

(NLP2005)

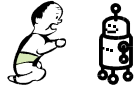
因果関係タグ付きコーパスの構築と分析

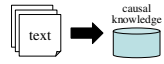
2005/3/15
乾 孝司
東京工業大学 精密工学研究所
COE21「大規模知識資源の体系化と活用基盤構築」
奥村 学
東京工業大学 精密工学研究所

20050315

背景

- 知的な応用処理
 - 例) 対話システム
 - 推論機構の実現が不可欠
- 推論規則(因果関係知識)の自動獲得
 - 大規模テキストを知識資源 [Girju02, 鳥澤03]



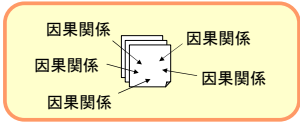


if 大雨が降る then 洪水が起こる

20050315

背景

- テキスト中には多様に因果関係が存在
- 因果関係の出現特性を把握することが望ましい



⇒テキスト中での因果関係の出現特性が不明

20050315

目的

- テキスト中に含まれる因果関係の出現特性を明らかにする

調査手順

- 因果関係タグ付きコーパスの構築
- 因果関係タグ付きコーパスの分析
 - 付与情報を基にした出現特性の定量的調査

20050315

因果関係タグ付きコーパスの構築

20050315

調査項目と方針

- 調査項目
 - 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
- タグ付けの方針
 - 表現形式に関する制約を設けない (できる限り網羅的に)
 - 昨日大雨が降ったため, 洪水が起こった
 - 大雨が降って洪水が起こった
 - 大雪で列車が止まる
 - 飲み終わった缶を捨てる

20050315

因果関係情報を付ける際のポイント

- 発見的なタグ付け
 - 発見過程の作業付加と見落とし
 - ⇒ 提示段階で作業者を誘導
- 因果関係のタグ付け基準
 - 作業者の主観だけでは信頼性に欠ける
 - ⇒ 言語テンプレートに基づく基準

20050315

因果関係タグ

- 構成要素
 - 出来事 (原因, 結果)
 - 因果関係
- 因果関係タグ
 - 出来事 *head, mod*
 - 因果関係 *causal_rel*
 - 2つの格フレーム情報を連結したような形

20050315

タグ付け例

原因: GWIに入る
結果: 遠方からの観光客が増える

遠方からの 観光客が GWIに 入って 増える

mod mod mod head head

原因: OOさん方で出火
結果: △△平方を全焼

文1: OOさん方で 出火

文2: △△平方を 全焼

mod head head

20050315

作業の流れ

1. テキストを提示
2. 因果関係情報を付与
3. タグ部分を抽出、修正

20050315

作業の流れ

1. テキストを提示
2. 因果関係情報を付与
3. タグ部分を抽出、修正

20050315

タグ付け基準

- 言語テンプレートに基づく判断
 - 主観+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

《頻度を表す副詞》『原因側出来事』という状況の時である。

《頻度を表す副詞》= しばしば | 大抵 | 常に | φ

20050315

タグ付け基準

- 言語テンプレートに基づく判断
主観＋言語的な判断の拠り所を与える

判断の手順

- 判断したい2つの出来事表現を用意する。

起きたら晴れだった。
眠いけれど洗濯物を干した。

洗濯物を干す
晴れる

『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

- 言語テンプレートに基づく判断
主観＋言語的な判断の拠り所を与える

判断の手順

- 判断したい2つの出来事表現を用意する。
- 出来事表現をスロットに埋め込む。
(テンプレートが文となる)

起きたら晴れだった。
眠いけれど洗濯物を干した。

洗濯物を干す
晴れる

『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

- 言語テンプレートに基づく判断
主観＋言語的な判断の拠り所を与える

判断の手順

- 判断したい2つの出来事表現を用意する。
- 出来事表現をスロットに埋め込む。
(テンプレートが文となる)

起きたら晴れだった。
眠いけれど洗濯物を干した。

洗濯物を干す
晴れる

『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

- 言語テンプレートに基づく判断
主観＋言語的な判断の拠り所を与える

判断の手順

- 判断したい2つの出来事表現を用意する。
- 出来事表現をスロットに埋め込む。
(テンプレートが文となる)
- テンプレート文の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。

起きたら晴れだった。
眠いけれど洗濯物を干した。

洗濯物を干す
晴れる

『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

- 言語テンプレートに基づく判断
主観＋言語的な判断の拠り所を与える

判断の手順

- 判断したい2つの出来事表現を用意する。
- 出来事表現をスロットに埋め込む。
(テンプレートが文となる)
- テンプレート文の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。

起きたら晴れだった。
眠いけれど洗濯物を干した。

洗濯物を干す
晴れる

『洗濯物を干す』ということをするのは
大抵『眠い』という状況の時である。

×

タグ付け基準

- 頻度を表す副詞(しばしば, 大抵, 常に)の有無
 - 因果関係の強さを区別する
 - 3つの副詞のいずれかをもつテンプレートの場合
⇒ 蓋然(強い)
 - 副詞をもたないテンプレートの場合
⇒ 偶然(弱い)

