自然言語処理 83－3
（1991. 5. 16）

文書の整理に関する一考察

福永 博信
NTT ヒューマンインタフェース研究所
fuku@ntthli.ntt.jp

概要

自然言語で書かれた文書は、流通性が非常に良い反面、冗長で曖昧であるという欠点をもつ。また通常、一行目から順次読んでいくというストリーム的な利用を要求される。従って、文書から必要な情報を取り出すためには、常に目的を達成するまで「読む」という作業が必要である。この「読む」作業を軽減し、自然言語で書かれた文書から必要な情報を利用しやすい形態に加工することを「文書の整理」と呼ぶことにする。
本稿では、文書中に記述された知識・情報の抽出，保存の過程をモデル化し、自然言語処理技術を用いて「文書の整理」をおこなう一つの方法を示した。

A Study of Text Pigeonhole

FUKUNAGA,Hironobu
NTT Human Interface Laboratories.
1-2356 Take Yokosuka-Shi Kanagawa 238-03 JAPAN
fuku%ntthli.ntt@relay-cs.net

Abstract

Texts are increasing more and more. It is not so ease to use texts for intelligent works because we must "read" them. To "read" consists of three steps: natural language analysis, interpretation, and memory merging. In this paper, I discussed a text pigeonholing method as a support of human's intelligent activities. In short, the method is to pigeonhole acquired knowledge from texts using abstract knowledge and a thesaurus. Both acquired knowledge and abstract knowledge are presented in semantic network model. A method of making abstract knowledge from texts and a thesaurus is discussed.

1 はじめに

自然言語で書かれた文書は、その記述法自身がその言語を使用する人の間では「常識」であるため、非常に優れた流通性をもつ。しかし、そこから情報を得ることを目的とする場合、一行目から順次読んでいくというストリーム的な利用を要求される。従って、目的を達成するまで冗長な文書を読み続けるという作業を行なわなければならない。

一方、近年の新しいメディアや計算機ネットワークの普及とともに電子化文書の流通量は大きく増加している。しかし現状では、計算機を使用して蓄積したものを検索して提示するという原始的な利用しかされていない。すなわち、大量の文書中から情報を獲得するのに必要な材料を選び出すことは支援するが、「読む」作業は人間側の仕事として残されている。

本稿では、人間が情報の取得の際に、文書を「読む」作業を軽減するために「文書の整理」について考え、これを自然言語処理技術を使って実現する一つの方法を示す。2節では、文書の利用のされ方について考察し、3節では知識のモデルについて、4,5節では文書からの情報(知識)の獲得・整理とそれに用いる上位知識の獲得について述べ、6節では、4節の獲得・整理に必要な自然言語処理について示した。

2 文書の利用

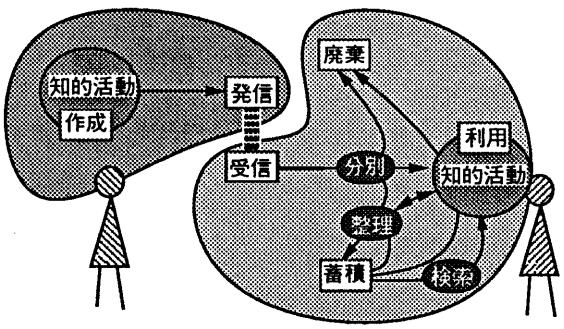

図 1: 文書のライフサイクル

文書は、人間相互間の情報伝達の手段として用いられるものである。文書のライフサイクルを下記の6つの状態の遷移で考えることができる。(図1参照)

作成 元側の知的活動の結果として作成される。

発信 元側から受側に向かって発信される。

受信 受側で受信される。

蓄積 受信されたものが蓄積される。受側で利用された後、再編集して蓄積される場合も考えられる

廃棄 受信あるいは蓄積されたものが廃棄される。受側で利用された後、廃棄される場合も考えられる

利用 受信あるいは蓄積されたものが受側の知的活動に利用される。

受側では、文書は利用されてはじめて有効なものとなる。すなわち、受信から利用までの間の文書は役に立っていない。そこで、その間の人間の作業を支援するために次のような研究が行なわれている。

