大規模言語資源時代の意味談話処理

乾 健太郎

東北大学

©2012 Kentaro Inui

感謝します

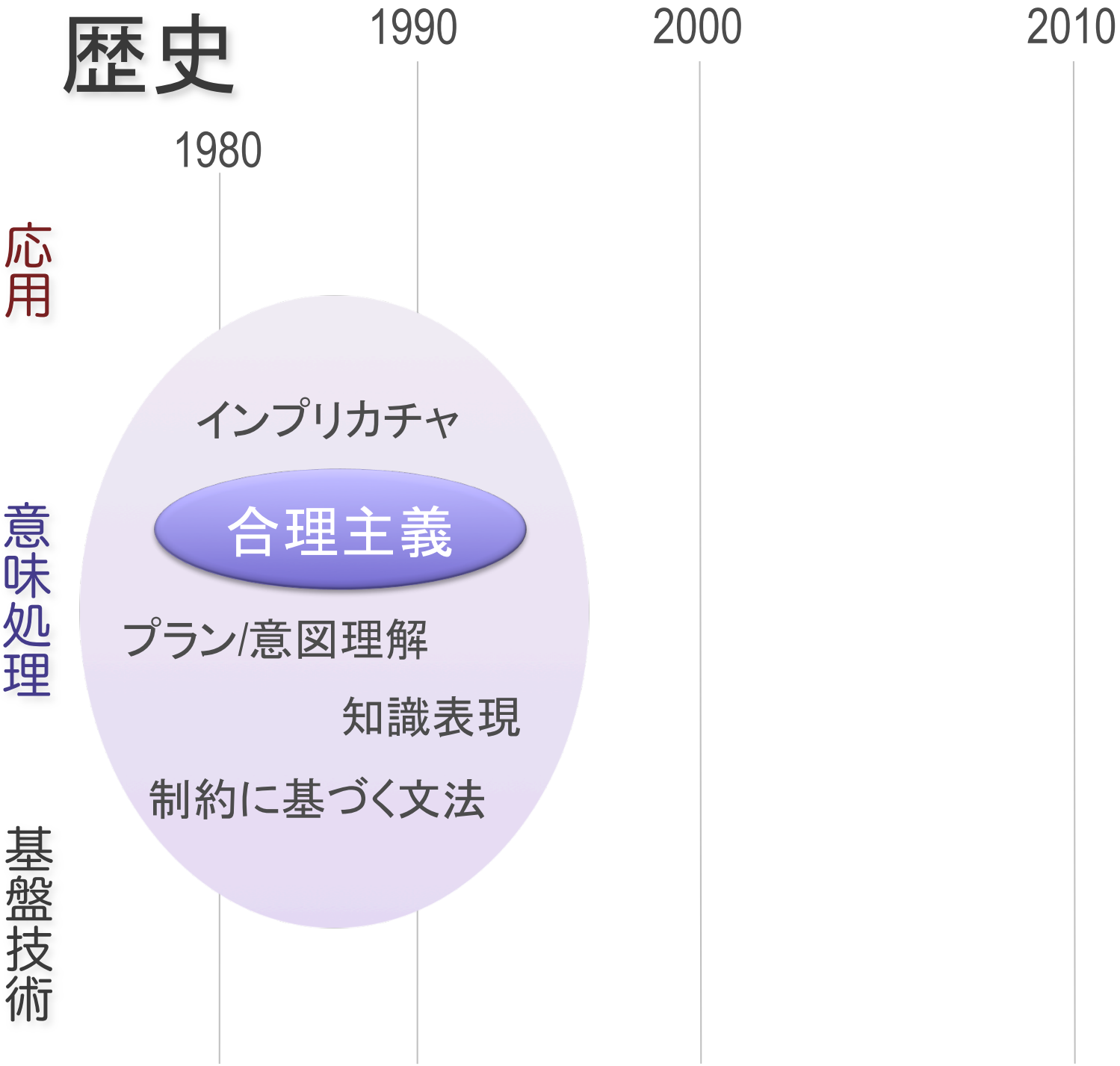
本講演の機会を与えてくださったプログラム委員の諸氏，講演内容について多くの貴重な助言をいただいた黒橋禎夫氏（京都大学），宮尾祐介氏（国立情報学研究所），小町守氏（奈良先端科学技術大学院大学），東北大学乾・岡崎研究室の諸氏に深く感謝いたします．

自然言語の理解

週末に母とやなぎに行きました。
私は評判のきつねうどんを頼みました。期待通りの味に大満足。

やなぎはうどん屋の名前
やなぎに行ったのは私
頼んだは注文したと同じ
私はきつねうどんを食べた
味はきつねうどんの味
大満足は美味しかったってこと

歴史


1980
1990
2000
2010
応用
意味処理
基盤技術
インプリカチャ
合理主義
プラン/意図理解
知識表現
制約に基づく文法

行為の意図を読む

庭に洗濯物を全部干し終わったところに，雨が降ってきた

ガッカリ

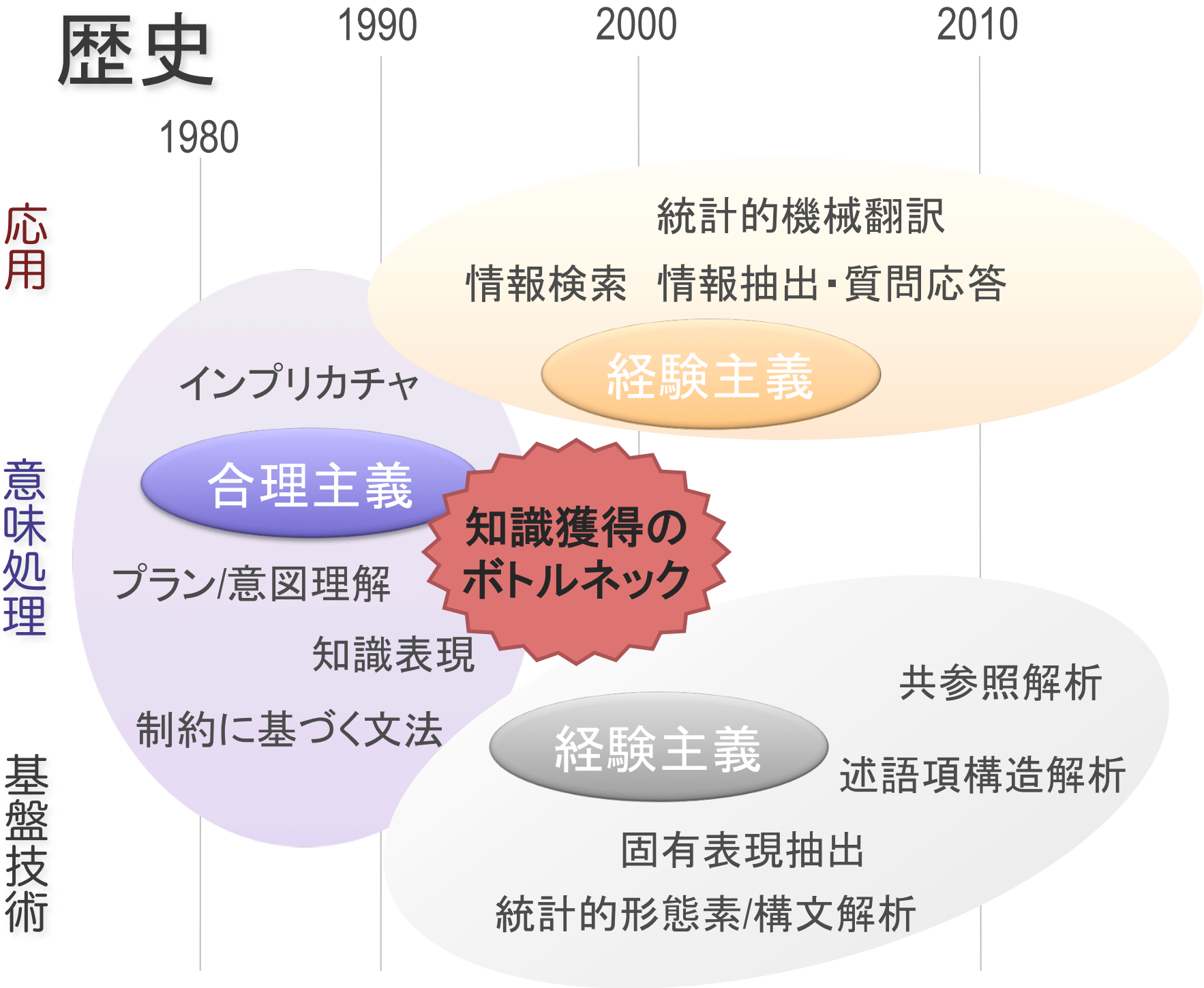
庭に洗濯物を干す

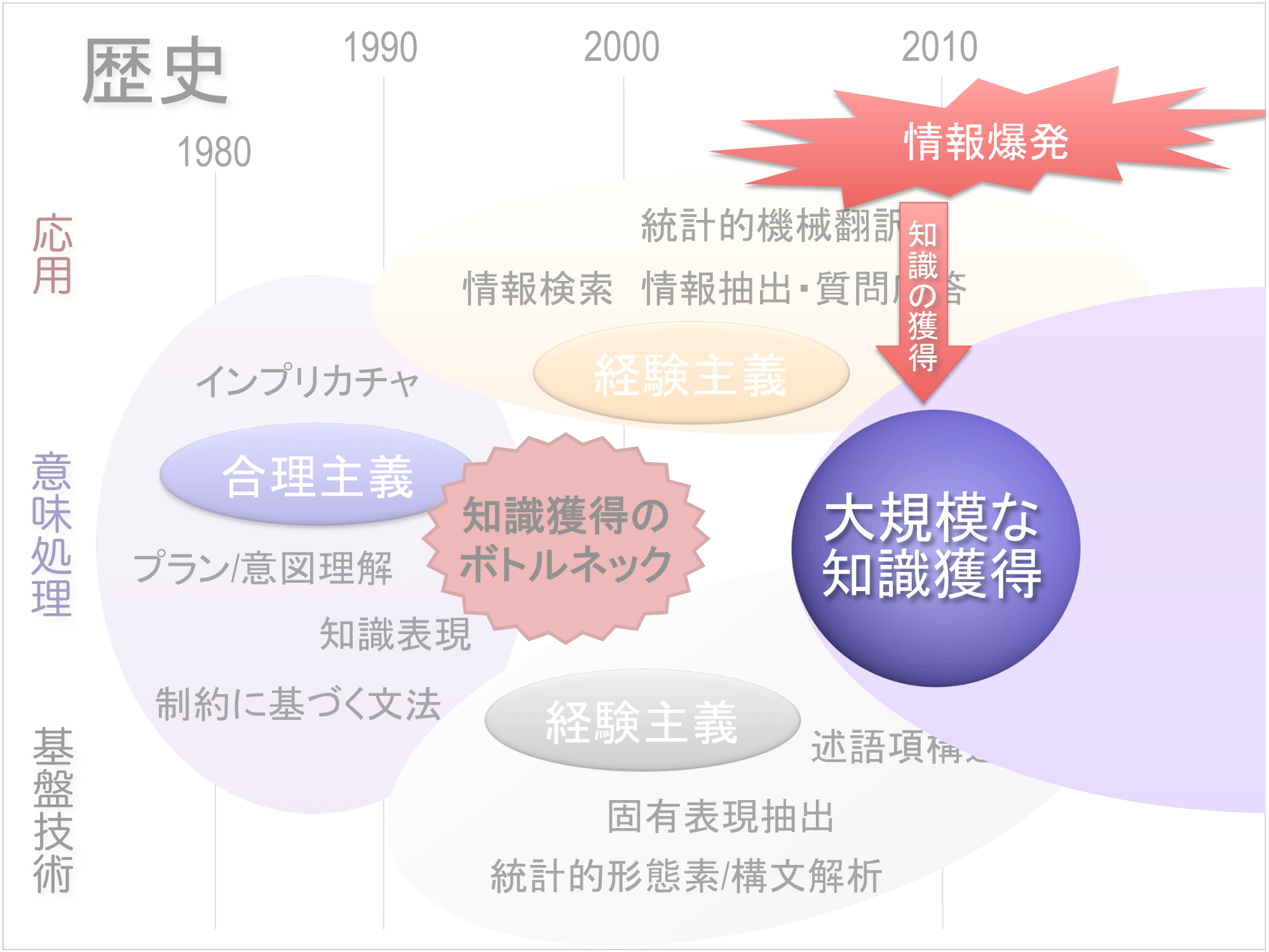
雨が降る

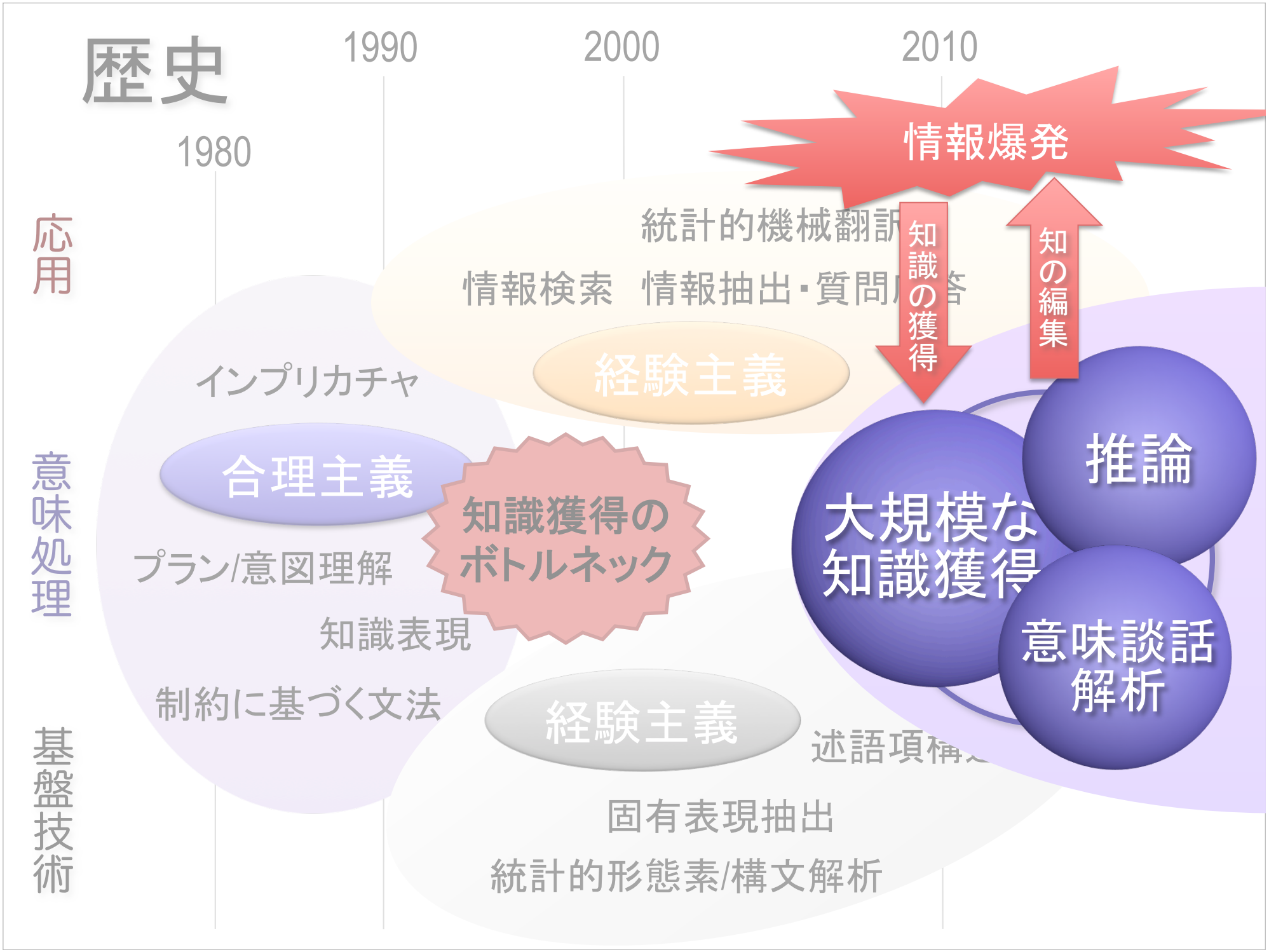
〈洗濯物を干す〉 —効果→ 〈洗濯物が乾く〉

〈晴れ〉 —原因結果（条件）→ 〈洗濯物が乾く〉

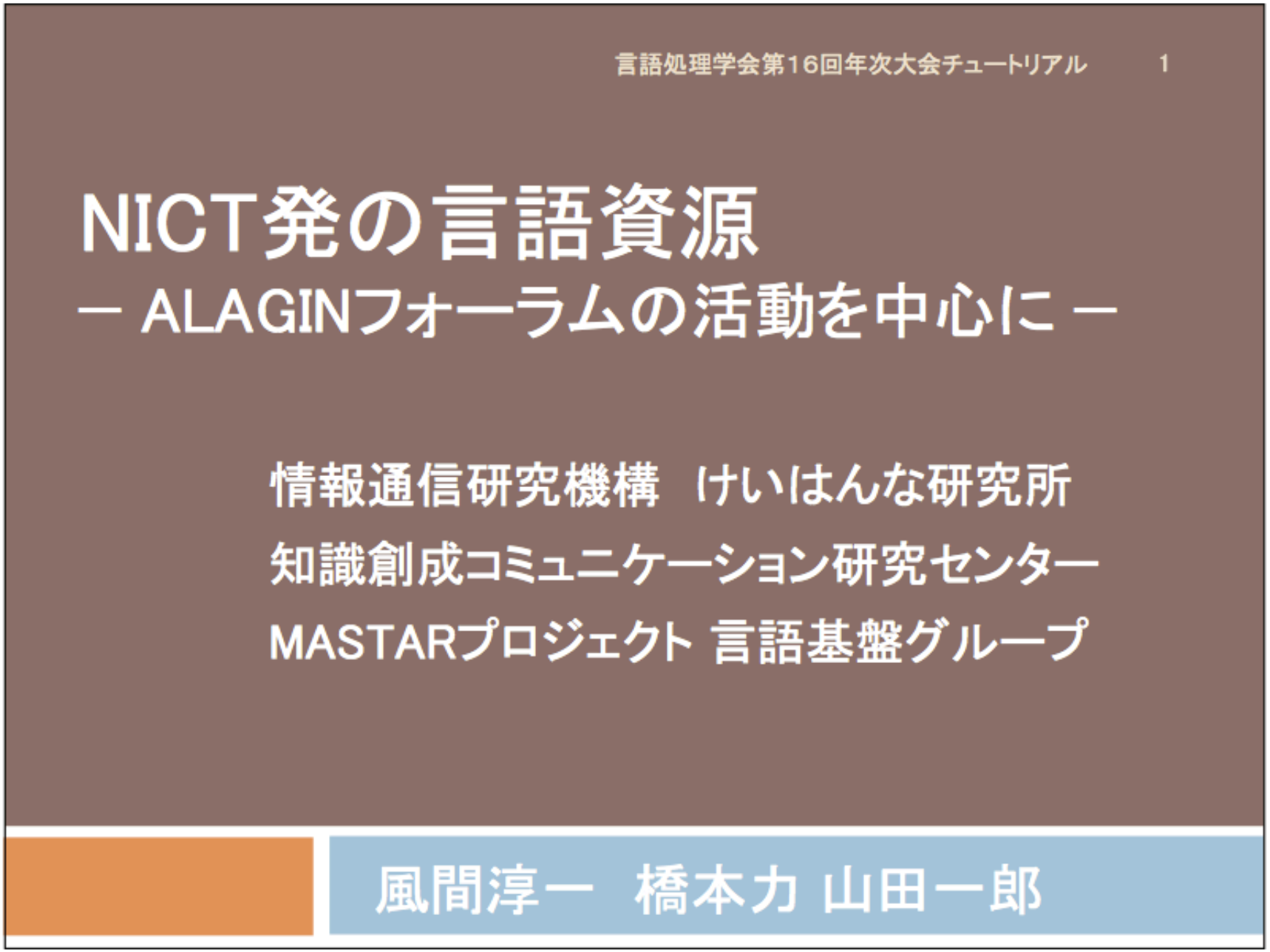
常識的な知識


歴史
1980
1990
2000
2010
応用
意味処理
基盤技術
統計的機械翻訳
情報検索
情報抽出・質問応答
経験主義
インプリカチャ
合理主義
知識獲得の
ボトルネック
プラン/意図理解
知識表現
制約に基づく文法
共参照解析
経験主義
述語項構造解析
固有表現抽出
統計的形態素/構文解析


歴史
1980
1990
2000
2010
情報爆発
知識の獲得
応用
意味処理
基盤技術
統計的機械翻訳
情報検索
情報抽出・質問応答
経験主義
インプリカチャ
合理主義
知識獲得のボトルネック
大規模な知識獲得
プラン/意図理解
知識表現
制約に基づく文法
経験主義
述語項構造
固有表現抽出
統計的形態素/構文解析


歴史
1980
1990
2000
2010
情報爆発
知識の獲得
知の編集
応用
統計的機械翻訳
情報検索
情報抽出・質問応答
経験主義
意味処理
インプリカチャ
合理主義
プラン/意図理解
知識表現
制約に基づく文法
知識獲得のボトルネック
大規模な知識獲得
推論
意味談話解析
基盤技術
経験主義
述語項構造
固有表現抽出
統計的形態素/構文解析

関根聡（2009）自然言語処理のための知識獲得．言語処理学会第15回年次大会チュートリアル資料．



風間淳一，橋本力，山田一郎（2010）NICT発の言語資源ーALAGINフォーラムの活動を中心にー．言語処理学会第16回年次大会チュートリアル資料．

ALAGINフォーラム

http://alagin.jp/

言語資源・音声資源サイト
高度言語情報融合フォーラム

ホーム | ニュース | 資源・サービス | オープンソース | ダウンロード | 関連サイト | お問合わせ

現在の場所: 資源・サービス > NICT/MASTAR 音声言語資源 > 言語資源・サービス > 各言語資源・サービスの概要

各言語資源・サービスの概要

契約書の入手及び記載方法等につきましては、右下の「ALAGIN会員」よりログイン頂く事で、右メニューより「入手の方法（契約書含む）」「契約書の記載方法・代理作成」がご覧になれます。

（A-1）文脈類似語データベース

100万語の名詞に対して、約1億ページのWeb文書上での文脈が類似している名詞を類似度とともに順に最大500個列挙したものです。（今年度中にさらに大規模な約6億ページのWebデータをもとに構築し直し、語彙数も200万まで増強する予定です）

例（数字は類似度）	
「ルパン三世」の文脈類似語（アニメタイトルが上位に集まっている）	ルパン３世 (-0.229) 名探偵コナン (-0.259) 宇宙戦艦ヤマト (-0.265) ケロロ軍曹 (-0.28) 鉄腕アトム(-0.282) ガッチャマン (-0.287) デビルマン (-0.289) サイボーグ００９ (-0.294) 新世紀エヴァンゲリオン (-0.295) ヤッターマン (-0.305) 聖闘士星矢 (-0.308) セーラームーン (-0.308) ...
「チャイコフスキー」の文脈類似語（有名作曲家が上位に集まっている）	ブラームス (-0.152) シューマン (-0.163) メンデルスゾーン (-0.166) ショスタコーヴィチ(-0.178) シベリウス (-0.18) ハイドン (-0.181) ヘンデル (-0.181) ラヴェル (-0.182) シューベルト (-0.187) ベートーヴェン (-0.19) ドヴォルザーク (-0.192) ラフマニノフ (-0.193) バルトーク (-0.198) ....

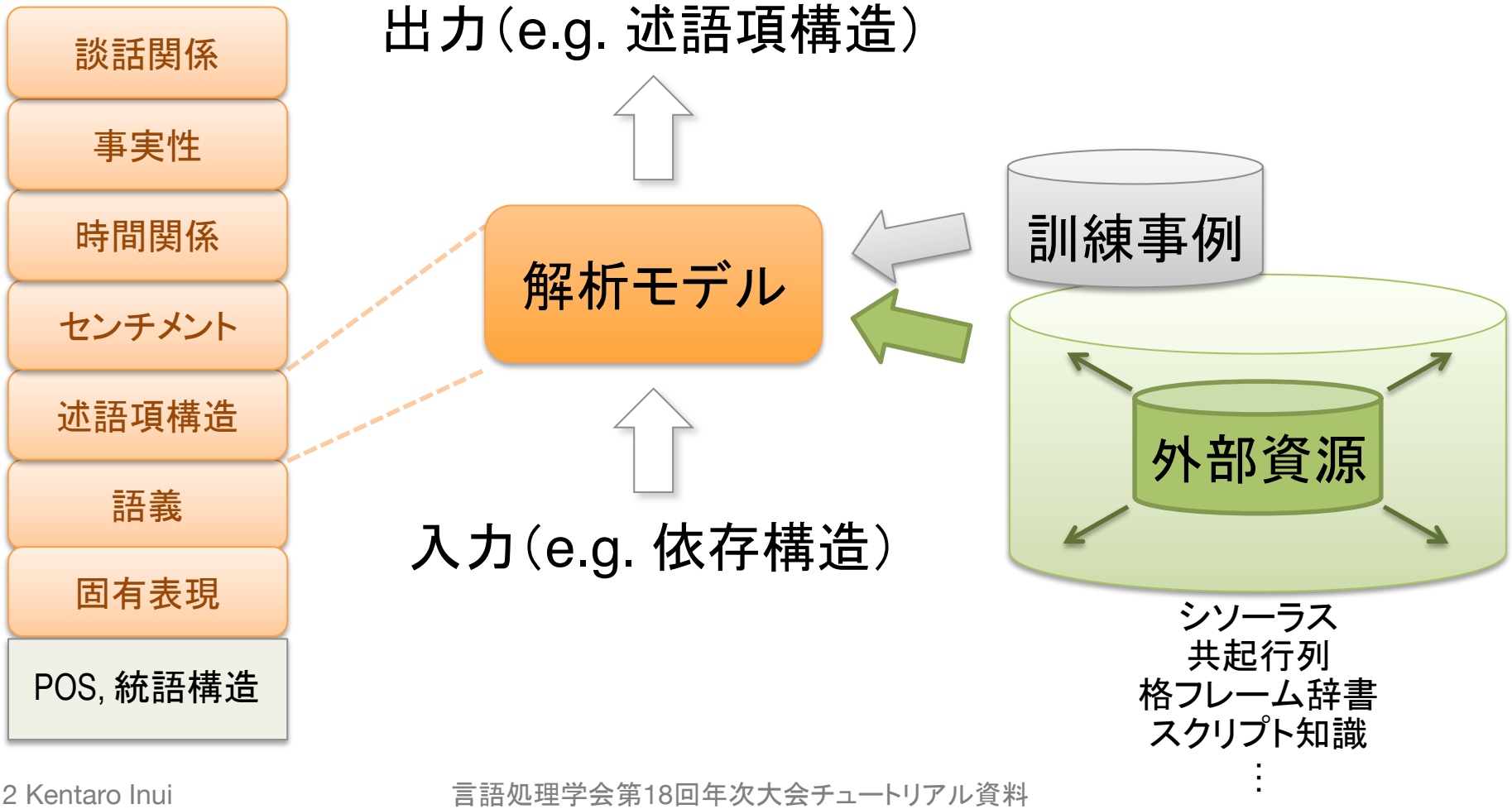
メニュー
- ホーム
- ニュース
- 資源・サービス
  - NICT/MASTAR 音声言語資源
    - 言語資源・サービス
      - 言語資源サービス一覧
      - 各言語資源・サービスの概要
    - 音声資源
  - 楽天データセット
- オープンソース

本講演では

- 大規模言語資源時代の意味談話処理
  - 意味談話処理 → むちゃくちゃ広い
  - 全体をカバーするのは困難
- 大規模言語資源のインパクトに焦点

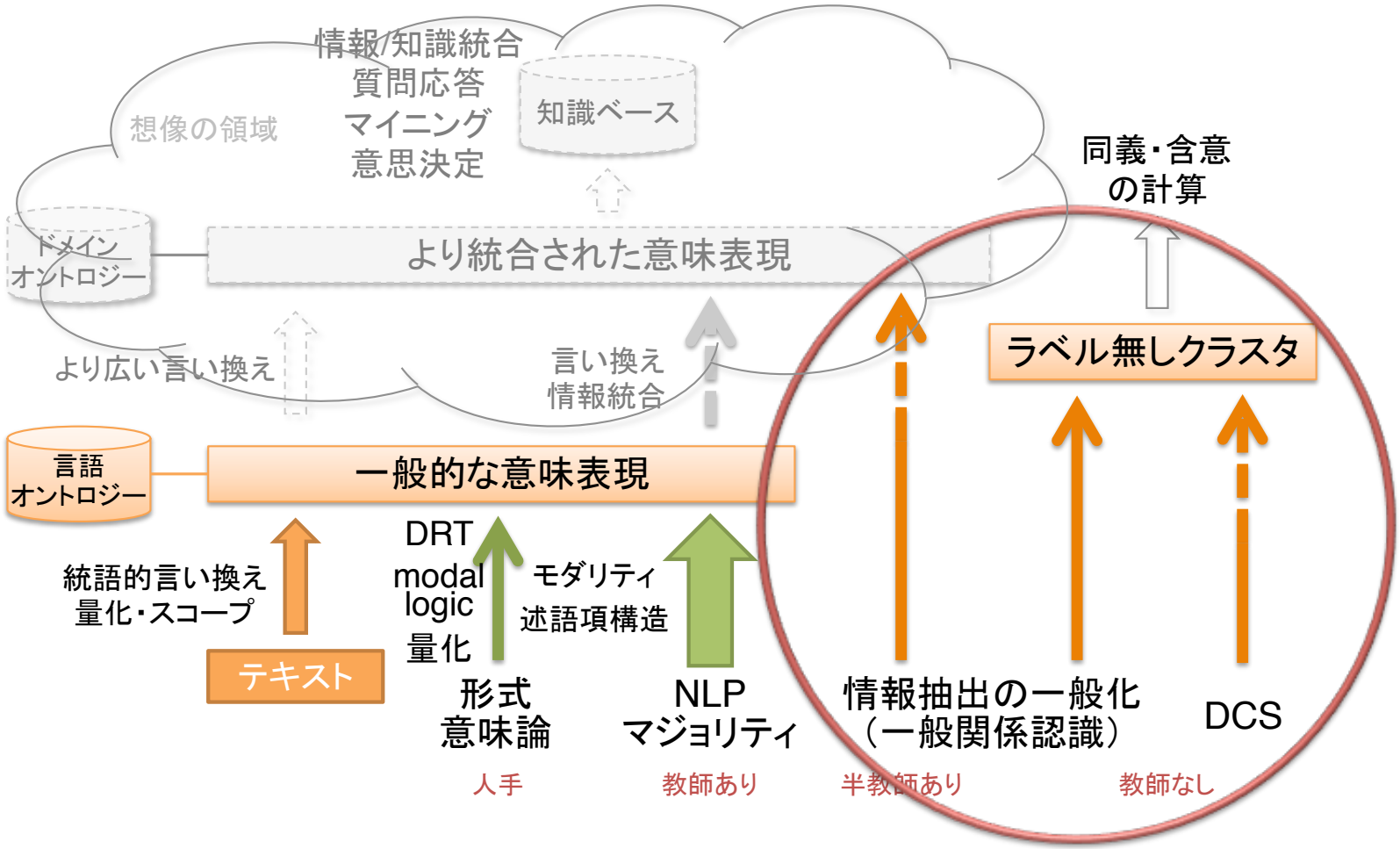
大規模言語資源のインパクト①

個々の部分問題における外部資源の利用



大規模言語資源のインパクト②

意味解析モデル構築と知識獲得の融合



大規模言語資源のインパクト③

実応用の広がり

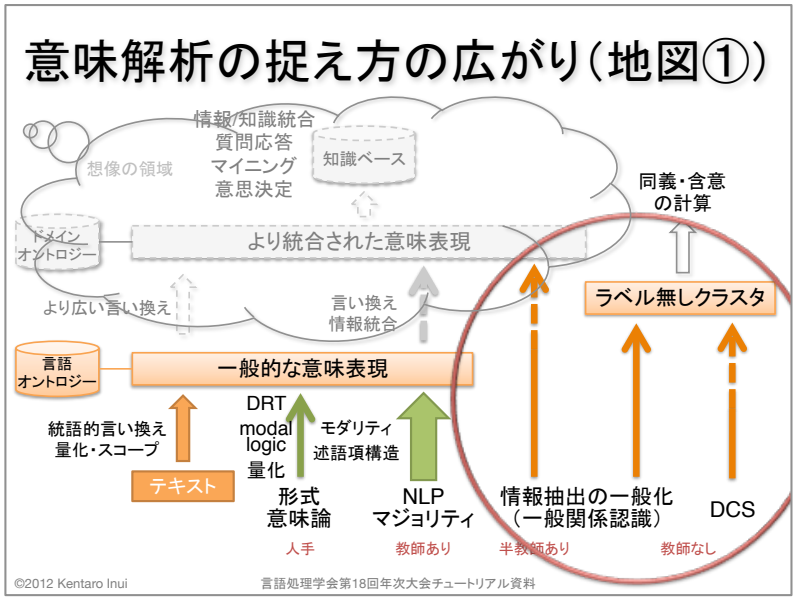
- テキストマイニング
- 意見/評判分析
- 質問応答
- 意思決定支援
- 機械翻訳
- その他いろいろ

