2005/5/27 〈NL167-2005〉

文書内に現れる因果関係の出現特性調査

乾 孝司
日本学術振興会 特別研究員
東京工業大学
奥村 学
東京工業大学

背景と目的

- 対話，質疑応答システム⇒推論機構
  - ◆推論規則（因果関係知識）が必要
- 大規模文書から獲得 [Girju02，鳥澤03，Inui04]
  - ◆現状：因果関係の出現特性が不明
  - ◆知識獲得の精度，効率の向上を妨げ
- 目的：文書内に現れる
  因果関係の出現特性を調査

調査手順

1. 因果関係タグ付きコーパスの構築
   - 文書内の因果関係にタグを付与
2. 出現特性調査
   - 付与情報を基にして出現特性を定量的に調査

調査手順

1. 因果関係タグ付きコーパスの構築
   - 文書内の因果関係にタグを付与

- 調査項目とタグづけ方針
- タグ（例による説明）
- タグ付与基準
- 作成したコーパスの概要

調査項目

- 手がかり標識の有無
  - ◆手がかり標識：「ため」,「ので」など
  - ◆手がかり標識が，どれぐらい伴うか？
- 出来事表現（原因, 結果）の統語カテゴリ
  - ◆ＶＰかＮＰか？
- 出来事表現（原因, 結果）の出現位置
  - ◆文末に多いか？
  - ◆原因と結果の相対的な位置関係は？

タグづけの方針

- 表現形式に関する制約を設けず，
  網羅的に，因果関係にタグを付与する

大雨が降ったため、川が増水した。 （明示的）

大雨が降り、川が増水した。 （非明示的）

大雨で川が増水した。 （出来事が名詞句）

大雨が降った。洪水が起こった。 （文をまたぐ）

調査手順

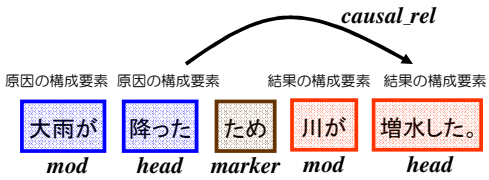
1. 因果関係タグ付きコーパスの構築
   - 文書内の因果関係にタグを付与

- 調査項目とタグづけ方針
- タグ（例による説明）
  - ***head***
  - ***mod***
  - ***causal_rel*** — 因果関係
  - ***marker*** — 手がかり標識

  （***head***, ***mod*** ：出来事）
- タグ付与基準
- 作成したコーパスの概要

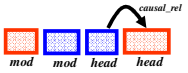
タグ付け例

元文：大雨が降ったため川が増水した。
原因：大雨が降る
結果：川が増水する



タグ付け（整理）

- 文節ごとに
  - 文節の組み合わせ＝出来事
  - 語順の影響を吸収



- 主辞要素を区別
  - 主辞要素の位置 で 出来事の位置を代表させる
- 手がかり標識に ***marker*** タグ

調査手順

1. 因果関係タグ付きコーパスの構築
   - 文書内の因果関係にタグを付与

- 調査項目とタグづけ方針
- タグ（例による説明）
- タグ付与基準
- 作成したコーパスの概要

タグ付け基準

- 言語テンプレートに基づく判断基準
  - 言語的な判断の拠り所を与える
- 言語テンプレート
  - ２つのスロットをもつ文

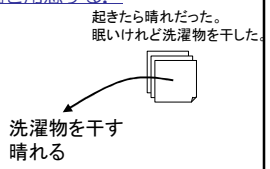
スロット

『結果側出来事』ということをするのは
大抵『原因側出来事』という状況の時である。

タグ付け基準

- 判断の手順

1. 判断したい２つの出来事表現を用意する。

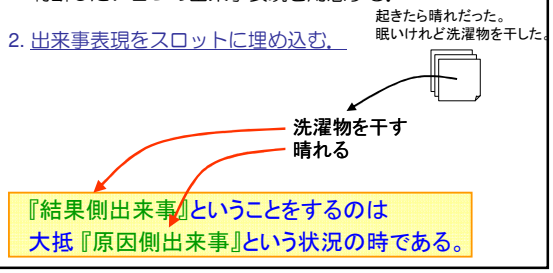


『結果側出来事』ということをするのは
大抵『原因側出来事』という状況の時である。

タグ付け基準

■判断の手順

1. 判断したい２つの出来事表現を用意する.
2. 出来事表現をスロットに埋め込む.