分別 受信した文書を、必要な文書がどうかを判断して廃棄するか蓄積するかを決定する作業が分別である。これを自動的に行なうのが filtering 技術である。

分類 文書の内容に応じて分類して蓄積する技術も研究されている。

検索 蓄積した文書の中から必要とするものを探す作業が検索である。これを精度・効率良く実行するための技術が文書検索技術である。

これらの技術によって、文書を利用するために行なわなければならない作業を軽減することは

できるが、最終的に人間に提供されるものは自然言語で書かれた文書である。従って、文書中に記述されている目的の情報に到達するためには、人間が文書を「読む」必要がある。文書の利用の第1ステップである「読む」作業は人間に残されているのである。次のような場合には、従来技術による支援の限界を越えている。

- 絞りきれない場合
  検索においては、どの手法を用いる場合でも適合率に限界がある。検索の条件によっては、適切な絞り込みが行なわれず多数の文書を出力してしまう場合がある。

- 比較を目的とする場合
  比較のための資料となる文書を取り出すことができた場合でも、個々の文書を読み、比較項目となる部分を人間の頭の中で整理しなくてはならない。

人間が文書を「読む」作業を次のように考える。(図2参照)

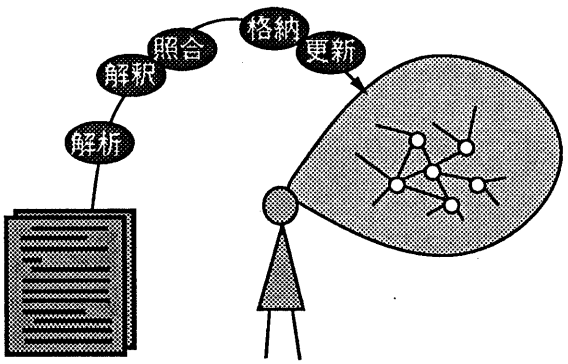

図 2: 文書を「読む」

人間は自然言語で書かれた文書を読んで解析し、自分の記憶構造[1]と照合しながら解釈し、その結果を自分の記憶構造中に知識として格納あるいは既存の知識を更新する。人間は自然言語から自分の記憶構造の中に投射した知識を利用している。

[1]人間の頭(心)の中のメンタルモデル，世界モデルを記憶構造と表すことにする。

本稿では自然言語処理技術を用いて、この「読む」作業を軽減する技術の研究を提案する。次節以降、人間が必要とする情報を獲得・蓄積する過程をモデル化し、それを計算機に支援させる方式について述べる。

3 概念・知識・記憶構造

人間は自分の経験を通して獲得した知識から構成される記憶構造を持っている。知識を概念相互間の関係ととらえ、記憶構造を概念を節とする網である意味ネットワークでモデル化すると、知識獲得は、記憶構造の網を拡張・修正することと考えることができる。その場合一つの知識は、数個の概念とそれらの関係であり、その知識を記憶構造に加える際の概念の新規導入・削除，概念間関係の追加・変更・削除が知識獲得である。獲得された知識は単一の知識として網に組み込まれ、その量の増加にともなって網は大きくなる。このモデルにおいては、知識相互間の関係は別の知識によってのみ言及されるため、類似事項の推論・類推は全て処理系に負わせることになる。

一方人間が情報処理に用いるのは、通常、文書中に記載されている知識の一部分である。また、複数の文書中から条件に合った類似の知識を抜きだしてそれらを対比するような場合も多い。そこで、必要な知識を記述するための抽象的な知識を規定し、それに従って知識を整理することを考える。

