

日本語述語句の統語的異形間の 言い換えらしさの推定

藤田 篤

名古屋大学大学院工学研究科
電子情報システム専攻

言い換え生成の技術的課題

■ 知識獲得

- パターンを書く
- 既存の辞書、シソーラスからの獲得
- コーパス／Webからの獲得

X wrote Y => X is the author of Y
自己記録を上回る => 自己記録を更新する

■ 知識の適用（生成）

- 言い換え元の表現の多義性解消
- 与えられた文脈における言い換えらしさの判定
 - 正しい言語表現であるかどうか（文法的適格性）
 - 意味が保存されているかどうか（同義性）

■ タスク指向の適切さの評価

- 分かりやすさ， 翻訳しやすさ， 曖昧性の少なさ

取り組むべき課題

■ 所与の文脈における知識の適用条件の特定

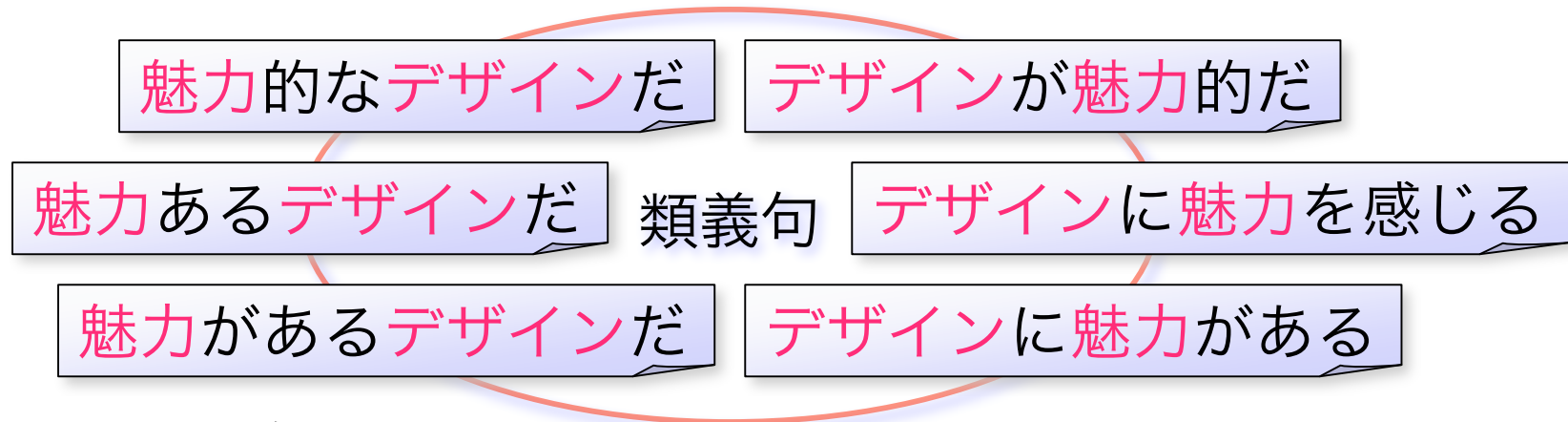
- 語や語のクラスに基づく制約の記述（内包的）
[Mel'cuk & Polguere, 1987][Fujita+, 2005;2006]
- 適用可能な例を事前に列挙する（外延的）
 - cf. 生成物を言語モデルで評価 [Pantel+, 2007]

■ 問い

- 事前に列挙しなくても動的に判定できないか？
- 構成的な言い換えの場合はどれくらいの精度？
- 語でなくもうちょっと複雑な句の場合に
 - データスパースネスをどう回避するか？
 - 品詞が変わっても分布類似度は通用するか？

句の統語的異形の生成器

■ フレーザル・シソーラス [Fujita+, 2007]



- どんな句？
 - 動詞句 ～ いわゆる節／単文
 - 「格要素名詞句 + 格助詞 + 述語文節」
- どんな言い換え？
 - 構成的な言い換え
 - 少数のパターンで動的に生成

今回の取り組み

■ 統語的異形間の言い換えらしさを計算する

- 入力：自動生成した日本語述語句の統語的異形対

- 言い換え元の表現 (s)

- 言い換え先の表現 (t)

確認を急ぐ

急いで確認する

- 出力：言い換えらしさ (0~1の実数値)

- 次の3つを満たすときに言い換え可能である

- t が文法的に適格である

- s が表すことが真であるならば

- t が表すことも真である (意味的に等価・含意)