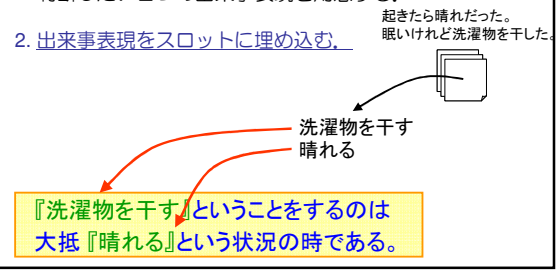


『結果側出来事』ということをするのは
大抵『原因側出来事』という状況の時である。

タグ付け基準

■判断の手順

1. 判断したい２つの出来事表現を用意する.
2. 出来事表現をスロットに埋め込む.

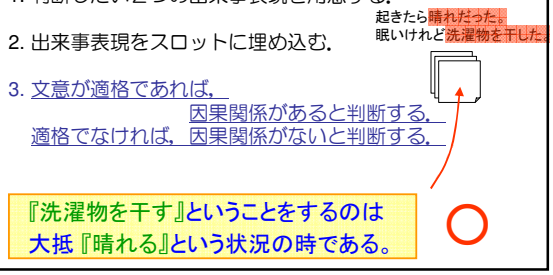


『洗濯物を干す』ということをするのは
大抵『晴れる』という状況の時である。

タグ付け基準

■判断の手順

1. 判断したい２つの出来事表現を用意する.
2. 出来事表現をスロットに埋め込む.
3. 文意が適格であれば,
   因果関係があると判断する.
   適格でなければ, 因果関係がないと判断する.



『洗濯物を干す』ということをするのは
大抵『晴れる』という状況の時である。

○

タグ付け基準

■１８種類のテンプレートを使用
- ◆それぞれに４つのバリエーション
- ◆因果関係の強さを区別する（強い，弱い）
  - ■「大抵」,「しばしば」,「常に」⇒強い因果（蓋然）
  - ■（副詞なし） ⇒弱い因果（偶然）

『結果側出来事』ということをするのは
大抵『原因側出来事』という状況の時である。

調査手順

1. 因果関係タグ付きコーパスの構築
   – 文書内の因果関係にタグを付与

- ■調査項目とタグづけ方針
- ■タグ（例による説明）
- ■タグ付与基準
- ■作成したコーパスの概要

作成したコーパスの概要

■データ：毎日新聞，社会面から７５０記事
■作業者：３名

■記事を読み,
因果関係を発見次第，タグ付け

因果関係が認められた例

原因側		結果側	
mod	*head*	*mod*	*head*
冬型の 気圧配置と	なる		冷え込み
	逮捕する		取り調べ
	急性心不全		死去
交信が	途絶える	安否が	気遣う
頭の　骨を	折る		死亡
線路に	転落	電車	停止
韓国大統領	訪日	交通	規制
琴　調べ	響く		うっとり
東海道新幹線	遅れる	約二十一万人に	影響する
酒を	飲む		酔いつぶれる

（語順は修正）

総数

作業者	総数	記事あたり
A	2014	2.7
B	1587	2.1
C	1048	1.4

- 1000〜2000件の因果関係が認められた
- 作業者間で総数が２倍近く異なる
  - ◆原因１：判断結果が異なる
  - ◆原因２：判断の試行回数が異なる
    - ⇒特に，非明示的な表現では判断自体がされない傾向がある

作業者間の一致度

- 作業者の判断が一致
  - ＝ *head* タグの位置が一致
- ２人以上一致
  - ◆蓋然：0.36（699/1952）
  - ◆偶然：0.24（314/1311）
- 母比率の差の検定
  - ◆H_0「母比率の差がd％である」
  - ◆H_0を棄却
    （$d \leqq 7$ のとき p値$\leqq 0.00805$）
  - ◆「蓋然」の強さをもつ事例の方が信頼できる