知識のモデルには、概念シンボルの節と、概念間関係シンボルを付与された弧で表現される意味ネットワークを採用する。従って、知識は次のように表すことができる。

$$\text{知識 } K = < C, R, \Gamma >$$

C: 知識を構成する概念シンボルの集合

R: 知識を構成する概念間関係シンボルの集合

Γ: 知識を構成する2つの概念とその間の概念間関係で表される単位知識の集合

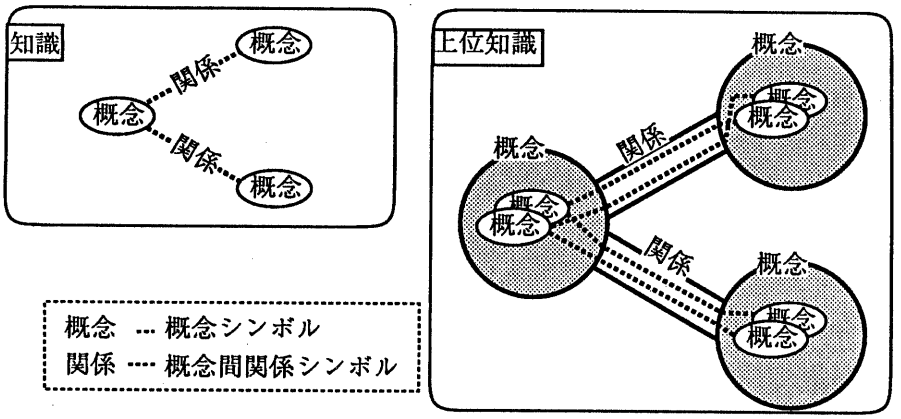

図 3: 知識・上位知識

概念には、いわゆるシソーラス的な上位・下位の体系がある。すなわち、概念を feature の集合で記述できたとすると、概念 A，概念 B の feature 集合をそれぞれ F_A,F_B で表す時、$F_A \subset F_B$ の関係がある場合に、概念 A は概念 B の上位概念，逆に概念 B は概念 A の下位概念である。同様に、概念間関係にも上位・下位の体系を考えることができる。知識の上位・下位は、その構成部品である概念と概念間関係に上位・下位の体系に従って次のように考えることができる。

> 知識 α=$< C_\alpha, R_\alpha, \Gamma_\alpha >$, 知識 β=$< C_\beta, R_\beta, \Gamma_\beta >$ において、次の条件を満たす時、知識 α は知識 β の上位知識である。
>
> 1. $c_{\alpha i} \preceq c_{\beta i}$
> $(c_{\alpha i} \in C_\alpha, c_{\beta i} \in C_\beta;$
> $1 \leq i \leq NUM(C_\alpha))$
> 2. $r_{\alpha i} \preceq r_{\beta i}$
> $(r_{\alpha i} \in R_\alpha, r_{\beta i} \in R_\beta;$
> $1 \leq i \leq NUM(R_\alpha))$
> 3. $\gamma_{\alpha i} = (c_{\alpha j}, r_{\alpha k}, c_{\alpha l})$ に対して $\gamma_{\beta i} = (c_{\beta j}, r_{\beta k}, c_{\beta l})$ である。
> $(1 \leq i \leq NUM(\Gamma_\alpha))$
> 4. $c_{\alpha i} \prec c_{\beta i}$ または $r_{\alpha j} \preceq r_{\beta j}$ であるシンボルの対応が少なくとも 1 つは存在する
> 5. $NUM(C_\alpha) = NUM(C_\beta)$,
> $NUM(R_\alpha) = NUM(R_\beta)$,
> $NUM(\Gamma_\alpha) = NUM(\Gamma_\beta)$
>
> ただし、A $\prec$ B は A が B の上位，A $\preceq$ B は A が B に等しいあるいは上位のシンボルであることを意味し、NUM(A) は集合 A の要素の数を意味する。