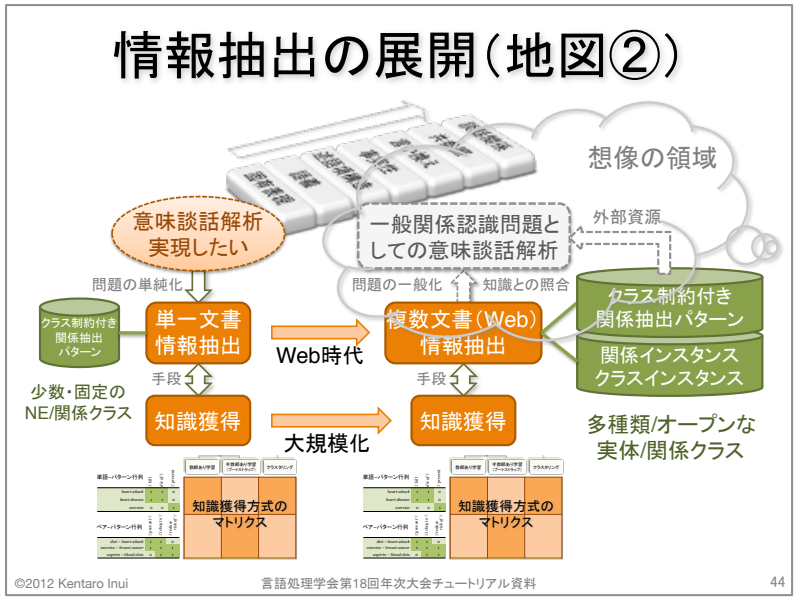
本講演では

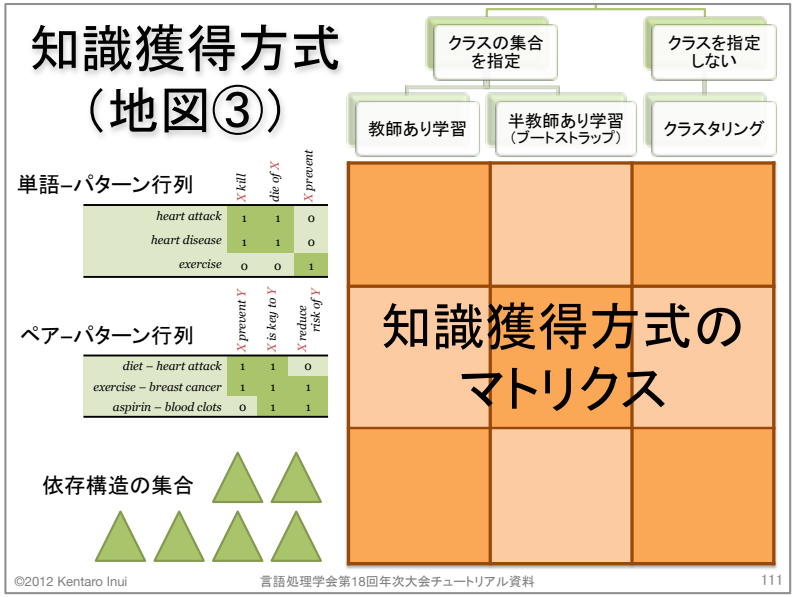
- 大規模言語資源のインパクトに焦点
  ①個々の部分問題における外部資源の利用
  → 方法論的には従来の延長，あまり話しません
  ②意味解析モデル構築と知識獲得の融合
  → 大きな動きあり，少し詳しく話します
  ③実応用の広がり
  → 話しません
- 動向を俯瞰する視座 → 地図を用意しました
  – 個別の技術ではなく，地図の見方を解説します

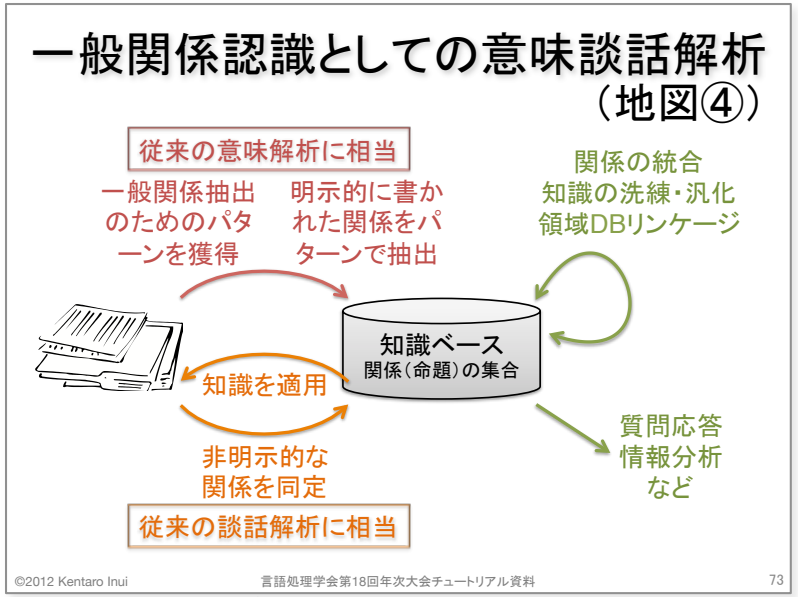
©2012 Kentaro Inui

意味談話解析研究の地図









本講演の構成

1. Annotation layer cake
2. 情報抽出と知識獲得の今
3. Making the implicit explicit
4. やわらかい記号の世界
5. ルールを作るところから

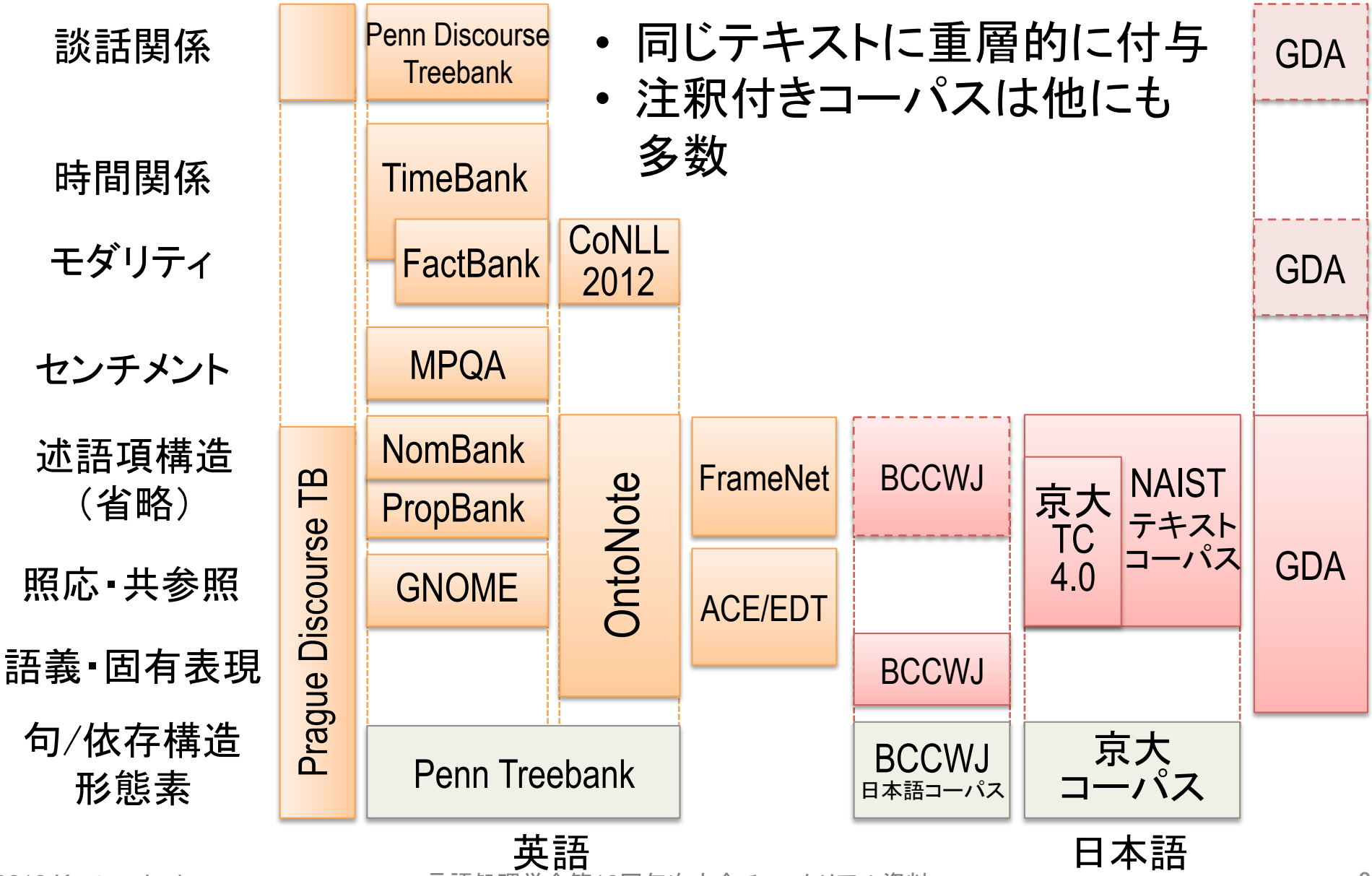
意味談話解析：テキスト→意味表現

① 漸進的（水平）アプローチ
- Corpus annotation layer cake
- 各レイヤの入出力を設計

② 応用指向的（垂直）アプローチ
- 問題を限定し，出口を設定
- 情報（関係）抽出

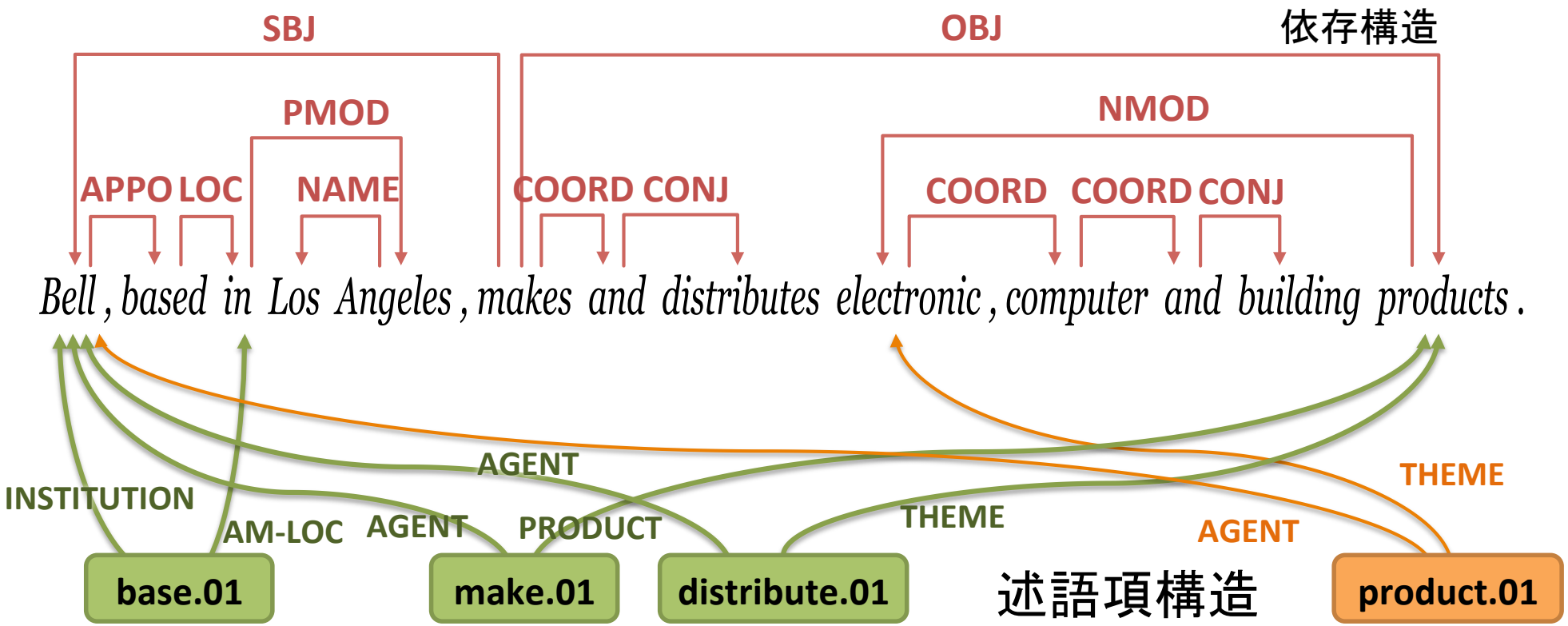
Corpus Annotation Layer Cake

- 同じテキストに重層的に付与
- 注釈付きコーパスは他にも多数



述語項構造

- PropBank（Palmer+ 05）: 動詞の項
- NomBank（Meyers+ 04）: 名詞の項



述語項構造

イベントの事実性

- FactBank（Sauri & Pustejovsky 09）

	Positive	Negative	Underspecified
Certain	Fact: <CT,+>	Counterfact: <CT,−>	Certain but unknown output: <CT, u>
Probable	Probable: <PR,+>	Not probable: <PR,−>	(NA)
Possible	Possible: <PS,+>	Not certain: <PS,−>	(NA)
Underspecif.	(NA)	(NA)	Unknown or uncommitted: <U,u>

R. Sauri. FactBank 1.0 Annotation Guidelines, 2008

イベントの事実性

Newspaper reports have said Amir was infatuated with Har-Shefi and may have been trying to impress her by killing the prime minister.

Event (ID):	Source (ID):	Fact. value:
said (e22)	*author* (s_0)	CT+
infatuated (e23)	*reports_author* (s_2_s_0)	CT+
	author (s_0)	Uu
trying (e24)	*reports_author* (s_2_s_0)	PS+
	author (s_0)	Uu
impress (e25)	*reports_author* (s_2_s_0)	Uu
	author (s_0)	Uu
killing (e26)	*reports_author* (s_2_s_0)	Uu
	author (s_0)	Uu

R. Sauri. FactBank 1.0 Annotation Guidelines, 2008

共参照関係

- OntoNotes（Pradhan+ 11）
  - *She had a good suggestion and it was unanimously accepted by all.*
  - *Japan's domestic sales of cars, trucks and buses in October rose 18% from a year earlier to 500,004 units, a record for the month, the Japan Automobile Dealers' Association said. The strong growth followed year-to-year increases of 21% in August and 12% in September.*

S. Pradhan, L. Ramshow, M. Marcus, M. Palmer, R.Weischedel, and N. Xue（2011）
CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes, CoNLL Shared Task.

テキストから意味表現へ

共参照

事実性

述語項構造

固有表現

set-up(e1), factuality(e1, CT+),
agent(e1, x1), theme(e1, x2),
co-agent(e1, x3)
Bridgestone-Sports-Co.(x1),
joint-venture(x2),
local-concern(x3),

Bridgestone Sports Co. said Friday it has ***set up*** *a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.*

テキストから意味表現へ

capitalize(e2),
factuality(e2, CT+),
theme(e2, x2),
Bridgestone-Sports-Taiwan-Co.(x2),
amount(e2, NT$2,000,000)

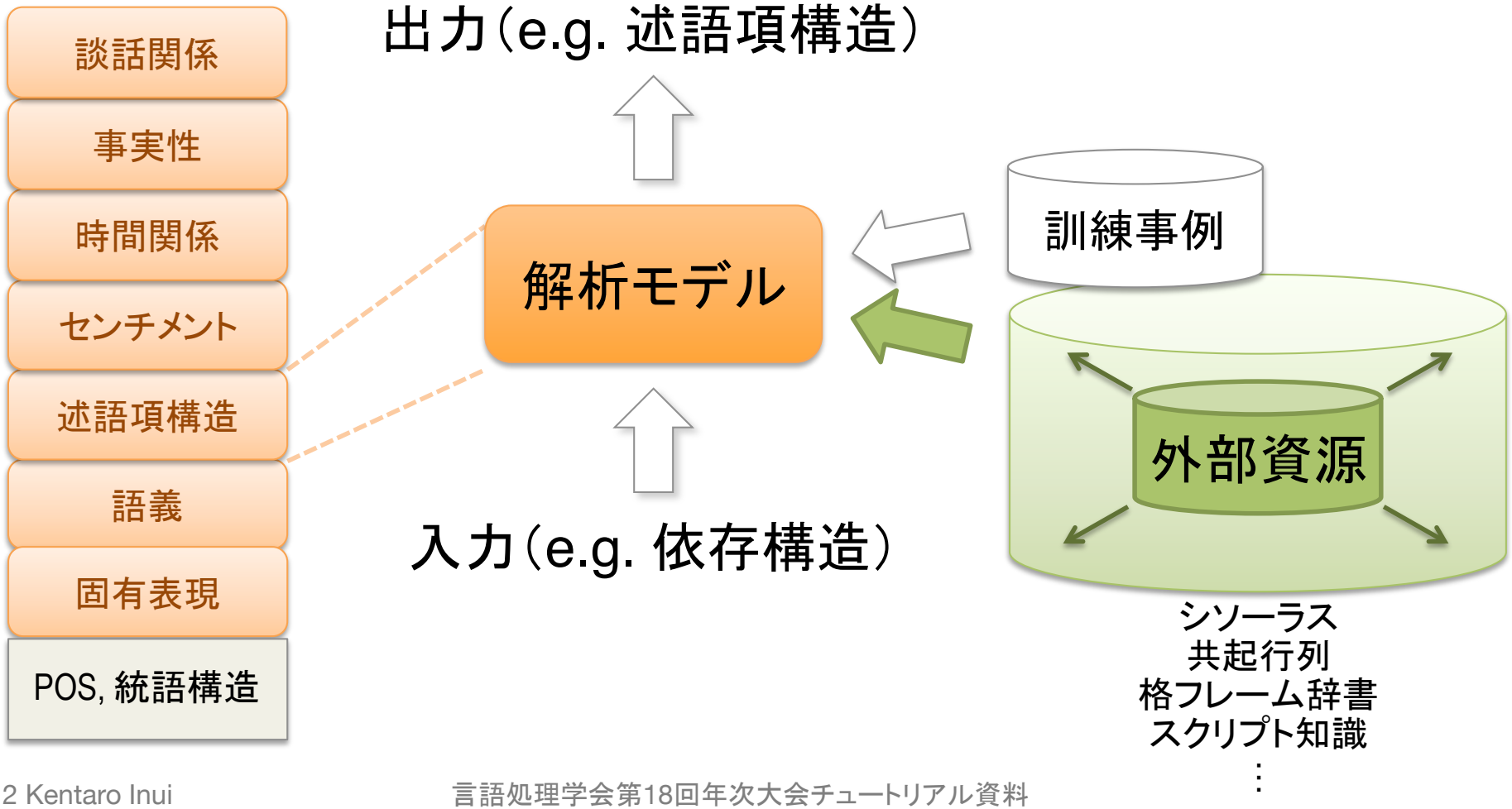
set-up(e1), factuality(e1, CT+),
agent(e1, x1), theme(e1, x2),
co-agent(e1, x3)
Bridgestone-Sports-Co.(x1),
joint-venture(x2),
local-concern(x3),

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.

*The joint venture, Bridgestone Sports Taiwan Co., **capitalized** at 20 million new Taiwan dollars, will start production in January 1990.*

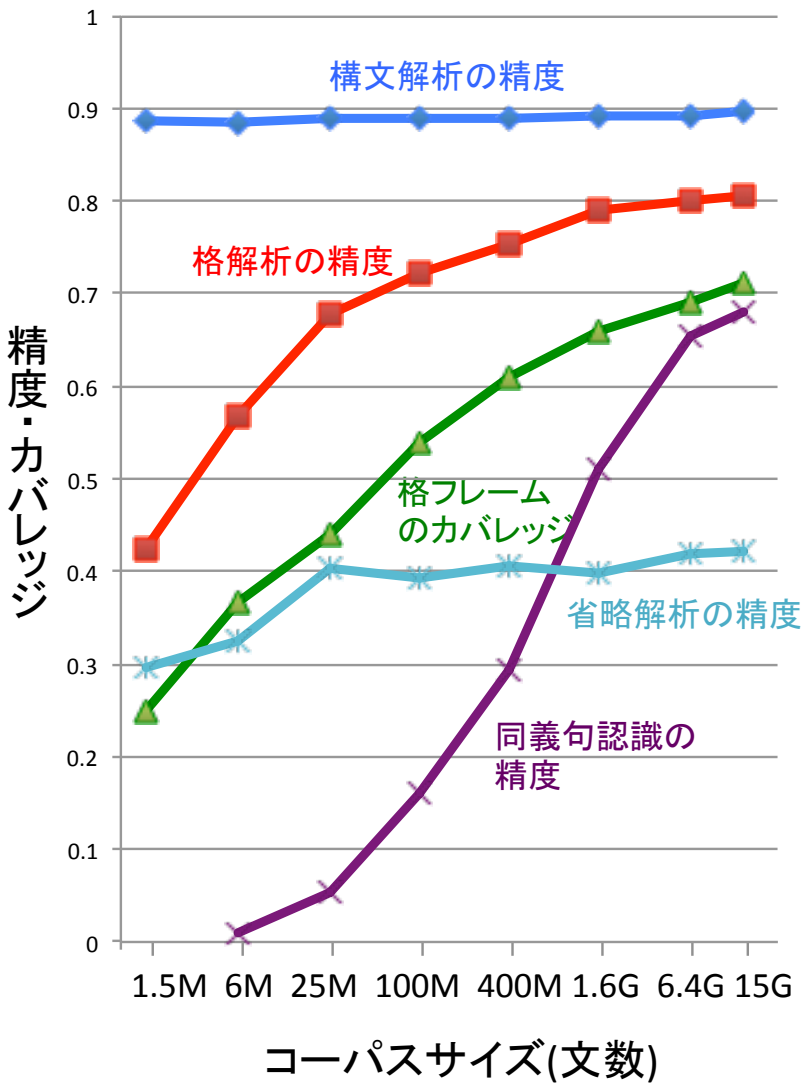
個別の部分問題での外部資源の利用

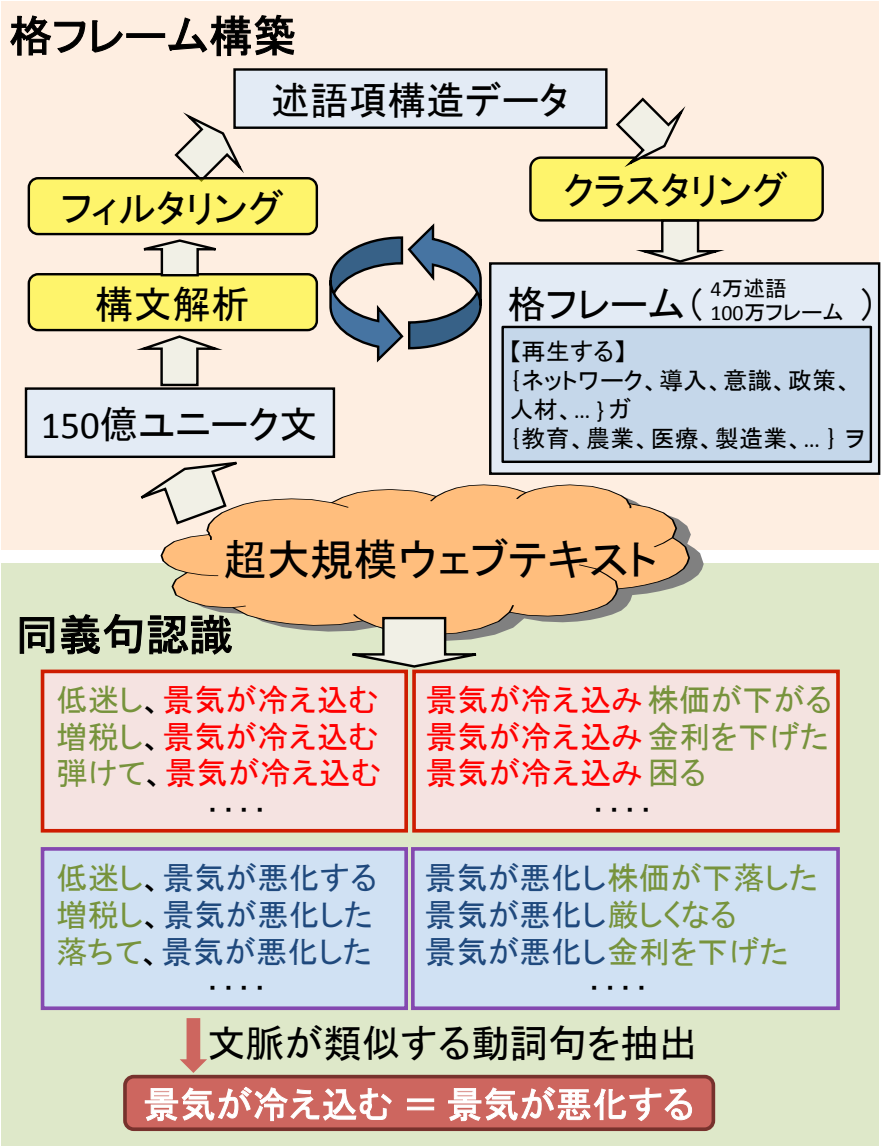
研究多数（NLPマジョリティ）



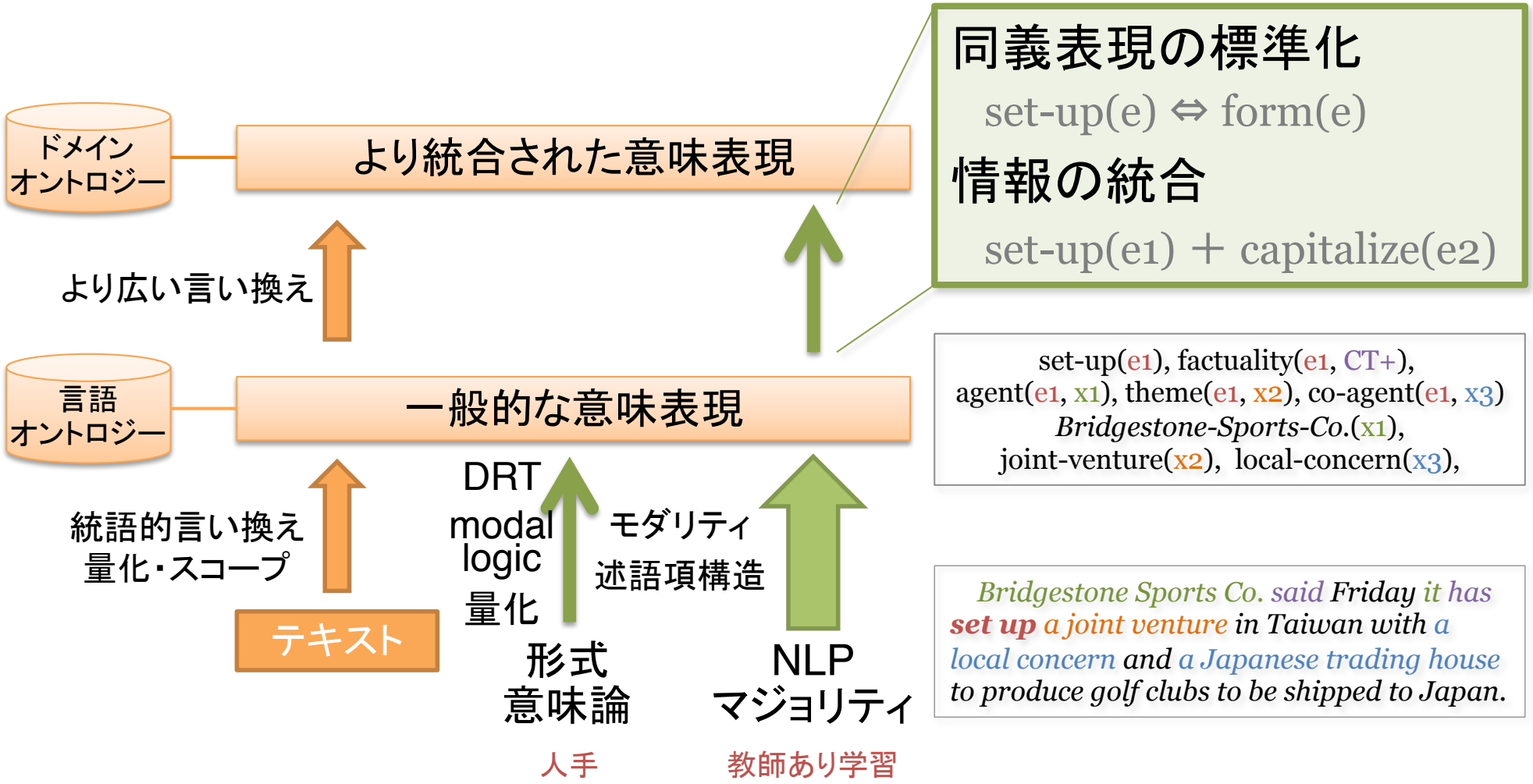
大規模外部資源のインパクト

（Kawahara & Kurohashi 06; Sasano+ 2010）





テキストから意味表現へ



©2012 Kentaro Inui

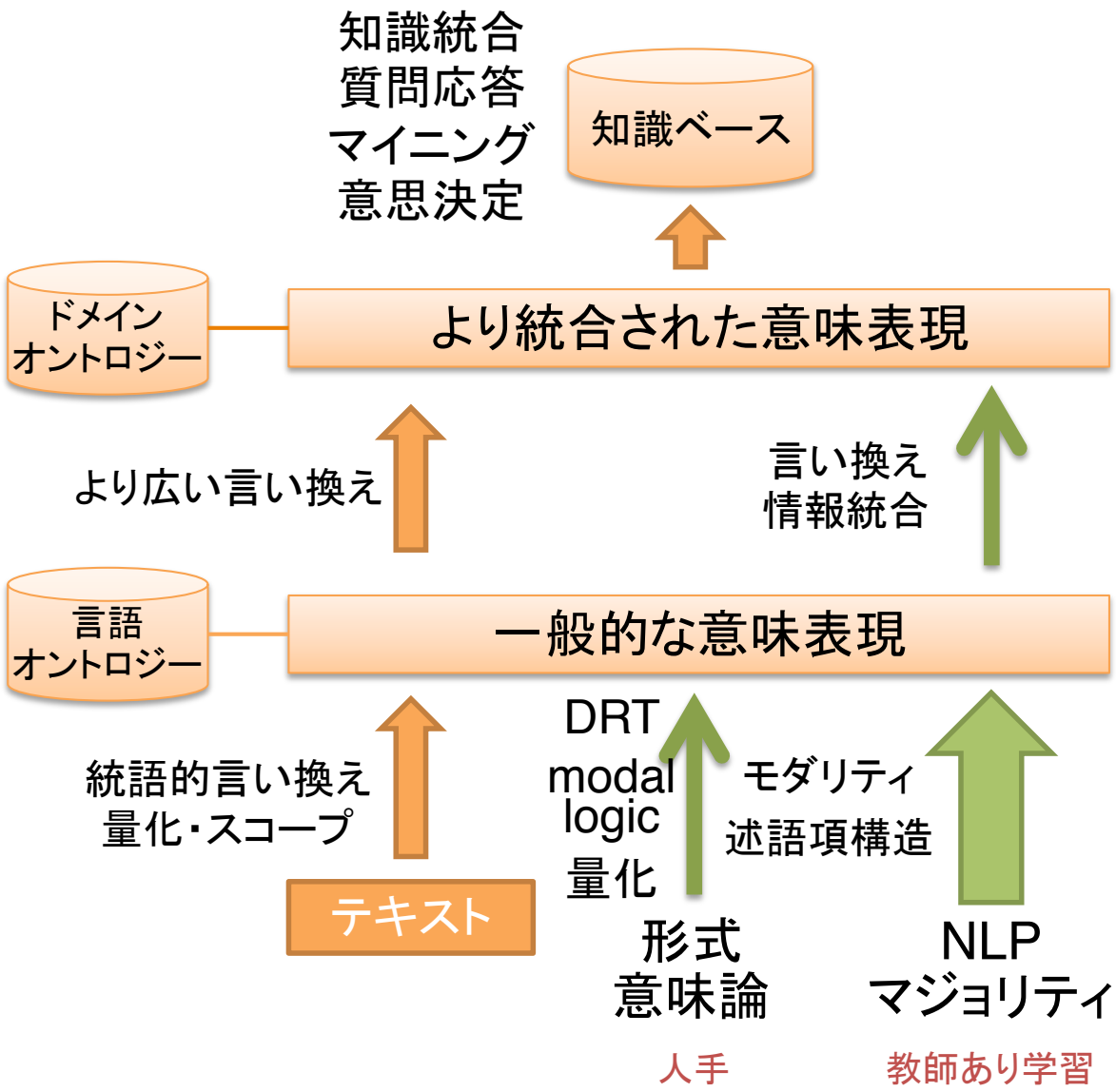
ドメインオントロジーへの対応付け

TIE-UP-1	
Relationship:	**TIE-UP**
Entities:	*"Bridgestone Sports Co."*
	"a local concern"
	"a Japanese trading house"
Joint Venture Company:	*"Bridgestone Sports Taiwan Co."*
Amount:	NT$2,000,000

capitalize(e2),
factuality(e2, CT+),
theme(e2, x2),
Bridgestone-Sports-Taiwan-Co.(x2),
amount(e2, NT$2,000,000)

set-up(e1), factuality(e1, CT+),
agent(e1, x1), theme(e1, x2),
co-agent(e1, x3)
Bridgestone-Sports-Co.(x1),
joint-venture(x2),
local-concern(x3),

テキストから意味表現へ



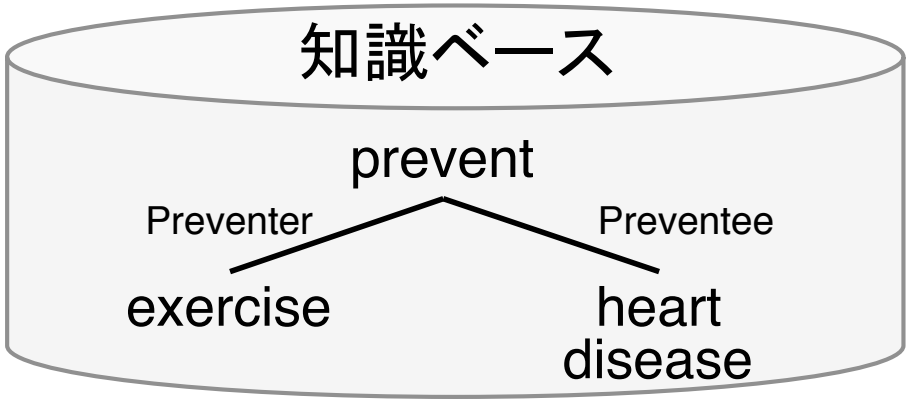
TIE-UP-1	
Relationship:	TIE-UP
Entities:	*"Bridgestone Sports Co."*
	"a local concern"
	"a Japanese trading house"
Joint Venture:	*"Bridgestone Sports Taiwan Co."*
Amount:	NT$2,000,000

set-up(e1), factuality(e1, CT+),
agent(e1, x1), theme(e1, x2), co-agent(e1, x3)
Bridgestone-Sports-Co.(x1),
joint-venture(x2), local-concern(x3),

Bridgestone Sports Co. said Friday it has ***set up*** *a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.*

