDOO PROMOTE).

コレ ハ ICHOOKAN , TANKAN , NYOOROKEI ニ センタクテキニ サヨウ スル. ソノケッカ KEIREN ニヨル イタミ オ カイゼン シ, ZENDOO ノ コウシン オ ヨクセイ スル.

図 1　テストに使用した代表的な例文. (a), (b) それぞれの上段は入力文, 下段は出力文を示す.

Fig. 1　Data of sample runs. In each sample (a) and (b), the upper is the input data and the lower the output.

最初の例文は,「口腔・咽頭の痛み, 腫脹, 炎症を緩解する.」である. もっとも単純な例の一つであるが, いくつかの点に注意を向けなければならない. まず, 原文には主語が省略されているが, 当然,「この薬剤は」が主語であることは明らかである. 次に, 名詞がたくさん並んでおり, 相互の関係を明確にしておく必要がある. つまり, (口腔と咽頭) における (痛みと腫脹と炎症) を緩解する, という文脈を入力文中に明らかにしておかなければならない. 図1 (a)の上段にこの例の入力文を, 下段に出力文を示してある. 出力文では, 字句の多少の相違は当然見られるものの, 原文の意図する内容が完全に再現されている.

第2の例文は「胃腸管, 胆管, 尿路系に選択的に作用し, けいれん性痛みを緩解するとともに, 高進した運動機能を抑制する.」であるが, ごく自然な日本文に翻訳されている (図1 (b)).

5. 検　索

文章をコード化することの第1のメリットは, その論理的な検索が可能になることである. 検索方法自体は, 本研究の目的とするところでないが, コード化によって構文の論理性がどの程度向上するかを最後に検討する.

まず, 異なるメーカより販売されている2種類の糖尿病用薬の説明を選び, 以下の実験を行った. 原文は次の通りである.

a) 膵臓のβ細胞に直接作用して, インシュリンの分泌を促進することにより血糖を下げる.

b) 膵臓とは無関係に血糖低下作用を現わす.

二つの文章より明らかなように, a) と b) の薬剤は作用が異なっている. もし, 作用の記述から前者の薬剤だけを選び出すとすれば, そのキーワードをどのように設定すればよいかを考える. ここでは, 膵臓に「作用する」というフレーズが重要であるが, 実際にはこれらの単語以外にも無数の同義語が述語として使われており,「膵臓に作用しない」ような糖尿病用薬と区別することは簡単でない.

そこで, この二つの例文を 12 名の被験者にコード化してもらい, 結果を検討した. コード化の過程では, 入力文を電算機に入力するたびにその復元処理を行い, 出力文を見ながら必要な修正を加えるという方法をとった. a) の例文に対しては, SUIZO の述語として EFFECT, ACTIVATE のいずれかの単語だけが使われ, b) の例文に対しては, EFFECT, STIMULATE のいずれかと否定を表わす NOT が使われた. したがって, コード化された文章について, a) の薬剤を検索するためのキーワードは {SUIZO ∧ (EFFECT ∨ ACTIVATE) ∧ $\overline{\text{NOT}}$} とすればよく, b) に対しては, {SUIZO ∧ (EFFECT ∨ STIMULATE) ∧ NOT} でよいことになる. キーワードがいく分冗長になってしまうのは, 動詞や名詞として定義した辞書中に同義語がいくつか含まれているからであるが, 復元に際して自然な日本語を得るために, ある程度の同義語は必要である.

次に, 日常よく使われる糖尿病用薬 10 種類 84 品目を対象に,「膵臓に直接作用しない」薬剤を選び出す作業を実際に行った. その結果, 原文のままではキーワードの設定そのものが困難であったのに対し, コード化した場合には前述 b) のキーワードで完全な検索が可能であった.

6. 考察と結論

Waltz[2] や, Hendrix[3] らは, 自然言語 (英語) によるデータベースへの問合せシステムを報告しているが, いずれも特定の構造を持つ文章だけを想定したものである. これに対して, 本法は, 問合せ言語ではなく, データそのもののコード化を目的としたものであることと, 特定の文章構造を想定していない点で, 彼らの研究とは本質的に異なっている. 一方, 医療データのコード化については, Helder[4] ら, Buckley[5] による報告があるが, 前者は使用するコードがすべて数値であるため, コード化の作業が簡単ではない. また, 後者は入力文のコード化を半自動的に行っている

が，自然言語の syntactic および semantic な理解について十分な検討がなされておらず，すぐに実用化する方法ではない．

当コード化規則は，第5章で述べた実験中 12 名の被験者全員が1時間の説明と練習で完全に使えるようになり，実用的な方法であることは証明できた．しかし，入力文の作製時に会話型で修正を行い，正しくコード化されたデータだけをファイルに登録するという本法の性質上，その精度を正解率などの定量的な形で表現することはできなかった．なお，本研究で使用したプログラムは FORTRAN サブルーチンの形をとっている．

（本研究の一部は，昭和54年度新潟県医師会学術助成金によった．）

参 考 文 献

1) 日本医薬情報センター編：日本医薬品集，薬業時報社，東京 (1978)．
2) Waltz, D. L.: An English Language Question Answering System for a Large Relational Database, Commun. ACM, Vol. 21, No. 7, pp. 526-539 (1978).
3) Hendrix, G. G., Sacerdoti, E. D., Sagalowicz, D. and Slocum, J.: Developing a Natural Language Interface to Complex Data, ACM Trans Database Systems, Vol. 3, No. 2, pp. 105-147 (1978).
4) Helder, J. C. and Verweij, H.: Semi-automatic Encoding and Report Generating of Medical Information, MEDINFO 80, pp. 708-712, North-Holland, Amsterdam (1980).
5) Buckley, J. D.: Narrative Report Generation from Numerically Coded Data, Comput. Biomed. Res., Vol. 11, No. 6, pp. 525-536 (1978).

（昭和56年6月3日受付）
（昭和56年9月7日採録）