上位の知識は、下位の知識を抽象化した知識である。従って、上位知識に従って下位知識を整理することができる。

4 知識獲得・整理

3節に従うと、知識は数個の概念とそれら相互間の関係であるから、文書から知識を獲得する作業は、言語処理技術を適用して

(1) 文章から概念を抽出する

(2) 文章から概念間の関係を抽出する

の 2 つを行なうことと規定される。従って、この種の知識獲得器の能力は、

1.1 取り扱える概念の種類

1.2 文書中の表現と概念との照合精度

2.1 取り扱える概念間関係の種類

2.2 文書中の表現と概念間関係との照合精度

の4つで測ることになる。

知識整理は、獲得した知識を上位知識で束ねることである。その作業は、

(3) 上位概念・上位関係との照合

(4) 上位知識の枠で整形する

の2つを行なうことと規定される。従って、この種の知識整理器の能力は、

3.1 概念，概念間関係体系の充実度

3.2 上位の概念，概念間関係との照合精度

4 上位知識と人間の知識整理モデルとの間の距離

の3つで測ることになる。

上述の知識獲得・整理器の動作手順は次のようになる。(図4参照)

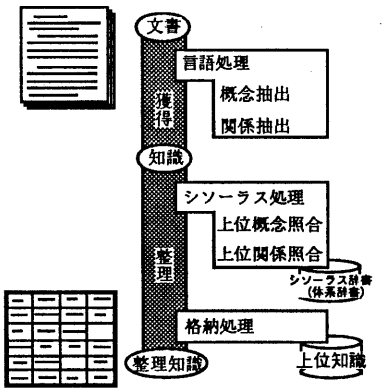

図 4: 知識獲得整理手順

1. 言語処理
   文書に対して言語処理を施し、知識 ($=< C, R, \Gamma >$) の集まりを抽出する。

2. シソーラス処理
   シソーラス辞書を参照して抽出した概念，概念間関係の上位概念，上位関係を得て各知識の上位知識を同定する。

3. 格納処理
   上位知識ごとに、知識を分けて格納する。

一般にシソーラス辞書とは、「言葉」の体系を収録した辞書であり、厳密には「概念」の体系とは異なるが、近似的に「概念」の体系として用いる。*1.1*, *1.2*, *2.1*, *2.2*は言語処理系、*3.1*は概念体系、*3.2*は概念体系と上位知識、*4*は上位知識の完成度に依存する。

次に知識獲得・整理の例を示す。例で使用する文書，シソーラス辞書，上位知識を次に示す。[2]

文書:

文書1	…A社が電話器を開発した。…
文書2	…B社が小型無線機を実用化した。…

シソーラス辞書:

概念 (単語)	上位語
A社, B社	製造会社
実用化, 開発	作る
電話器, 小型無線機	通信機器

上位知識:κ

C(C_κ)	{ 製造会社, 作る, 通信機器 }
R(R_κ)	{ 動作主, 対象 }
F(Γ_κ)	{(作る, 動作主, 製造会社), (作る, 対象, 通信機器)}

上記の文書，シソーラス辞書，上位知識をもとに次の手順で知識を獲得・整理する。

1. 言語処理:
   文書1より知識 α:

C(C_α)	{A社, 電話器, 開発 }
R(R_α)	{ 動作主, 対象 }
Γ(Γ_α)	{(開発　動作主　A社), (開発　対象　電話器)}

   文書2より知識 β:

C(C_β)	{B社, 小型無線機, 実用化 }
R(R_β)	{ 動作主, 対象 }
Γ(Γ_β)	{(実用化　動作主　B社), (実用化　対象　小型無線機)}