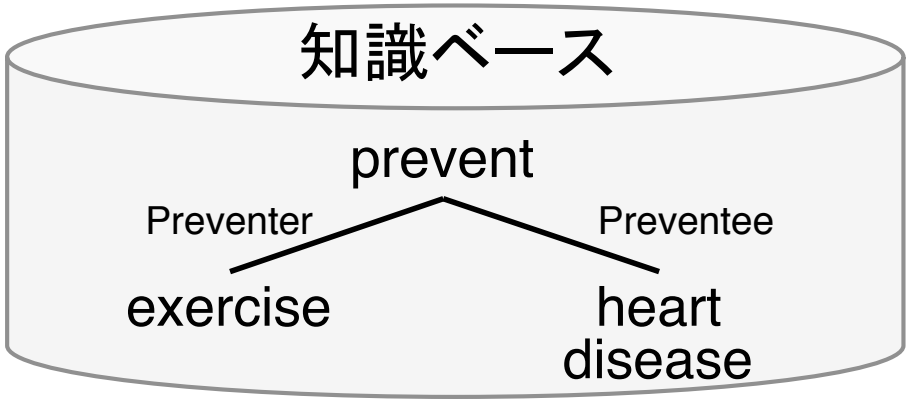
©2012 Kentaro Inui

多様な言い換えの存在



exercise { *prevents* / *protects against* } *heart disease*

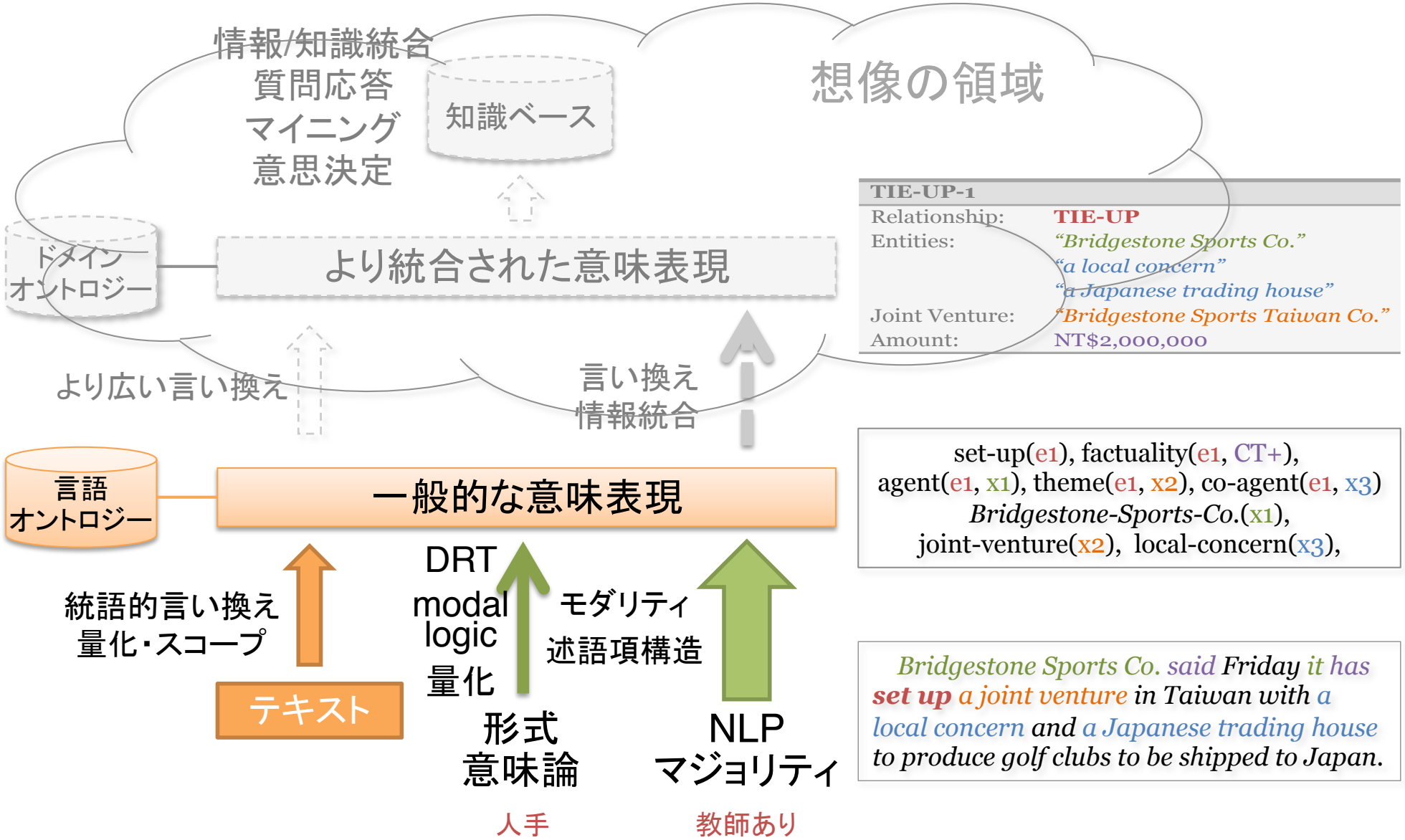
多様な言い換えの存在



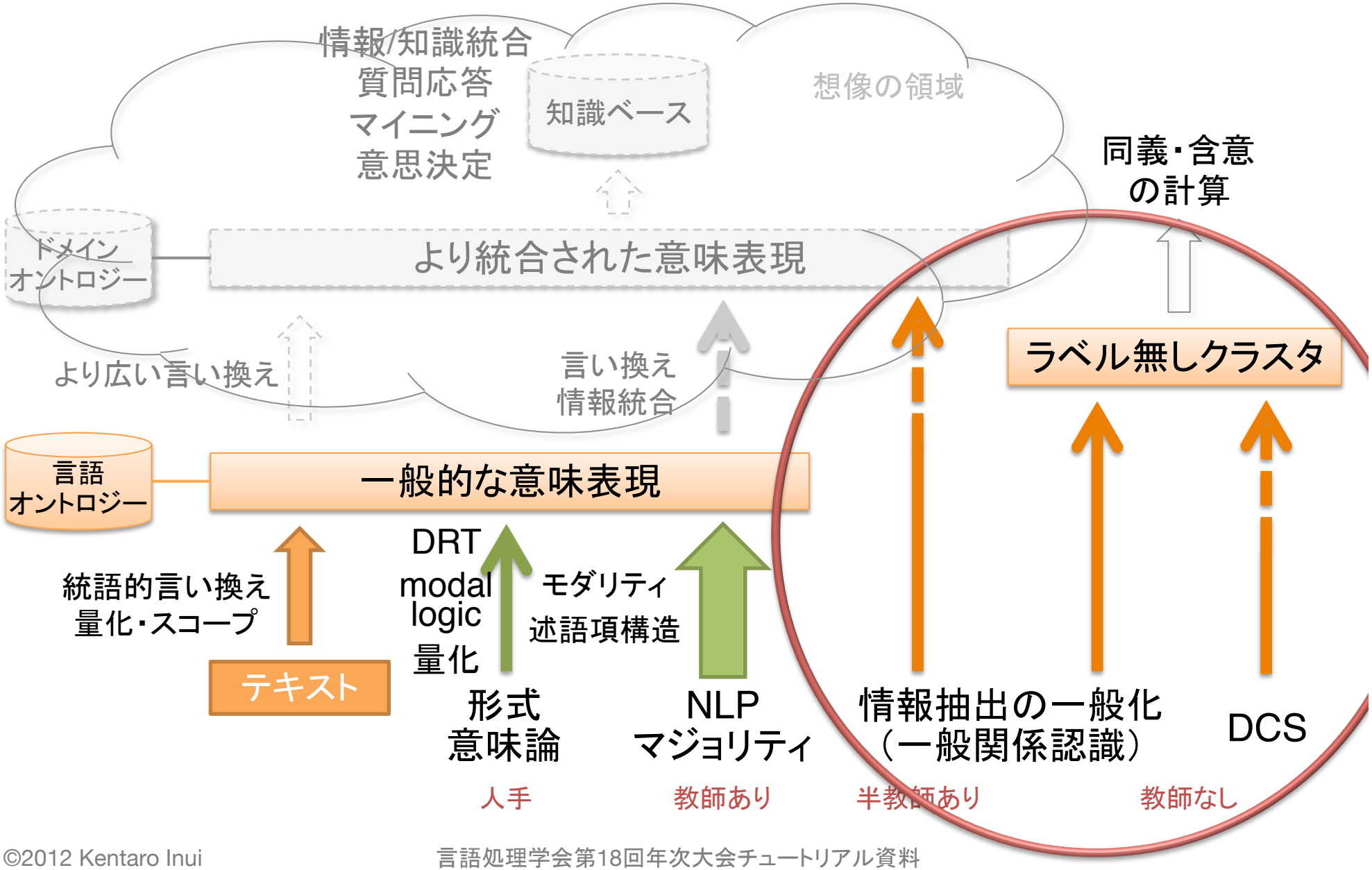
大問題！

exercise { *prevents*
protects against
has a special importance in the prevention of
plays a significant role in preventing
reduces the risk of
reduces your overall risk of developing
is associated with a lower risk of } *heart disease*

テキストから意味表現へ



意味解析の捉え方の広がり（地図①）



©2012 Kentaro Inui

本講演の構成

1. Annotation layer cake
2. 情報抽出と知識獲得の今
3. Making the implicit explicit
4. やわらかい記号の世界
5. ルールを作るところから

©2012 Kentaro Inui

情報抽出の展開（地図②）

意味談話解析
実現したい

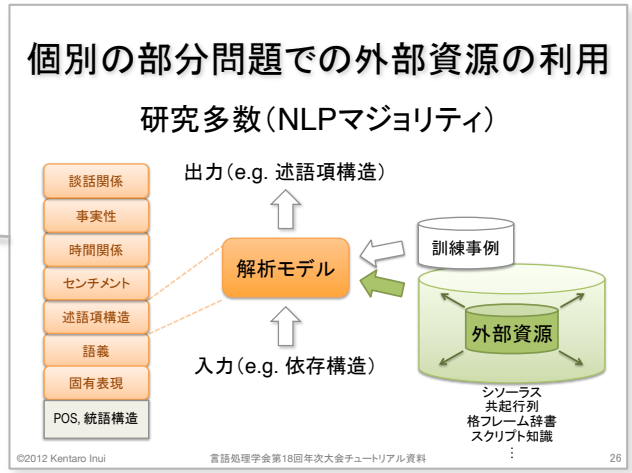
問題の単純化

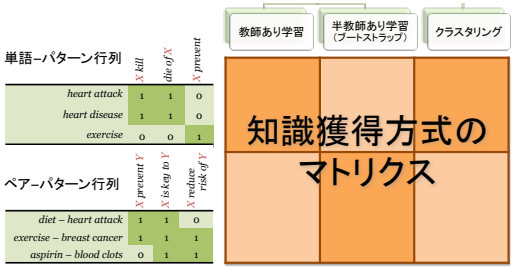
少数・固定の
NE/関係クラス

単一文書
情報抽出

手段

知識獲得





Example from MUC-5（1993）

Web誕生の年

テンプレート

TIE-UP-1	
Relationship:	**TIE-UP**
Entities:	*"Bridgestone Sports Co."*
	"a local concern"
	"a Japanese trading house"
Joint Venture Company:	*"Bridgestone Sports Taiwan Co."*
Amount:	NT$2,000,000

Bridgestone Sports Co. said Friday it has ***set up*** *a joint venture in Taiwan* ***with*** *a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.*

The joint venture, Bridgestone Sports Taiwan Co., ***capitalized at*** *20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.*

Onyshkevych, et al. Tasks, Domains, and Languages. Proc. of the 5th Message Understanding Conference, 1993.

パターンによる情報抽出

TIE-UP-1	
Relationship:	**TIE-UP**
Entities:	*"Bridgestone Sports Co."*
	"a local concern"
	"a Japanese trading house"
Joint Venture Company:	*"Bridgestone Sports Taiwan Co."*
Amount:	NT$2,000,000

意味クラス付きパターン

<Company> ***set up*** <Joint-Venture> ***with*** <Company>

<Company> ***capitalized at*** <Currency>

*Bridgestone Sports Co. said Friday it has **set up** a joint venture in Taiwan **with** a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.*

*The joint venture, Bridgestone Sports Taiwan Co., **capitalized at** 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.*

情報抽出の展開（地図②）

意味談話解析
実現したい

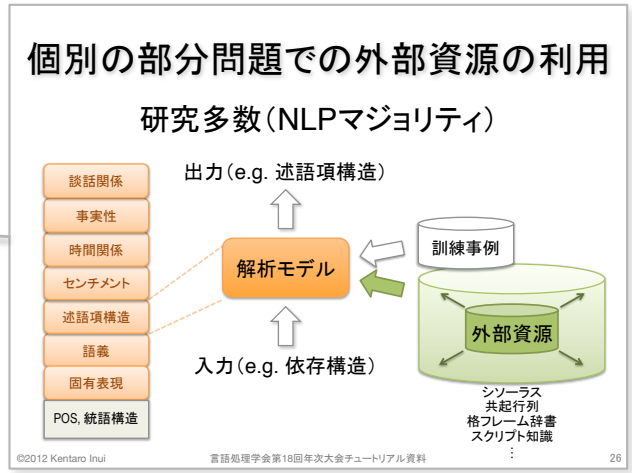
問題の単純化

少数・固定の
NE/関係クラス

単一文書
情報抽出

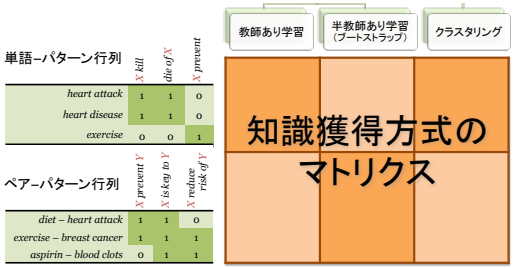
手段

知識獲得



パターンベースの情報抽出

- 情報抽出モデル＝抽出パターン集合
  → 解析モデルと知識獲得が一体
- 知識獲得方法に多様なアイデア



抽出のための知識

知識①

関係を表すパターン

<Company> ***set up*** <Joint-Venture> *with* <Company>

<Company> ***tie up*** <Joint-Venture> *with* <Company>

<Company> *and* <Company> ***enter into an agreement of*** <Joint-Venture>

知識②

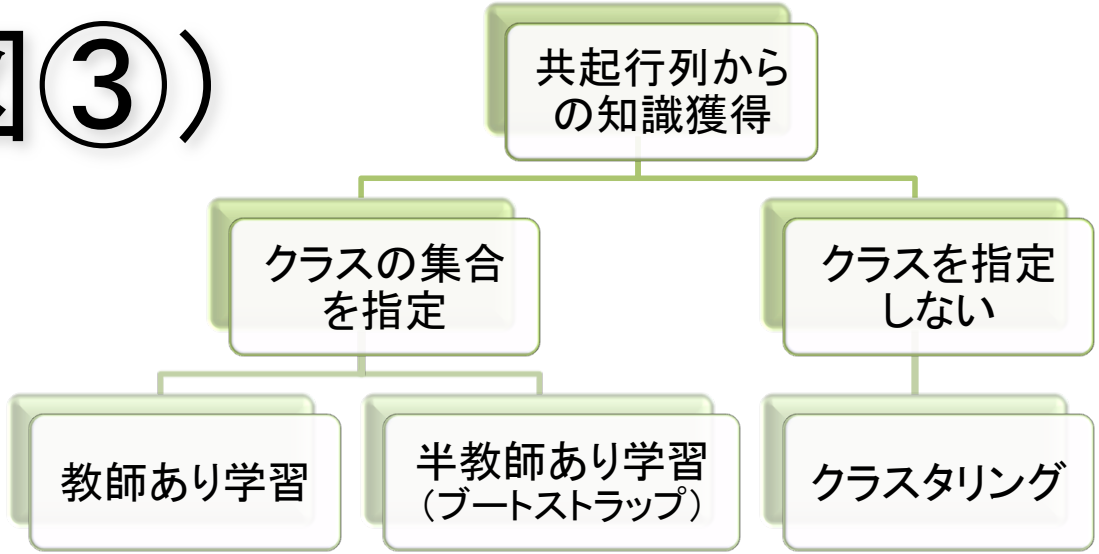
関係パターンの言い換え

***** Co.*
local concern
trading house
:

知識③

語の意味クラス

知識獲得（地図③）

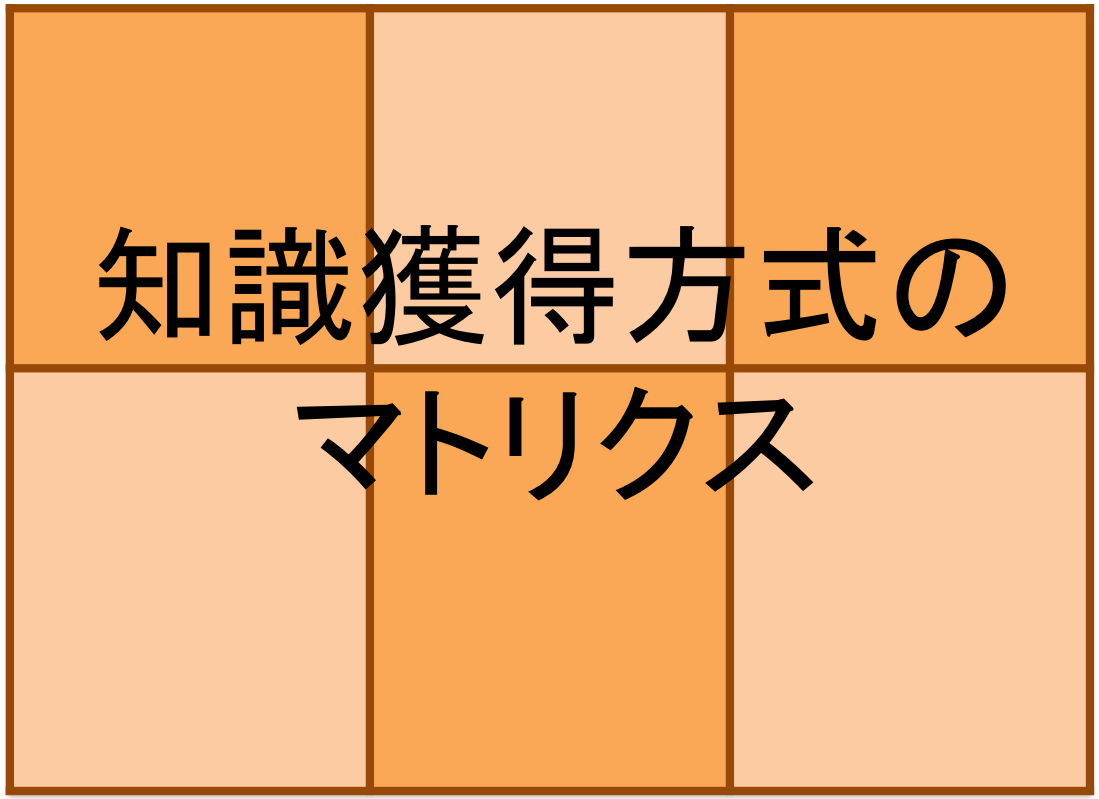


単語–パターン行列

	X kill	*die of X*	*X prevent*
heart attack	1	1	0
heart disease	1	1	0
exercise	0	0	1

ペア–パターン行列

	X prevent Y	*X is key to Y*	*X reduce risk of Y*
diet – heart attack	1	1	0
exercise – breast cancer	1	1	1
aspirin – blood clots	0	1	1



情報抽出の展開（地図②）

意味談話解析
実現したい

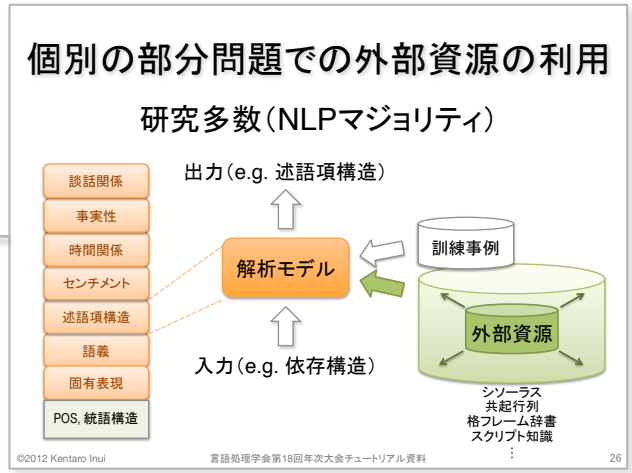
問題の単純化

少数・固定の
NE/関係クラス

単一文書
情報抽出

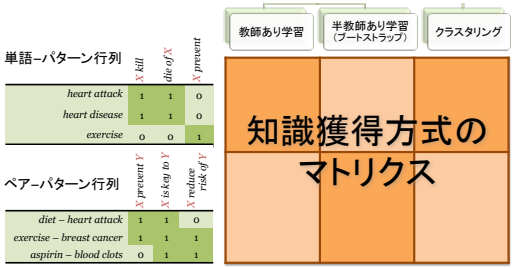
手段

知識獲得



パターンベースの情報抽出

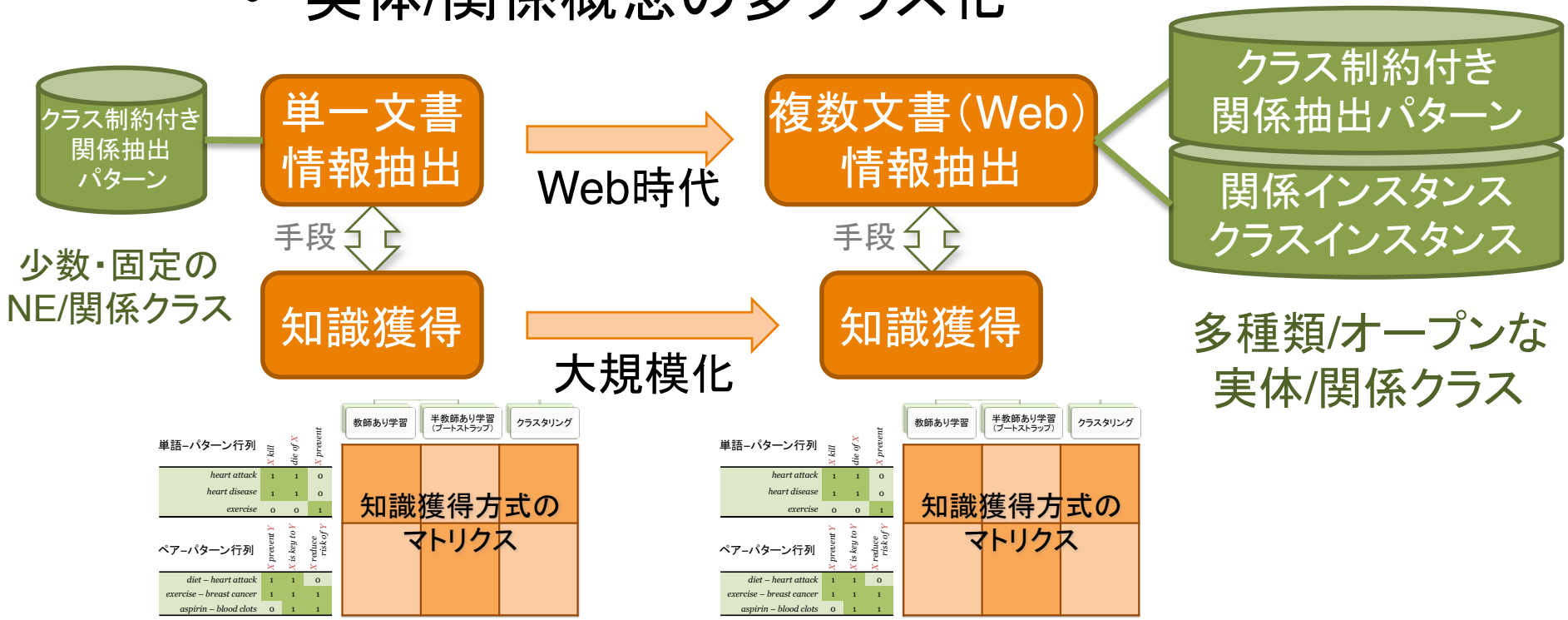
- 情報抽出モデル＝抽出パターン集合
  → 解析モデルと知識獲得が一体
- 知識獲得方法に多様なアイデア



情報抽出の展開（地図②）

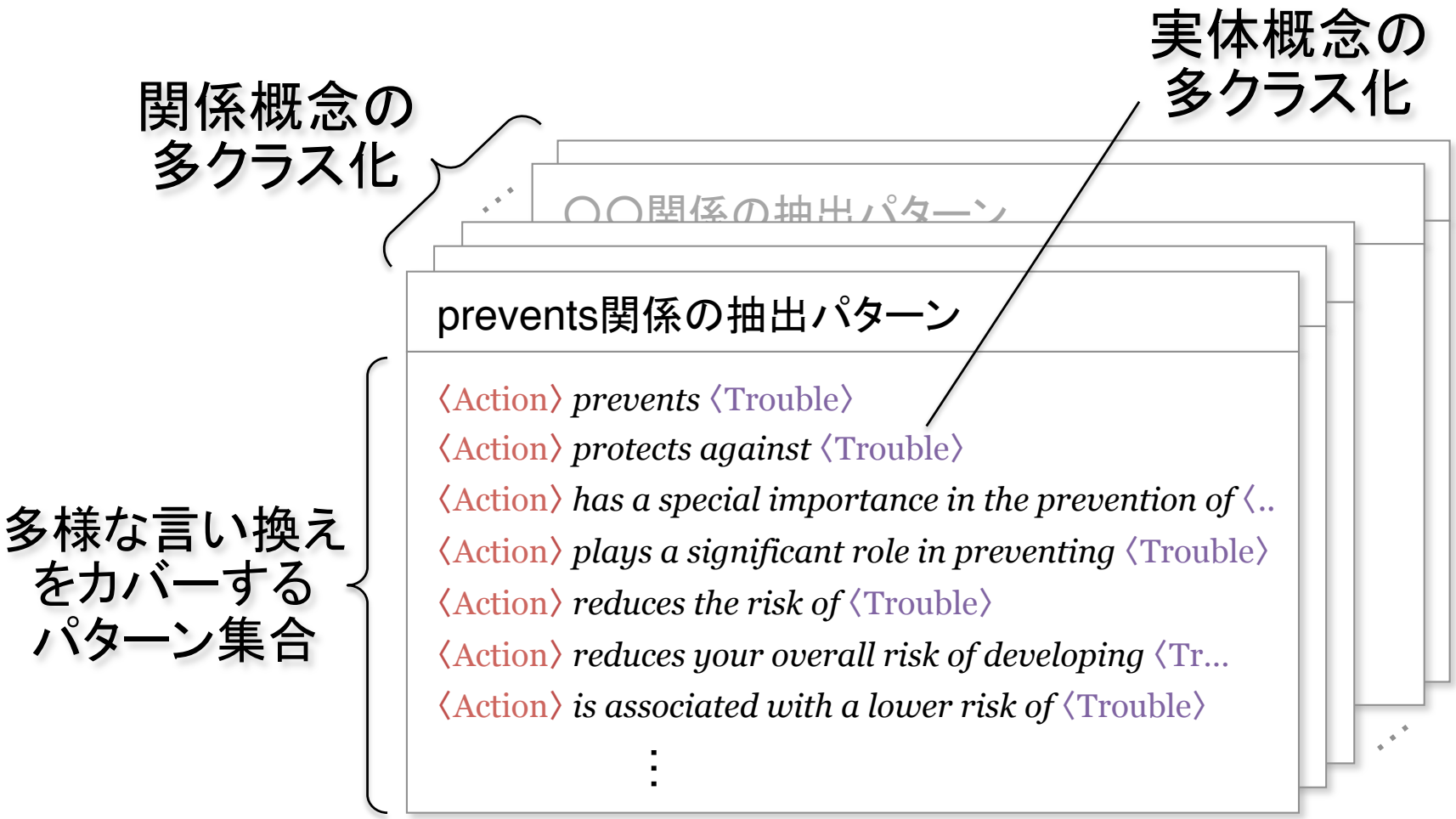
情報爆発時代へ

- 複数文書間の冗長性を利用（macro reading）
- 知識獲得が大規模化・実用化
- 実体/関係概念の多クラス化

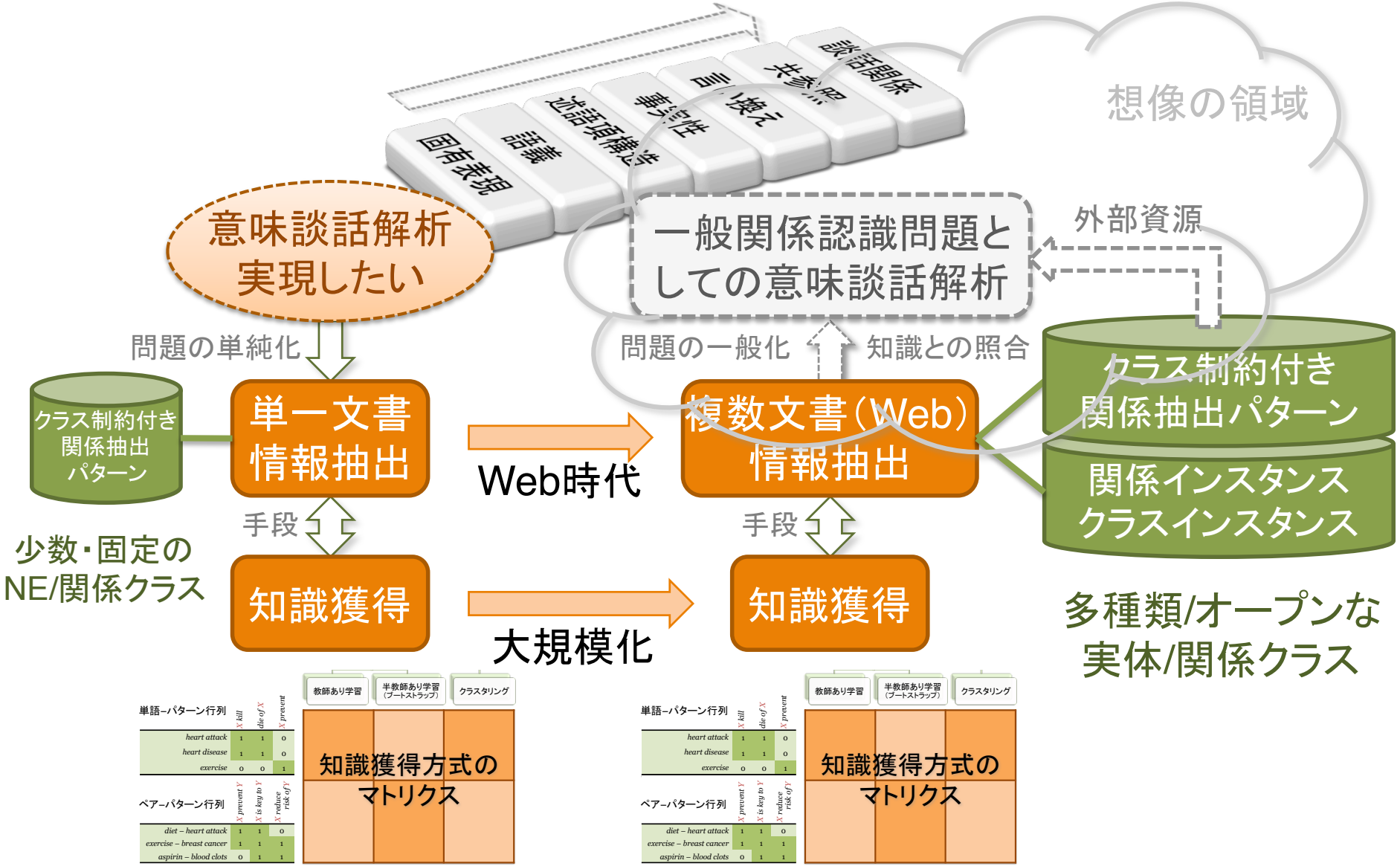


情報抽出の大規模化・多クラス化

意味解析の目的に接近

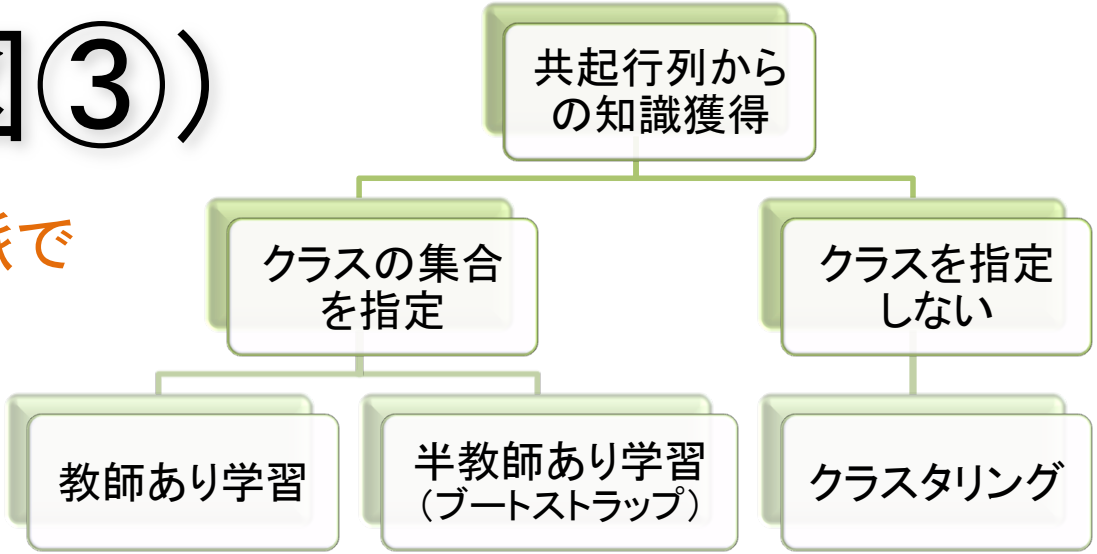


情報抽出の展開（地図②）


固有表現
語義
述語項構造
時間性
言い換え
共参照
談話関係
想像の領域
意味談話解析
実現したい
一般関係認識問題と
しての意味談話解析
外部資源
問題の単純化
問題の一般化
知識との照合
クラス制約付き
関係抽出
パターン
単一文書
情報抽出
Web時代
複数文書（Web）
情報抽出
クラス制約付き
関係抽出パターン
関係インスタンス
クラスインスタンス
手段
手段
少数・固定の
NE/関係クラス
知識獲得
大規模化
知識獲得
多種類/オープンな
実体/関係クラス
単語–パターン行列
X kill
die of X
X prevent
heart attack 1 1 0
heart disease 1 1 0
exercise 0 0 1
ペア–パターン行列
X prevent Y
X is key to Y
X reduce risk of Y
diet – heart attack 1 1 0
exercise – breast cancer 1 1 1
aspirin – blood clots 0 1 1
教師あり学習
半教師あり学習
（ブートストラップ）
クラスタリング
知識獲得方式の
マトリクス
単語–パターン行列
X kill
die of X
X prevent
heart attack 1 1 0
heart disease 1 1 0
exercise 0 0 1
ペア–パターン行列
X prevent Y
X is key to Y
X reduce risk of Y
diet – heart attack 1 1 0
exercise – breast cancer 1 1 1
aspirin – blood clots 0 1 1
教師あり学習
半教師あり学習
（ブートストラップ）
クラスタリング
知識獲得方式の
マトリクス

知識獲得（地図③）

情報抽出の一般化の文脈で

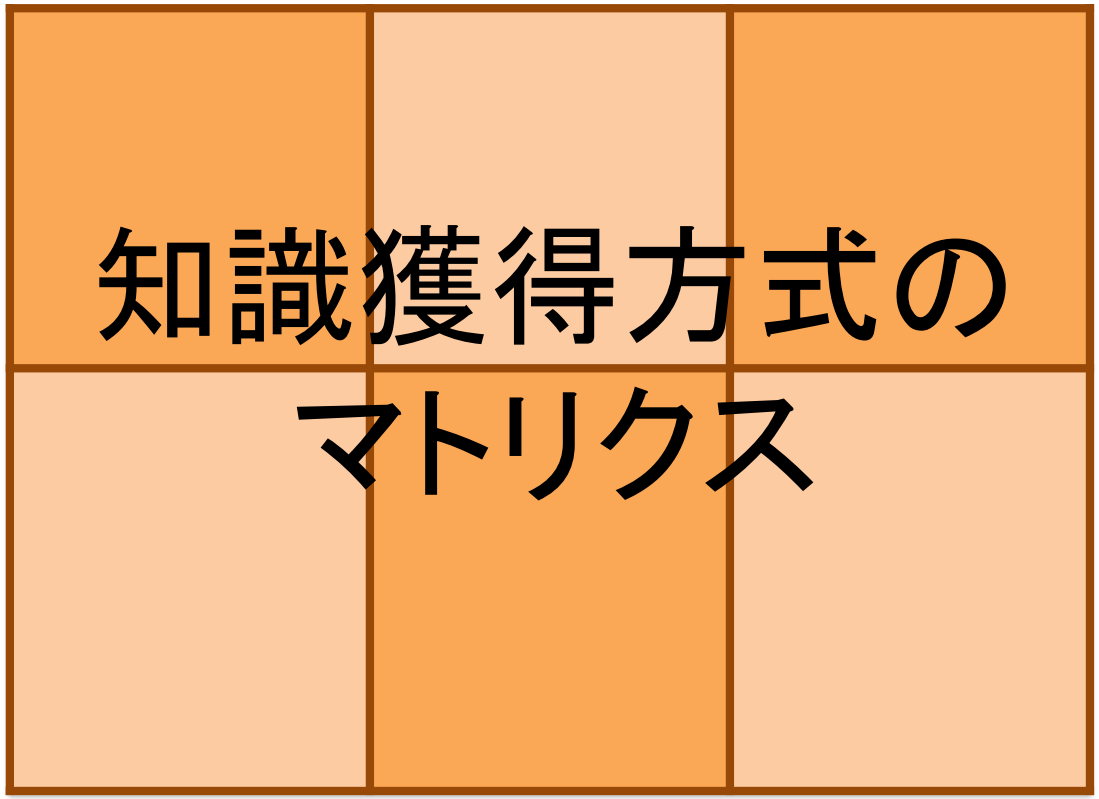


単語–パターン行列

	X kill	*die of X*	*X prevent*
heart attack	1	1	0
heart disease	1	1	0
exercise	0	0	1

ペア–パターン行列

	X prevent Y	*X is key to Y*	*X reduce risk of Y*
diet – heart attack	1	1	0
exercise – breast cancer	1	1	1
aspirin – blood clots	0	1	1



単語–パターン行列

- 単語の意味 ⇒ 出現文脈の分布から推測できる
  （Firth 57, Harris 64）
- 文脈の分布（文脈ベクトル）
  ＝ その単語の局所文脈に現れる表現の分布

	X kill	*X is for*	*X prevent*	*X be cause of*	*X be key to*	*X reduce risk*	*die of X*	*prevent X*	*reduce risk of X*
heart attack	1	0	0	1	0	0	1	1	1
heart disease	1	0	0	1	0	0	1	1	1
sunburn	0	0	0	1	0	0	0	1	1
diet	0	1	1	0	1	1	0	0	0
exercise	0	1	1	0	1	1	0	0	0

ペアーパターン行列

関係を表すパターン

関係インスタンス

	…	…	*X prevent Y*	…	*X is key to Y*	…	…	*X reduce risk of Y*	…
diet – heart attack	0	0	1	0	1	0	0	1	0
sunscreen – sunburn	0	0	1	0	1	0	0	0	0
…	1	0	0	0	0	1	0	0	0
exercise – breast cancer	0	0	1	0	0	0	0	1	0
…	0	1	0	0	0	0	1	0	0
aspirin – blood clots	0	0	0	0	1	0	0	1	0
…	0	0	0	0	0	0	1	0	1
safety belt – fatalities	0	0	0	0	0	0	0	1	1

ペアーパターン行列の作り方

- 依存構造（NVN, NPN, NVNPN, VCV, ...）
  - Sekine（2006）, Shinyama & Sekine（2006）, Abe+（2008）, Peñas & Hovy（2010）, Wu & Weld（2010）, Shibata & Kurohashi（2011）, etc.
- 述語項構造（APA, APAA, ...）
  - Christensen+（2011）, etc.
- 文中のすべての名詞ペア
  - De Saeger+（2009）, etc.
- 品詞列＋学習
  - Banko+（2007）（TextRunner）
  - Fader+（2011）（ReVerb）, etc.

