

4 A-7

出来事型情報の構造化*

稲垣 博人、中川 透†

NTTヒューマンインタフェース研究所‡

1 はじめに

我々は、人間や機械との間で行なわれる情報伝達行為において、伝達される情報構造の解明が、ヒューマンインタフェースの研究には必須であるとして、情報の構造化の研究を進めている[1, 2]。特に、言語表現から、現実世界に起きた(る)出来事の構造を得ることを目的として、言語情報の構造化を行ってきた。本稿では、このような出来事の構造を言語情報から抽出するための構造化アルゴリズムと、構造化で使用する言語知識の設計思想、および言語知識の評価について述べる。

2 出来事型情報

2.1 出来事型情報の構造

出来事型情報とは、現実世界における出来事の内容を表現するための構造である。

現実世界では、出来事は動作主体によって実行され、その動作により、ある対象が変化する。そして、動作(変化)は、ある時間と場所を支配すると考えると、出来事型情報における最小の要素は、動作主体、対象、動作、時間、場所の5つとなる。

出来事型情報の構造化では言語で表現された意味構造を出来事型情報に写像することにより、自然言語表現からこれら5つの出来事型情報の要素を決定する。

2.2 構造化アルゴリズム

構造化で用いる言語知識は、一般によく用いられる意味素性と、格パターンに相当するような意味的優先制約を用いる。また、言語知識は、すべての語に対して記述するのではなく、処理に必要な最小限の語について記述することにする。その場合、語が連結して、複合語、文節、文になる際に、言語知識の演算が必要となる。意味素性の場合、意味的合成演算を行なうが、この演算は、意味素性が同じ場合は、その値を演算結果として渡し、意味素性値が一致しない場合は、もっとも上位ノードの意味素性を優先して渡す。

構造化アルゴリズムでは、意味的合成演算や種々の演算が行なえるように、属性文法的枠組を用いた。

まず、入力された文(すでに出来事単位に分割された文)を形態素解析、係り受け解析を行ない、単一の

出来事文の木構造を取得する。素性構造で記述された言語知識は、その木構造の終端に位置する語に付与される。

構造化処理では、木構造の各ノードに対して、意味的合成演算を行ないながら各ノードの意味素性を決定し、出来事型情報の要素を意味的優先制約により決定する。

例えば、"大阪にあるABC社は..."という文を処理する場合、図1のような木構造を得る。初期状態で

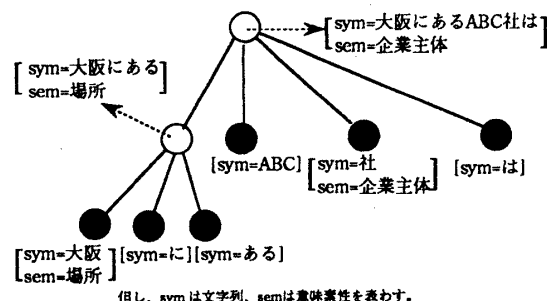


図1: 木構造の計算例

は、各ノード終端の語に対してのみ言語知識が付与される。文節"大阪にある"では、単語"大阪"が意味素性: 場所を持ち、文節全体の意味的合成演算により、文節"大阪にある"は、意味素性: 場所を持つことになる。"大阪にあるABC社は"の全体を計算する場合、"大阪にある"という文節と"ABC"、"社"、"は"の意味的合成演算を行なうことになる。この場合、すべてのノードの意味素性が一致していないので、上位ノードの意味素性優先規則により、全体の意味素性は、"企業主体"という値に設定される。このような演算をすべての木構造に対して行ない、各ノードの意味素性等を決定する。

2.3 構造化に必要な言語知識

出来事の構造化においては意味素性と意味的優先制約の2つの言語知識を用いる。

例えば、企業が動作主体となって生じさせた出来事型情報では、意味素性として、企業主体、企業実体(企業が動作を起こす場合の実体)、資本(金銭などの企業の経営を行なうために必要な財産)、人材、商品(企業が動作を起こした対象)と、企業動作、場所、時間を設定し、必要な単語に対して意味素性を付与する。

*Structural Analysis of Event-type Information

†INAGAKI, Hirohito and NAKAGAWA, Tohru

‡NTT Human Interface Laboratories

意味的優先制約は企業の起こす行動のパターン(企業の発生、消滅、存続等)により多少変動するが、以下の表のような対応関係を、意味的優先制約を行なう語(通常は、動作を表す動詞やサ変名詞)に対して記述する。