   が抽出される。

2. シソーラス処理
   1で得た知識の各概念シンボルの上位概念をシソーラス辞書から得て、それぞれの上位知識を探索する。知識 α, 知識 β は

[2]簡単のため概念間関係の体系は導入しない。

ともに κ を上位知識としてもつことがわかる。

3. 格納処理:
κ を上位知識とする各知識を κ の概念シンボルに対応づけて格納する。

作る	企業	通信機器
開発	A社	電話器
実用化	B社	小型無線機

5 上位知識獲得

上位知識は知識整理に必須であるが、それを準備することは比較的困難である。それは、次のようなことに起因する。

(1) 知識獲得系と同一の概念体系，概念間関係で記述する必要がある。

(2) 多層をなす概念の選択が困難。

そこで知識獲得に用いるものと同一の言語処理系とシソーラス辞書を用いて、次のような手順で構築する方法を提案する。(図5参照)

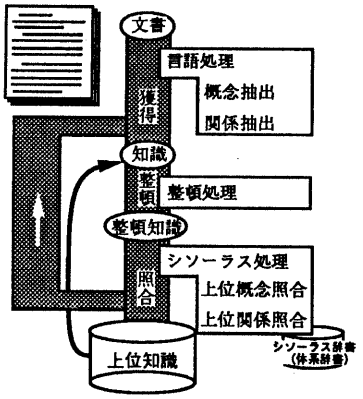

図 5: 上位知識獲得

1. 言語処理
文書に対して言語処理を施し、概念と関係で表される知識に分解する。

2. 整頓処理:
1で得た知識の集まりに対して、ある概念を中心に整頓する。

3. シソーラス処理:
シソーラス辞書を参照して、同一または共通の上位関係をもつ概念間関係にある概念群に適合する上位概念があれば、上位知識として獲得する。

4. 3で得た上位知識を文書から獲得した知識と同様に扱いながら、2,3を繰り返す。

前述のとおり、シソーラス辞書は概念体系の近似である。従って、同義語などの取り扱いなどに注意を要する。

次に上位知識獲得の例を示す。4節の例で使用した文書とシソーラス辞書をもとに、次の手順で上位知識を獲得する[3]。

1. 言語処理:
文書1より知識 α，文書2より知識 β を抽出する。(α と β は4節の例参照)

2. 整頓処理:
「作る」に着目し、その下位概念を含む知識を集める。「開発」($\in C_\alpha$)，「実用化」($\in C_\beta$) は「作る」の下位概念であり、知識 α, 知識 β は条件を満たす知識として集められる。以下、着目した概念(ここでは「作る」)を着目概念，着目概念の下位概念(「開発」「実用化」)を着目下位概念と表す。

3. シソーラス処理:
知識 α, 知識 β において着目下位概念に対して概念間関係「動作主」である「A社」「B社」は共通の上位概念「製造会社」をもつ。同様に概念間関係「対象」である「電話器」「小型無線機」は「通信機器」を共通の上位概念にもつ。従って、上位知識

C	{製造会社, 作る, 通信機器}
R	{動作主, 対象}
Γ	{(作る, 動作主, 製造会社), (作る, 対象, 通信機器)}

が獲得される。

[3]簡単のため、概念間関係の体系は導入しない。

6　知識整理と言語処理

日本語における言語処理は、形態素解析，係り受け解析，構文解析，意味解析をパイプライン的に接続して行なうものが多い。知識の抽出という観点から言語処理を考えると、形態素解析，係り受け解析，構文解析の各段の解析処理の結果から行なえる知識抽出処理が存在する。(図6参照)

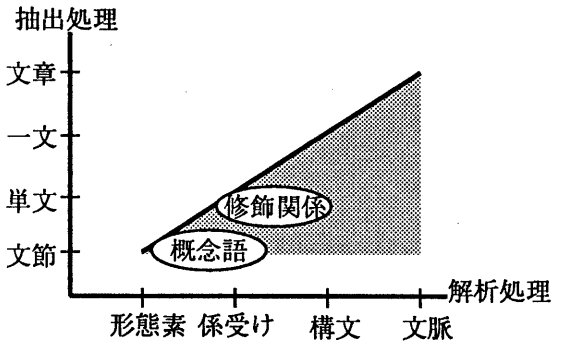

図 6: 知識抽出と言語処理

解析処理と抽出処理が可能範囲の対応は、形態素解析では文節内，係り受け解析では単文内，構文解析では一文内，文脈処理では文章内となる。抽出範囲と抽出可能な概念，概念間関係は次のようになる。