ReVerb Search

http://openie.cs.washington.edu/

（Fader+ 2011）

ReVerb took 0.10 seconds. (cache miss)

Displaying **186** of **186** results for Argument 1 containing "**exercise**" and Argument 2 containi

Grouping results by argument 1. Group by: ***argument 2*** | ***predicate***

Exercise (115 results)

Exercise reduces the risk of (27), helps prevent (42), can reduce the risk of (11), ***80 more...*** **heart disease**

Exercise can be an excellent tool in (2), is an excellent tool in (1), has a special importance in (1), has numerous other benefits in (1) the prevention of **heart disease**

Regular , moderate **exercise** helps preventing (1), plays a significant role in preventing (1) **heart** and blood vessel **disease**

Exercise protects against (1), can help reduce your overall risk of developing (1) **heart disease** and stroke

Exercise reduces (3) **heart disease** risk

physical **exercise** is associated with (2) a lower risk of coronary **heart disease**

exercise can be a key part of managing (1) high blood pressure and **heart disease**

Exercise improves (1) blood pressure and **heart disease** risk

Regular **exercise** prevent (1) cancer & **heart disease**

exercise helps (1) control **heart disease**

ReVerb Search

http://openie.cs.washington.edu/

（Fader+ 2011）

ReVerb took 22.89 seconds. (cache miss)

Displaying **300** of **6554** results for Predicate containing "**prevent**"

Grouping results by predicate. Group by: ***argument 2*** | ***argument 1***

prevents (300 results)

Vaccines (15), Immunisation (12), Condoms (10), ***8 more...*** can **prevent** disease

Condoms (18), emergency contraception (13), Contraception (8), ***7 more...*** **prevents** pregnancy

Fiber (28), Exercise (8), Water (5) helps **prevent** constipation

Sturdy base (21), Sturdy construction (3), Stable design (3), controlled turning radius (3) **prevents** tipping

vaccine (16), Vaccination (4), Good hygiene (4), ***2 more...*** can **prevent** infection

prior planning (15), Proper prior planning (8), Proper preparation (4), 'Proper preparation (6) **prevents** poor performance

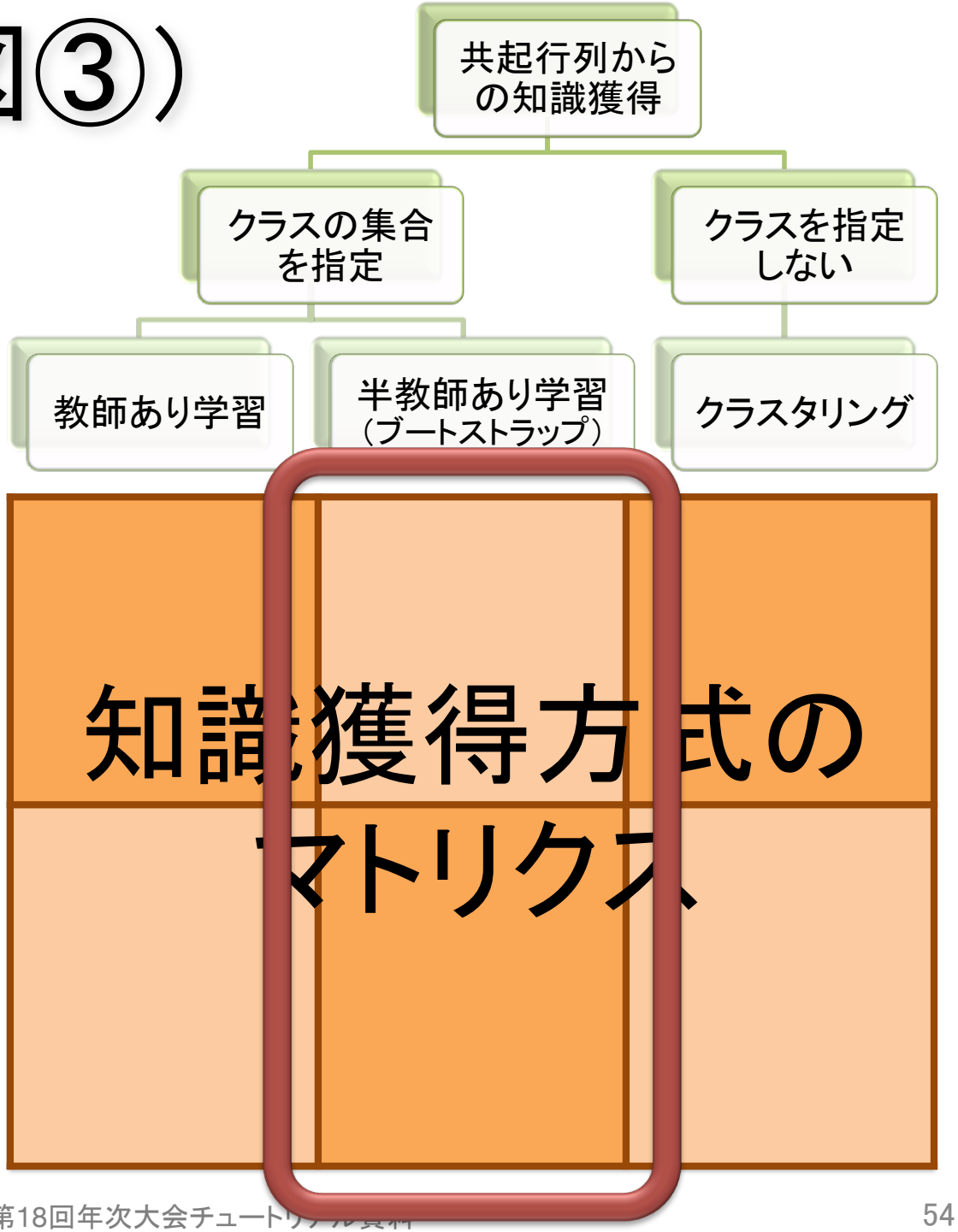
Exercise (11), diet (6), Physical activity (5), ***2 more...*** helps **prevent** heart disease

Weight-bearing exercise (9), Calcium (7), Estrogen (4), Vitamin D (3) helps **prevent** osteoporosis

vaccine (9), Removing polyps (4), Plant-based diets (4), Green Tea (3) **prevents** cancer

treatment (6), Proper dental care (4), method (3), ***2 more...*** can **prevent** future problems

知識獲得（地図③）



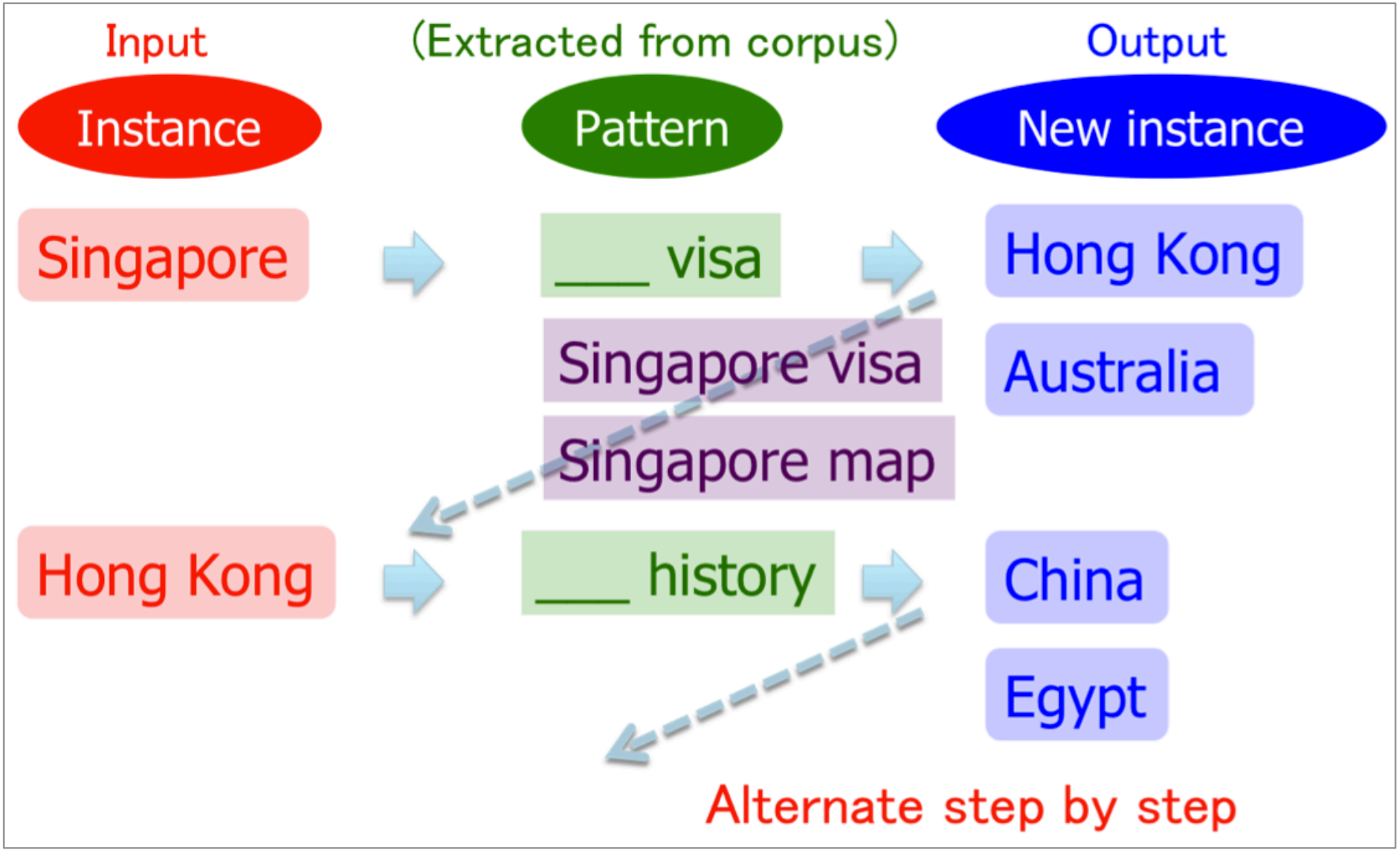
単語–パターン行列

	X kill	*die of X*	*X prevent*
heart attack	1	1	0
heart disease	1	1	0
exercise	0	0	1

ペア–パターン行列

	X prevent Y	*X is key to Y*	*X reduce risk of Y*
diet – heart attack	1	1	0
exercise – breast cancer	1	1	1
aspirin – blood clots	0	1	1

半教師あり学習（ブートストラップ）



Slide from M. Komachi (2010) Graph-Theoretic Approaches to Minimally-Supervised Natural Language Learning, Ph.D. Thesis.

半教師あり学習（ブートストラップ）

Riloff & Jones（1999）

シード

bolivia, city, colombia, district, guatemala, honduras, region, nicaragua, neighborhood, town

Best pattern	"headquartered in <x>" (F=3,N=4)
Known locations	*nicaragua*
New locations	*san miguel, chapare region, san miguel city*
Best pattern	"gripped <x>" (F=2,N=2)
Known locations	*colombia, guatemala*
New locations	none
Best pattern	"downed in <x>" (F=3,N=6)
Known locations	*nicaragua, san miguel*, city*
New locations	*area, usulutan region, soyapango*
Best pattern	"to occupy <x>" (F=4,N=6)
Known locations	*nicaragua, town*
New locations	*small country, this northern area, san sebastian neighborhood, private property*
Best pattern	"shot in <x>" (F=5,N=12)
Known locations	*city, soyapango**
New locations	*jauja, central square, head, clash, back, central mountain region,*

©2012 Kentaro Inui

半教師あり学習（ブートストラップ）

Espresso（Pantel+ 2006）

パターンの信頼度

インスタンスの信頼度

$$r_\pi(p) = \frac{\sum_{i \in I} \left(\frac{pmi(i,p)}{max_{pmi}} \times r_\iota(i) \right)}{|I|}$$

$$r_\iota(i) = \frac{\sum_{p \in P} \left(\frac{pmi(i,p)}{max_{pmi}} \times r_\pi(p) \right)}{|P|}$$

インスタンスの信頼度

パターンの信頼度

ペアーパターン行列でも同じ

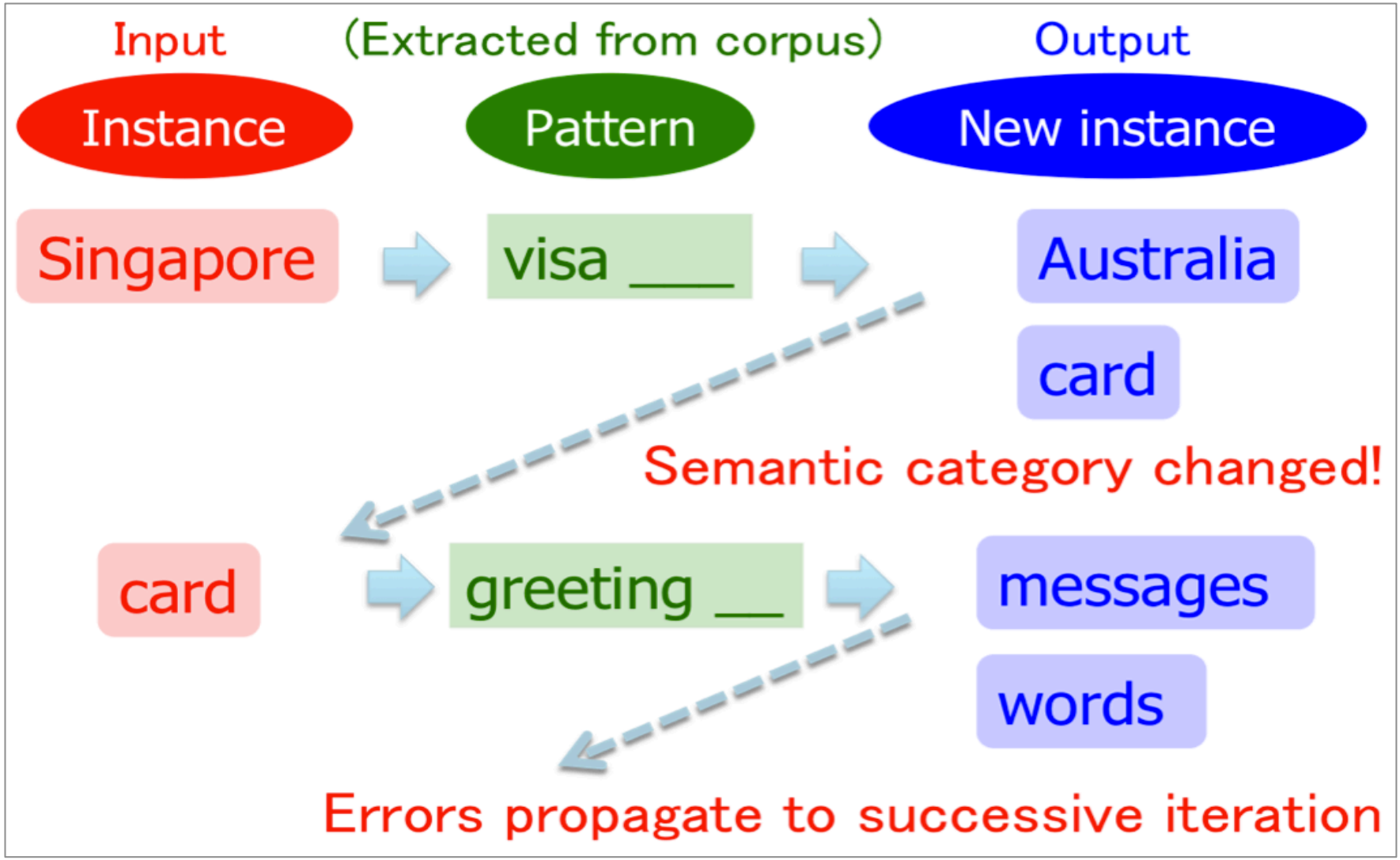
		…	…	Prevents *X prevent Y*	…	Prevents *X is key to Y*	…	…	Prevents *X reduce risk of Y*	…
Prevents	*diet – heart attack*	0	0	1	0	1	0	0	0	0
Prevents	*sunscreen – sunburn*	0	0	1	0	1	0	0	0	0
	…	1	0	0	0	0	1	0	0	0
Prevents	*exercise – breast cancer*	0	0	1	0	1	0	0	1	0
	…	0	1	0	0	0	0	1	0	0
Prevents	*aspirin – blood clots*	0	0	1	0	1	0	0	1	0
	…	0	0	0	0	0	0	0	0	1
Prevents	*safety belt – fatalities*	0	0	0	0	0	0	0	1	1

TextRunner（Banko+ 07）の出力から

Acqusition	Birthplace	InventorOf	WonAward
X.* acquire .*Y	X.* be born in .*Y	X.* invent .*Y	X.* win .*Y
X.* to acquire .*Y	X.* would be born in .*Y	X.* create .*Y	X.* have win .*Y
X.* to buy .*Y	X.* have be born in .*Y	X.* develop .*Y	X.* will win .*Y
X.* have acquire .*Y	X.* be born in of .*Y	X.* have invent .*Y	X.* to win .*Y
X.* purchase .*Y	X.* who be born in .*Y	X.* who invent .*Y	X.* receive .*Y
X.* have buy .*Y	X.* to be born in .*Y	X.* introduce .*Y	X.* win in .*Y
X.* is buying .*Y	X.* will be born in .*Y	X.* patent .*Y	X.* can win .*Y
X.* merge with .*Y	X.* marry be born in .*Y	X.* is inventing .*Y	X.* leave .*Y
X.* have purchase .*Y	X.* be born in before .*Y	X.* demonstrate .*Y	X.* would win .*Y
X.* will acquire .*Y	X.* must be born in .*Y	X.* bring .*Y	X.* is winning .*Y
X.* buy out .*Y		X.* have create .*Y	X.* may have win .*Y
X.* is acquiring .*Y		X.* may have invent .*Y	X.* could win .*Y
X.* pay for .*Y		X.* have bring .*Y	X.* will win in .*Y

M. Banko (2009) Open Information Extraction from the Web. Ph.D. Thesis.

意味ドリフト（意味の曖昧性）



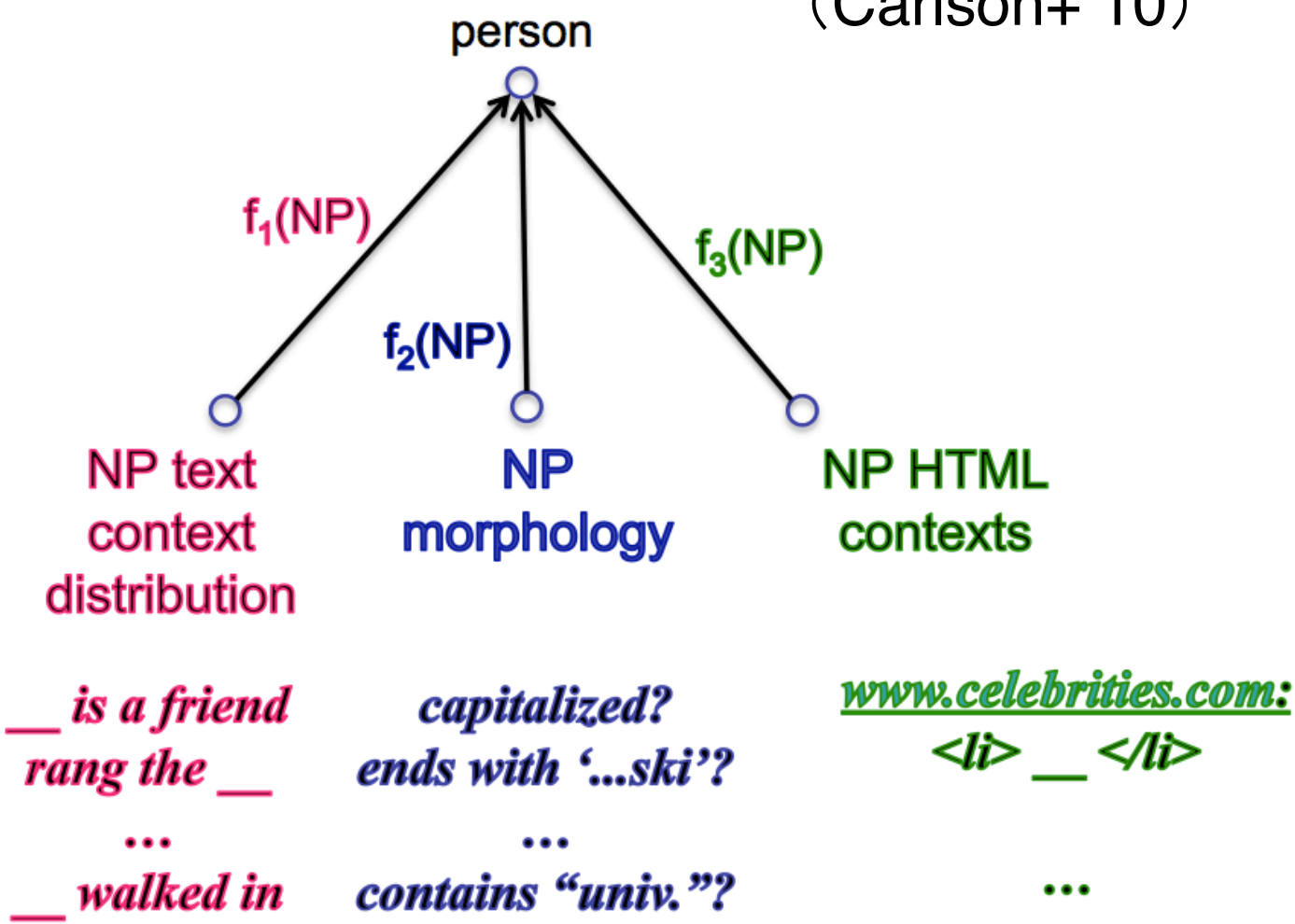
M. Komachi (2010) Graph-Theoretic Approaches to Minimally-Supervised Natural Language Learning, Ph.D. Thesis.

意味ドリフト対策（曖昧性解消）

- 多様な手がかりを併用
  - Carlson+（2010）, etc.
- クラスの排他性を利用
  - Riloff and Jones（1999）, Yangarber（2003）, etc.
- パターンに意味クラス制約を導入
  - Pasca+（2006）, Rosenfeld+（2007）,
    De Saeger+（2009）, Carlson+（2009）, etc.
- クラス獲得と関係獲得を統合
  - Carlson+（2009）: Coupled pattern learning
- 途中で人間の判断を追加

多様な手がかりを併用

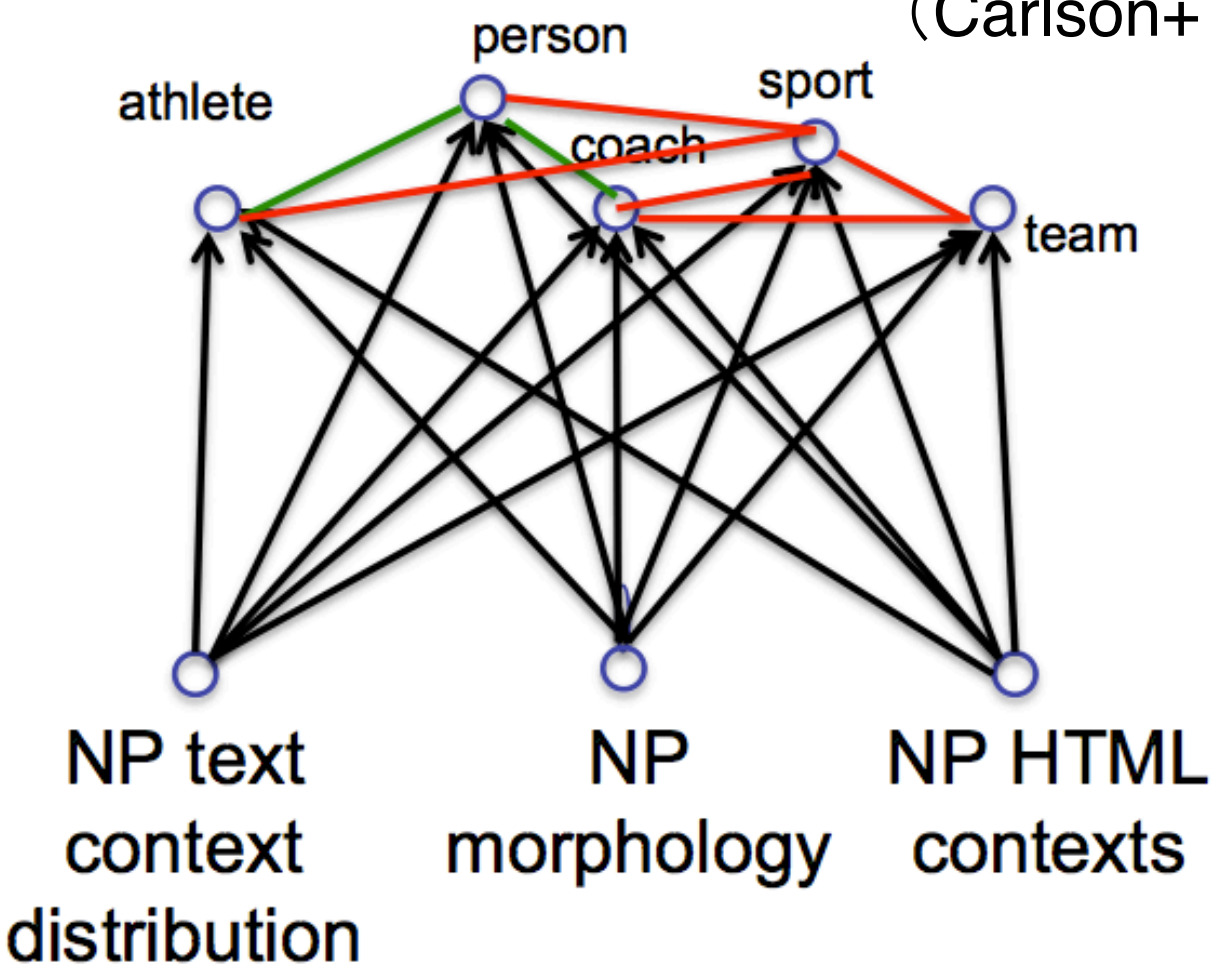
（Carlson+ 10）



Slides from Tom Mitchell's invited talk in the Univ. of Washington CSE Distinguished Lecture Series, October, 2010.
http://rtw.ml.cmu.edu/slides/Mitchell_NELL_UW_Oct_21_2010.pdf

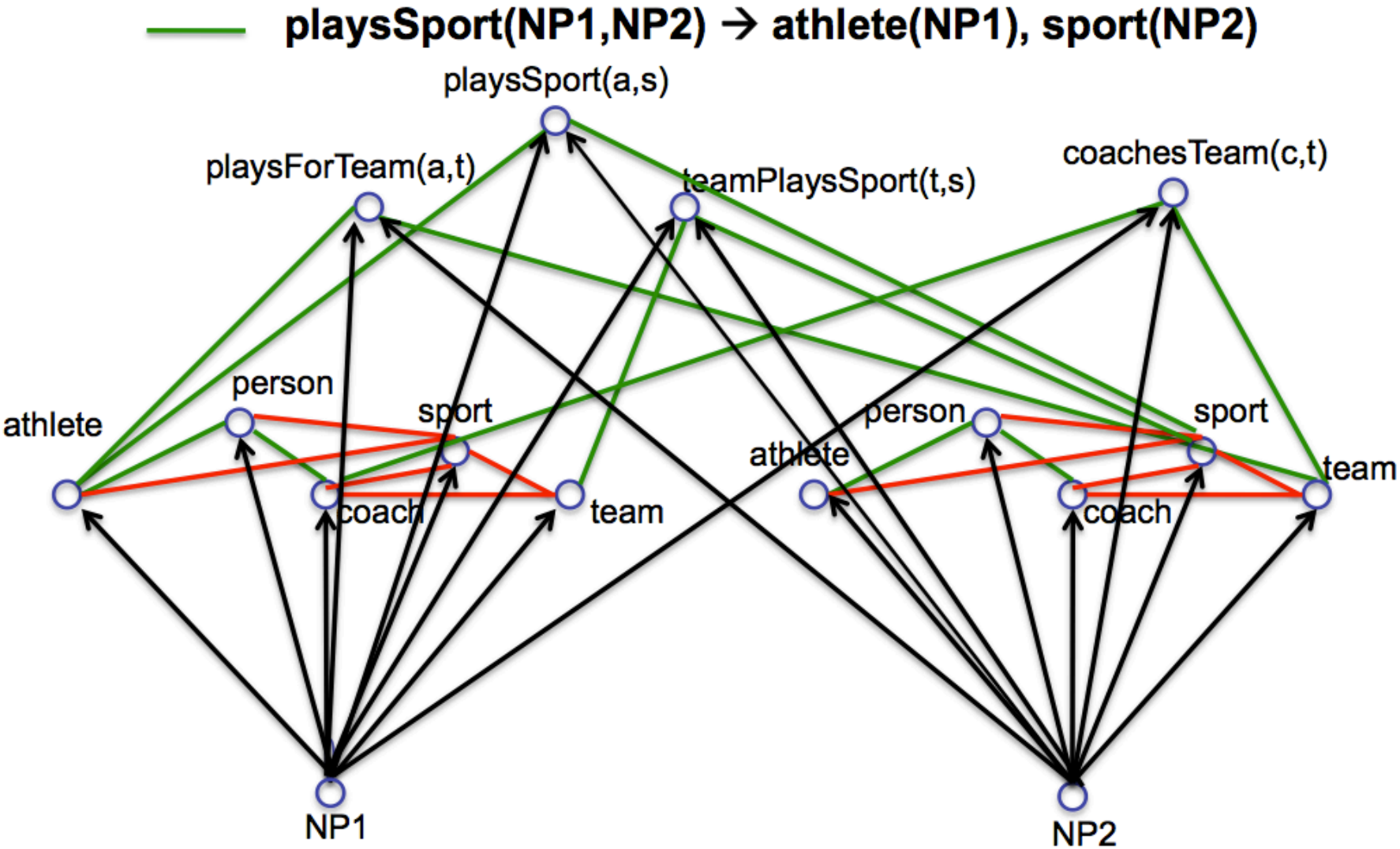
意味クラス間の排他制約

（Carlson+ 10）



Slides from Tom Mitchell's invited talk in the Univ. of Washington CSE Distinguished Lecture Series, October, 2010.
http://rtw.ml.cmu.edu/slides/Mitchell_NELL_UW_Oct_21_2010.pdf

