〈COE-LKR2005〉

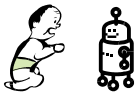
因果関係知識の獲得を目的とした タグ付きコーパスの構築と分析

2005/3/1

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」

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

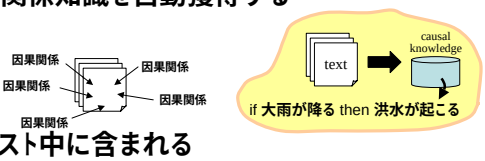


- 推論規則 if 大雨が降る then 洪水が起こる

AI領域 言語処理領域
- 推論規則の獲得 ＝ 因果関係知識の獲得

背景

- 大規模テキストデータを知識資源として
  因果関係知識を自動獲得する



- テキスト中に含まれる
  因果関係を把握することが望まれる
- 先行研究：取り扱い対象を限定
  ⇒テキスト中での因果関係の出現特性が不明
  ⇒大規模データを効果的に利用できていない

本研究の目的

- 大規模データを効果的に利用する基礎として
  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
- 手順
  1. 因果関係タグ付きコーパスの構築
     (因果関係情報を付与する)
  2. 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)

調査項目

	本研究での調査項目	先行研究での対象
(1)	手がかり標識の有無	標識あり
(2)	出来事表現の統語カテゴリ	動詞句
(3)	出来事表現の出現位置	

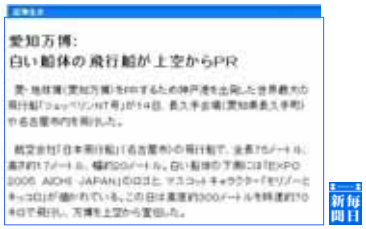
表現形式に関する制約を設けずタグを付与

大雨が降った<u>ため</u>，川が増水した　先行研究
大雨が降り，川が増水した
大雨で川が増水した　本研究

目次

コーパス

- 生コーパス
  - 電子化されたテキストデータ
  - 例） 新聞記事　論文　Webページ


愛知万博:
白い船体の飛行船が上空からPR

コーパス

- タグ付きコーパス
  - 生コーパスに情報付与
  - 例) 品詞タグ付きコーパス

GW	に	入って	遠方	から	の	観光	客	が	増える	。
名詞-一般	助詞-格助詞	動詞-自立	助詞-接続助詞	名詞-一般	助詞-格助詞	助詞-連体化	名詞-サ変	名詞-一般	助詞-格助詞	動詞-自立	記号-句点

コーパス

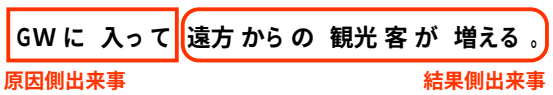
- 因果関係タグ付きコーパス
  - 原因側出来事，結果側出来事，因果関係の情報を付与

目次

0. 因果関係タグ付きコーパスについて
1. 因果関係タグ付きコーパスの構築
   - ひとつの因果関係情報の付け方
   - 作業全体の流れ
2. 因果関係タグ付きコーパスの分析