- ある文脈で s を t に置換可能である

言い換えらしさ計算の概要

■ 言い換えらしさの計算手順

- 1. s と t の各々に対してWebスニペットを取得
- 2. s と t の各々の素性をWebスニペットから抽出
- 3. s と t の素性ベクトルの類似度を
 s と t の言い換えらしさとして出力

■ 特徴

- Webスニペットから各表現の素性を抽出
 - 文法的適格性の判定
 - データスパースネスの軽減
 - 言い換え元の述語句の関連語の利用
- 分布類似度を述語句の間の類似度計算に利用

1. Webスニペットの収集

- 各句表現に対するスニペット（とヒット数）
 - Yahoo! JAPAN Web-search API
<http://developer.yahoo.co.jp/search/>
 - 上位500件（平均0.75秒/クエリ）

The screenshot shows the Yahoo! Japan search interface. At the top, the Yahoo! JAPAN logo is on the left, followed by a search bar containing the text "急いで確認する". To the right of the search bar are buttons for "検索" (Search) and "検索オプション" (Search Options). Below the search bar, a header bar indicates the search results: "ウェブ検索結果 (検索結果の見方)" on the left and "\"急いで確認する\" で検索した結果 1~10件目 / 約269件 - 0.03秒" on the right. The main content area displays two search results, each with a title, a snippet of text, and a URL. The first result is titled "1. 熊本城攻防戦一転一(3)" and the second is "2. ポケモンノベル". Both snippets are enclosed in red dotted boxes, highlighting the phrase "急いで確認する". To the right of the search results, there is a light blue box with the text "スポンサーサイト 掲載について".

YAHOO! JAPAN 検索 "急いで確認する" 検索 検索オプション

ウェブ検索結果 (検索結果の見方) "急いで確認する" で検索した結果 1~10件目 / 約269件 - 0.03秒

1. [熊本城攻防戦一転一\(3\)](#)
視界の隅で何かが動いた気がして、**急いで確認する**。見れば子猫が二匹歩いていた。「なんだ猫かよ。にゃーと茂みから聞こえ、2匹の猫はその声に釣られたように茂みの中に消えた。しばらくして、ひょこっと城壁の上に何かが乗った気がした。また猫か?と思い視線を上げる。...」
www.geocities.jp/darts2035/ss_11_tenn_3.html - [キャッシュ](#)

2. [ポケモンノベル](#)
第六十六話/光と闇 「く・・・(ここは・・・) ——大丈夫ですか?—— 「大丈夫です。でもここは・・・」 ポケモンセンターではない。(畜生・・・) ポケモンを**急いで確認する**。幸い全員いたが、皆、瀕死状態まで痛め付けられていた。「畜生 ...」
www7.plala.or.jp/mewtwo777/novel/137/18-66.html - [キャッシュ](#)

[スポンサーサイト 掲載について](#)

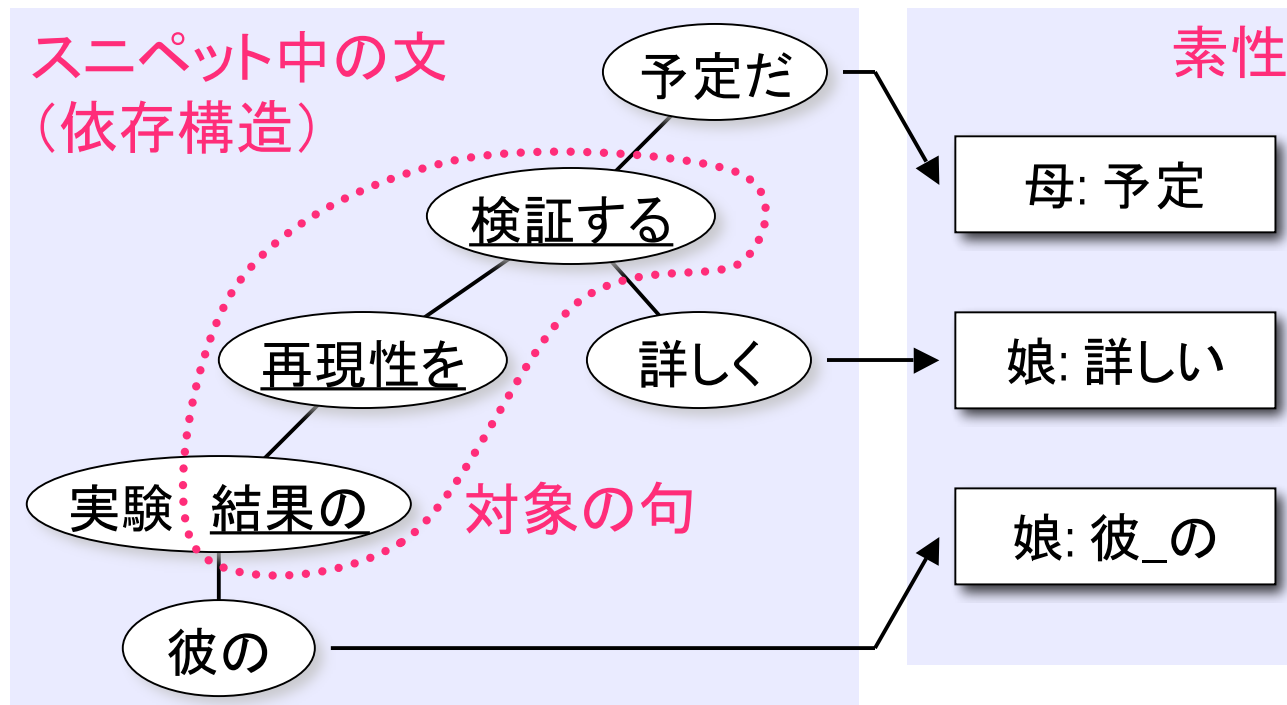
2. Webスニペットからの素性抽出 (1/2)