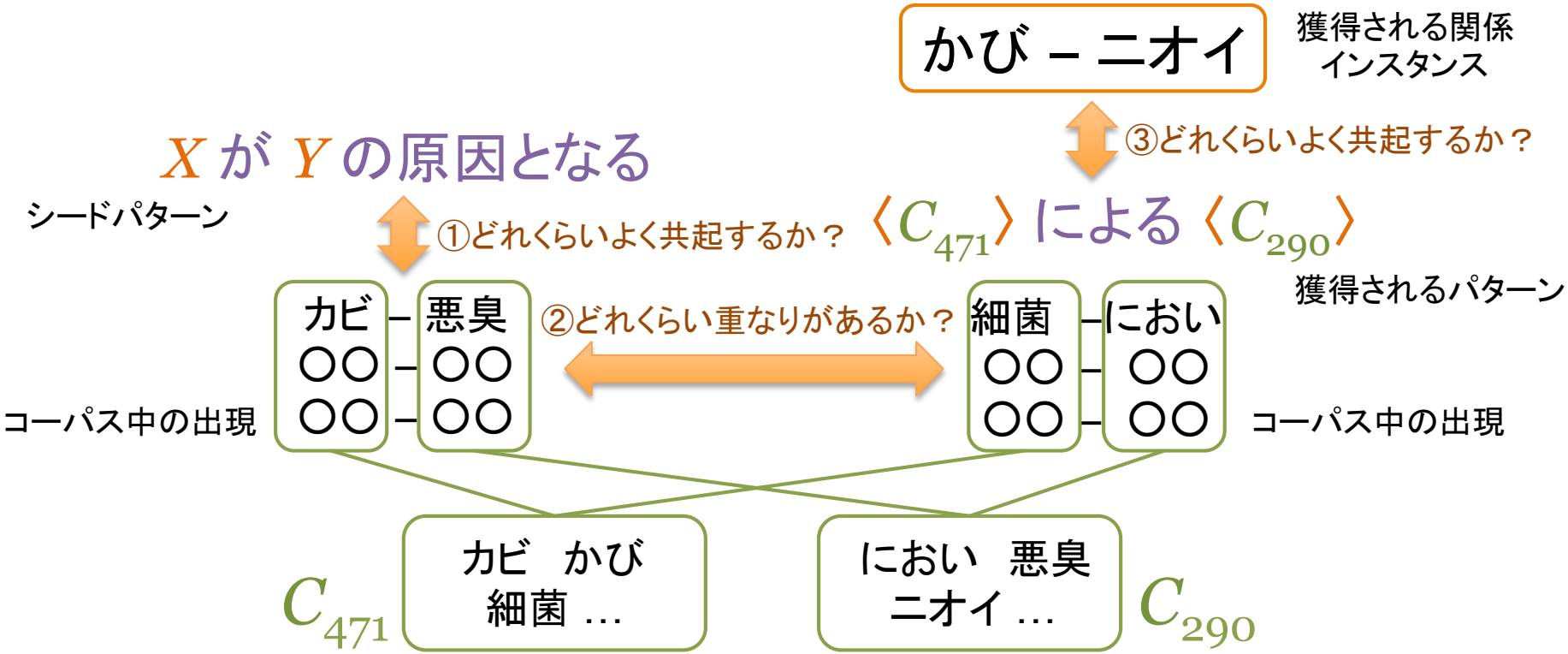
意味クラスの制約（Carlson+ 10）



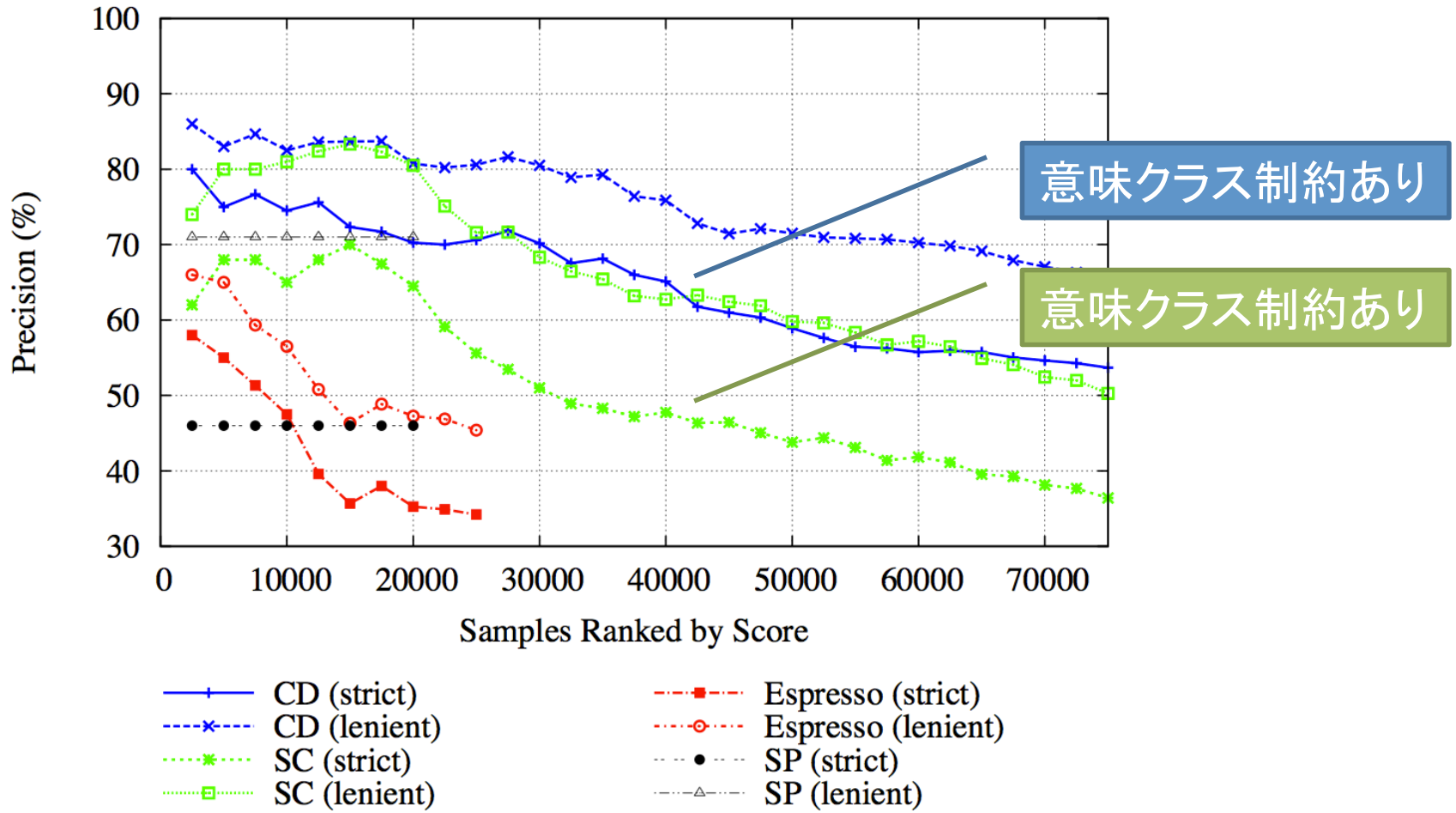
Slides from Tom Mitchell's invited talk in the Univ. of Washington CSE Distinguished Lecture Series, October, 2010.
http://rtw.ml.cmu.edu/slides/Mitchell_NELL_UW_Oct_21_2010.pdf

意味クラス制約つきパターンの学習

- De Saeger+（2009）
  - 500クラスにクラスタリングされた名詞50万語（Kazama & Torisawa 08）を利用



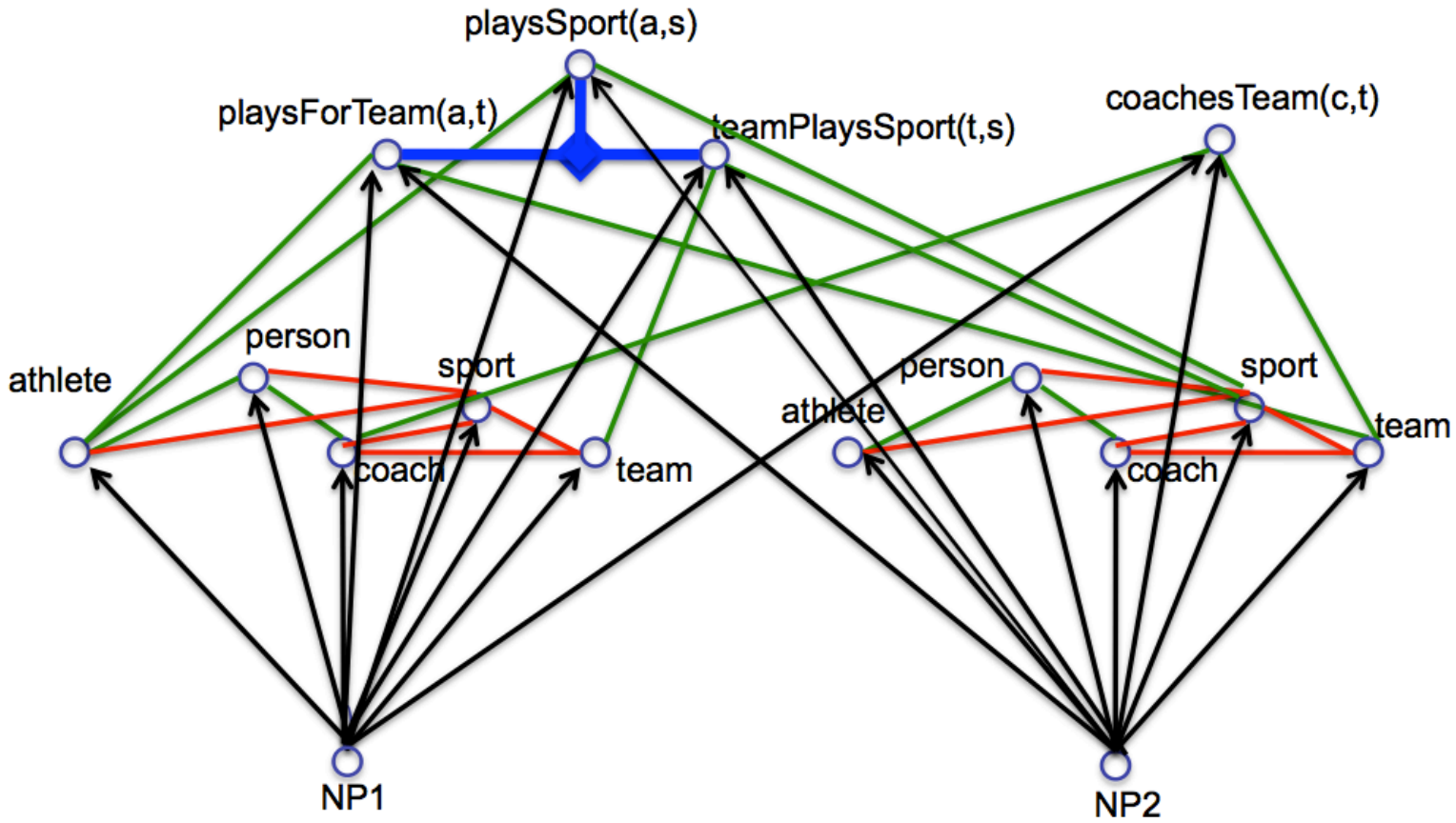
意味クラス制約つきパターンの学習



S. De Saeger, K. Torisawa, J. Kazama, K. Kuroda, and M. Murata (2009)
Large Scale Relation Acquisition Using Class Dependent Patterns, ICDM.

関係間の依存関係

0.93 playsSport(?x,?y) ← playsForTeam(?x,?z), teamPlaysSport(?z,?y)

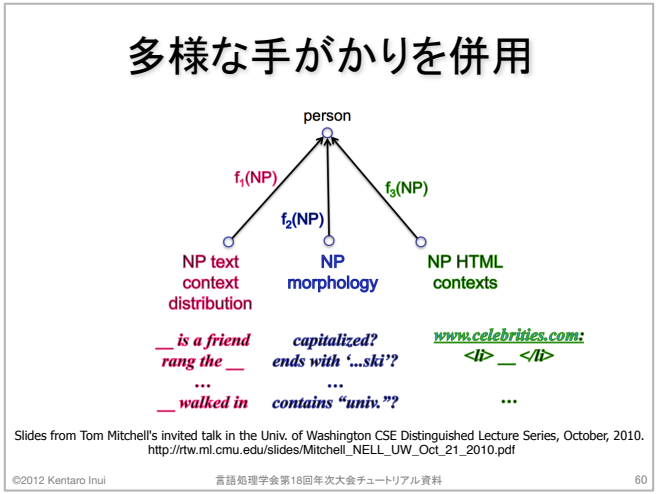


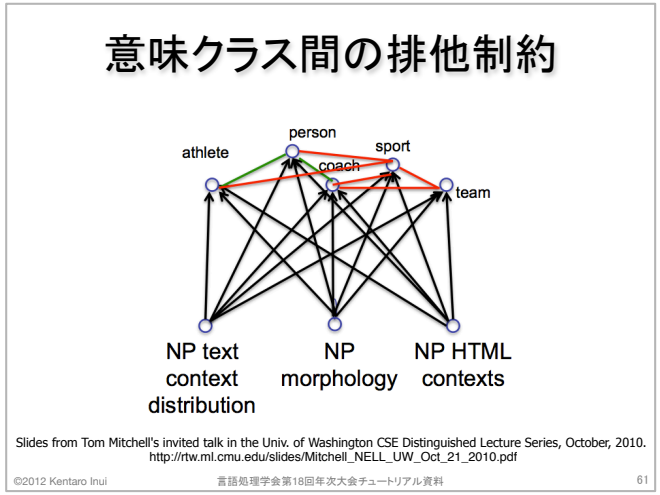
Slides from Tom Mitchell's invited talk in the Univ. of Washington CSE Distinguished Lecture Series, October, 2010.
http://rtw.ml.cmu.edu/slides/Mitchell_NELL_UW_Oct_21_2010.pdf

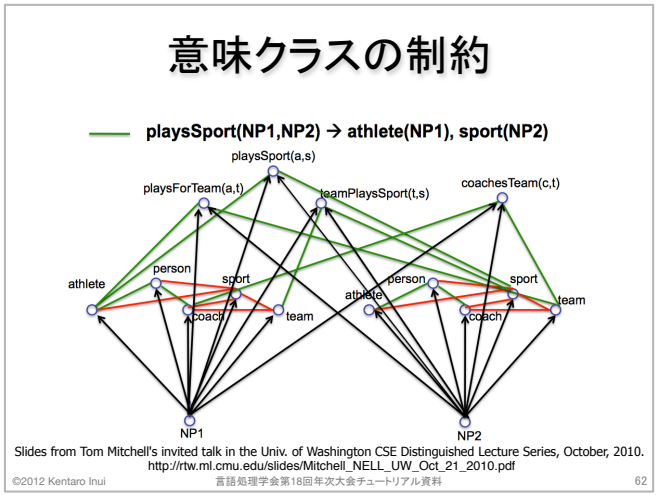
Never-Ending Language Learner

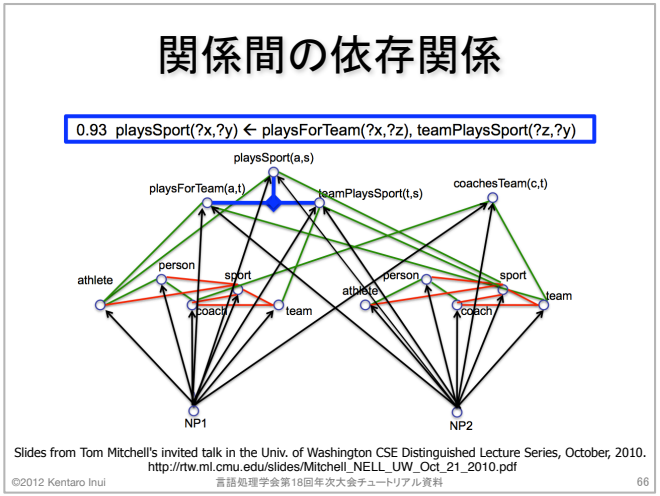
（Carlson+ 10）

多重の意味ドリフト対策







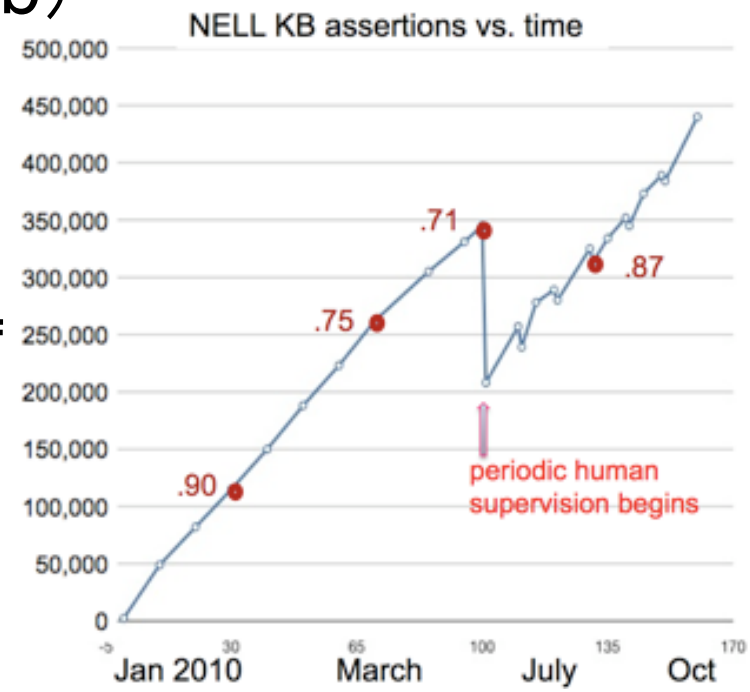


Never-Ending Language Learner

（Carlson+ 10）

- 入力
  - オントロジ
    - 実体/関係のクラス総数 約500
    - 定義はシードで与える
  - Web文書 約5億ページ（Clueweb）

    http://boston.lti.cs.cmu.edu/Data/clueweb09/
- タスク
  - クラス＆関係インスタンスの獲得
  - ブートストラップを24時間×毎日
  - 定期的に人手でチェック
- 約30万個以上の“知識”を獲得（2010年1～7月）



http://rtw.ml.cmu.edu/rtw/overview

Never-Ending Language Learner

http://rtw.ml.cmu.edu/rtw/

NELL Knowledge Base Browser
CMU Read the Web Project

Search
log in | preferences | help/instructions | feedback

categories | relations

- relatedto
  - agentcontributedtocreativework
    - actorstarredinmovie
    - writerwrotebook
    - directordirectedmovie
  - athleteplayssport
  - creativeworkcontributedtobyagent
    - bookwriter
    - moviestaractor
    - moviedirectedbydirector
  - coachwontrophy
  - companyeconomicsector
  - countrycurrency
  - currencycountry
  - dateofevent
  - economicsectorcompany
  - equipmentusedbysport
  - atdate
    - eventdate
      - sportsgamedate
    - personbirthdate
    - persondeathdate
    - locationexistsatdate
    - itemexistsatdate

languageofcountry

(relation: domain language, range country)

Typical language spoken in a country (official or otherwise)

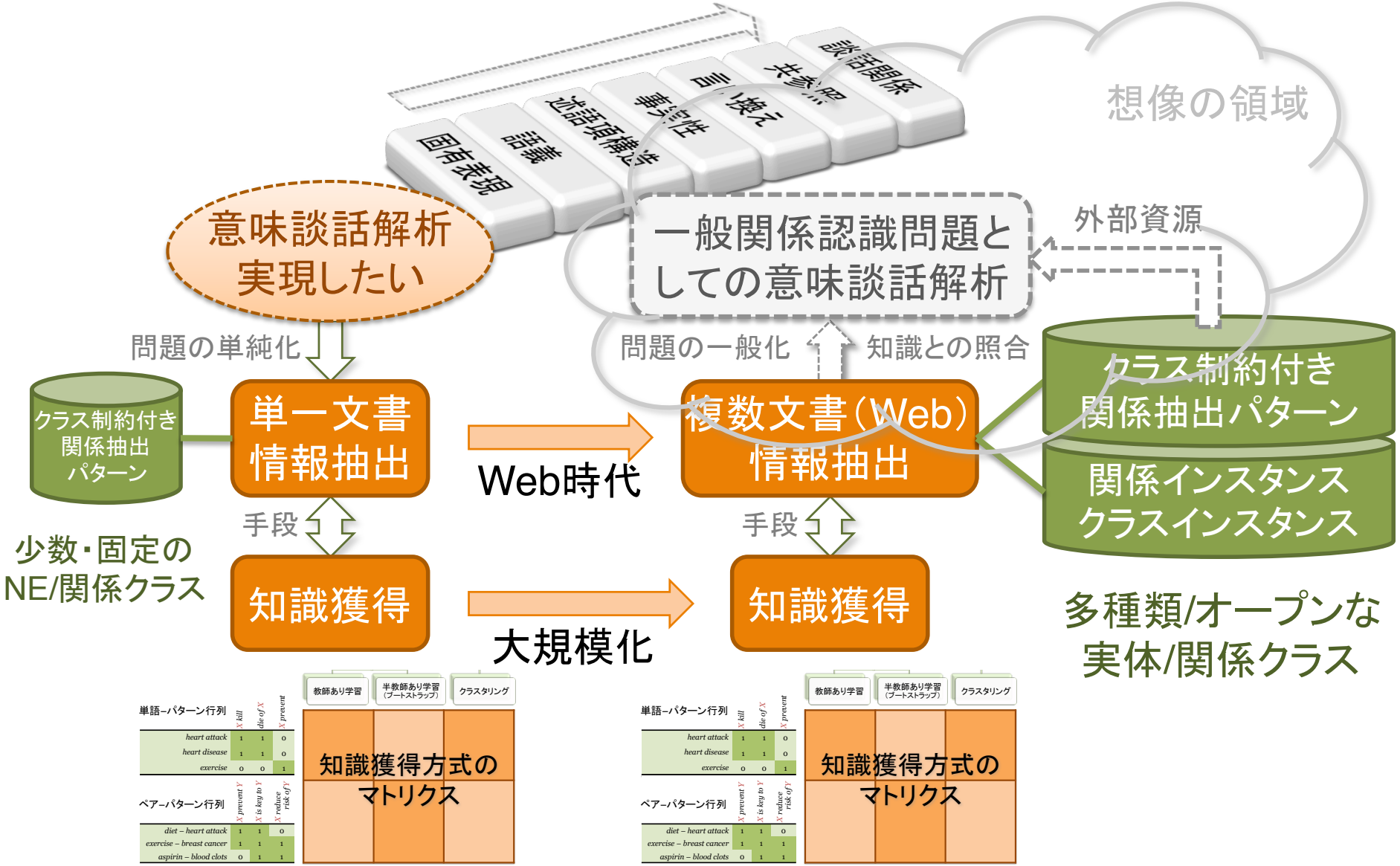
See metadata for languageofcountry
151 instances, 1 page

instance	iteration	date learned	confidence
arabic, egypt	487	15-jan-2012	(Seed) 100.0
chinese, singapore	511	15-feb-2012	(Seed) 100.0
english, britain	523	02-mar-2012	100.0
english, singapore	517	23-feb-2012	100.0
english, thailand	511	15-feb-2012	100.0
english, new_zealand	524	05-mar-2012	100.0
english, uk	524	05-mar-2012	100.0
german, switzerland	406	08-sep-2011	(Seed) 100.0
greek, greece	500	02-feb-2012	(Seed) 100.0
polish, poland	511	15-feb-2012	(Seed) 100.0
russian, russia	526	08-mar-2012	100.0
arabic, morocco	511	15-feb-2012	100.0
english, malaysia	511	15-feb-2012	100.0
arabic, algeria	454	17-nov-2011	100.0

ibm:
generalizations = {company}
candidateValues = {conference, company, product}
headquarteredIn = armonk
candidateValues = {armonk}
producesProduct = {pc}
candidateValues = {domino, thinkpad_line, ibm_e_business_logo, first_pcs, powerpc, internet, ibm_pc, iseries, rational, first_pc, quickplace, first_ibm_pc, vga_controller, original_pc, at_computer, wsfl_specification, selectric, pc, pc_convertible, workplace_client_technology, workplace, ids, opteron_server, linux_strategy, very_interesting_study, video_graphics_array, business_partner_emblem, ibm, …}
acquired = {iss, cognos, informix}
candidateValues = {spi, watchfire, telelogic, daksh, lotus, iss, internet_security_systems, gluecode, cognos, sequent, tivoli, diligent, informix, webify_solutions, geronimo, rational, information_laboratory, meiosys, webify, …}
acquiredBy = lenovo_group
candidateValues = {lenovo_group, lenovo, china, arsenal}
competesWith = {sun, texas_instruments, samsung, hewlett_packard, apple, novell, oracle, microsoft, ricoh, hp, amazon}
companyEconomicSector = {software}
hasOfficeInCountry = {united_states, canada, usa, germany, england, uk, france}
candidateValues = {san_jose, dallas, cambridge, europe, boca_raton, boulder,

Slides from Tom Mitchell's invited talk in the Univ. of Washington CSE Distinguished Lecture Series, October, 2010.
http://rtw.ml.cmu.edu/slides/Mitchell_NELL_UW_Oct_21_2010.pdf

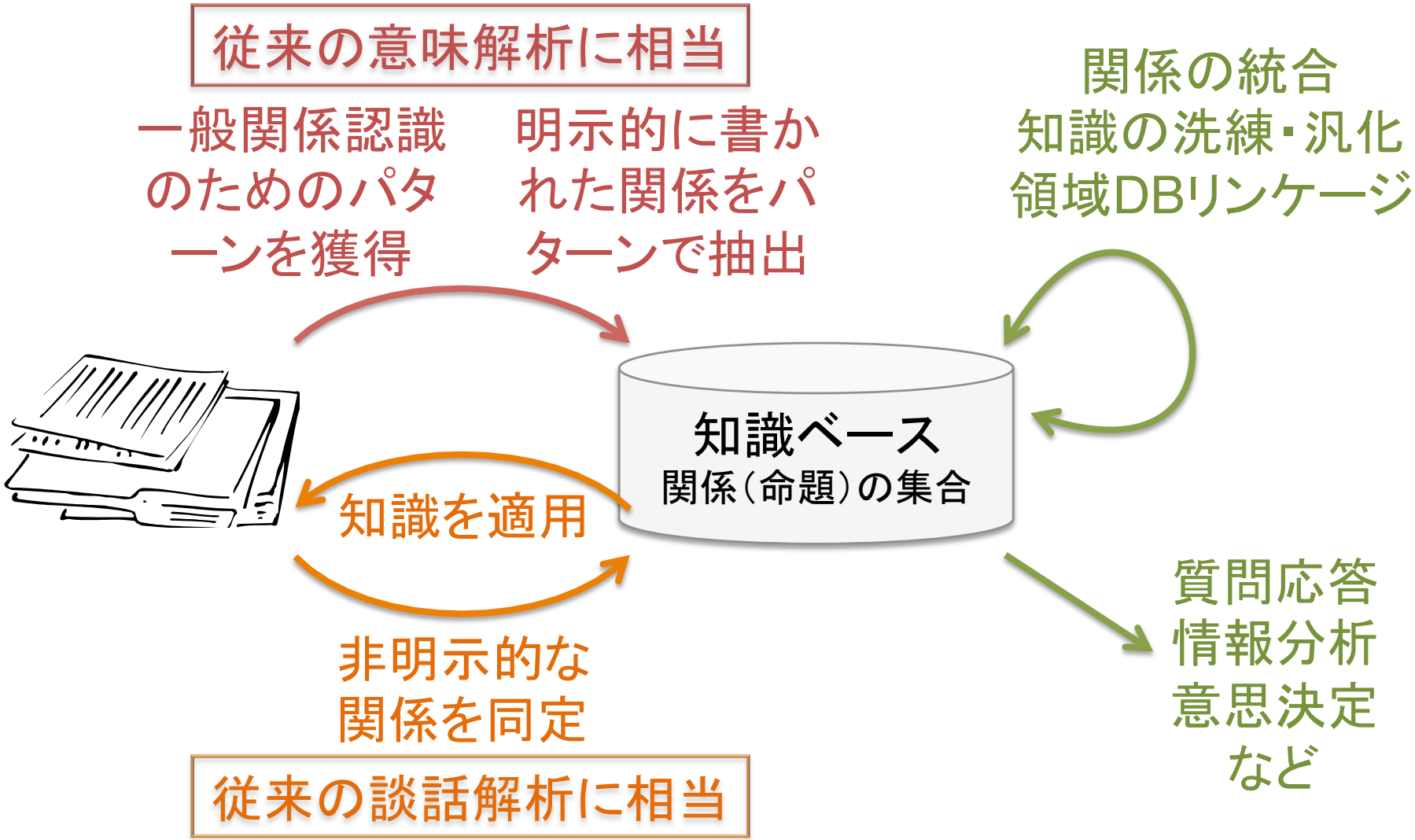
情報抽出の展開（地図②）


固有表現
語義
述語項構造
時間性
言い換え
共参照
談話関係
想像の領域
意味談話解析
実現したい
一般関係認識問題と
しての意味談話解析
外部資源
問題の単純化
問題の一般化
知識との照合
クラス制約付き
関係抽出
パターン
単一文書
情報抽出
Web時代
複数文書（Web）
情報抽出
クラス制約付き
関係抽出パターン
関係インスタンス
クラスインスタンス
手段
手段
少数・固定の
NE/関係クラス
知識獲得
大規模化
知識獲得
多種類/オープンな
実体/関係クラス
単語–パターン行列
X kill
die of X
X prevent
heart attack 1 1 0
heart disease 1 1 0
exercise 0 0 1
ペア–パターン行列
X prevent Y
X is key to Y
X reduce risk of Y
diet – heart attack 1 1 0
exercise – breast cancer 1 1 1
aspirin – blood clots 0 1 1
教師あり学習
半教師あり学習
（ブートストラップ）
クラスタリング
知識獲得方式の
マトリクス
単語–パターン行列
X kill
die of X
X prevent
heart attack 1 1 0
heart disease 1 1 0
exercise 0 0 1
ペア–パターン行列
X prevent Y
X is key to Y
X reduce risk of Y
diet – heart attack 1 1 0
exercise – breast cancer 1 1 1
aspirin – blood clots 0 1 1
教師あり学習
半教師あり学習
（ブートストラップ）
クラスタリング
知識獲得方式の
マトリクス

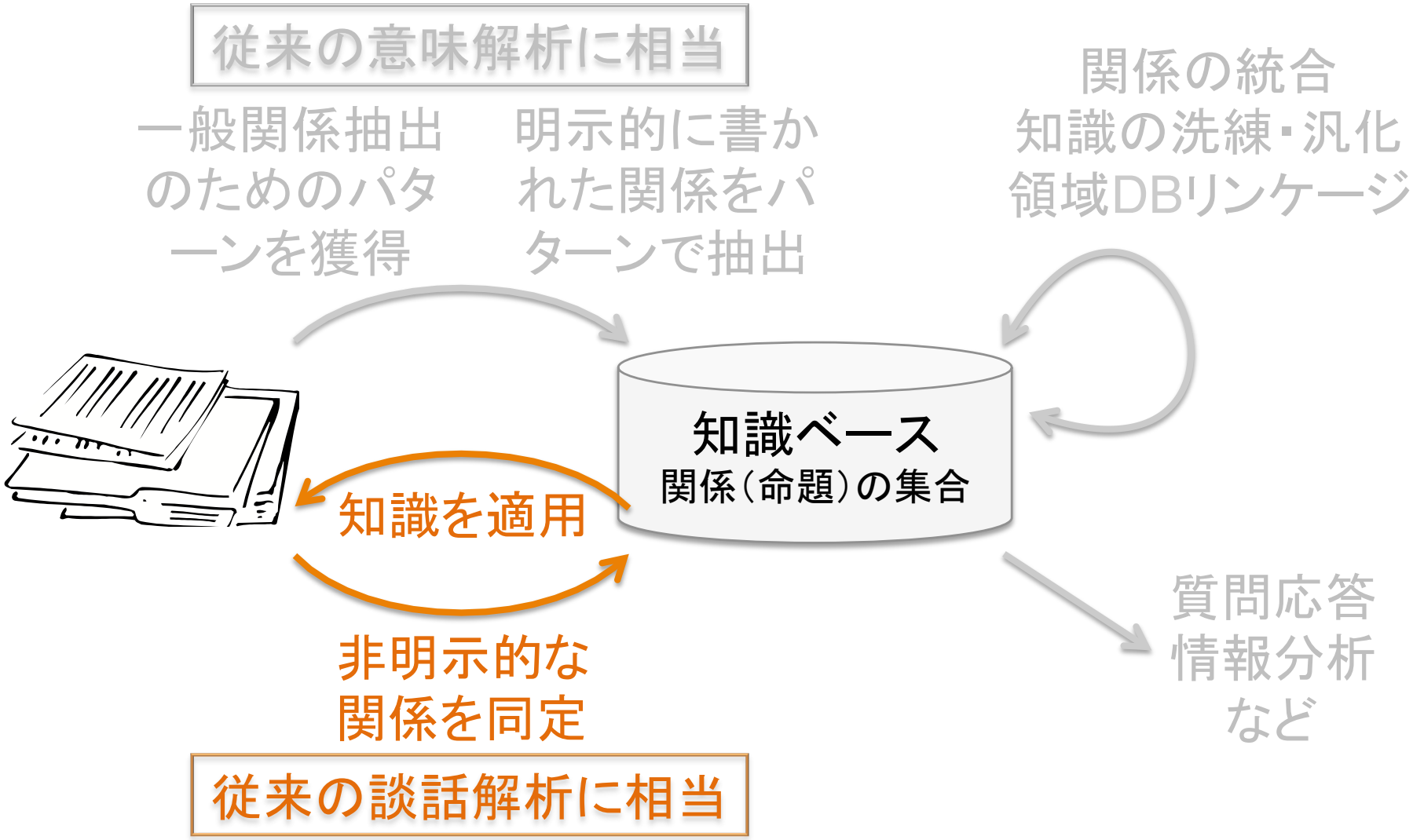
情報抽出と知識獲得：まとめ

- 情報抽出
  - 応用指向的，ドメイン限定，垂直的
  - 解析モデルと知識獲得が一体
- Web時代
  - 文書間の冗長性（macro reading）
  - オープンドメイン，大規模化，超多クラス化
  - 意味解析の目的に別の角度から接近
- 課題
  - 抽出対象は明示的な関係のみ
  - モダリティ（事実性），時間，量化等との統合

一般関係認識としての意味談話解析（地図④）



一般関係認識としての意味談話解析（地図④）



本講演の構成

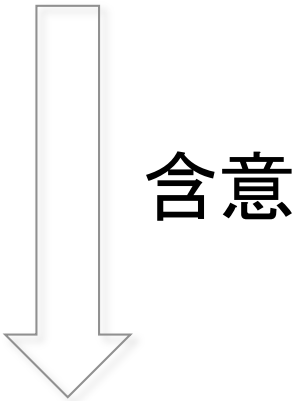
1. Annotation layer cake
2. 情報抽出と知識獲得の今
3. Making the implicit explicit
4. やわらかい記号の世界
5. ルールを作るところから

©2012 Kentaro Inui

非明示的な関係の認識

含意関係認識で

[t] *Harriet Lane served as Buchanan's White House hostess.*

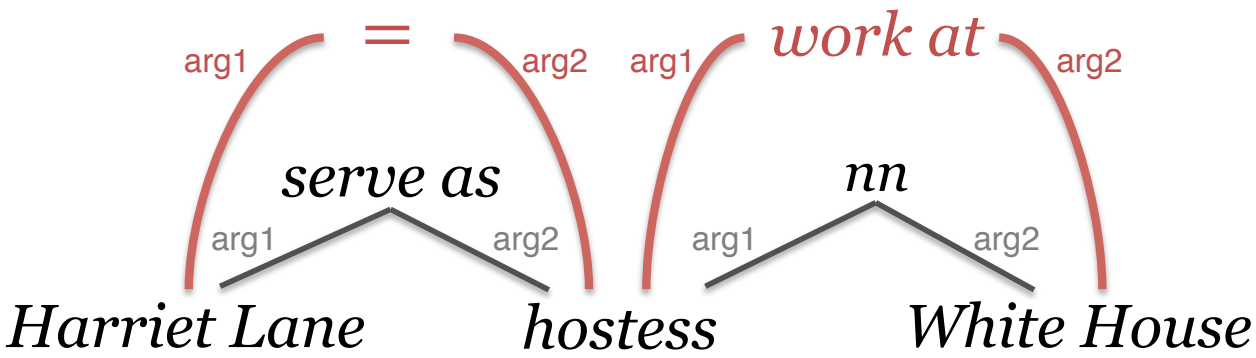


[h] *Harriet Lane worked at the White House.*

非明示的な関係の認識

含意関係認識で

[t] *Harriet Lane served as Buchanan's White House hostess.*



[h] *Harriet Lane worked at the White House.*

非明示的な関係の認識

- Hickl & Bensley（07）
  - Discourse commitment を網羅的に抽出できれば，含意関係認識のかなりの問題が解ける

Text: A Revenue Cutter, the ship was named for Harriet Lane, niece of President James Buchanan, who served as Buchanan's White House hostess.