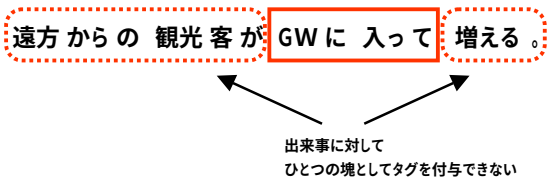
ひとつの因果関係情報の付け方

- 出来事情報の付け方
  - 語順変化への柔軟な対応



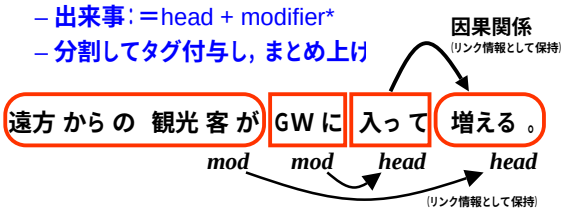
ひとつの因果関係情報の付け方

- 出来事情報の付け方
  - 語順変化への柔軟な対応



ひとつの因果関係情報の付け方

- 出来事情報の付け方
  - 出来事：＝head + modifier*
  - 分割してタグ付与し，まとめ上げ



- 因果関係情報の付け方
  - 2つの *head* に付ける

原因側出来事
結果側出来事

ひとつの因果関係情報の付け方

- 因果関係があるかないか？
- 因果関係の判断基準
  - 基準を定めることは非常に困難
  - 共通した認識は存在するが本来的に主観に依存
- 先行研究：主観判断
- 本研究：言語テンプレートに基づく判断の基準を提案

ひとつの因果関係情報の付け方

- 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

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

言語テンプレート

スロット

ひとつの因果関係情報の付け方

- 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

1. 判断したい2つの出来事表現を用意する．

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
『原因側出来事』という状況の時である。

ひとつの因果関係情報の付け方

- 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

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．
   (テンプレートが文となる)

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
『原因側出来事』という状況の時である。

ひとつの因果関係情報の付け方

- 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

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．
   (テンプレートが文となる)

『洗濯物を干す』ということをするのは
『晴れる』という状況の時である。

ひとつの因果関係情報の付け方

- 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
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.
   (テンプレートが文となる)
3. 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.

『洗濯物を干す』ということをするのは
『晴れる』という状況の時である。 ○

ひとつの因果関係情報の付け方

- 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
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.
   (テンプレートが文となる)
3. 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.

『洗濯物を干す』ということをするのは
『雨が降る』という状況の時である。 ×

ひとつの因果関係情報の付け方

- 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
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.
   (テンプレートが文となる)
3. 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.

『洗濯物を干す』ということをするのは
『眠い』という状況の時である。 ×

ひとつの因果関係情報の付け方

- 言語テンプレートに基づく判断
- 因果関係に強さの概念を導入
  – 因果関係の中には
    推論規則として不適当な事例も存在

因果関係の強さ	例
蓋然 (強い:しばしば共起する)	原因側:電気スタンドを付ける 結果側:電気スタンドが付く
偶然 (弱い:たまたま共起する)	原因側:宝くじを買う 結果側:宝くじに当たる

ひとつの因果関係情報の付け方

- 言語テンプレートに基づく判断
- 因果関係に強さの概念を導入
  – 先行研究:主観による区別は困難
  – 本研究:テンプレートを拡張(副詞を挿入)

『結果側出来事』ということをするのは
《adv》『原因側出来事』という状況の時である。
《adv》:= しばしば | 大抵 | 常に | φ

- いずれかの副詞込みで文意が適格⇒蓋然
  φで文意が適格⇒偶然

ひとつの因果関係情報の付け方

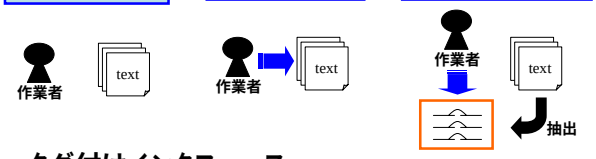
- 言語テンプレートに基づく判断
- 因果関係に強さの概念を導入
  – 強さは因果関係の属性情報として保持

目次

0. 因果関係タグ付きコーパスについて
1. 因果関係タグ付きコーパスの構築
   - ひとつの因果関係情報の付け方
   - 作業全体の流れ
2. 因果関係タグ付きコーパスの分析

作業全体の流れ

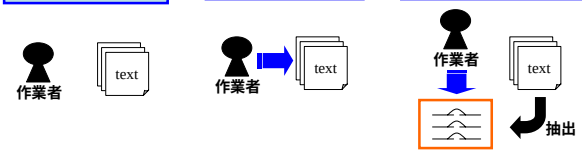
1. テキストを提示 (動詞など候補をハイライト表示)
2. 因果関係情報を付与
3. タグ部分を抽出し確認，修正