一致	A	B	C	蓋然	偶然
1人	1	0	0	632	535
	0	1	0	487	255
	0	0	1	134	207
2人	1	1	0	230	90
	1	0	1	92	77
	0	1	1	107	83
3人	1	1	1	270	64

調査手順

1. 因果関係タグ付きコーパスの構築
   - 文書内の因果関係にタグを付与
2. 出現特性調査
   - 付与情報を基にして出現特性を定量的に調査

対象データ：699件（２人以上が一致 &「蓋然」）

- ■手がかり標識の有無
- ■出来事表現（原因, 結果）の統語カテゴリ
- ■出来事表現（原因, 結果）の出現位置

手がかり標識の有無

- *marker* タグを伴う割合を調査
- ■約７割が標識なし
  - ◆非明示的な表現は判断されない傾向にある
  - ◆標識なし：さらに増える

手がかり標識	頻度
あり	219
なし	480
計	699

出来事表現の統語カテゴリ

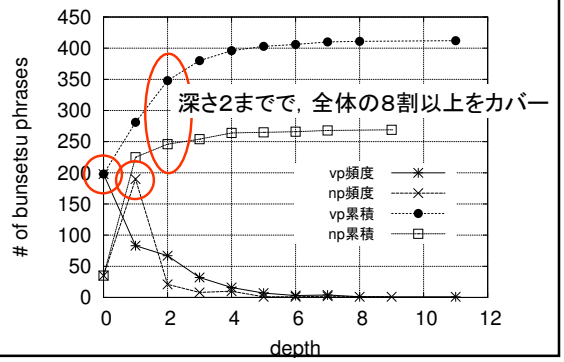
カテゴリ	***head*** の例	原因側	結果側
ＶＰ	焼く 難しい	365	412
ＮＰ	停電 火災	322	269
その他	うっとり	12	18

- ■カテゴリ（ＶＰ，ＮＰ）の割合を調査
- ■原因，結果ともに，ＶＰが過半数
- ■ＮＰもＶＰと同等数存在

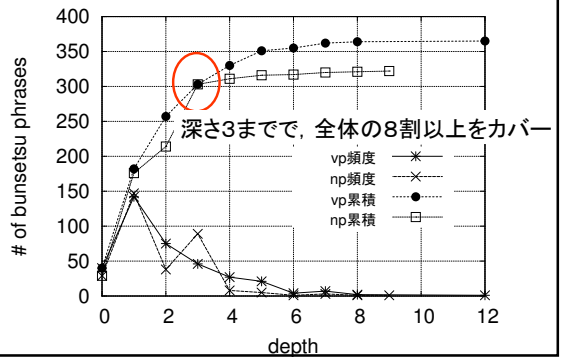
出来事表現の出現位置

- 元文の係り受け木を考える
  （文末文節が根，深さ＝０）
- *head* タグを含む文節の根からの深さ
- 統語カテゴリごとに，深さの出現分布を調査

出来事表現の出現位置（結果側）

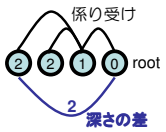


出来事表現の出現位置（原因側）



原因-結果の相対的な出現位置

- 元文の係り受け木を考える



- 原因と結果の
  head タグを含む文節が位置する深さの差

原因-結果の相対的な出現位置

		頻度
文内	深さの差＝１	259
	＝２	152
	＞２	33
	係り受けなし	72
文間		141

（原因が結果よりも深い場合）

- 原因が結果に直接係る場合が最も多い
- ただし，それ以外の位置にも存在

出現傾向調査（まとめ）

- 約７割が標識なしで出現
- 過半数がＶＰ（ＮＰもほぼ同数）
- 原因（文末－１） ⇒ 結果（文末）
- 原因が結果に直接係る

- 既存の知識獲得のカバーする範囲
  - ◆標識あり＋ＶＰ