# シソーラスと概念間の距離に基づいた意味内容の抽象化*

永松健司　田中英彦†
東京大学大学院 工学系研究科

## 1　はじめに

高度な自然言語処理を行う場合、常に問題となるのが、対象ドメインに対する知識をいかに得、利用するかという問題である。要約処理においては、スクリプトや物語文法、解説文など特定の文章形態に固有の構造知識を処理の前提とすることによって、より一般的な処理を指向する反面で、適用対象となるテキストの範囲を限定してしまっている。また、これらの知識は人間が前もって抽出したものを与えておく必要もある。

この問題に対して、大規模な知識ベースを構築しようとするプロジェクトも存在するが、まだ現実世界の問題に適用するには不十分と言わざるを得ない。この一方で、辞書、シソーラス、コーパスは着実に整備が進み、現在においても十分な質のものが利用できる状態になってきている。

我々の目的は、現段階で利用可能なこれらのリソースのみを用いて、必要とされる検索・要約処理を行うということであり、本稿では、コーパスから抽出した知識に対してシソーラスに基づく概念間距離を用いることで、要約処理に必要な意味内容の抽象化を試みる。

## 2　シソーラス・概念間距離と抽象化

### 2.1　抽象化による要約

要約 (抄録) 過程で必要な処理には次の 2 つがある。

1. **有用な情報の選択と不要な情報の削除**

   例「それに比べ役人の意識は、人びとの心にわき上がる隣人愛に柔軟に応じるには、遅れているように思われる。」[1]

   ⇒『役人の意識は隣人愛に応じるには遅れている。』

2. **複数の語句の意味を 1 つで言い表す語句へと変換する抽象化**

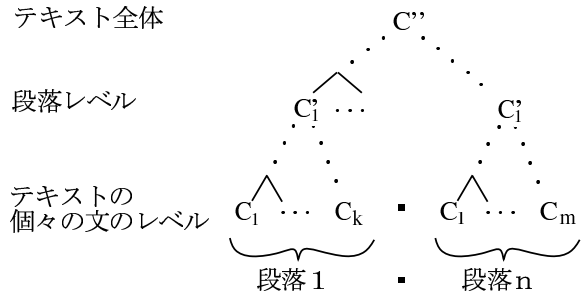


図 1: 再帰的な抽象化による、テキストを端的に表現する格フレームへのまとめ上げ

例「がれきの中に潜って犬と一緒に生存者を捜すスイスやフランスの救助隊員がいた。ヘルメット姿のカナダ大使は現地で避難住民のためのテント張りを指揮した。」

⇒『ヨーロッパからの人々が救助活動をした。』

この内、抽象化とは記述のレベルを 1 つ上げることを意味し、要約の質や要約率を高めようとする場合には欠かせない処理である。

また、この抽象化処理を再帰的に行うことにより、個々の節 (文) に対応する格フレームから順にまとめ上げられ、(究極的には) そのテキスト全体の意味を端的に表す一つの格フレームが取り出すことができる (図 1)。もちろん、実際のテキストでは、抽象化する処理には高度な知識・判断を要する場合が多く、そこでは本稿で述べる手法では不十分なことも多いと思われるが、本稿の目的は現在利用可能なリソースでどれだけ適切な処理が可能かを見ることであるため、ここでは考慮しない。

### 2.2　コーパスから抽出する知識

単にシソーラスだけを利用して、入力となる複数の概念から、その共通上位概念を取り出すという抽象化を行うのでは、あまりにも抽象的な (字義通り) 概念のみが得られることになり、全く意味を成さなくなる。

**例「(捜索犬を連れた救助隊は、) 地震発生の日から現地入りした東京消防庁の救助隊ほどの成果はあがらなかったが、捜索犬の能力を存分に見せてくれた。」**

---

*A definition of the distance between two semantic structures and an evaluation of it

†Kenji Nagamatsu　Hidehiko Tanaka
{naga,tanaka}@mtl.t.u-tokyo.ac.jp
Faculty of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo, 113, Japan

[1]本稿で引用している例文はすべて [2] から採ったものである。

に含まれる 2 つの格構造[2]の抽象化を考えると、

$$\left(\begin{array}{l} \text{上げる}\ (3cee1a) \\ agt:\text{救助隊} \\ obj:\text{成果}\ (f922b) \end{array}\right),\quad \left(\begin{array}{l} \text{見せる}\ (30f86f) \\ agt:\text{救助隊} \\ obj:\text{能力}\ (faf70) \end{array}\right)$$