『結果側出来事』ということをするのは
《頻度を表す副詞》『原因側出来事』という状況の時である。

《頻度を表す副詞》= しばしば | 大抵 | 常に | φ

作成したコーパスの概要

- データ: 新聞記事750記事 (10文以内で構成)
- 作業員: 3名, 並列に独立に
- 言語テンプレート: 18個
- タグ付けの対象の制約
 - 原因と結果のheadが隣接する2文以内にある場合
- メタ作業員による修正は行わない

20050315

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

20050315 (語順は修正)

タグ総数, 重なりと強さの関係

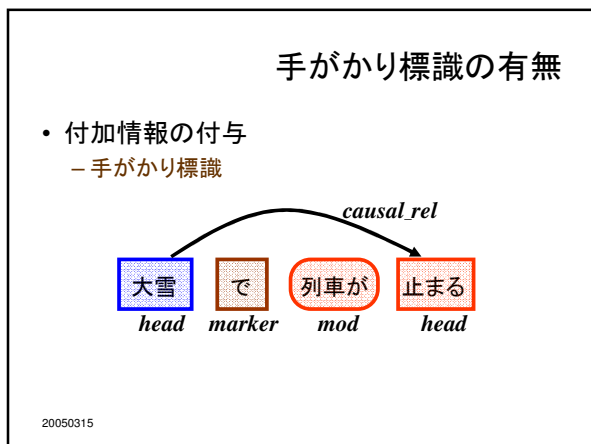
	重なり	A	B	C	蓋然	偶然
総数		1	0	0	632	535
A	1人	0	1	0	487	255
B		0	0	1	134	207
C		1	1	0	230	90
	2人	1	0	1	92	77
		0	1	1	107	83
	3人	1	1	1	270	64

20050315

調査項目

	重なり	A	B	C	蓋然	偶然
(1) 手がかり標識の有無		1	0	0	632	535
(2) 出来事表現の統語カテゴリ	1人	0	1	0	487	255
(3) 出来事表現の出現位置		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

20050315



手がかり標識の有無

手がかり標識	頻度
あり	219
なし	480
計	699

見落とし分 +

内訳	頻度
ため	120
で	35
結果 ので と	5
場合 ば ことから	4
から	3
理由で 目的で 影響で より	2
ように よう として ところが	2
背景には 際に ており	1
ことによって ...	

20050315

原因-結果の 相対的な出現位置

- 文末文節を根とする係り受け木の深さ
- 原因と結果の head を含む文節が位置する深さの差

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

20050315 (原因が結果よりも深い場合)

まとめ

- 因果関係タグ付きコーパスの構築
 - テキスト中に含まれる因果関係の出現特性を調査
 - (1) 手がかり標識の有無
 - (2) 出来事表現の統語カテゴリー
 - (3) 出来事表現出現位置

今後の課題

- 規模
- その他の情報付与
 - 共参照情報など

20050315

- 1 『e1』(という) ことが起こるその結果として、<<adv>>『e2』(という) ことが起こる。
 - 2 『e1』(という) 状態になれば、それに伴い、<<adv>>『e2』(という) 状態になる。
 - 3 『e1』(という) 状態になれば、それに伴い、<<adv>>『e2』(という) 状況になる。
 - 4 『e1』(という) 状態であると、<<adv>>『e2』(という) 状態である。
 - 5 『e1』(という) 状態であると、<<adv>>『e2』(という) 状況である。
 - 6 『e1』(という) ことをする結果、『e2』(という) ことが<<adv>> 起こる。
 - 7 『e1』(という) ことをすると、<<adv>>『e2』(という) 状態になる。
 - 8 『e1』(という) ことをすると、<<adv>>『e2』(という) 状況になる。
 - 9 『e1』(という) ことをすると、<<adv>>『e2』(という) 状態を保つ。
 - 10 『e2』(という) ことをするのは、<<adv>>『e1』(という) 状態の時である。
 - 11 『e2』(という) ことをするのは、<<adv>>『e1』(という) 状況の時である。
 - 12 『e1』(という) 状態になる場合、<<adv>>『e2』(という) ことをする。
 - 13 『e1』(という) 状況になる場合、<<adv>>『e2』(という) ことをする。
 - 14 『e1』(という) 状態では、<<adv>>『e2』(という) ことをする。
 - 15 『e1』(という) 状況では、<<adv>>『e2』(という) ことをする。
 - 16 『e1』(という) ことが起こらなければ、<<adv>>、『e2』(という) ことができない。
 - 17 X が『e2』(という) ことを実現する手段として、<<adv>> X が『e1』(という) ことを行なう。
 - 18 X が『e1』(という) ことをすることによって、<<adv>> X が『e2』(という) ことができる。
- 20050315

出来事表現の統語カテゴリー

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

20050315

原因-結果の 相対的な出現位置

20050315

作成したコーパスの概要

- 作業員間の判断の一致度
- 判断が一致とは、
 - 付与されたタグは範囲が異なる
 - 原因側、結果側の head 部分に注目
 - head が作業員間で共に同一の文節内に含まれる

20050315