- 測りたいこととそれを近似する素性の検討
 - 文法的適格性 (HITS)
 - Webでヒット数するならばもっともな表現である
 - 意味的等価性 (BOW)
 - その表現の周辺に出現する語の分布が似ていれば意味的に等価である
 - 置換可能性 (MOD)
 - その表現を修飾したり, その表現に修飾される表現の分布が似ていれば, 置換可能である

係り受け文節素性MOD

■ 1文節 → 1素性

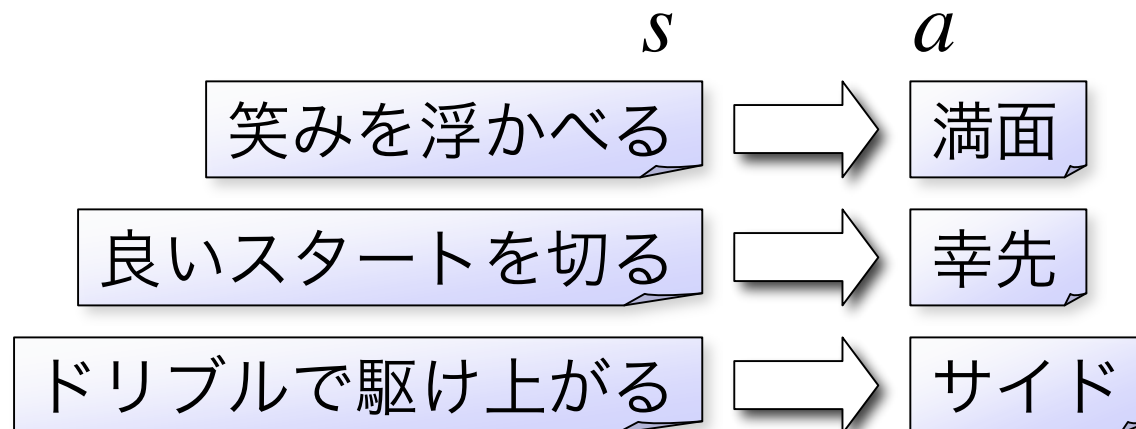
- 対象の句に直接係るまたは直接係られる文節
- 主辞の基本形
- 名詞に後続する格助詞, 動詞に後続する助動詞



2. Webスニペットからの素性抽出 (2/2)

■ より適切なスニペットの取得 (Anc)

- 1. s と最も関連する語 a を特定
 - 今回は s を修飾する回数が最も高い名詞
 - MOD素性から選別
- 2. $s \text{ AND } a$, $t \text{ AND } a$ に対するスニペットを取得
- 3. s と t 各々の素性を $s \text{ AND } a$, $t \text{ AND } a$ に対するスニペットから抽出



3.素性ベクトル間の類似度計算

■ Linの類似度 [Lin & Pantel, 2001]

$$Par_{Lin}(s \Rightarrow t) = \frac{\sum_{f \in F_s \cap F_t} (w(s, f) + w(t, f))}{\sum_{f \in F_s} w(s, f) + \sum_{f \in F_t} w(t, f)}$$