出来事型情報の要素	意味素性
動作主体	企業主体
対象	企業実体 商品 人材 資本
動作	企業動作
場所	場所
時間	時間

(但し、記号|は選言を表す。)

3 言語知識の設計

実用性の高い処理を実現するためには、保守すべき語数を少なくし、普遍性のある言語知識を用意する必要がある。ここでは、語基よりも大きな単位の基本的な物事概念を表す語(基本概念語と呼ぶ。例えば、“車”、“会社”、“建設”等)に対して、言語知識を記述し、基本概念語の組合せで、言語知識が付与されていない語があっても処理できるようにした。

また、固有語については、常識性の高い固有語と常識性の低い固有語に分け、常識性の高い固有語(例えば、地名、国名、会社名)については、最初から言語知識に登録し、常識性の低い固有語(例えば、ある特定の商品名等)については、文章の特徴的表現を用いて推定する。

3.1 基本概念語の記述

基本概念語とは、現実世界に存在する物(動作)に与えられた文字表現の中で最も基本的な概念を表す語を意味する。例えば、“車”は、丸いタイヤをエンジン等の原動機で回すことによって移動するものを表す基本概念語であり、“牽引車”、“新車”などは、その基本概念語の派生語として捉えた。基本概念語以外の物の概念を修飾するような語に対しては言語知識には登録しなかった。

動詞、名詞、接頭語、接尾語の約6万語の中から企業を動作主体とする場合の基本概念語を抽出したところ、約8000語(企業動作関係が約1000語)が抽出された。

3.2 固有語の推定

言語知識には、常識性の高い固有語(例えば、地名、国名、企業名等)は最初から知識として登録した。一方、商品名などのように常識性が低く、出現頻度の低い固有語については、文章中から固有語を抽出し、意味素性を決定した。

新聞のようなマスメディアでは、常識性の低い固有語は特徴的なレトリックを用いて表現されることが多い。そこで、意味素性として企業主体、商品を持つ

単語を、特徴的なレトリックから推定した。例えば、意味素性: 企業主体を表す企業名であれば、企業名の後に括弧書きで、企業の所在地、社長名などが記述される。意味素性: 商品を表す商品名であれば、鈎括弧で括られた表現が用いられる。

3.3 固有語の推定

企業主体と商品の意味素性として持つ固有語を、文章中から自動的に推定できる割合を求めた。新聞記事400件に対して、最初から保持する言語知識(約2万語、企業名と地名)、記事中から抽出した言語知識、基本概念語により実行時に判別した情報に分けてそれぞれの固有語推定率(計算機で抽出した語/人間が抽出した語)を求めた。

意味素性が企業主体の語に対して、レトリックに着目した処理で57%の語を自動的に推定した。また、基本概念語により、実行時に29%の語を推定し、全体として、約91%の企業主体の固有語を適切に抽出できた。

意味素性が商品の語に対しては、レトリックを用いた手法により50%を自動的に推定し、基本概念語により34%の語を推定することができた。

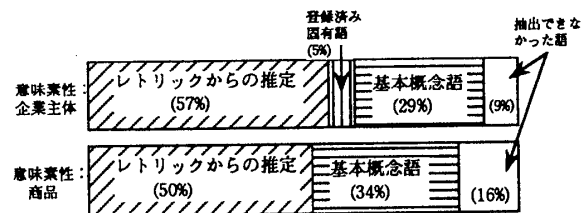


図2: 固有語推定率

4 まとめ

本稿では、入力として自然言語文、出力として出来事の構造を得るような構造化処理における、構造化アルゴリズム、言語知識の構築方法について検討した。知識構築手法として、基本的概念語からの意味素性の合成演算と、常識性の低い固有語に対するレトリックに着目した手法を提案した。この枠組を用いて、400件の記事について固有語推定率を求めた結果、意味素性が企業主体の語について9割、意味素性が商品の語について8割の精度で、自動的に抽出することができた。

ここで提案した知識構築手法は、出来事型の構造化処理だけでなく、一般の格解析等の文章解析にも用いることができる。

参考文献

- [1] 稲垣. 事象解析による要約情報の抽出. 情報処理学会自然言語研究会, NL84-3, 1991.
- [2] 稲垣, 中川. 情報、処理の部分性を考慮した文章解析の実現. 情報処理学会第43回大会, 5H-5, 1991.