T1. A Revenue Cutter is a ship.
T2. The ship was named for Harriet Lane.
T3. Harriet Lane was the niece of President James Buchanan.
T4. The niece of Buchanan served as Buchanan's White House hostess.
T5. A Revenue Cutter was named for Harriet Lane.
T6. A Revenue Cutter was named for the niece of President James Buchanan.
T7. A Revenue Cutter was named for Buchanan's White House hostess.
T8. A Revenue Cutter was named for a White House hostess.
T9. A Revenue Cutter was named for a hostess.
T10. The niece of a President served as Buchanan's White House hostess.
T11. The niece of a President served as Buchanan's hostess.
T12. The niece of a President served as a White House hostess.
T13. The niece of a President served at the White House.
T14. The niece of a President had occupation hostess.
T15. The niece of a President served as a hostess.

T16. Harriet Lane was related to President James Buchanan.
T17. Harriet Lane was the niece of a President.
T18. Harriet Lane was related to a President.
T19. Harriet Lane was related to James Buchanan.
T20. James Buchanan had title of President.
T21. James Buchanan had a White House hostess.
T22. James Buchanan had a hostess.
T23. James Buchanan was associated with the White House.
T24. James Buchanan had a niece.
T25. Harriet Lane served as Buchanan's White House hostess.
T26. Harriet Lane served as Buchanan's hostess.
T27. Harriet Lane served as a White House hostess.
T28. Harriet Lane served at the White House.
T29. Harriet Lane had occupation hostess.
T30. Harriet Lane served as a hostess..

Selected Commitment

Hyp(34): Harriet Lane owned a Revenue Cutter.

Negative Instance of Textual Entailment

Hyp(36): Harriet Lane worked at the White House.

Positive Instance of Textual Entailment

Selected Commitment

A. Hickl and J. Bensley（2007）A Discourse Commitment-Based Framework for Recognizing Textual Entailment, ACL-07 Workshop on Textual Entailment and Paraphrasing.

非明示的な関係の認識

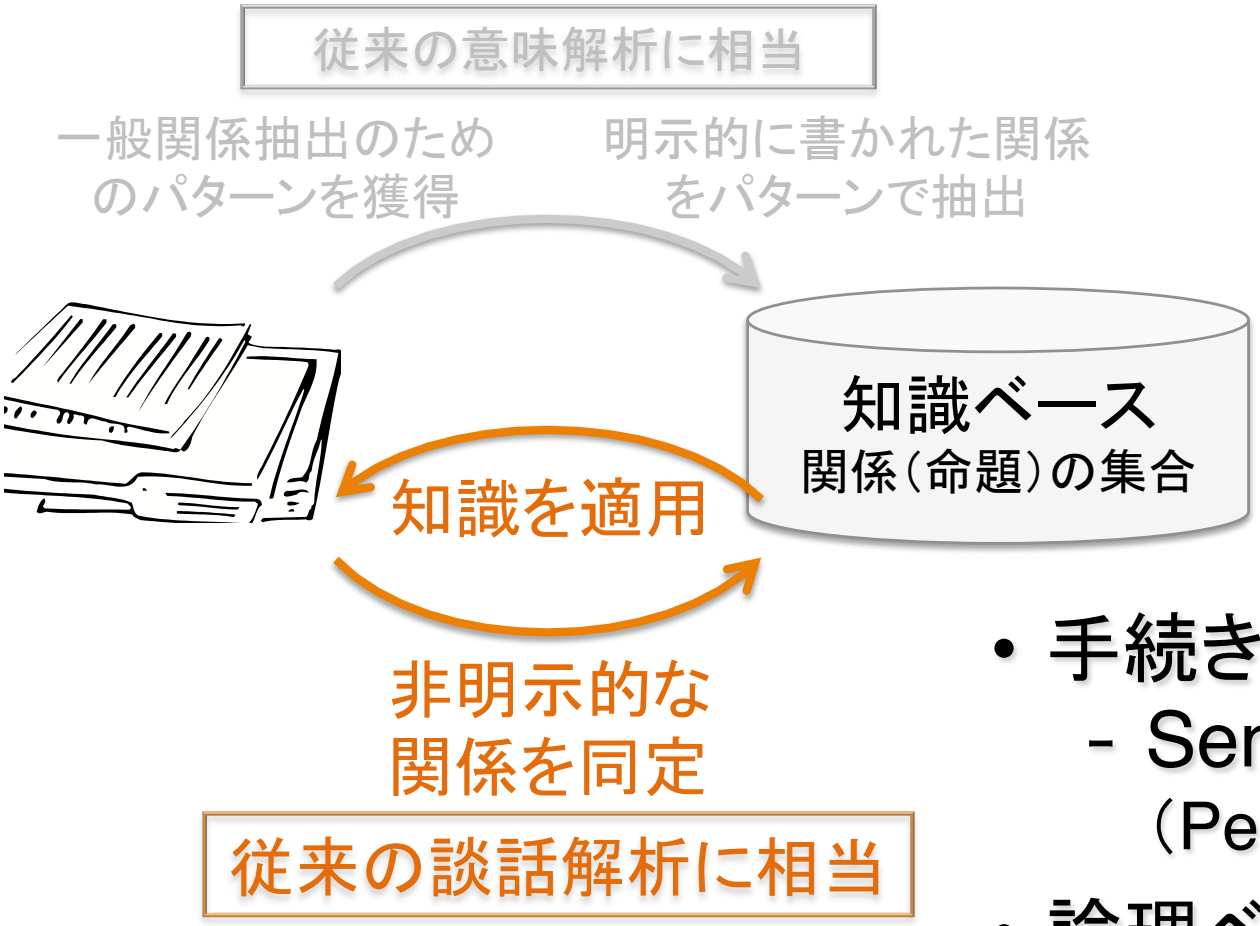
（Peñas and Hovy 2010）

Example: *San Francisco's Eric Davis intercepted a Steve Walsh pass on the next series to set up a seven-yard Young touchdown pass to Brent Jones.*

Implicit	(More) explicit
San Francisco's Eric Davis	Eric Davis plays for San Francisco
Eric Davis intercepted pass	—
Steve Walsh pass	Steve Walsh threw pass Steve Walsh threw interception
Young touchdown pass	Young completed pass for touchdown
touchdown pass to Brent Jones	Brent Jones caught pass for touchdown

E. Hovy. A New Semantics: Merging Propositional and Distributional Approaches. IWCS, 2010.
http://www.meaningfactory.com/iwcs2011/hovy.pdf

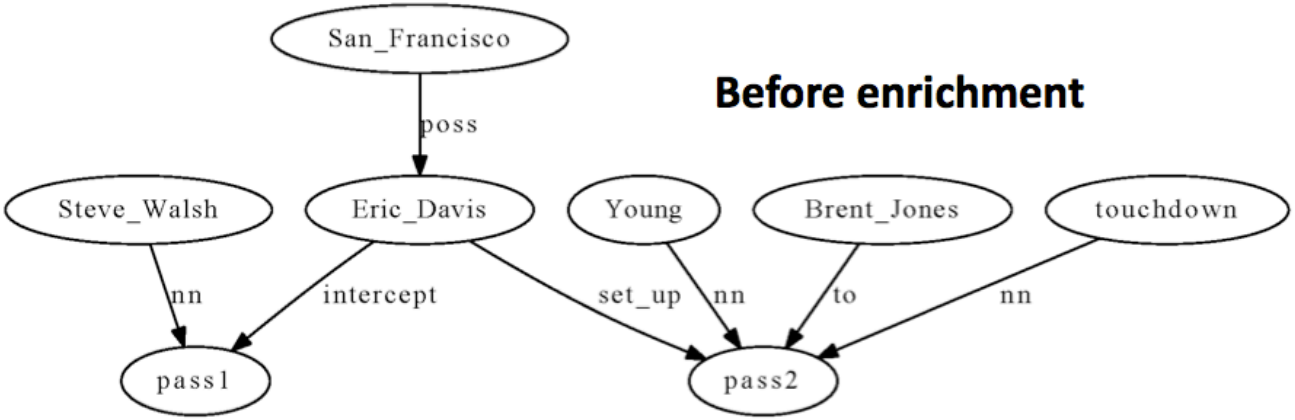
一般関係認識としての意味談話解析（地図④）

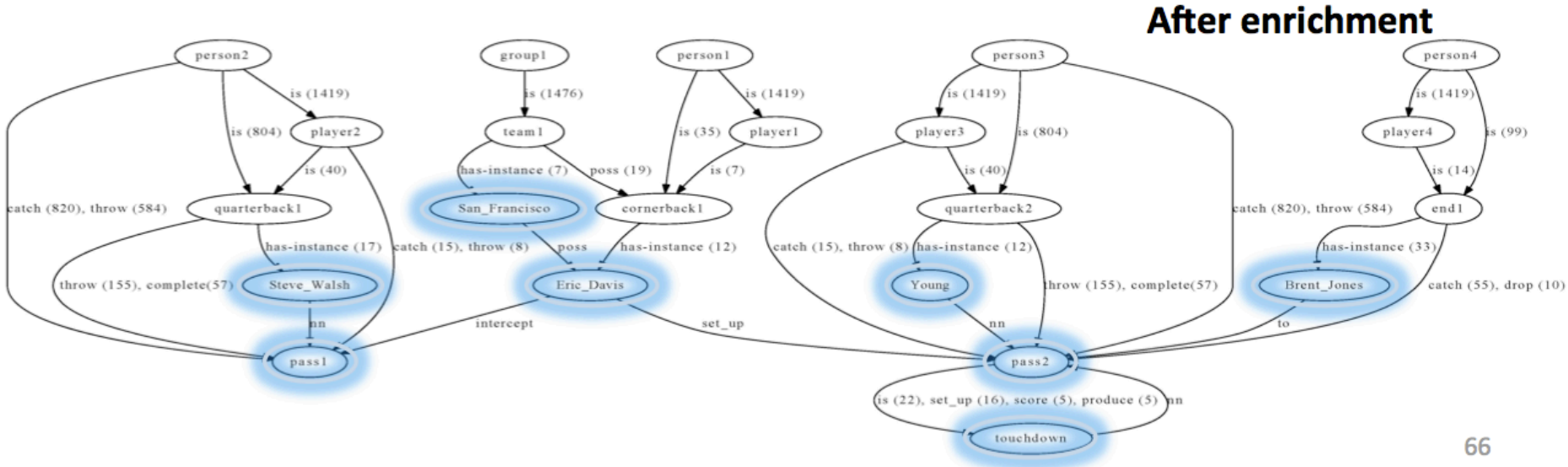


- 手続き的アプローチ
  - Semantic enrichment（Peñas & Hovy 2010）
- 論理ベースのアプローチ
  - アブダクション

Semantic Enrichment（Peñas & Hovy 10）

San Francisco's Eric Davis intercepted a Steve Walsh pass on the next series to set up a seven-yard Young touchdown pass to Brent Jones.

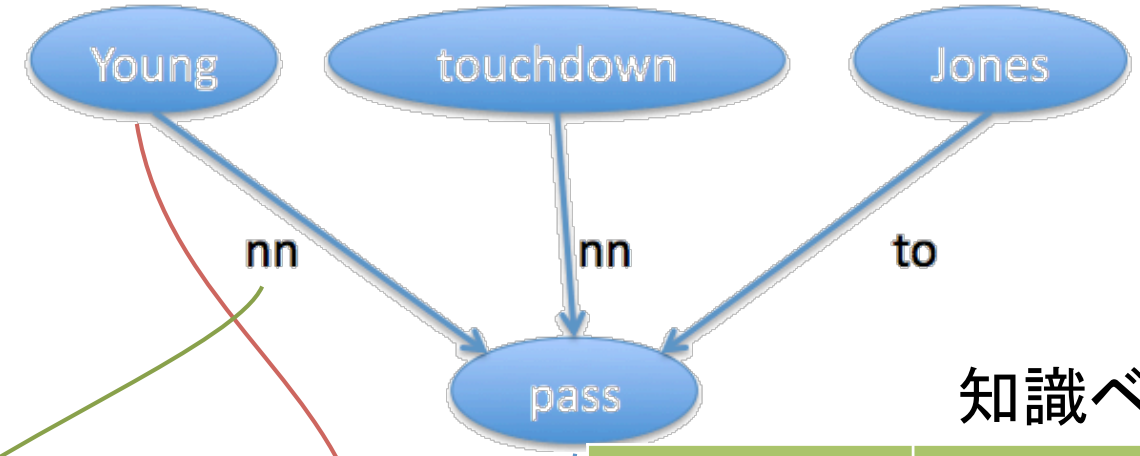




E. Hovy. A New Semantics: Merging Propositional and Distributional Approaches. IWCS, 2010.
http://www.meaningfactory.com/iwcs2011/hovy.pdf

Semantic Enrichment（Peñas & Hovy 10）

...to set up a 7-yard Young touchdown pass to Brent Jones



Young pass

?> X:has-instance:Young

X=quarterback

?> NVN:quarterback:X:pass

X=throw

X=complete

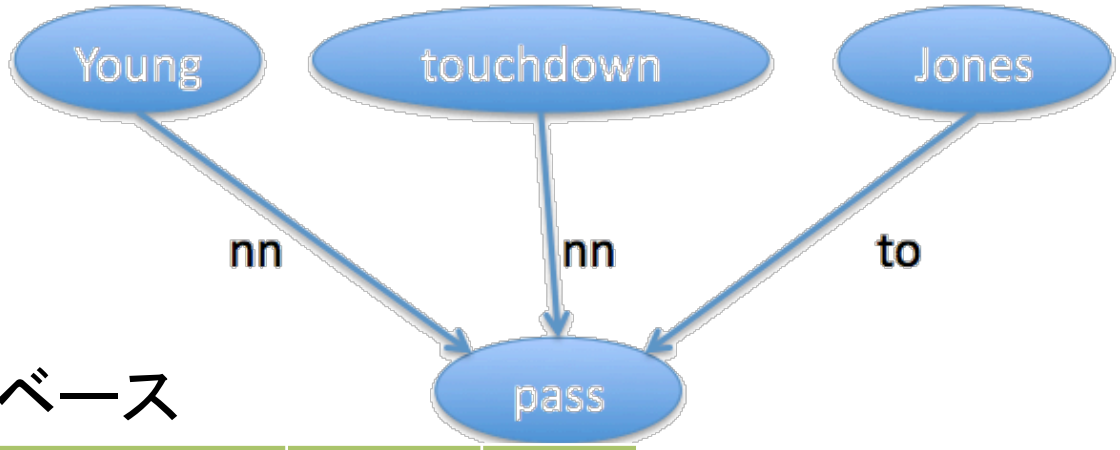
知識ベース

arg 1	relation	arg 2	freq
quarterback	has-instance	Young	12
end	has-instance	Jones	33
quarterback	throw	pass	98
quarterback	complete	pass	27
end	catch	pass	28
end	drop	pass	6

E. Hovy. A New Semantics: Merging Propositional and Distributional Approaches. IWCS, 2010.
http://www.meaningfactory.com/iwcs2011/hovy.pdf

Semantic Enrichment（Peñas & Hovy 10）

...to set up a 7-yard Young touchdown pass to Brent Jones



知識ベース

arg 1	relation	arg 2	freq
quarterback	has-instance	Young	12
end	has-instance	Jones	33
quarterback	throw	pass	98
quarterback	complete	pass	27
end	catch	pass	28
end	drop	pass	6

Pass to Jones

?> X:has-instance:Jones

X=end

?> NVN:end:X:pass

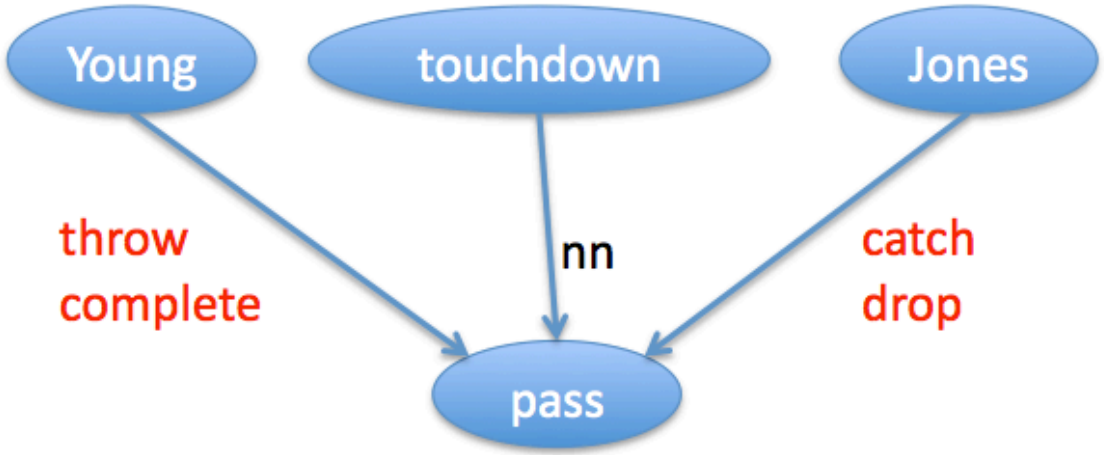
X=catch

X=drop

E. Hovy. A New Semantics: Merging Propositional and Distributional Approaches. IWCS, 2010.
http://www.meaningfactory.com/iwcs2011/hovy.pdf

Semantic Enrichment（Peñas & Hovy 10）

...to set up a 7-yard Young touchdown pass to Brent Jones



touchdown pass

?> NVN touchdown:X:pass

False

?> NPN pass:X:touchdown

X=for

E. Hovy. A New Semantics: Merging Propositional and Distributional Approaches. IWCS, 2010.
http://www.meaningfactory.com/iwcs2011/hovy.pdf

Semantic Enrichment（Peñas & Hovy 10）

...to set up a 7-yard Young touchdown pass to Brent Jones

?> NVNPN NAME:X:pass:for:touchdown

X=complete

X=catch

E. Hovy. A New Semantics: Merging Propositional and Distributional Approaches. IWCS, 2010.
http://www.meaningfactory.com/iwcs2011/hovy.pdf

Semantic Enrichment（Peñas & Hovy 10）

...to set up a 7-yard Young touchdown pass to Brent Jones

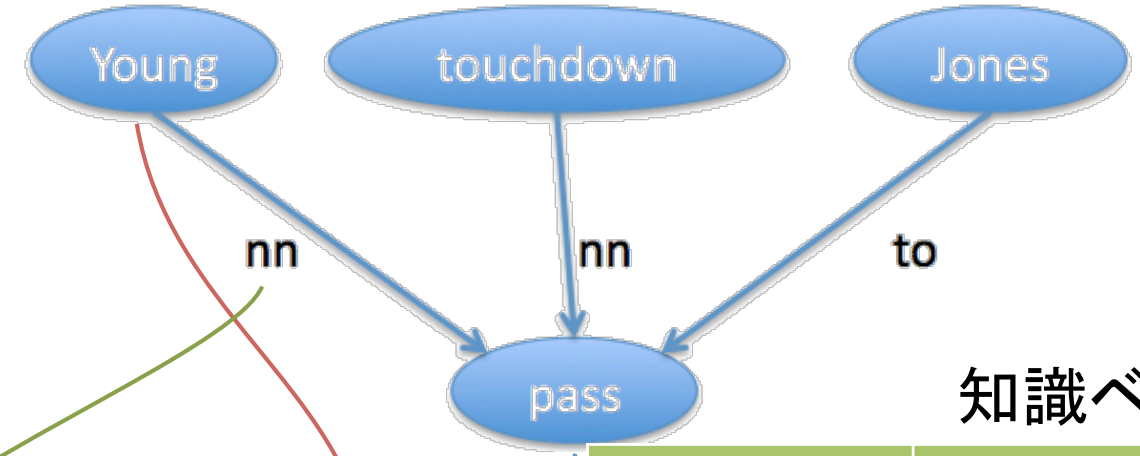
⇒ Young complete pass for touchdown

⇒ Jones catch pass for touchdown

E. Hovy. A New Semantics: Merging Propositional and Distributional Approaches. IWCS, 2010.
http://www.meaningfactory.com/iwcs2011/hovy.pdf

Semantic Enrichment（Peñas & Hovy 10）

...to set up a 7-yard Young touchdown pass to Brent Jones



Young pass

?> X:has-instance:Young

X=quarterback

?> NVN:quarterback:X:pass

X=throw

X=complete

知識ベース

arg 1	relation	arg 2	freq
quarterback	has-instance	Young	12
end	has-instance	Jones	33
quarterback	throw	pass	98
quarterback	complete	pass	27
end	catch	pass	28
end	drop	pass	6

E. Hovy. A New Semantics: Merging Propositional and Distributional Approaches. IWCS, 2010.
http://www.meaningfactory.com/iwcs2011/hovy.pdf

論理ベースのアプローチ

知識ベース

quarterback(Young)

quarterback(x) ⇒ human(x)

throw(x1, x2) ⇒ nn(x1, x2)

quarterback(x1) ∧ pass(x2) ⇒ throw(x1, x2)

説明（仮説）

quarterback(Young)

quarterback(x1) ∧ pass(x2) ⇒ throw(x1, x2)

quarterback(x) ⇒ human(x)

throw(x1, x2) ⇒ nn(x1, x2)

入力（観察）

human(Young)　nn(Young, x2)　pass(x2)

論理ベースのアプローチ

- テキストに埋め込まれた非明示的な関係
  ＝明示的な情報（観察）に対する説明（仮説）
- アブダクション（仮説推論）
  - 背景知識 B と観察 O とが与えられたとき，観察を説明できる尤もらしい仮説 H を見つける推論

$H \cup B \models O$ （背景知識 B のもとで観察 O を導く仮説を選べ）

$H \cup B \not\models \bot$ （背景知識 B に矛盾しない仮説を選べ）

尤もらしい仮説 （なるべく小さい（事後確率の高い）仮説を選べ）

- Interpretation as abduction（Hobbs+ 1993）
  - *“Interpreting sentences is to find the best explanation to the sentences.”*

Interpretation as Abduction

（Hobbs+ 1993, etc.）

背景知識

$$\mathrm{loan}(y2) \Rightarrow (\exists e2, y1, y3)\ \mathrm{issue}(e2, y3, y2, y1) \wedge \mathrm{financial_inst}(y3)$$

$$\mathrm{issue}(e2, x2, x3, x1) \Rightarrow \mathrm{go}(e2, x1, x2)$$

$$\mathrm{financial_inst}(x2) \Rightarrow \mathrm{bank}(x2)$$

$$\mathrm{issue}(e2, y3, y2, y1) \Rightarrow \mathrm{get}(e2, y1, y2)$$

観察

$$(\exists e1,e2,x1,x2,y1,y2)\ \mathrm{John}(x1) \wedge \mathrm{go}(e1, x1, x2) \wedge \mathrm{bank}(x2) \wedge \mathrm{he}(y1) \wedge \mathrm{get}(e2, y1, y2) \wedge \mathrm{loan}(y2)$$

入力 *John went to the bank. He got a loan.*

Interpretation as Abduction

（Hobbs+ 1993, etc.）

John and *he* are coreferent

went to the bank is the purpose of *got a loan*

bank refers to *a financial bank*

loan(y2) ⇒ (∃e2, y1, y3) issue(e2, y3, y2, y1) ∧ financial_inst(y3)

hypothesized

説明

x1=y1

x2=y3

issue(e2, x2, x3, x1) ⇒ go(e2, x1, x2)

financial_inst(x2) ⇒ bank(x2)

issue(e2, y3, y2, y1) ⇒ get(e2, y1, y2)

観察

(∃e1,e2,x1,x2,y1,y2)

John(x1) ∧ go(e1, x1, x2) ∧ bank(x2) ∧ he(y1) ∧ get(e2, y1, y2) ∧ loan(y2)

hypothesized

入力 *John went to the bank. He got a loan.*

Interpretation as Abduction

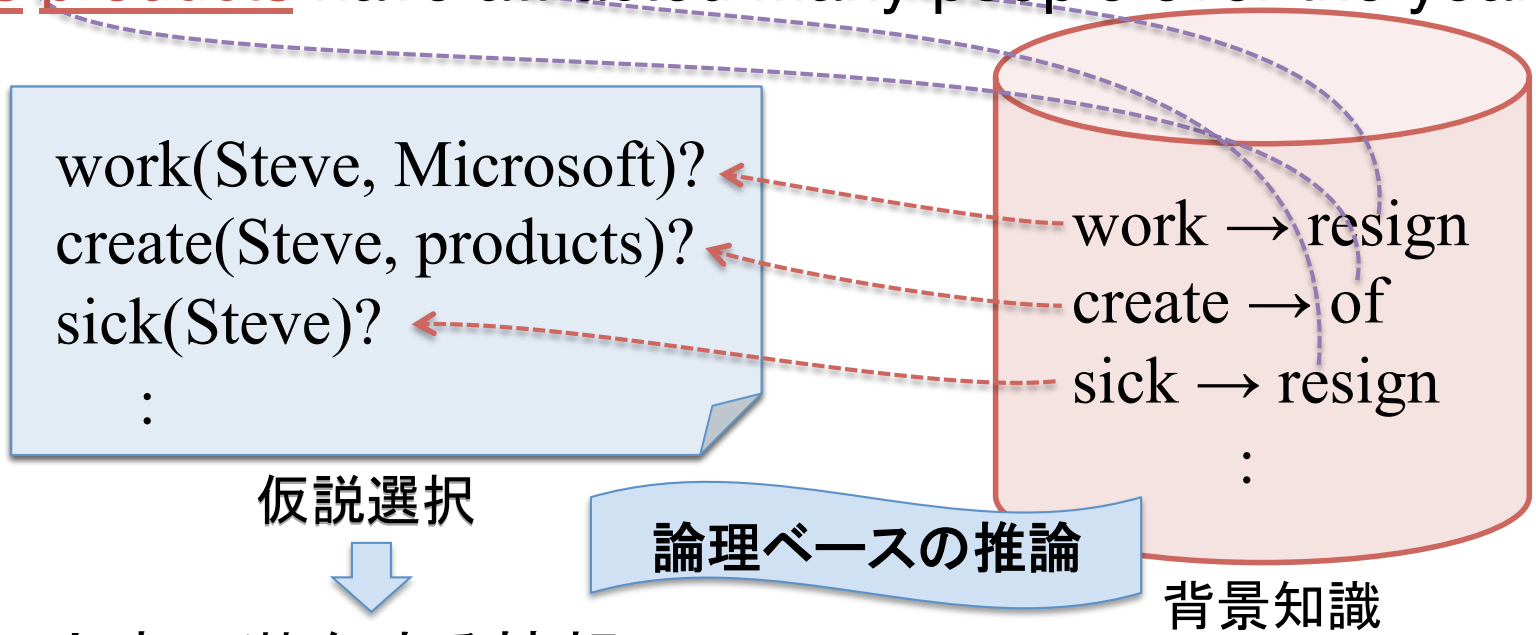
- 知識の使い方に原理的な指針
- スケーラブルな推論エンジン
  - Markov Logicによる実現（Blythe+ 2011）
  - 整数線形計画法による高速化（近似解，並列化も可）（井之上&乾 2012）
- 仮説の評価関数も（おそらく）学習可能

スケーラブルな推論エンジン

（井之上&乾 12）

入力 (文章):

Steve will resign from Microsoft next week.
His products have attracted many people over the years.



出力 (文章に潜在する情報):

- Steve works for Microsoft
- His refers to Steve
- Steve creates products

スケーラブルな推論エンジン

（井之上&乾 12）

入力 (文章):

Steve will resig eek.
His products have attracted many people over the years.

20リテラル

work(Steve, Microsoft)?
:

1,000 個のリテラルの組
み合わせ最適化問題

work → resign
sick → resign
:

37万の推論規則

仮説選択

論理ベースの推論

背景知識

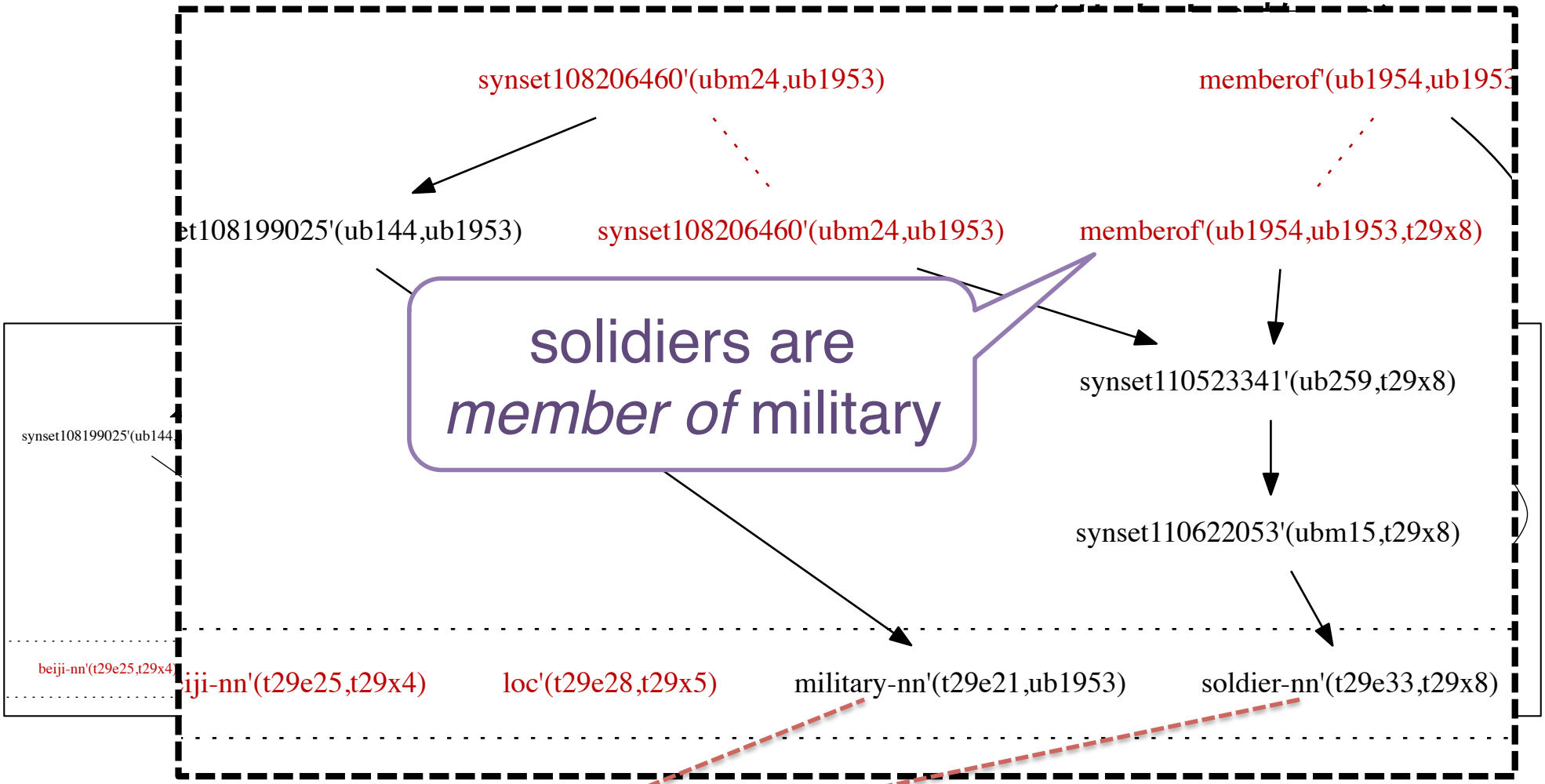
出力

提案手法の推論時間: 約2.6秒

- Markov Logic Networks (Richardson & Domingos 05) による実装 (Blythe+ 11): 7分
- Mini-TACITUS (Mulkar-Mehta+ 07): 30分以上

©2012 Kentaro Inui

スケーラブルな推論エンジン



観測
(入力)

Also Friday, five Iraqi soldiers were killed and nine wounded in a bombing, targeting their convoy near Beiji, 150 miles north of Baghdad, the Iraqi military said.