- F_s, F_t : s および t の素性
- $w(x, f)$: 素性集合 F_x における素性 f の重み

■ α -skew divergence [Lee, 1999]

$$Par_{skew}(s \Rightarrow t) = \exp(-d_{skew}(t, s))$$

$$d_{skew}(t, s) = D(P_s \| \alpha P_t + (1 - \alpha)P_s)$$

- P_s, P_t : s および t の素性ベクトルの確率分布版
- α : KL divergenceにどれだけ近似するかの変数

評価セット作成 (1/2)

■ 言い換え元表現の収集

- 6タイプの句（詳細は下表）
- タイプごとに毎日新聞15年分で最頻出の約1,000句

■ 網羅的に統語的異形を生成

- フレーザル・シソーラス [Fujita+, 2007] の修正版

Phrase type	# of tokens	# of types	th types	Cov.(%)	Output	Ave.
$N : C : V$	20,200,041	4,323,756	1,000 1,014	10.7	1,536 (489)	3.1
$N_1 : N_2 : C : V$	3,796,351	2,013,682	107 1,005	6.3	88,040 (966)	91.1
$N : C : V_1 : V_2$	325,964	213,923	15 1,022	12.9	75,344 (982)	76.7
$N : C : Adv : V$	1,209,265	923,475	21 1,097	3.9	8,281 (523)	15.7
$Adj : N : C : V$	378,617	233,952	20 1,049	14.1	128 (50)	2.6
$N : C : Adj$	788,038	203,845	86 1,003	31.4	3,212 (992)	3.2
Total	26,698,276	7,912,633	6,190		176,541 (4,002)	44.1

評価セット作成 (2/2)

■ 言い換えらしさを計算できた対の数

Phrase type	HITS			BOW			MOD		
	Output		Ave.	Output		Ave.	Output		Ave.
$N : C : V$	1,405	(489)	2.9	1,402	(488)	2.9	1,396	(488)	2.9
$N_1 : N_2 : C : V$	9,544	(964)	9.9	9,249	(922)	10.0	8,652	(921)	9.4
$N : C : V_1 : V_2$	3,769	(876)	4.3	3,406	(774)	4.4	3,109	(762)	4.1
$N : C : Adv : V$	690	(359)	1.9	506	(247)	2.0	475	(233)	2.0
$Adj : N : C : V$	45	(20)	2.3	45	(20)	2.3	42	(17)	2.5
$N : C : Adj$	1,459	(885)	1.6	1,459	(885)	1.6	1,399	(864)	1.6
Total	16,912	(3,593)	4.7	16,067	(3,336)	4.8	15,073	(3,285)	4.6

Anc.HITS			Anc.BOW			Anc.MOD			Mainichi		
Output		Ave.	Output		Ave.	Output		Ave.	Output		Ave.
1,368	(488)	2.8	1,366	(487)	2.8	1,360	(487)	2.8	1,103	(457)	2.4
7,437	(897)	8.3	7,424	(894)	8.3	6,795	(891)	7.6	3,041	(948)	3.2
2,517	(697)	3.6	2,497	(690)	3.6	2,258	(679)	3.3	1,156	(548)	2.1
342	(174)	2.0	339	(173)	2.0	322	(168)	1.9	215	(167)	1.3
41	(18)	2.3	41	(18)	2.3	39	(16)	2.4	14	(7)	2.0
1,235	(809)	1.5	1,235	(809)	1.5	1,161	(779)	1.5	559	(459)	1.2
12,940	(3,083)	4.2	12,902	(3,071)	4.2	11,935	(3,020)	4.0	6,088	(2,586)	2.4

評価結果 (Ev.Gen)

■ 入力に対する1位出力の精度

Precision	HITS	BOW		MOD	
		Lin	skew	Lin	skew
Lenient	120 (60%)	123 (62%)	125 (63%)	130 (65%)	132 (66%)
Strict	84 (42%)	82 (41%)	86 (43%)	91 (46%)	92 (46%)