- 文節内
  言葉(概念語)が直接表す概念，あるいは付属語によって屈折した概念が抽出可能である。概念間関係は文章中での共起という関係が抽出可能である。

- 単文内
  直接の修飾関係による概念間関係の抽出が可能である。

- 一文内，文章内
  一文内，文章内の修飾関係による概念間関係の抽出が可能となり、さらに省略補完や照応処理による同一概念の同定により、より多くの概念間関係を得ることができる。

概念系の方から言語処理を概観すると次のようになる。

- 数量
  数詞処理と上下限や幅表現の処理が問題となる。数詞の取り扱いは文節内で処理可能であるが、上下限・幅表現の処理は、単文以上の処理が必要である。

  数詞表現:　数十メートル
  幅表現:　3～4,3から4まで
  上下限表現:　6未満

- 時間
  上述の数量の問題にあわせてテンス・アスペクトを含む時間軸上の処理が問題となる。テンズ・アスペクト処理のためには、文節内の付属語(助動詞，補助動詞)の処理が必須である。この他、時間軸上での配置を決定するためには、一文内あるいは文章内での時間表現の処理が必要である。

  テンス:　読む, 読んでいた(文節内)
  アスペクト:読みはじめる(文節内)
  　読むことを始める
  時間軸表現:～後、…(一文内)
  　～。その前に…(文章内)

- 空間
  上述の数量の問題にあわせて、方向，起終点などの空間表現の処理が問題となる。文節内での助詞の処理で対応できるものと単文以上での空間表現が必要なものとがある。

  方向:　上方, 上の方(文節内)
  起終点:　地面から(文節内)

- その他
  因果や意図の表現などにも、単文内，一文内，文章内それぞれで取り扱える範囲がある。

  因果:　雨により中止する(単文内)
  　走ったので苦しい(一文内)
  　～。その結果、…(文章内)
  意図:　早起きのため早寝する(単文内)
  　早く起きるために～。(一文内)
  　～。その目的は…(文章内)

対象とする文書の分野，概念によって、知識獲得・整理器が必要とする言語処理は異なる。実用システムを考える際には、システム構築と処理の経済性を考慮して言語処理の対象範囲を決定しなくてはならない。

7 おわりに

計算機で取り扱える文書データは増大し、比較的容易に入手できるようになっている。しかし、文書を利用した人間の純粋な知的活動の前に行なわなければならない「読む」作業は、人間にとって比較的大きな負荷である。これを軽減することができなければ、大量の文書を利用できるようになってもそれを有効に活用することはできない。そこで、人間の整理モデルと近い状態に整理された形態で文書を提示・保存すれば人間はより高速にあるいは高度な知的活動ができるという発想のもとに文書整理に関する研究を提案した。また、知識獲得・整理器の能力を評価するためのポイントと言語処理側で解決すべき問題との対応を示した。今後は、その評価の尺度を考えるとともに、本モデルの有効性を確かめるための実験を行う予定である。

謝辞

有意義な議論を頂いた 中川透主幹研究員をはじめとするヒューマンインタフェース方式研究部の方々に感謝いたします。

参考文献

[1] 長尾真:「言語工学」(昭晃堂)

[2] 長尾真監修:「日本語情報処理」(電子通信学会)

[3] Harry Tennant: Natural Language Processing,1981. 森健一他訳「自然言語処理入門」産業図書

[4] 水谷静夫他:「文法と意味I」(朝倉書店)

[5] 日本語教育学会:「日本語教育辞典」(大修館書店)