タグ付けインタフェース
- 簡単なマウス，キーボード操作でタグ付与が可能

作業全体の流れ

1. テキストを提示 (動詞など候補をハイライト表示)
2. 因果関係情報を付与
3. タグ部分を抽出し確認，修正



本作業では判断箇所は，与えられていない
作業者が発見する

id	原因側	結果側	強さ
940211192	しけ	陸揚げ　延期	蓋然
940211202	冬型の 気圧配置と なる	冷え込み	蓋然
940323183	逮捕する	取り調べ	蓋然
940322113	急性心不全	死去	蓋然
940425251	交信が 途絶える	安否が 気遣う	蓋然
940701300	頭の 骨を　折る	死亡	蓋然
940701304	線路に 転落	電車　停止	蓋然
940323185	韓国 大統領　訪日	交通　規制	蓋然
940211205	池に　落ちる	水死	偶然
940425250	琴　調べ 響く	うっとり	偶然
940211202	東海道 新幹線 遅れる	約二十一万人に 影響する	偶然
940701300	酒を 飲む	酔いつぶれる	偶然

目次

0. 因果関係タグ付きコーパスについて
1. 因果関係タグ付きコーパスの構築
2. 因果関係タグ付きコーパスの分析

作成したコーパスの概要

- データ：新聞記事，７５０記事
- 作業者：3名，並列に独立に
- 言語テンプレート：１８個 (予稿表1参照)

- タグ付けの対象
  - 表現形式に関する制約は無し

大雨が降ったため，川が増水した
大雨が降り，川が増水した
大雨で川が増水した

  - 2つの出来事が隣接する2文以内にある場合のみ

作成したコーパスの概要

- タグ総数（強さの内訳）

作業者	総数	（記事あたり）	強さ		
			蓋然	偶然	未付与
A	2014	2.7	1224	766	24
B	1587	2.1	1094	492	1
C	1048	1.4	603	431	14

作成したコーパスの概要

- 作業者間の判断の一致度
- 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* が作業者間で共に同一の文節内に含まれる

遠方 から の 観光 客 が　GW に　入っ て　増える 。

mod　*mod*　*head*　*head*

作成したコーパスの概要

- 作業者間の判断の一致度

	A	B	C	蓋然	偶然
1人のみ	1	0	0	632	535
	0	1	0	487	255
	0	0	1	134	207
2人以上一致	1	1	0	230	90
	1	0	1	92	77
	0	1	1	107	83
3人一致	1	1	1	270	64

判断が一致しない　因果関係が弱いほど

作成したコーパスの概要

- 作業者間の判断の一致度

	A	B	C	蓋然	偶然
1人のみ	1	0	0	632	535
	0	1	0	487	255
	0	0	1	134	207
蓋然の強さ 2人以上が一致				230	90
				92	77
				107	83
699件を調査対象とする				270	64

判断が一致しない　因果関係が弱いほど

調査項目

(1) 手がかり標識の有無
(2) 出来事表現の統語カテゴリ
(3) 出来事表現の出現位置

- 先行研究：手がかり標識がある場合を考慮
- 標識の有無の割合を調査

- タグ付け作業時に
  手がかり標識にもタグを付けている

(1) 手がかり標識の有無

手がかり標識	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 ように よう として ところ が	2
背景には 際に ており ことによって ...	1

(1) 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

- 半数以上が手がかり標識なし
- 因果関係知識を抽出する際，被覆率を向上させるには標識を伴わない場合をうまく考慮する必要がある

調査項目

(1) 手がかり標識の有無
(2) 出来事表現の統語カテゴリ
(3) 出来事表現の出現位置

- 先行研究：動詞句の場合を考慮

- *head* 部分の末尾形態素の品詞に注目
- 動詞or形容詞なら動詞句 (vp)
  名詞なら名詞句 (np)