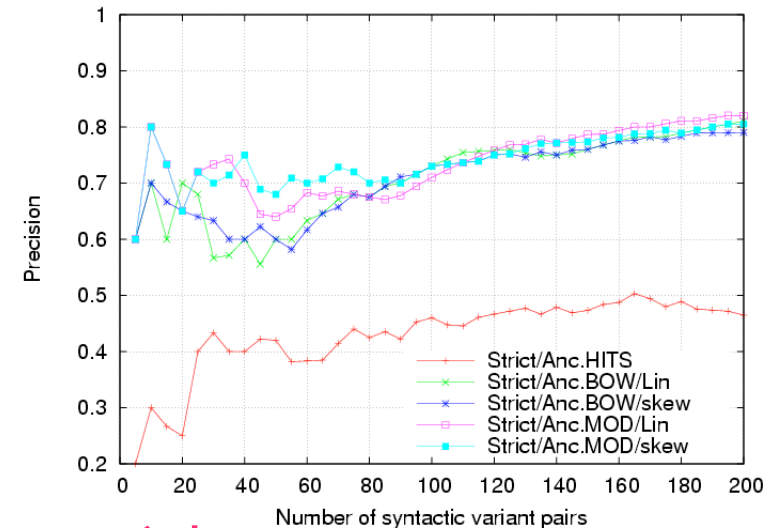
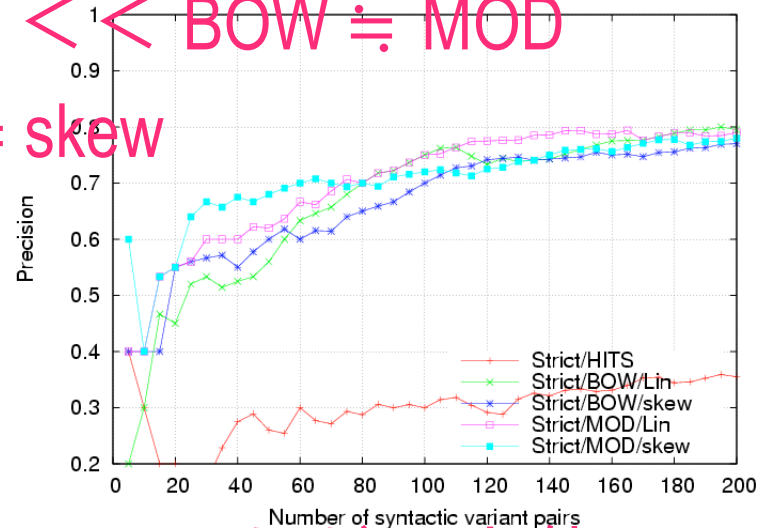
Precision	Anc.HITS	Anc.BOW		Anc.MOD	
		Lin	skew	Lin	skew
Lenient	119 (60%)	124 (62%)	124 (62%)	131 (66%)	130 (65%)
Strict	83 (42%)	85 (43%)	87 (44%)	91 (46%)	90 (45%)

- HITS < BOW < MOD (ただし統計的有意差なし)
 - 句同士を比較することには意味がある
- Lin $\doteq$ skew
- X $\doteq$ Anc.X
 - スニペットの取り方を工夫した効果はない
 - 複数の内容語からなる句の多義性は元々小さい

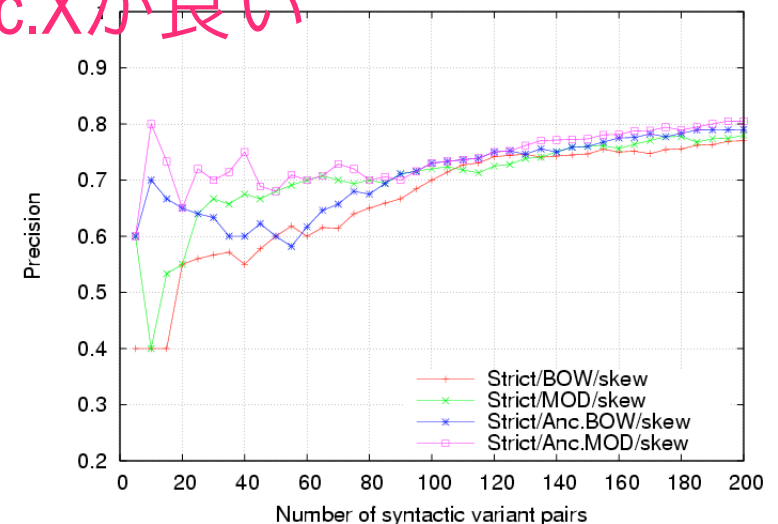
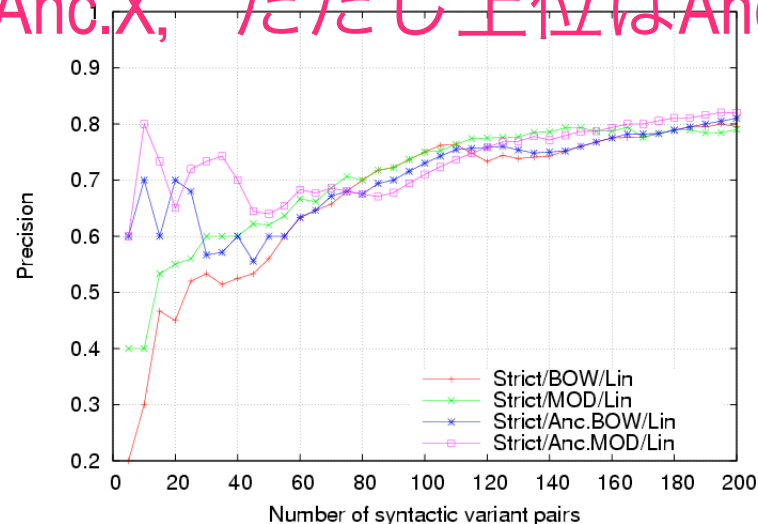
評価結果 (Ev.Rec)