Making the Implicit Explicit: まとめ

- 問題: 非明示的な情報
  - パターンの漏れ
  - 関係の組み合わせ
  - 文をまたぐ関係
- 対策
  - 既知の知識で「行間」を埋める推論
- 現状はまだまだ
  - Machine Reading (DARPA)
  - 文章全体で整合する解釈の探索
  - アブダクション

http://www.darpa.mil/Our_Work/I2O/Programs/Machine_Reading.aspx

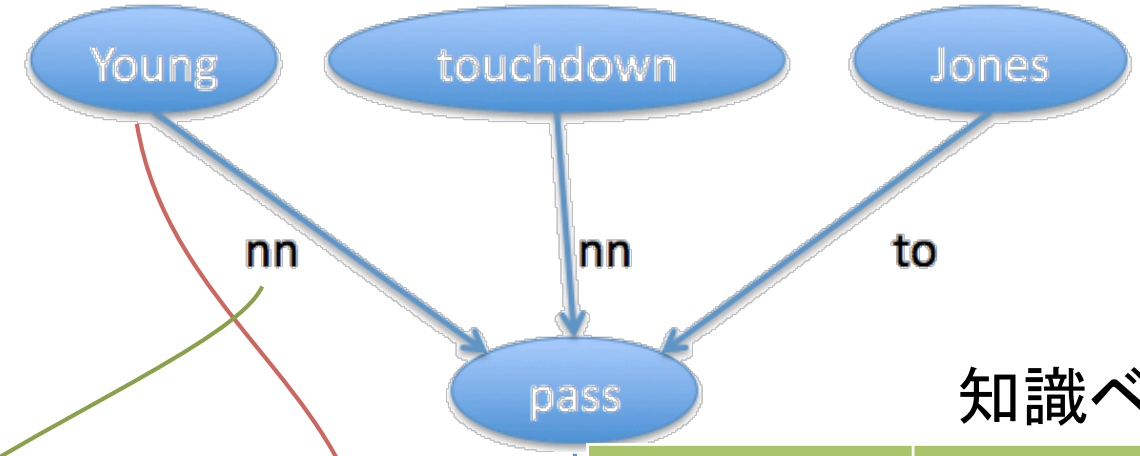
本講演の構成

1. Annotation layer cake
2. 情報抽出と知識獲得の今
3. Making the implicit explicit
4. やわらかい記号の世界
5. ルールを作るところから

©2012 Kentaro Inui

実体/関係のクラスをどう決めるか

...to set up a 7-yard Young touchdown pass to Brent Jones



Young pass

?> X:has-instance:Young

X=quarterback

?> NVN:quarterback:X:pass

X=throw

X=complete

知識ベース

arg 1	relation	arg 2	freq
quarterback	has-instance	Young	12
end	has-instance	Jones	33
quarterback	throw	pass	98
quarterback	complete	pass	27
end	catch	pass	28
end	drop	pass	6

E. Hovy. A New Semantics: Merging Propositional and Distributional Approaches. IWCS, 2010.
http://www.meaningfactory.com/iwcs2011/hovy.pdf

実体/関係のクラスをどう決めるか

① 固定のセット

② クラスを自動選択

③ クラスタリング

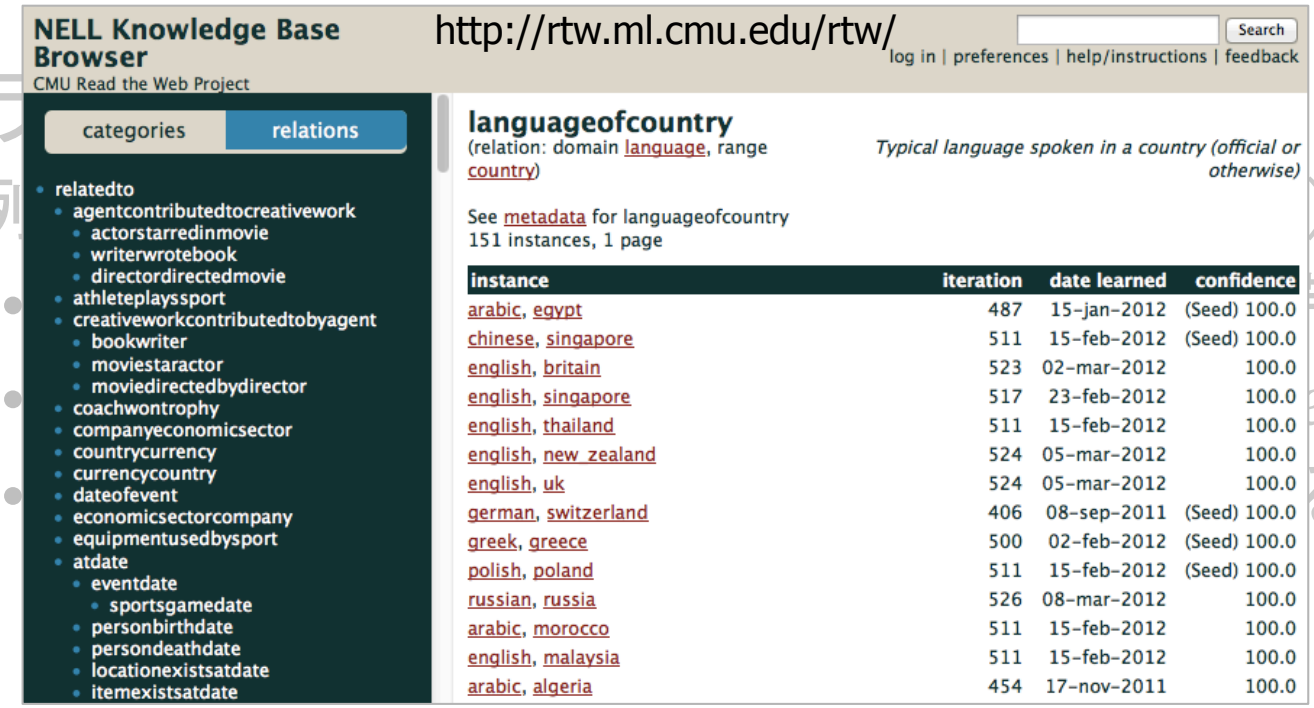
④ クラスタも作らない

実体/関係のクラスをどう決めるか

① 固定のセット

– 例：NELL（Carlson+ 10）

- 実体/関係のクラス総数 約500
- 人間が設計，シードで指定（ドメインごとに？）



知識獲得（地図③）

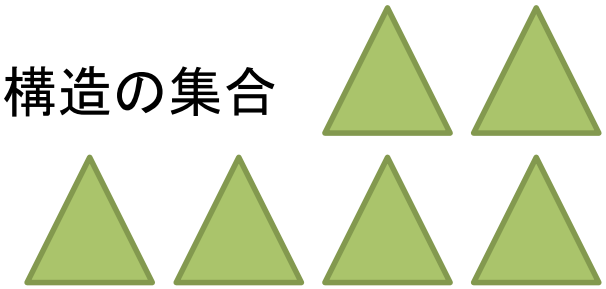
単語–パターン行列

	X kill	*die of X*	*X prevent*
heart attack	1	1	0
heart disease	1	1	0
exercise	0	0	1

ペア–パターン行列

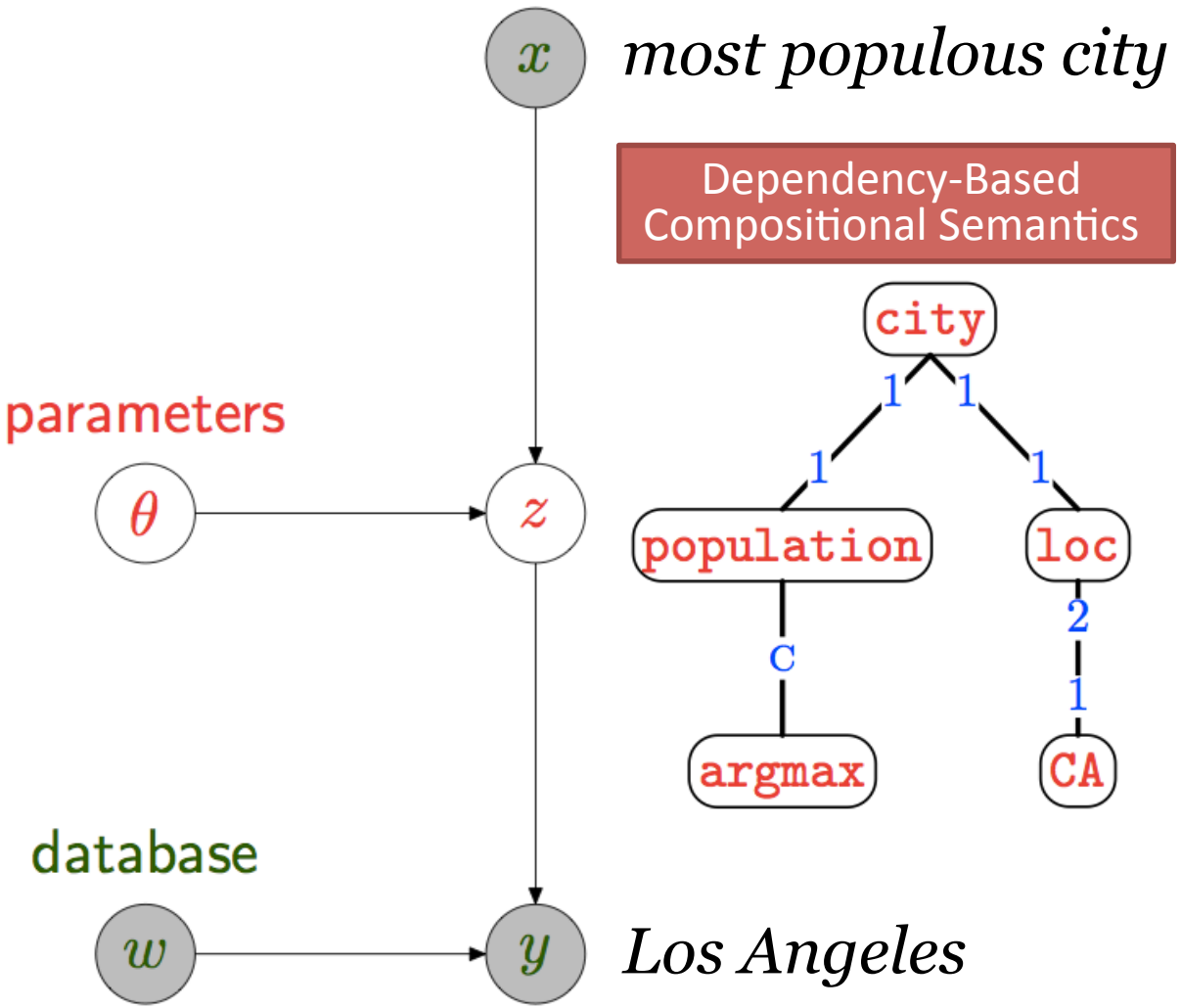
	X prevent Y	*X is key to Y*	*X reduce risk of Y*
diet – heart attack	1	1	0
exercise – breast cancer	1	1	1
aspirin – blood clots	0	1	1

依存構造の集合



量化を含む意味表現の学習

（Liang+ 11）



Semantic Parsing

$$p(z \mid x, \theta)$$

（確率的）

Interpretation

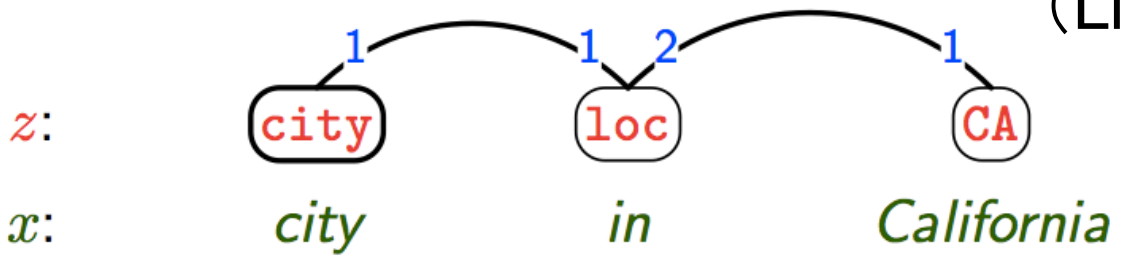
$$p(y \mid z, w)$$

（決定的）

P. Liang, et al. (2011), Learning dependency-based compositional semantics, ACL.
http://cs.stanford.edu/~pliang/papers/dcs-acl2011-talk.pdf

量化を含む意味表現の学習

（Liang+ 11）



$$\text{features}(x, z) = \begin{pmatrix} \textit{in} \cdots\cdots \texttt{loc} \\ \texttt{city} \text{-}1\text{—}1\text{-} \texttt{loc} \\ \cdots \end{pmatrix}$$

$$\text{score}(x, z) = \text{features}(x, z) \cdot \theta$$

$$p(z \mid x, \theta) \propto e^{\text{score}(x,z)}$$

P. Liang, et al. (2011), Learning dependency-based compositional semantics, ACL.
http://cs.stanford.edu/~pliang/papers/dcs-acl2011-talk.pdf

量化を含む意味表現の学習

（Liang+ 11）

Objective:

$$\max_{\theta} \sum_{z} p(y \mid z, w)\, p(z \mid x, \theta)$$

Interpretation **Semantic parsing**

EM-like Algorithm:

parameters θ

$(0.2, -1.3, \ldots, 0.7)$

enumerate/score DCS trees →

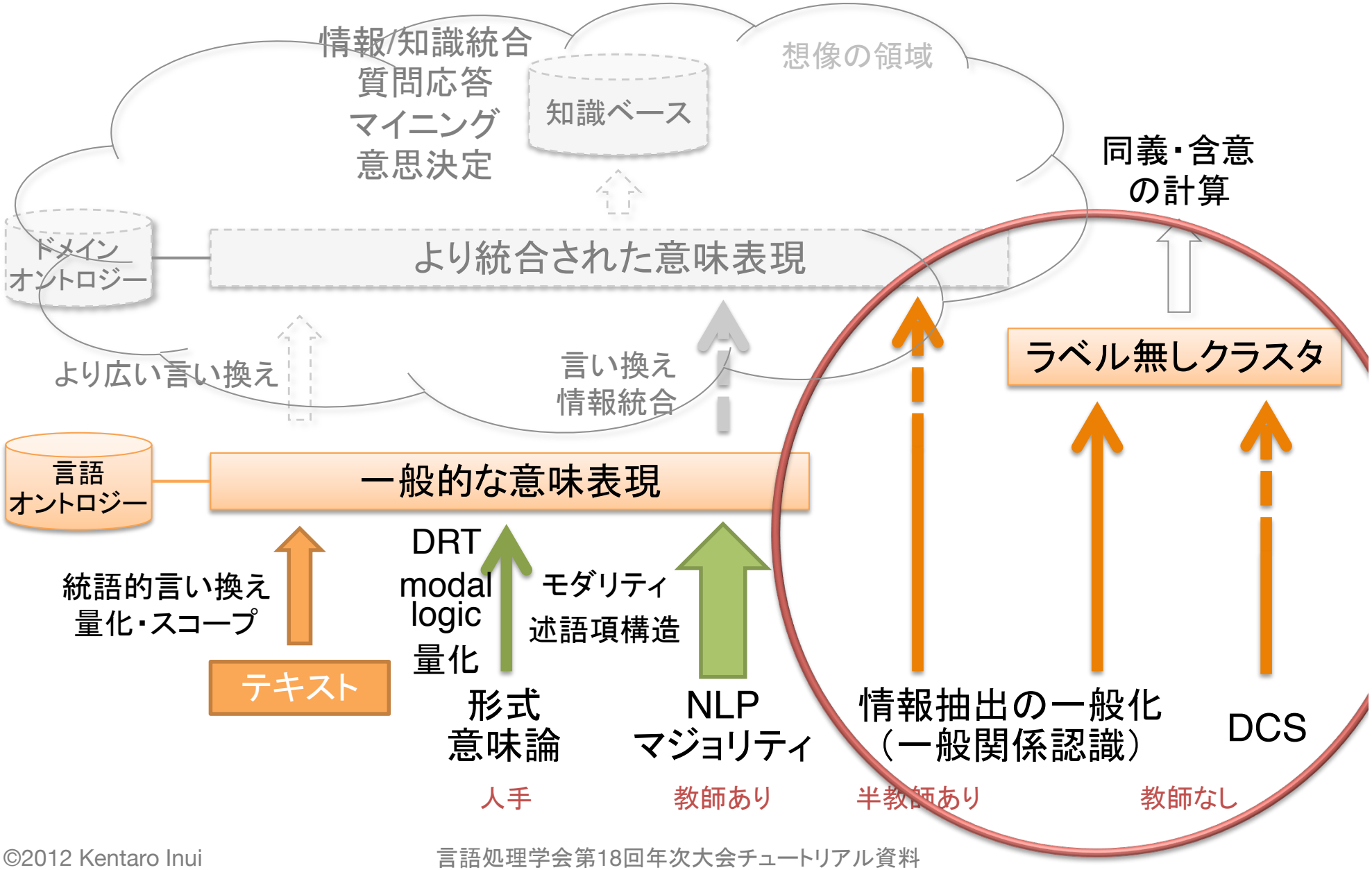
← numerical optimization (L-BFGS)

k-best list

tree1 ✗
tree2 ✗
tree3 ✓
tree4 ✗
tree5 ✗

P. Liang, et al. (2011), Learning dependency-based compositional semantics, ACL.
http://cs.stanford.edu/~pliang/papers/dcs-acl2011-talk.pdf

意味解析の捉え方の広がり（地図①）



©2012 Kentaro Inui

実体/関係のクラスをどう決めるか

① 固定のセット

– 例：NELL（Carlson+ 10）

- 実体/関係のクラス総数 約500
- 人間が設計，シードで指定（ドメインごとに？）

② クラス粒度を自動選択

– 例：Proposition Store（Peñas & Hovy 10）

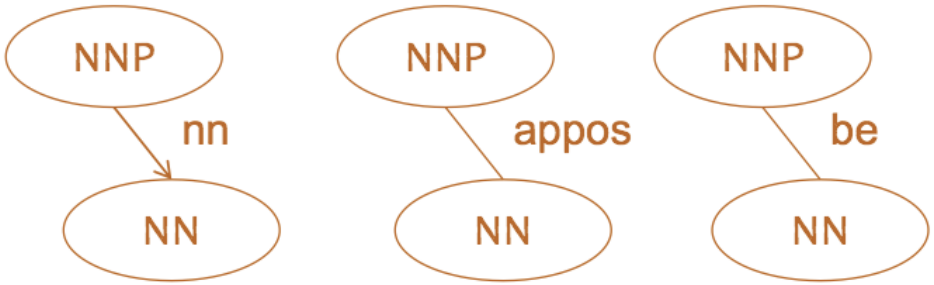
- フットボールドメインでは*person*は粒度が荒すぎる
- 適当な粒度は，*quarterback, leader, veteran, etc.*
- ドメインでインスタンスがどう分類されているか

  → 適当な粒度のクラスを獲得

human? player? quarterback?

（Peñas & Hovy 10）

〈クラス–インスタンス〉のパターン



ドメインコーパスから抽出

334:has_instance:[quarterback:n, ('Kerry':'Collins'):name].
306:has_instance:[end:n, ('Michael':'Strahan'):name].
192:has_instance:[team:n, 'Giants':name].
178:has_instance:[owner:n, ('Jerry':'Jones'):name].
151:has_instance:[linebacker:n, ('Jessie':'Armstead'):name].
145:has_instance:[coach:n, ('Bill':'Parcells'):name].
139:has_instance:[receiver:n, ('Amani':'Toomer'):name].
...

http://www.mavir.net/docs/ReadingMachine-apenas.pdf

human? player? quarterback?

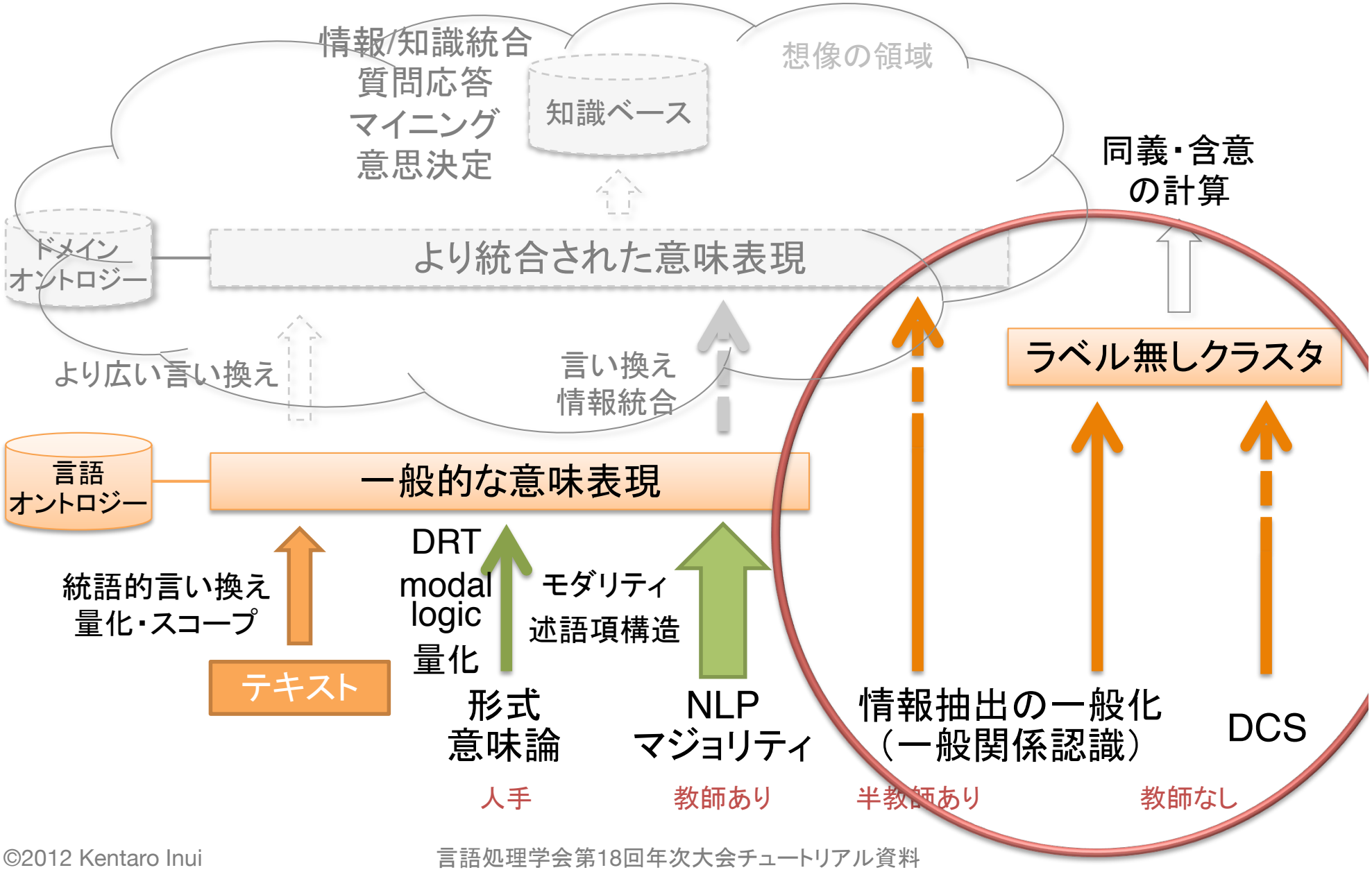
（Peñas & Hovy 10）

頻出クラス

15457 quarterback
12395 coach
7865 end
7611 receiver
6794 linebacker
4348 team
4153 coordinator
4127 player
4000 president
3862 safety
3722 cornerback
3479 (wide:receiver)
3344 (defensive:end)
3265 director
3252 owner
2870 (tight:end)
2780 agent
2291 guard
2258 pick
2177 manager
2138 (head:coach)
2082 rookie
2039 back
1985 (defensive:coordinator)
1882 lineman
1836 (offensive:coordinator)
1832 tackle
1799 center
1776 (general:manager)
1425 fullback
1366 (vice:president)
1196 backup
1193 game
1140 choice
1102 starter
1100 spokesman
1083 (free:agent)
1006 champion
990 man
989 son
987 tailback

http://www.mavir.net/docs/ReadingMachine-apenas.pdf

意味解析の捉え方の広がり（地図①）



©2012 Kentaro Inui

実体/関係のクラスをどう決めるか

③ 文脈類似度によるクラスタリング

- 各表現はクラスタに確率的に所属
- クラスタはラベルを持たない

④ クラスタも作らない

- ベクトル空間モデル
  - 語の意味を文脈ベクトルで表現
- Compositional distributional semanticsへ

知識獲得（地図③）

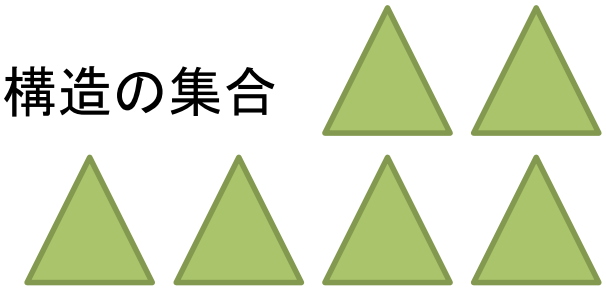
単語–パターン行列

	X kill	*die of X*	*X prevent*
heart attack	1	1	0
heart disease	1	1	0
exercise	0	0	1

ペア–パターン行列

	X prevent Y	*X is key to Y*	*X reduce risk of Y*
diet – heart attack	1	1	0
exercise – breast cancer	1	1	1
aspirin – blood clots	0	1	1

依存構造の集合



文脈ベクトルのクラスタリング

- 教師なし学習
  - 意味クラス集合を用意しない
  - 意味クラスはラベルを持たない

似たベクトルを持つ語をグルーピング

	X kill	*X is for*	*X prevent*	*X be cause of*	*X be key to*	*X reduce risk*	*die of X*	*prevent X*	*reduce risk of X*
heart attack	1	0	0	1	0	0	1	1	1
heart disease	1	0	0	1	0	0	1	1	1
sunburn	0	0	0	1	0	0	0	1	1
diet	0	1	1	0	1	1	0	0	0
exercise	0	1	1	0	1	1	0	0	0

文脈ベクトルのクラスタリング

- ソフトクラスタリング，共クラスタリング
- LSA, pLSI, LDA, etc.
- 研究多数
  - Dhillon (2001)
  - Hodge & Austin (2002)
  - Pereira+ (1993)
  - Dhillon & Guan (2003)
  - Kazama & Torisawa (2008)

Class 791	Class 2760
ウィン ダ ム (WINDOM)	マリン/スタジアム (Chiba Marine Stadium [abb.])
カ ム リ (CAMRY)	大阪/ドーム (Osaka Dome)
ディア マ ン テ (DIAMANTE)	ナ ゴ/ド (Nagoya Dome [abb.])
オ デッセ イ (ODYSSEY)	福岡/ドーム (Fukuoka Dome)
インスパイア (INSPIRE)	大阪/球場 (Osaka Stadium)
ス イ フ ト (SWIFT)	ハマ/スタ (Yokohama Stadium [abb.])

J. Kazama and K. Torisawa (2008) Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. ACL.

関係とパターンの共クラスタリング

	...	...	*X prevent Y*	...	*X is key to Y*	...	...	*X reduce risk of Y*	...
diet – heart attack	0	0	1	0	1	0	0	1	0
sunscreen – sunburn	0	0	1	0	1	0	0	0	0
...	1	0	0	0	0	1	0	0	0
exercise – breast cancer	0	0	1	0	0	0	0	1	0
...	0	1	0	0	0	0	1	0	0
aspirin – blood clots	0	0	0	0	1	0	0	1	0
...	0	0	0	0	0	0	1	0	1
safety belt – fatalities	0	0	0	0	0	0	0	1	1

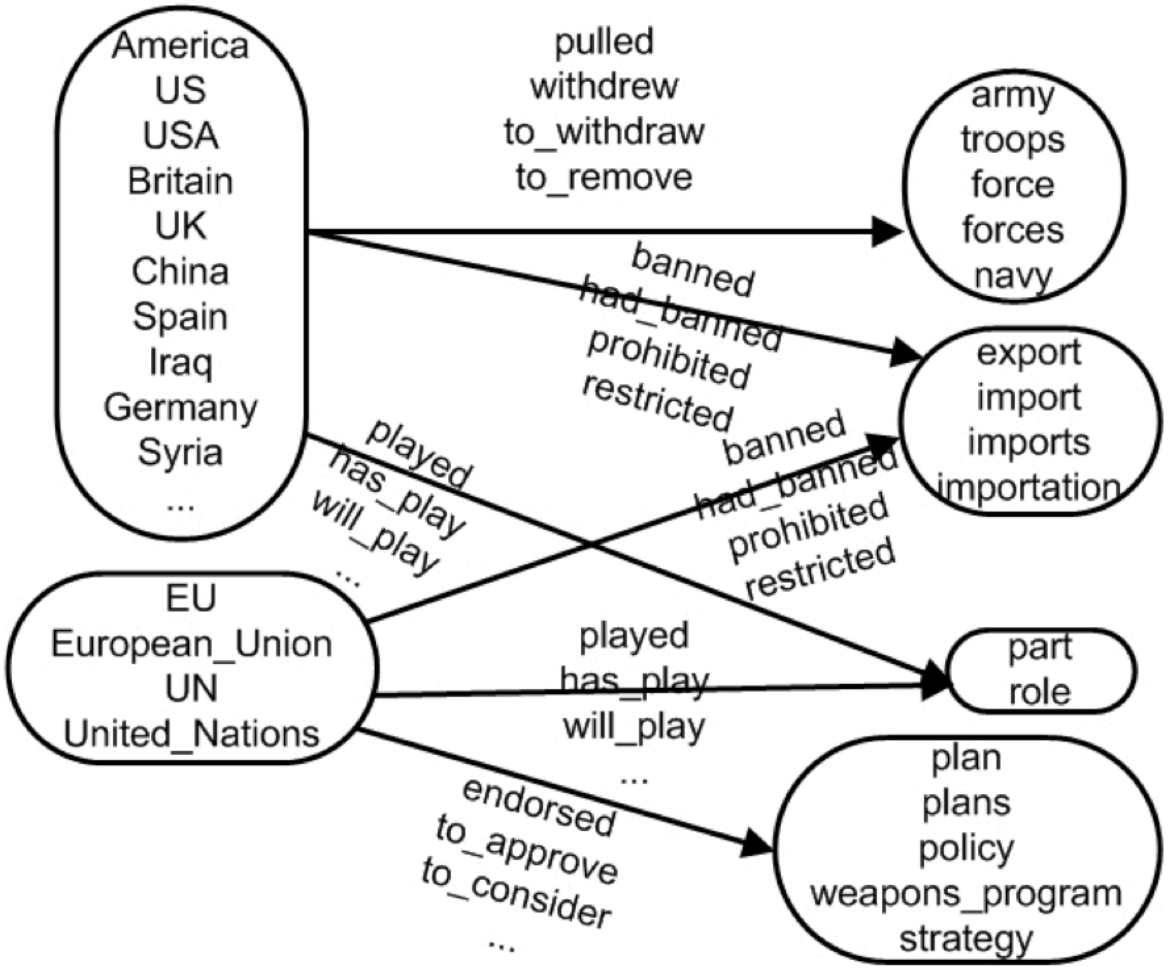
関係とパターンの共クラスタリング

例. pLSI

$$P(r,p) = \sum_c P(r \mid c) P(p \mid c) P(c)$$

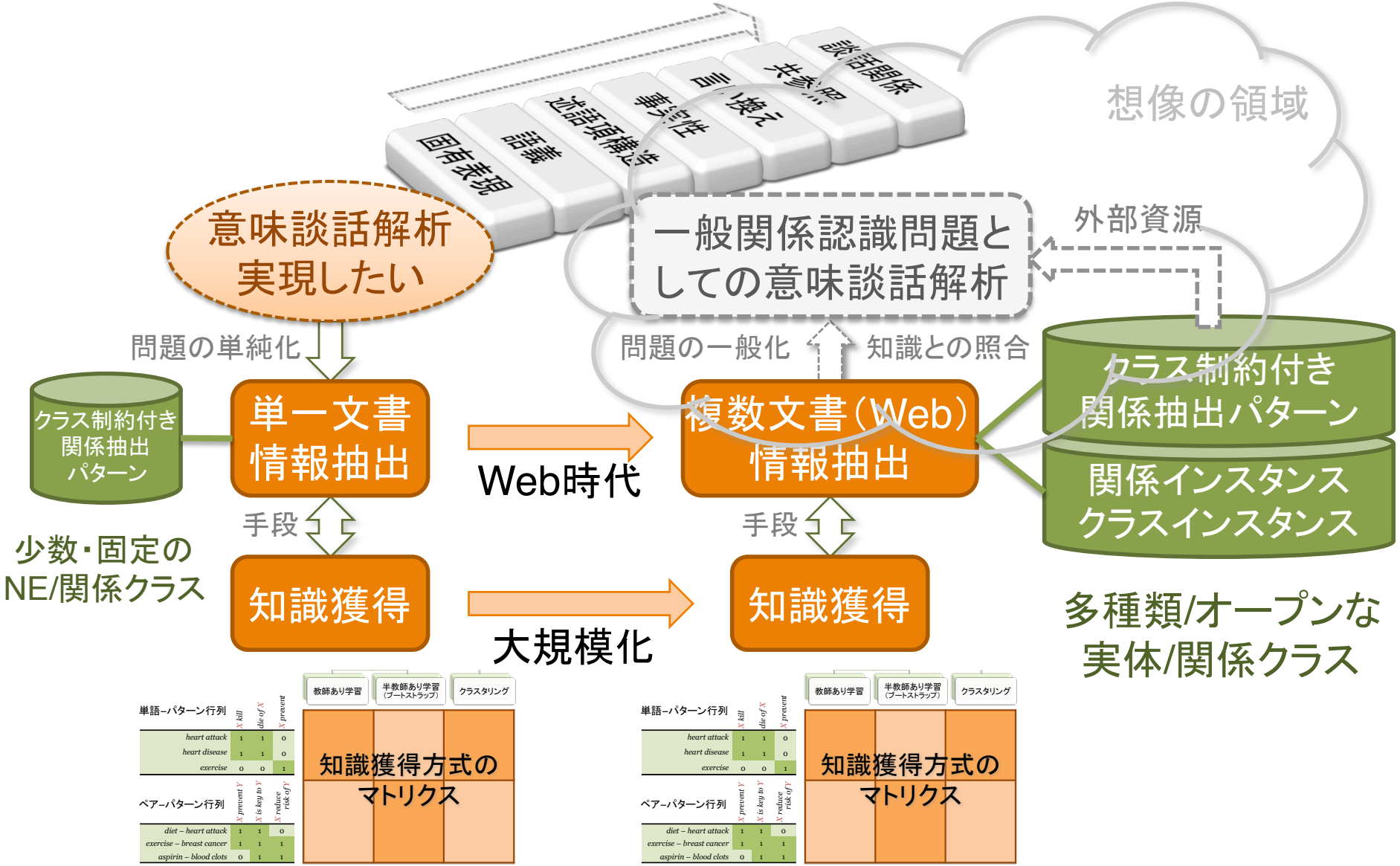
	…	…	*X prevent Y*	*X is key to Y*	*X reduce risk of Y*	*X protect against Y*	…	…	…
…	1	1	0	0	0	1	0	0	0
diet – heart attack	0	0	1	1	1	0	0	0	0
sunscreen – sunburn	0	0	1	1	0	0	0	0	0
exercise – breast cancer	0	0	1	0	1	0	0	0	0
aspirin – blood clots	0	0	0	1	1	0	0	0	0
safety belt – fatalities	0	0	0	0	1	1	0	0	0
…	0	0	0	0	0	0	1	1	1
…	0	0	0	0	0	0	1	0	1