(2) 出来事表現の統語カテゴリ

カテゴリ	例	原因側	結果側
vp	動詞-自立 (焼く) 形容詞-自立 (難しい)	365	412
np	名詞-サ変接続 (停電) 名詞-一般 (火災)	322	269
	その他 (うっとり)	12	18

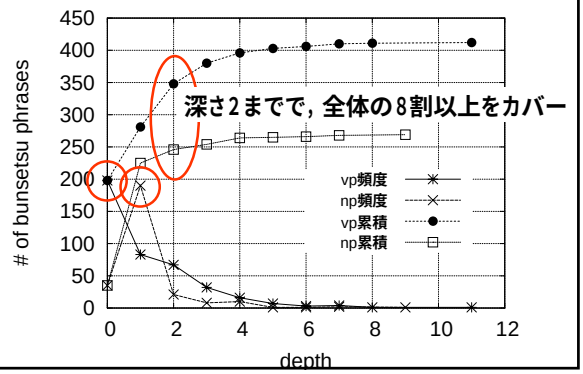
- 動詞句が過半数．名詞句 (np) も存在
- 因果関係知識を抽出する際，名詞句への対応も必要

調査項目

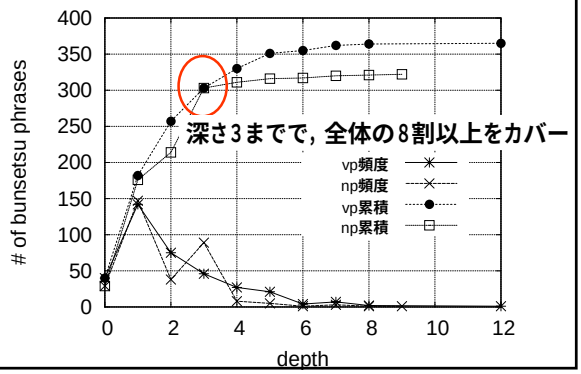
(1) 手がかり標識の有無
(2) 出来事表現の統語カテゴリ
(3) 出来事表現の出現位置

- 文＝文末文節を根とする係り受け木
- *head* を含む文節が位置している深さを調査

(3) 出来事表現の出現位置 (結果側)

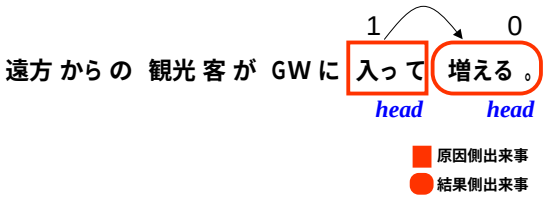


(3) 出来事表現の出現位置 (原因側)



(3) 出来事表現の出現位置

- 原因側，結果側の間の相対的な出現位置関係
- 深さの差に注目
- 原因側深さから結果側深さを引く（1－0＝1）



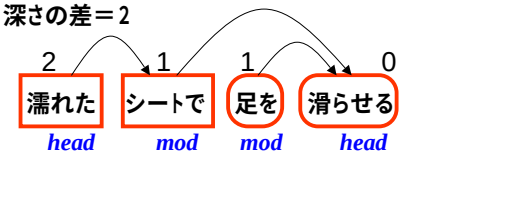
(3) 出来事表現の出現位置

- 原因側，結果側の間の相対的な出現位置関係

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

(3) 出来事表現の出現位置

- 原因側，結果側の間の相対的な出現位置関係
- 連体修飾の被修飾名詞を介する場合
  - 深さの差＝2



まとめ

- 因果関係知識獲得に大規模テキストデータを効果的に利用する
  - テキスト中に含まれる
    因果関係の出現特性を明らかにした
    (1) 手がかり標識の有無
    (2) 出来事表現の統語カテゴリ
    (3) 出来事表現出現位置
  - 言語学からの知見を定量的に検討
- 得られた知見を基に知識獲得へ