

ビッグデータと対話する 自然言語処理

乾 健太郎

東北大学 情報科学研究科

inui@ecei.tohoku.ac.jp

@inuikentaro



コミュニケーション あらゆる社会活動の根幹

最も重要なメディアは
言語（ことば）





Barack Obama @BarackObama

11月7日

This happened because of you. Thank you.

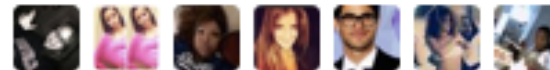
[閉じる](#) [返信](#) [リツイート](#) [★ お気に入りに登録](#)

250,392

件のリツイート

79,179

FAVORITES



<https://twitter.com/>



Barack Obama @BarackObama

11月7日

Four more years. pic.twitter.com/bAJE6Vom

🖼️ 画像を非表示にする ← 返信 ↻ リツイート ★ お気に入り
に giriş に登録



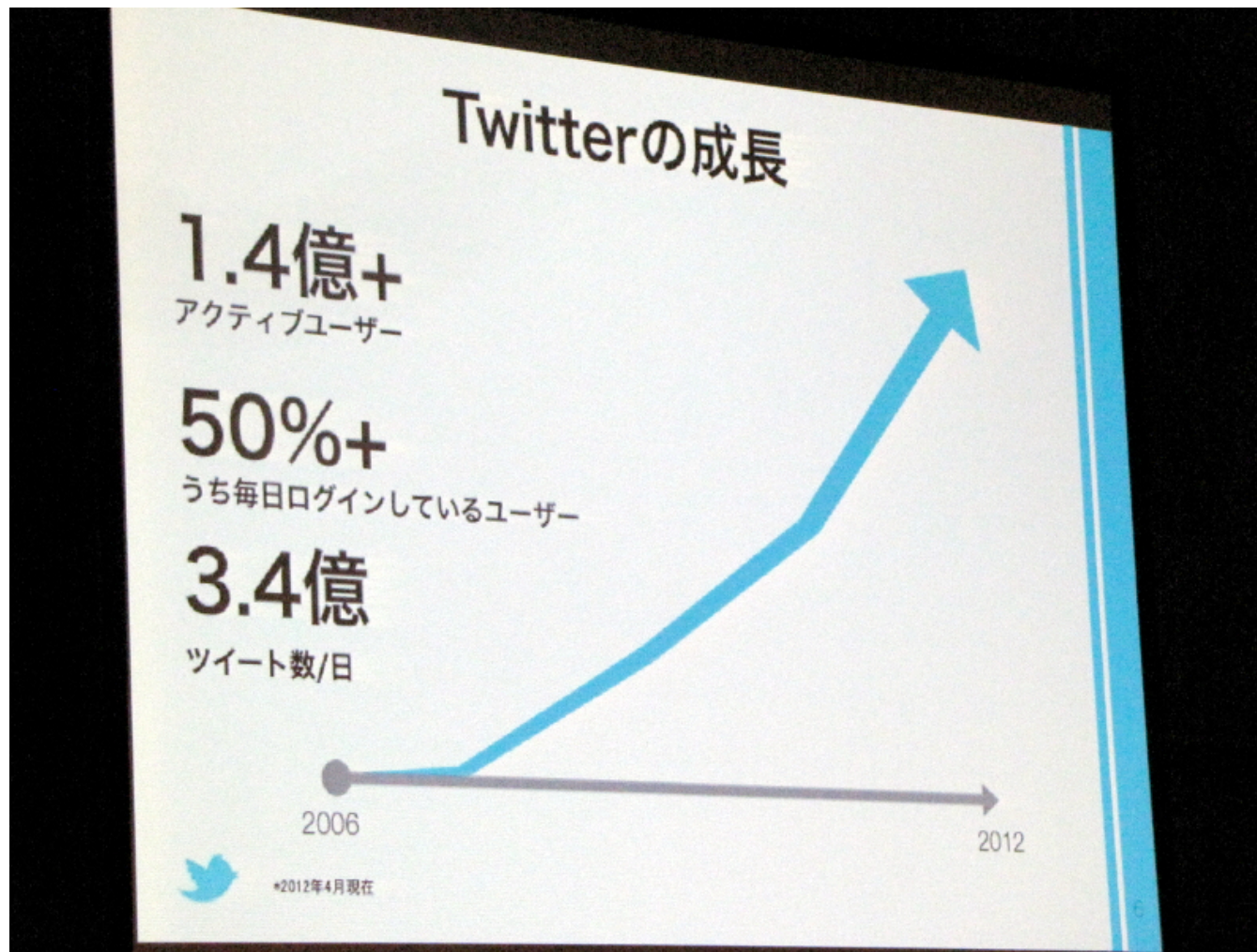
811,021
件のリツイート

294,075
FAVORITES



<https://twitter.com/>

ソーシャルメディアの爆発的普及



http://internet.watch.impress.co.jp/docs/event/itweek12/20120511_532162.html

自然言語処理

人工知能

言葉がわかる賢いコンピュータを作る

言葉 (= 自然言語) で伝達される情報を
理解・検索・抽出・翻訳・整理・分析し、

地球規模のコミュニケーションを**支援**するソフトウェア技術



言語意味解析による情報の編集

信憑性分析

根拠検索

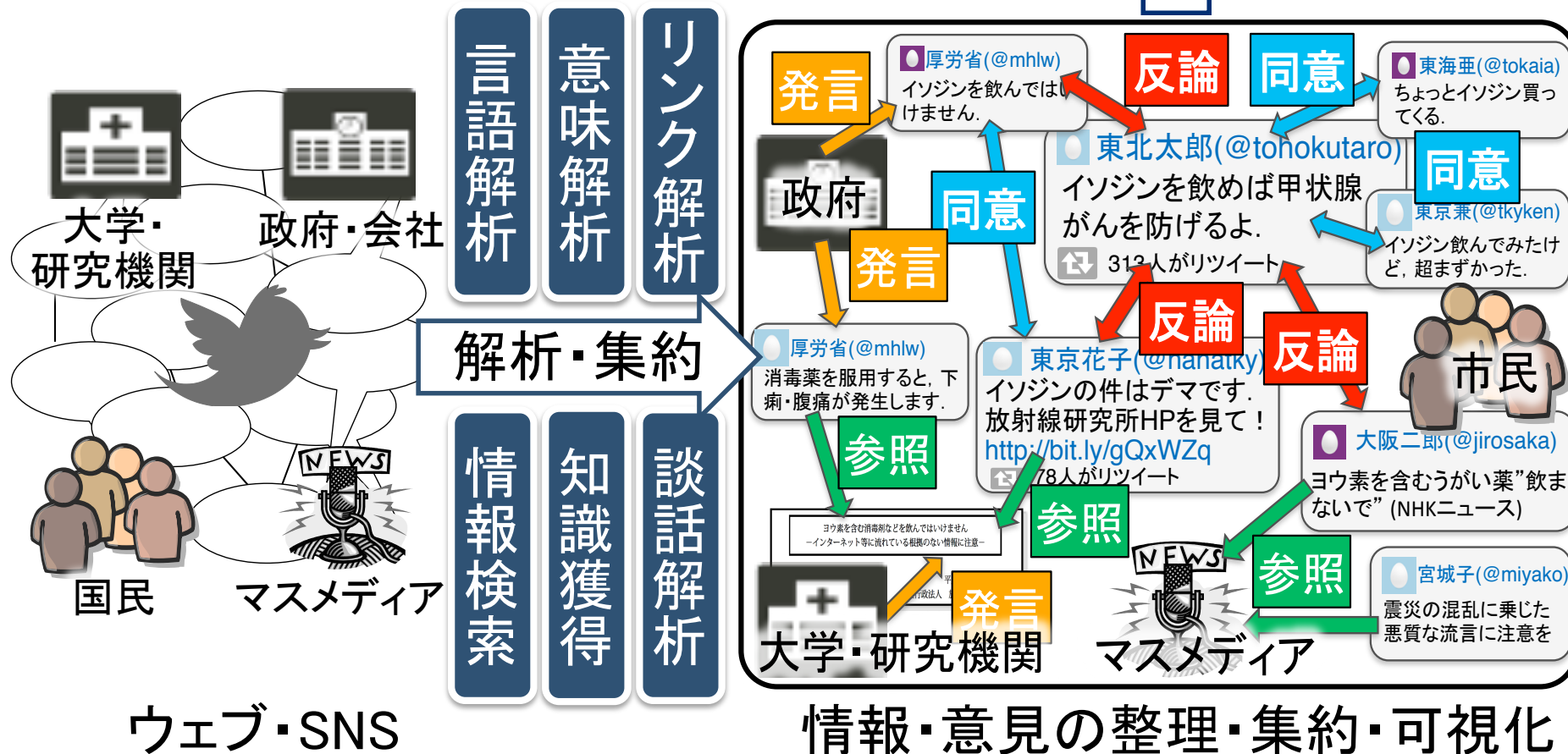
健康・危機管理

デマ検出

行動決定支援

情報リテラシー

アプリケーション

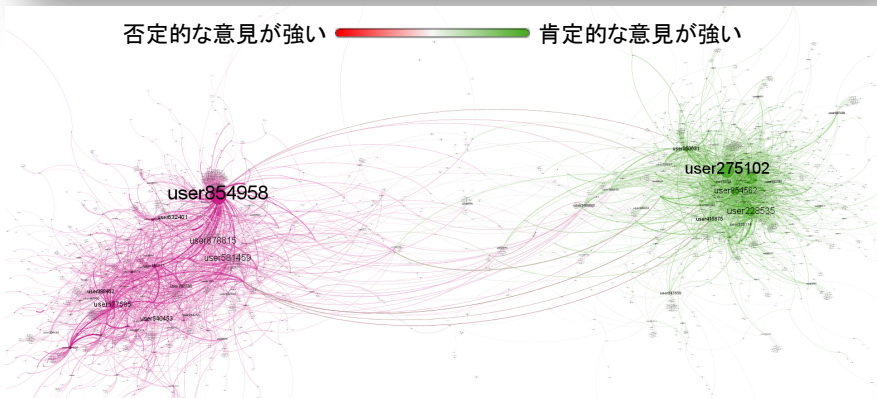


ニーズは社会のどこにもある

震災時のデマ拡散・収束の分析

トピック	スコア	命題
田尻智さん	18.7	海外で、ポケモンの生みの親の田尻智さんが亡くなった
コスモ石油千葉製油所	11.8	@市原市のコスモ石油千葉製油所LPGタンクの爆発により、千葉県、近隣圏に在住の方に有害物質が雨などと一緒に飛散する
女性暴行	10.4	@「阪神大震災の際には女性暴行が増えた」
有毒物質	8.6	コンビニート火災に関し『有毒物質が発生し、雨に混じって降ってくるので肌をさらさないように』
尾田栄一郎先生	5.3	尾田栄一郎先生が15億円を寄付
円資産	5.2	池田経済産業副大臣は、「日本企業が円資産を買い戻している
空中投下	3.0	『物資の空中投下が日本で認められていない』
サーバルーム	1.9	サーバルーム、ラックで腹部を怪我した

コスモ石油: 5時間で訂正, 24時間で収束



参院選の世論分析

風評被害の分析



東北大学 自然言語処理研究室

研究スタッフ



乾健太郎
教授



岡崎直観
准教授



渡邊陽太郎
助教



松林優一郎
特任助教



水野淳太
(NICT研究員)