に対して、

$$Abs \Rightarrow \left(\begin{array}{l} \text{行為}\ (444dd9) \\ agt:\text{救助隊} \\ obj:\text{状態}\ (3aa96a) \end{array}\right)$$

となり、「救助隊が状態を行為した」では何も言っていないに等しい。つまり、シソーラスは概念間に本質的な関係のみを記述した知識ベースであり、それらの概念間の手続き的な知識、例えば「打つ」と「走る」に対して「野球」の一場面であることを示す、スクリプト的な知識を得ることはできない。

これに対して、前もって以下のような形の情報をコーパスから抽出しておく。

(格フレーム、または概念)$^+$, 例示語句
$\rightarrow$ 格フレーム、または概念

これは、例示語句 (「など」、「という」 etc.) の後の語句は、それ以前の複数の語句を何らかの観点でまとめた内容を表すからである。

現在、EDR コーパスを利用して、この知識情報を抽出することを考えている。

### 2.3 意味内容の抽象化

本稿で述べる「抽象化」の処理は次の通りである。

1. $C_1,\ldots,C_n$と、コーパスから得られた知識情報の前件部を構成する格フレームとの概念的距離[3]をそれぞれ求め、その距離が最も近い知識を選択し、抽象化の結果としてその後件部の格フレーム (または概念) を与える。ただし、$C_1,\ldots,C_n$と前件部の語句との対応関係および、前件部と後件部の語句の対応関係に従って、出力される格フレーム $C'$内の語句は $C_1,\ldots,C_n$の対応する語句で置き換えられる。
2. $C_1,\ldots,C_n$がそれぞれ単一の概念の場合、その抽象化 $Abs(C_1,\ldots,C_n)$ はすべての概念の共通上位概念 $C'$を与える。

$$Abs(C_1,\ldots,C_n) \Rightarrow C'$$

3. $C_1,\ldots,C_n$がそれぞれ格フレームの場合、その抽象化 $Abs(C_1,\ldots,C_n)$ は、それぞれの格フレーム内で共通する格要素について、抽象化した結果を格要素とする格フレームを与える。

$$Abs(C_1,\ldots,C_n) \Rightarrow \left(\begin{array}{c} Abs(V_1,\ldots,V_n) \\ \text{格}_1 : Abs(C_1^1,\ldots,C_n^1) \\ \vdots \\ \text{格}_k : Abs(C_1^k,\ldots,C_n^k) \end{array}\right)$$

ここで、$C_i = (V_i,\quad \text{格}_1 : C_i^1,\ldots,\text{格}_k : C_i^k,$
共通でない格 : $C_i^{k+1},\ldots)$

この内、2 と 3 は、入力情報の意味が近い場合、共通上位概念を求めるだけでも十分な結果が得られる可能性があること、入力とコーパスから抽出した知識との関連が薄い場合への対処を考慮した。

また、余りにかけ離れた上位概念に抽象化されることを避けるために、概念間距離のスレッショルド $T_D$ を設定し、上のそれぞれの場合において、

1. $\max_i D(C',C_i) \geq T_D$ のとき
2. $D(C',C_i) \geq T_D$ のとき
3. $Abs(V_1,\ldots,V_n)$ の結果が棄却されたとき

には、その結果を棄却し、抽象化は失敗するものとする。さらに、

- 動詞概念 $Abs(V_1,\ldots,V_n)$ の取りうる格要素と、$Abs(C_1^j,\ldots,C_n^j)$ の間で選択制限違反が生じたとき

には、格要素 $Abs(C_1^j,\ldots,C_n^j)$ は結果から削除する。

## 3 おわりに

本稿では、要約処理において求められる意味内容の抽象化処理を、対象領域の知識に依らず、シソーラスとそれを用いた概念間距離だけで行う方法を提案した。

現在は、コーパスの解析など、定量的な評価を行う準備を行っている。さらに、今後は、人間が許容できる抽象化の範囲を調べ、概念間距離のスレッショルドとの関係を求める必要がある。妥当な結果が得られるようなスレッショルドを定めることで、従来の要約処理ではドメイン知識に依存していた抽象化の処理をより一般的な範囲へと広げることができると思われる。

## 参考文献


[1] 永松, 田中. 視点情報を前提とした意味構造間距離の定義とその評価. 情報処理学会全国大会, pp. 3.11–3.12, 3 1996.

[2] 朝日新聞. 社説「世界に広がる隣人意識」, 1995 年 2 月 7 日分.


[2]ここでは簡単のため、埋め込み文、名詞句の係り受け、時制・態 etc. は無視する。

[3]格フレーム間の距離関係としては [1] に基づいている。