■ 類似度上位の言い換え対の精度

- $HITS < BOW \doteq MOD$
- $Lin \doteq skew$



- $X \doteq Anc.X$, ただし上位はAnc.Xが良い

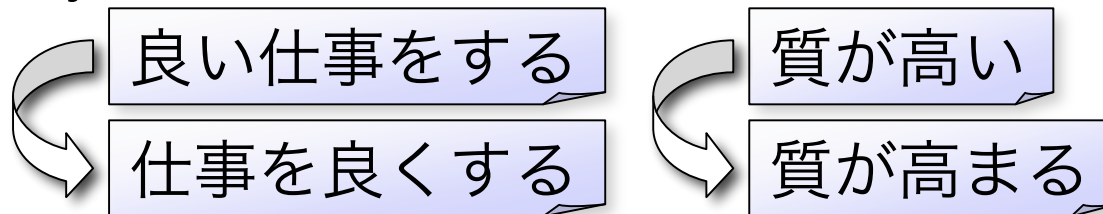


評価結果 (Ev.Ling)

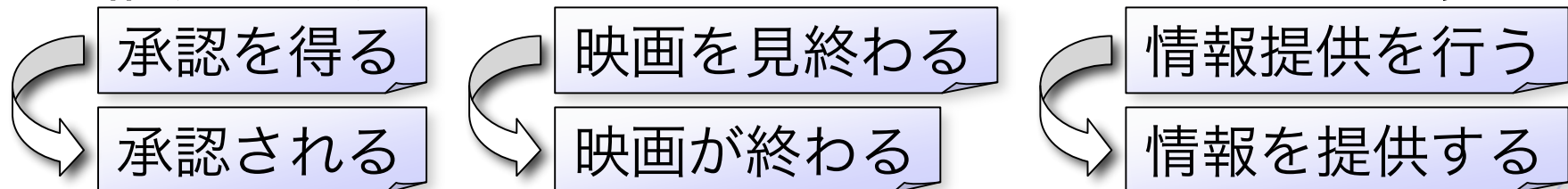
■ 言い換え元の句タイプごとの精度

Phrase type	Lenient	Strict
$N : C : V$	69/ 98 (70%)	52/ 98 (53%)
$N_1 : N_2 : C : V$	64/ 72 (89%)	51/ 72 (71%)
$N : C : V_1 : V_2$	58/ 84 (69%)	41/ 84 (49%)
$N : C : Adv : V$	44/ 61 (72%)	33/ 61 (54%)
$Adj : N : C : V$	4/ 25 (16%)	0/ 25 (0%)
$N : C : Adj$	38/ 73 (52%)	18/ 73 (25%)
Total	277/413 (67%)	195/413 (47%)

- Adjを含む句の言い換えに間違い多数



- 精度が良いタイプの句ではありきたりな例が多い



まとめ

■ 統語的異形間の言い換えらしさの計算

- 対象：自動生成した日本語述語句の統語的異形対
 - 品詞が変わるものも含む
 - 文法的に不適格な表現も含む
- 言い換えらしさを近似

■ シンプルだがそれなりに高精度

- 割と文法的に適格なものを残せる
- スコアと言い換え可能性の相関を確認
 - 閾値を設けることで精度良く収集可能
- 句の複雑さがある程度まではスケールする
 - 少なくとも新聞（のみ）を使うよりは