実体と関係の共クラスタリング



S. Kok and P. Domingos (2008) Extracting semantic networks from text via relational clustering, ECML.

情報抽出の展開（地図②）


固有表現
語義
述語項構造
時間性
言い換え
共参照
談話関係
想像の領域
意味談話解析
実現したい
一般関係認識問題と
しての意味談話解析
外部資源
問題の単純化
問題の一般化
知識との照合
クラス制約付き
関係抽出
パターン
単一文書
情報抽出
Web時代
複数文書（Web）
情報抽出
クラス制約付き
関係抽出パターン
関係インスタンス
クラスインスタンス
手段
手段
少数・固定の
NE/関係クラス
知識獲得
大規模化
知識獲得
多種類/オープンな
実体/関係クラス
単語–パターン行列
X kill
die of X
X prevent
heart attack 1 1 0
heart disease 1 1 0
exercise 0 0 1
ペア–パターン行列
X prevent Y
X is key to Y
X reduce risk of Y
diet – heart attack 1 1 0
exercise – breast cancer 1 1 1
aspirin – blood clots 0 1 1
教師あり学習
半教師あり学習
（ブートストラップ）
クラスタリング
知識獲得方式の
マトリクス
単語–パターン行列
X kill
die of X
X prevent
heart attack 1 1 0
heart disease 1 1 0
exercise 0 0 1
ペア–パターン行列
X prevent Y
X is key to Y
X reduce risk of Y
diet – heart attack 1 1 0
exercise – breast cancer 1 1 1
aspirin – blood clots 0 1 1
教師あり学習
半教師あり学習
（ブートストラップ）
クラスタリング
知識獲得方式の
マトリクス

知識獲得方式（地図③）

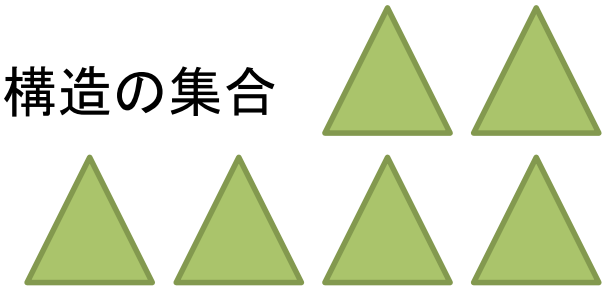
単語–パターン行列

	X kill	*die of X*	*X prevent*
heart attack	1	1	0
heart disease	1	1	0
exercise	0	0	1

ペア–パターン行列

	X prevent Y	*X is key to Y*	*X reduce risk of Y*
diet – heart attack	1	1	0
exercise – breast cancer	1	1	1
aspirin – blood clots	0	1	1

依存構造の集合



再帰構造のクラスタリング

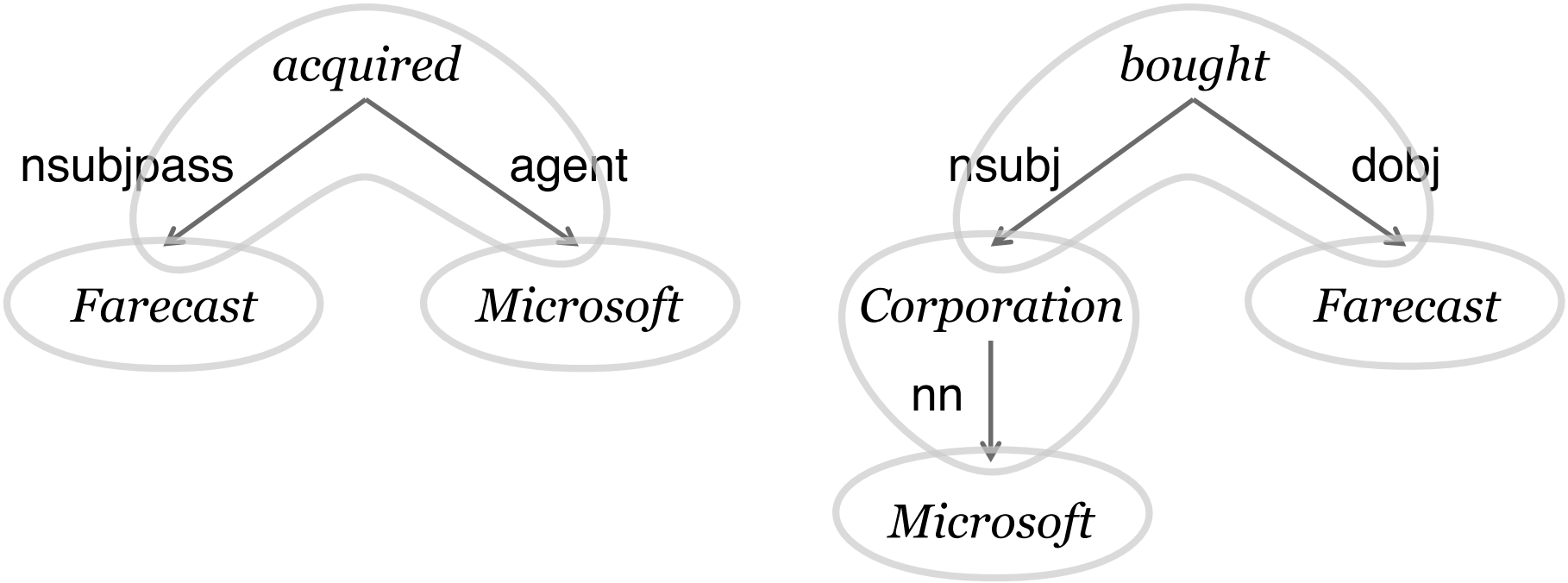
- Poon and Domingos（2009）
  - Unsupervised semantic parsing
  - 依存構造の部分木を確率的にクラスタリング
  - Markov logicで確率モデルを表現

部分木のクラスタリング

（Poon & Domingos 2009）

クラスタリングの初期状態

Farecast was acquired by Microsoft.　*Microsoft Corporation bought Farecast.*



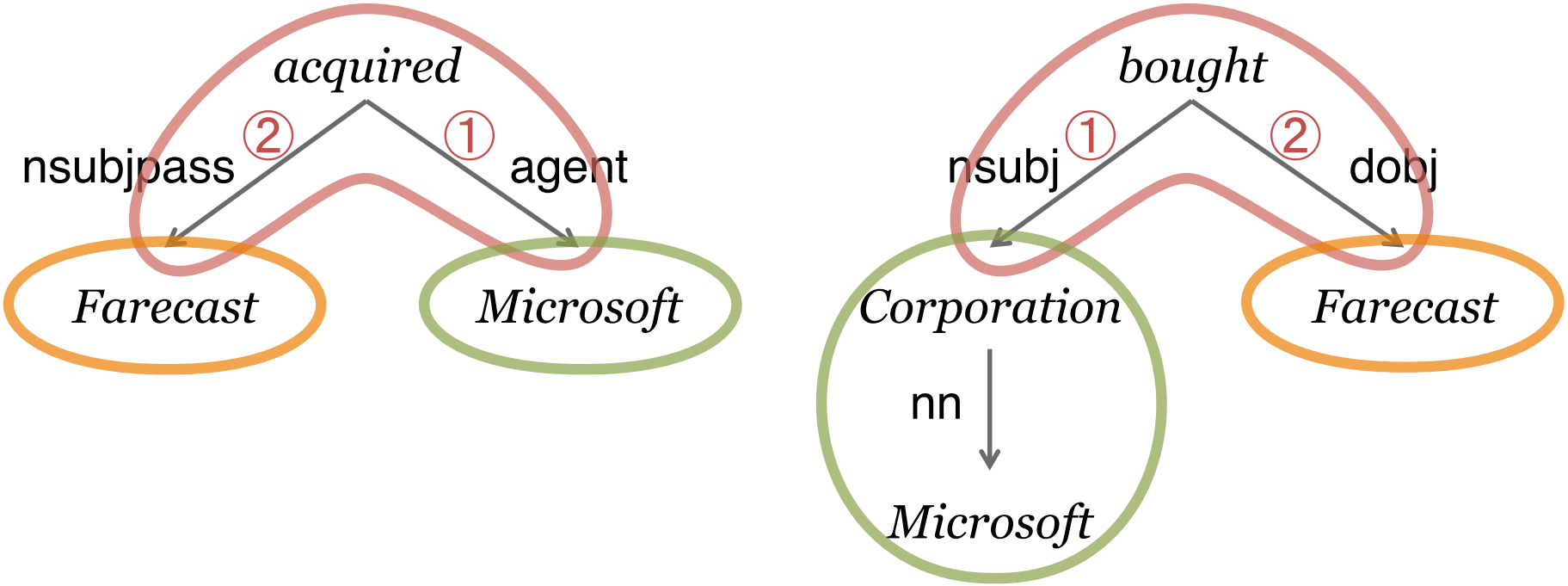
H. Poon (2011), Markov Logic for Machine Reading. Ph.D. Thesis.

部分木のクラスタリング

（Poon & Domingos 2009）

推定されるクラスタ

Farecast was acquired by Microsoft.　*Microsoft Corporation bought Farecast.*



H. Poon (2011), Markov Logic for Machine Reading. Ph.D. Thesis.

実体/関係のクラスをどう決めるか

③ 文脈類似度によるクラスタリング

– 各表現はクラスタに確率的に所属

– クラスタはラベルを持たない

④ クラスタも作らない

– ベクトル空間モデル

  • 語の意味を文脈ベクトルで表現

– Compositional distributional semanticsへ

Compositional Distributional Semantics

- 語の意味を文脈ベクトルで表現（ベクトル空間モデル; VSM）
- フレーズの文脈ベクトルを単語の文脈ベクトルから構成的に計算
- フレーズ間の意味的類似度 ＝ 文脈ベクトルの近さ

例. Thater+（2010）

$$\begin{array}{ccccc} \textit{acquire} & \times & \textit{knowledge} & = & \textit{acquire knowledge} \\ \begin{pmatrix} 15_{(\mathrm{OBJ},\ \mathrm{OBJ}-1,\ \textit{gain})} \\ 6_{(\mathrm{OBJ},\ \mathrm{CONJ}-1,\ \textit{skill})} \\ 6_{(\mathrm{OBJ},\ \mathrm{OBJ}-1,\ \textit{buy-back})} \\ 42_{(\mathrm{OBJ},\ \mathrm{OBJ}-1,\ \textit{purchase})} \\ \ldots \end{pmatrix} & & \begin{pmatrix} 5_{(\mathrm{OBJ}-1,\ \textit{gain})} \\ 2_{(\mathrm{CONJ}-1,\ \textit{skill})} \\ 0_{(\mathrm{OBJ}-1,\ \textit{buy-back})} \\ 0_{(\mathrm{OBJ}-1,\ \textit{purchase})} \\ \ldots \end{pmatrix} & & \begin{pmatrix} 75_{(\mathrm{OBJ},\ \mathrm{OBJ}-1,\ \textit{gain})} \\ 12_{(\mathrm{OBJ},\ \mathrm{CONJ}-1,\ \textit{skill})} \\ 0_{(\mathrm{OBJ},\ \mathrm{OBJ}-1,\ \textit{buy-back})} \\ 0_{(\mathrm{OBJ},\ \mathrm{OBJ}-1,\ \textit{purchase})} \\ \ldots \end{pmatrix} \end{array}$$

Compositional Distributional Semantics

- 近年論文増加
  - Clark & Pulman (2007), Erk & Padó (2008), Widdows (2008), Preller & Prince (2010), Thater+ (2010), Reddy+ (2011), Clarke (2011), Mitchell (2011), etc.
  - Proc. of the ACL-HLT 2011 Workshop on Distributional Semantics and Compositionality (DiSCo'2011)
- 対象
  - 動詞×名詞，形容詞×名詞，文全体，etc.
  - 語順，多義性，連言，選言，否定
- 文の意味をどこまでVSMで表現できるかは今後の課題

やわらかい記号の世界: まとめ

- 問題
  - テキストと知識ベースの柔軟な照合が必要
  - 知識ベースの実体/関係のクラスをどう決めるか
- 対策
  - 固定多クラス(ラベルあり)から
    クラスタリング, DCS(ラベルなし)まで多様化
  - 構造を持つ意味表現を教師なし学習するアプローチも
- Open issues
  - ラベル無しの意味表現だけで自己完結できるか？
  - Propositional & Distributional Semanticsの融合？

本講演の構成

1. Annotation layer cake
2. 情報抽出と知識獲得の今
3. Making the implicit explicit
4. やわらかい記号の世界
5. ルールを作るところから

©2012 Kentaro Inui

多様なタスク、性質の見極めを

"Event extraction" represents a broad range of tasks. The differences among these tasks reflect themselves in differences in the applicable extraction procedures and differences in the effectiveness of particular learning methods. These factors must be recognized if we are to make continued progress in event extraction.

（Ralph Grishman: The Impact of Task and Corpus on Event Extraction Systems, LREC 2010.）

多様なタスク、性質の見極めを

We must also maintain an awareness of the characteristics of the corpora, particularly if the corpora for development or evaluation differ in some way from the final target application. Understanding the characteristics of the corpus is an inherent part of understanding the extraction task.

（Ralph Grishman: The Impact of Task and Corpus on Event Extraction Systems, LREC 2010.）

ルールを作っているのは誰か？

テンプレート

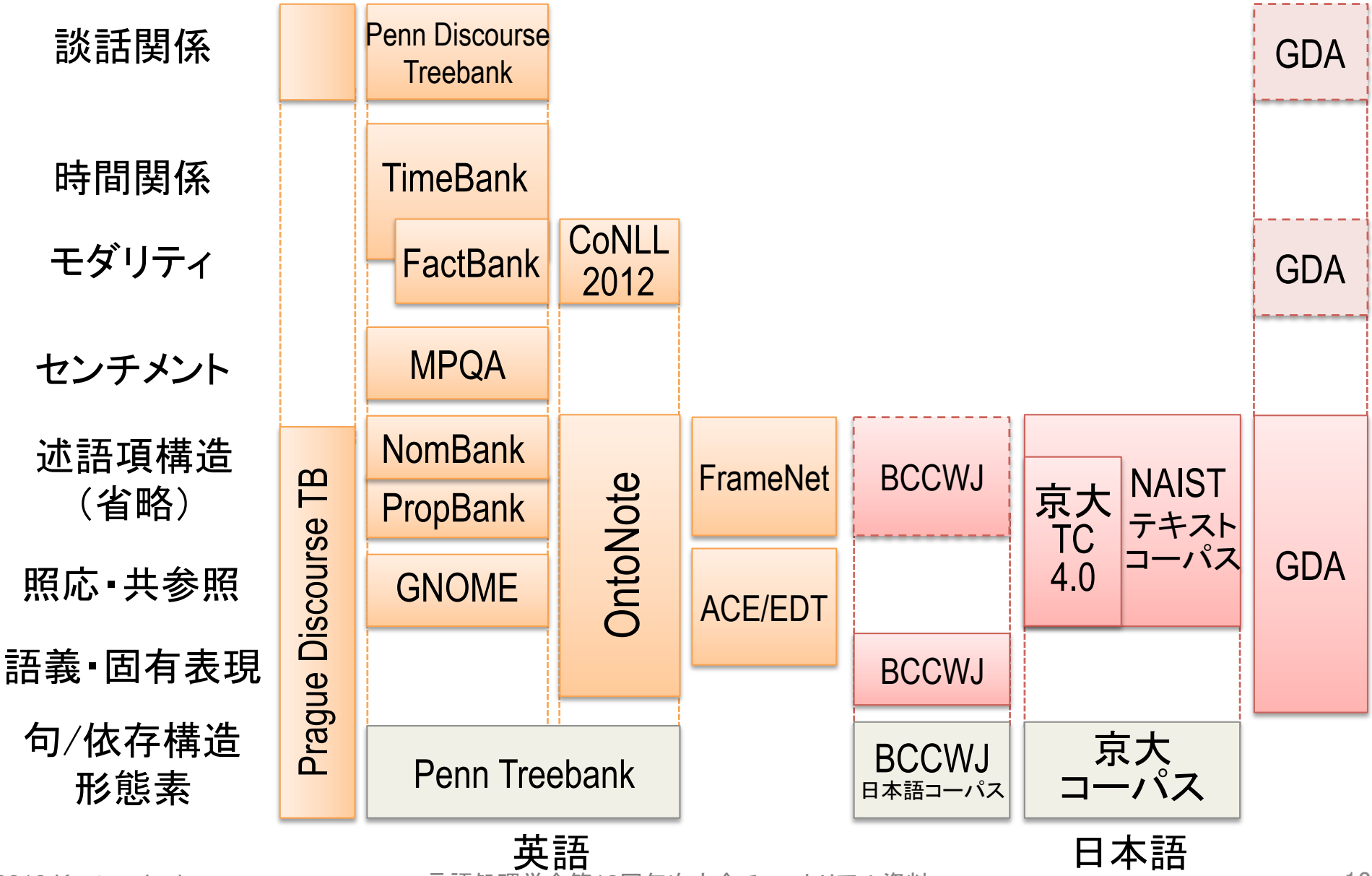
TIE-UP-1	
Relationship:	**TIE-UP**
Entities:	*"Bridgestone Sports Co."*
	"a local concern"
	"a Japanese trading house"
Joint Venture Company:	*"Bridgestone Sports Taiwan Co."*
Amount:	NT$2,000,000

*Bridgestone Sports Co. said Friday it has **set up** a joint venture in Taiwan **with** a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan.*

*The joint venture, Bridgestone Sports Taiwan Co., **capitalized at** 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.*

Onyshkevych, et al. Tasks, Domains, and Languages. Proc. of the 5th Message Understanding Conference, 1993.

ルールを作っているのは誰か？



OntoNote（CoNLL2011）の共参照タグ

generic nominal mention との共参照は代名詞と定名詞句のみ. *generic nominal mention* 同士の共参照は考えない.

- *Meetings are most productive when they are held in the morning.*（この例のみ共参照）
- *Allergan Inc. said it received approval to sell the PhacoFlex intraocular lens, the first foldable silicone lens available for *cataract surgery. The lens' foldability enables it to be inserted in smaller incisions than are now possible for *cataract surgery.*
- *Argentina said it will ask creditor banks to *halve its foreign debt of $64 billion – the third-highest in the developing world . Argentina aspires to reach *a reduction of 50% in the value of its external debt.*

S. Pradhan, L. Ramshow, M. Marcus, M. Palmer, R.Weischedel, and N. Xue（2011）
CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes, CoNLL Shared Task.

Penn Discourse Treebank

（Miltsakaki, Prasad, Joshi and Webber, 2004）

- 談話マーカで明示された談話関係
  → スコープを同定
- 非明示的な談話関係
  → 隣接する文の間の談話関係を同定

Michelle lives in a hotel room, *and* although **she drives a canary-colored Porsche**, *she hasn't time to clean or repair it*. (2402)

Also unlike Mr. Ruder, Mr. Breeden appears to be in a position to get somewhere with his agenda. Implicit = BECAUSE (CAUSE) **As a former White House aide who worked closely with Congress, he is savvy in the ways of Washington**. (0955)

Examples from Prasad, et al. (2004) Attribution and its annotation in the Penn Discourse TreeBank. L'objet.

計算言語学者が牽引

- C. Fillmore（認知言語学，フレーム意味論）
  - C. Baker, C. Fillmore, and J. Lowe. The Berkeley FrameNet project. COLING/ACL, 1998.
  - N. Ide, C. Baker, C. Fellbaum, C. Fillmore, and R. Passonneau. MASC: The manually annotated sub-corpus of American English. LREC, 2008.
- J. Pustejovsky（語彙意味論，生成語彙論）
  - J. Pustejovsky, A. Meyers, M. Palmer, and M. Poesio. Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference. Workshop on Frontiers in Corpus Annotation II, 2005.
  - R. Sauri, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky. TimeML Annotation Guidelines Version 1.2.1. 2006.
  - R. Sauri and J. Pustejovsky. FactBank: A corpus annotated with event factuality. Language Resources and Evaluation, 2009.

計算言語学者が牽引

- Prague Tectogrammatics

1. **Eva Hajicova**. (1993). Issues of Sentence Structure and Discourse Patterns. Prague: Charles University.
   Available in: BibTex item
2. **Eva Hajicová** (1998a). Movement Rules Revisited. In: Processing of Dependency-Based Grammars, *Proceedings from the Workshop COLING/ACL*, Montreal, ed. S. Kahane and A. Polguere, 49-57.
   Available in: BibTex item
3. **Eva Hajicova, Marketa Ceplova**. (2000). Deletions and Their Reconstruction in Tectogrammatical Syntactic Tagging of Very Large Corpora. In *Proceedings of COLING'2000*, pp. 228-284, Saarbruecken, Germany.
   Available in: BibTex item
4. **Eva Hajicova, Jarmila Panevova**. (1984). Valency (case) frames of verbs. In: Sgall (1984:147-188).
   Available in: BibTex item
5. **Eva Hajicova, Barbara Partee, Petr Sgall** (1998): *Topic-focus articulation, tripartite structures, and semantic content*. Amsterdam:Kluwer
   Available in: BibTex item
6. **Marcus M. P., Kim G., Marcinkiewicz M. A. et al.** (1994). The Penn Treebank: Annotating Predicate Argument Structure. Proceedings of the ARPA Human Language Technology Workshop. San Francisco: Morgan Kaufmann.
7. **Marcus M. P., Santorini B. and Marcinkiewicz M. A.** (1993). Building a Large Annotated Corpus of English: the Penn Treebank. Computational Linguistics, 19(2), 313-330.
8. **Jarmila Panevova**. 1974. "On verbal frames in Functional Generative Description". *Prague Bulletin of Mathematical Linguistics* 22:3-40; 23(1975):17-52.
   Available in: BibTex item
9. **Jarmila Panevova**. 1980. *Formy a funkce ve stavbe ceske vety*. [Forms and Functions in the Structure of the Czech Sentence]. Prague: Academia.
   Available in: BibTex item
10. **Vladimir Petkevic** (1987). A New Dependency Based Specification of Underlying Representations of Sentences. *Theoretical Linguistics* 14:143-172.
    Available in: BibTex item
11. **Vladimir Petkevic.** (1995). A New Formal Specification of Underlying Representations. *Theoretical Linguistics* 21:7-61.
    Available in: BibTex item
12. **Petr Sgall.** (1967). *Generativni popis jazyka a ceska deklinace*. [Generative Description of Czech and Czech Declension.] Prague: Academia.
    Available in: BibTex item
13. **Petr Sgall** ed. (1984). Contributions to Functional Syntax, Semantics and Language Comprehension. Amsterdam: Benjamins - Prague: Academia.
    Available in: BibTex item
14. **Petr Sgall.** 1992. Underlying Structure of Sentences and Its Relations to Semantics. *Wiener Slawistischer Almanach*. Sonderband 33. Ed. by T. Reuther. Wien: Gesellschaft zur Förderung slawistischer Studien, 273-282.
    Available in: BibTex item
15. **Petr Sgall.** (1997a). Valency and Underlying Structure. An Alternative View on Dependency. In: L. Wanner (ed.): *Recent Trends in Meaning-Text Theory*. Amsterdam/Philadelphia: Benjamins, 149-166.
    Available in: BibTex item
16. **Petr Sgall** (1997b). On the Usefulness of Movement Rules. In: Caron B. (ed.), *Actes du 16e Congres International des Linguistes* (Paris 20-25 juillet 1997), Oxford: Elsevier Sciences.
    Available in: BibTex item
17. **Petr Sgall** (in press). The Freedom of Language. To appear in *Prague Linguistic Circle Papers* 4.
18. **Petr Sgall, Eva Hajicova, Jarmila Panevova** (1986): *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, ed. by J. L. Mey, Dordrecht:Reidel - Prague: Academia.

http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/Doc/tect.html

ルール作りから①：大学入試問題

2009年度センター試験
世界史B

問 6　下線部⑥に関連して，兵制や兵士について述べた文として最も適当なものを，次の①～④のうちから一つ選べ。 6

兵制や兵士について述べた文として最も適当なものを，次のうちから一つ選べ。
① イェニチェリは，オスマン帝国の常備軍であった。
② フランク王国では，テマ制（軍管区制）の下で屯田兵制が行われた。

東地中海の強国―オスマン帝国
．．．イェニチェリ軍団は，軍楽隊，工兵隊，大砲隊，鉄砲隊などをそなえた皇帝直属の常備軍で，のちにヨーロッパで発展する近代的陸軍の先がけであった。

ビザンツ帝国
．．．7世紀前半のヘラクレイオス1世は，アラブ軍の侵入にそなえて地方の体制を整えるためにテマ（軍管区）制を採用した。兵士には軍役とひきかえに世襲地が与えられ（屯田兵制），自作農の増加によって農村社会は活力にあふれるようになった。

東京書籍　平成19年度　世界史B

宮尾祐介「自然言語の意味解析はなぜ難しいのか」，ALAGIN技術開発部会：自然言語処理分科会セミナー，2012年2月

ルール作りから①：大学入試問題

- 記憶していることと問われていることが意味的に一致しているかどうかを認識する
  → 含意関係認識

兵制や兵士について述べた文として最も適当なものを，次のうちから一つ選べ。
① イェニチェリは，オスマン帝国の常備軍であった。
② フランク王国では，テマ制（軍管区制）の下で屯田兵制が行われた。

イェニチェリ
＝イェニチェリ軍団

オスマン帝国の常備軍
← 皇帝直属の常備軍

東地中海の強国―オスマン帝国
．．．イェニチェリ軍団は，軍楽隊，工兵隊，大砲隊，鉄砲隊などをそなえた皇帝直属の常備軍で，のちにヨーロッパで発展する近代的陸軍の先がけであった。

宮尾祐介「自然言語の意味解析はなぜ難しいのか」，ALAGIN技術開発部会：自然言語処理分科会セミナー，2012年2月

ルール作りから②: 常識推論

- 文の implication を認識する
  - 文のEntailment: 文が真ならば，必然的に真である命題
  - 文のImplication: 文が真ならば，真であると期待できる命題

The service at the restaurant was slow.

[implication] *The restaurant was crowded.*

ルール作りから②: 常識推論

- Choice Of Plausible Alternatives（COPA）（Roemmele+ 11）
  - premise に対して最も plausible な implication を選択するタスク
  - 開発セット500問，テストセット500問
  - 多様なトピック，高い作業者間一致率
  - 共起指標を用いた単純な手法: 正解率58.8%

[Premise] The man fell unconscious. *What was the cause of this?*
[Alternative 1] The assailant struck the man in the head.
[Alternative 2] The assailant took the man's wallet.

M. Roemmele, C. A. Bejan, and A. S. Gordon (2011), Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning, AAAI-2011 Spring Symposium on Logical Formalizations of Commonsense Reasoning.

まとめ

- 大規模言語資源（コーパス，知識）が利用可能な時代を迎え，知識獲得のボトルネックが少なくとも部分的には解消されつつある
- しかし，大規模知識ベースなるミッシングピースをはめ込めば直ちに深い意味談話解析や推論が実現できる，といった単純なストーリーには今のところなっていない

まとめ

- 一方，応用指向的に問題を限定することで発展してきた情報抽出技術は，ウェブスケールの知識獲得との融合により，超多クラスの情報抽出と情報統合をねらう新しい領域を切り拓きつつあり，意味解析の目的に別の角度から接近している
- ただし，パターンベースの抽出が困難な非明示的関係を推論する談話解析相当問題については，頑健で柔軟な推論方式や意味の表現方法，モデルの評価方法など，今後取り組むべき課題が山積している

参考文献

Annotation layer cake

- D. Kawahara and S. Kurohashi (2006) A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis. HLT.
- A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman (2004), The Nombank Project: An interimreport, HLT-NAACL Workshop on Frontiers in Corpus Annotation.
- E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber (2004), The Penn Discourse Treebank, LREC.
- M. Palmer, D. Gildea, and P. Kingsbury (2005), The proposition bank: An annotated corpus of semantic roles, Computational Linguistics.
- R. Prasad, et al. (2004), Attribution and its annotation in the Penn Discourse Treebank. L'objet.
- R. Sasano, D. Kawahara and S. Kurohashi (2010) The Effect of Corpus Size on Case Frame Acquisition for Predicate-Argument Structure Analysis. IEICE TRANSACTIONS on Information and Systems, E93-D (6).
- R. Sauri and J. Pustejovsky. FactBank (2009) A corpus annotated with event factuality. LREC.

情報抽出と知識獲得の今

- S. Abe, K. Inui, and Y. Matsumoto (2008), Two-phased event relation acquisition: coupling the relation-oriented and argument-oriented approaches. COLING.
- M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni (2007), Open information extraction from the web. IJCAI.
- A. Carlson, J. Betteridge, E. R. Hruschka Jr. and T. M. Mitchell (2009) Coupling Semi-Supervised Learning of Categories and Relations. NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing.
- A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr. and T.M. Mitchell (2010) Toward an Architecture for Never-Ending Language Learning. AAAI.
- T. Chklovski and P. Pantel (2004), Verbocean: Mining the web for fine-grained semantic verb relations, EMNLP.
- J. Christensen, M. S. Soderland and O. Etzioni (2011), An Analysis of Open Information Extraction based on Semantic Role Labeling. International Conference on Knowledge Capture.

- T. Chklovski and P. Pantel (2004), Verbocean: Mining the web for fine-grained semantic verb relations, EMNLP.

- J. Christensen, M. S. Soderland and O. Etzioni (2011), An Analysis of Open Information Extraction based on Semantic Role Labeling. International Conference on Knowledge Capture.
- S. De Saeger, K. Torisawa, J. Kazama, K. Kuroda, and M. Murata (2009), Large Scale Relation Acquisition Using Class Dependent Patterns. ICDM.
- O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M. T. Center (2011), Open Information Extraction: The Second Generation, IJCAI.
- A. Fader, S. Soderland, and O. Etzioni (2011), Identifying Relations for Open Information Extraction. EMNLP.
- M. Hearst (1992), Automatic acquisition of hyponyms from large text corpora. COLING.
- M. Komachi (2010), Graph-Theoretic Approaches to Minimally-Supervised Natural Language Learning. Ph.D. Thesis.
- B. Onyshkevych, et al. (1993), Tasks, Domains, and Languages, Proc. of MUC-5.
- P. Pantel and M. Pennacchiotti (2006), Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, COLING.
- M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain (2006), Organizing and searching the world wide web of facts – step one: The one-million fact extraction challenge. AAAI.

- A. Peñas and E. H. Hovy (2010), Filling Knowledge Gaps in Text for Machine Reading, COLING, Poster.
- E. Riloff, and R. Jones (1999), Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. AAAI.
- B. Rosenfeld and R. Feldman (2007), Using corpus statistics on entities to improve semi-supervised relation extraction from the Web. ACL.
- S. Sekine (2006), On-Demand Information Extraction, COLING.
- T. Shibata and S. Kurohashi (2011), Acquiring Strongly-related Events using Predicate-argument Co-occurring Statistics and Case Frames. IJCNLP, poster.
- Y. Shinyama and S. Sekine (2006), Preemptive information extraction using unrestricted relation discovery, HLT-NAACL.
- F. Wu and D. S. Weld (2010), Open information extraction using Wikipedia. ACL.
- R. Yangarber (2003), Counter-training in discovery of semantic patterns. ACL.

Making the implicit explict

- J. Blythe, J. R. Hobbs, P. Domingos, R. J. Kate, and R. J. Mooney (2011), Implementing Weighted Abduction in Markov Logic. IWCS.

- A. Hickl and J. Bensley (2007), A Discourse Commitment-Based Framework for Recognizing Textual Entailment. ACL Workshop on Textual Entailment and Paraphrasing.
- J. R. Hobbs, M. Stickel, P. Martin, and D. Edwards (1993), Interpretation as abduction. Artificial Intelligence.
- E. Hovy (2010), A New Semantics: Merging Propositional and Distributional Approaches. IWCS.
- N. Inoue and K. Inui (2011), ILP-Based Reasoning for Weighted Abduction. AAAI Workshop on Plan, Activity and Intent Recognition.
- 井之上直也，乾健太郎，E. Ovchinnikova，J. R. Hobbs（2012），大規模世界知識を用いた仮説推論による談話解析の課題と対策．言語処理学会第18回年次大会．
- A. Peñas and E. H. Hovy (2010), Semantic Enrichment of Text with Background Knowledge. First International Workshop on Formalisms and Methodology for Learning by Reading.

やわらかい記号の世界

- A. Carlson, J. Betteridge, E. R. Hruschka Jr. and T. M. Mitchell (2009), Coupling Semi-Supervised Learning of Categories and Relations. NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing.

- D. Clarke (2011), A Context-theoretic Framework for Composi-tionality in Distributional Semantics, Computational Linguistics.
- S. Clark and S. Pulman (2007), Combining symbolic and distributional models of meaning, AAAI Spring Symposium on Quantum Interaction.
- I. S. Dhillon (2001), Co-clustering documents and words using Bipartite Spectral Graph Partitioning. KDD, 2001.
- I. S. Dhillon and Y. Guan (2003), Information theoretic clustering of sparse co-occurrence data. ICDM.
- K. Erk and S. Padó (2008), A structured vector space model for word meaning in context, EMNLP.
- V. J. Hodge and J. Austin (2002), Hierarchical word clustering – automatic thesaurus generation. Neurocomputing, 48 (1-4).
- J. Kazama and K. Torisawa (2008), Inducing Gazetteers for Named Entity Recognition by Large-scale Clustering of Dependency Relations. ACL.
- S. Kok and P. Domingos (2008), Extracting Semantic Networks from Text via Relational Clustering. ECML.
- P. Liang, M. I. Jordan, D. Klein (2011), Learning dependency-based compositional semantics, ACL.
- J. Mitchell (2011), Composition in distributional models of semantics, Ph.D. Thesis.

- A. Peñas and E. H. Hovy (2010), Semantic Enrichment of Text with Background Knowledge. First International Workshop on Formalisms and Methodology for Learning by Reading.
- F. Pereira, N. Tishby, and L. Lee (1993), Distributional clustering of English words. ACL.
- A. Preller and V. Prince (2010), Quantum logic unites compositional functional semantics and distributional semantic models. In Compositionality and Distributional Semantic Models, ESSLLI.
- H. Poon (2011) Markov Logic for Machine Reading. Ph.D. Thesis.
- H. Poon and P. Domingos (2009) Unsupervised semantic parsing. EMNLP.
- S. Reddy, I. P. Klapaftis, D. McCarthy, S. Manandhar (2011), Dynamic and Static Prototype Vectors for Semantic Composition, IJCNLP.
- S. Thater, H. Furstenau, and M. Pinkal (2010), Contextualizing Semantic Representations Using Syntactically Enriched Vector Models, ACL.
- D. Widdows (2008), Semantic vector products: Some initial investigations. AAAI Symposium on Quantum Interaction.
- Proc. of the ACL-HLT 2011 Workshop on Distributional Semantics and Compositionality (DiSCo’2011).