井之上直也
研究員



C. Kruengkrai
研究員

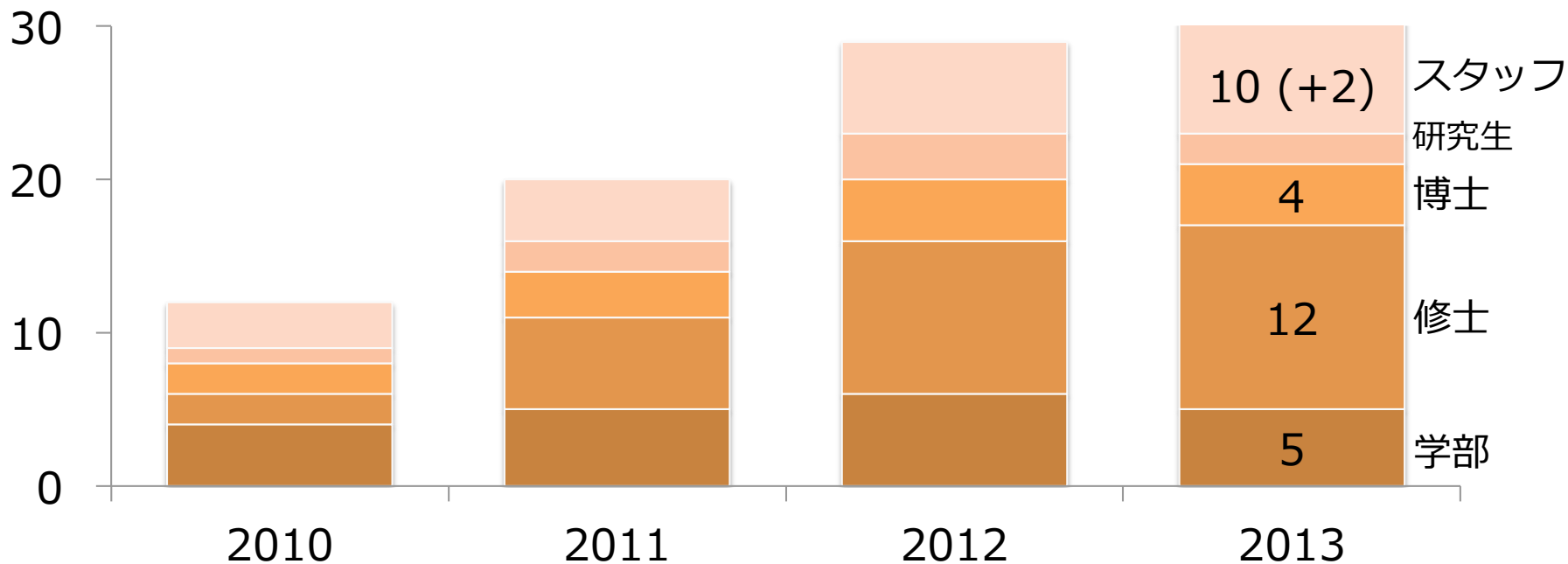


募集中



募集中

研究員
募集中



東北大学 自然言語処理研究室

- 2010年9月 言語処理学会 第16回年次大会 最優秀発表賞
- 2011年2月 CICLing-2011 Best Paper Award (First Place)
- 2011年3月 言語処理学会 2010年度 最優秀論文賞
- 2011年6月 日本経済新聞社 技術トレンド調査 総合3位
- 2011年9月 NLP若手の会 第6回シンポジウム 奨励賞
- 2012年5月 情報処理学会 自然言語処理研究会 学生奨励賞
- 2012年9月 情報処理学会 2012年度 山下記念研究賞
- 2012年9月 言語処理学会 第18回年次大会 優秀賞
- 2012年9月 NLP若手の会 第7回シンポジウム 奨励賞
- 2013年3月 東北大学 情報科学研究科長賞
- 2013年3月 情報処理学会 第75回全国大会 学生奨励賞
- 2013年3月 情報処理学会 第75回全国大会 学生奨励賞
- 2013年3月 情報処理学会 東北支部 学生奨励賞
- 2013年9月 言語処理学会 第19回年次大会 優秀賞
- 2013年11月 日本リスク研究学会 年次大会 優秀発表論文賞

本日の講演内容

我々は今どこにいるか

- ビッグデータ出現のインパクト

知の編集へ

- 情報信憑性判断の支援を例に

その先の言語理解研究へ

- 経験主義と合理主義の間で

自然言語処理

我々は今どこにいるか

ビッグデータ出現のインパクト

「言葉がわかる」とは？

単語を知っている、単語に分けられる

そのひとことで元気になった。

そのひとことで元気になった。

そのひと(人)こと(事)で元気になった。

そのひ(日)とことで元気になった。

そのひ(火)とこ(床)とで元気になった。

そのひとことで元気になっ(担っ)た。

そのひとことで元気になった。

「言葉がわかる」とは？

構文構造(修飾関係)がわかる

(名詞)と(名詞)に(動詞)

彼女 と 仙台 に 行きました

彼女と 仙台に 行きました

I visited Sendai with her.

盛岡 と 仙台 に 行きました

盛岡 と 仙台 に 行きました

I visited Mori. and Sendai.

基本戦略①：教師あり学習

例題と答え(教師データ)

言語 と は 、 コミュニケーション の た め の 記号 の 体系 で ある
名詞 助詞 助詞 記号 名詞 助詞 名詞 助詞 名詞 助詞 名詞 助動詞 助動詞

自然 言語 処理 は 、 人間 が 使っ て いる 言葉 を コンピュータ に
名詞 名詞 名詞 助詞 記号 名詞 助詞 動詞 助動詞 助動詞 名詞 助詞 名詞 助詞

落語 は 、 江戸 時代 の 日本 で 成立 し 、 現在 まで 伝承 さ れ
名詞 助詞 記号 名詞 名詞 助詞 名詞 助詞 サ変 動詞 記号 名詞 助詞 サ変 動詞 助動詞

都市 に 人口 が 集積 す る こ と に よっ て 芸能 と し て 成立 し た
名詞 助詞 名詞 助詞 サ変 動詞 名詞 助詞 動詞 助詞 サ変 助詞 動詞 助詞 サ変 動詞 助詞



統計解析(機械学習)

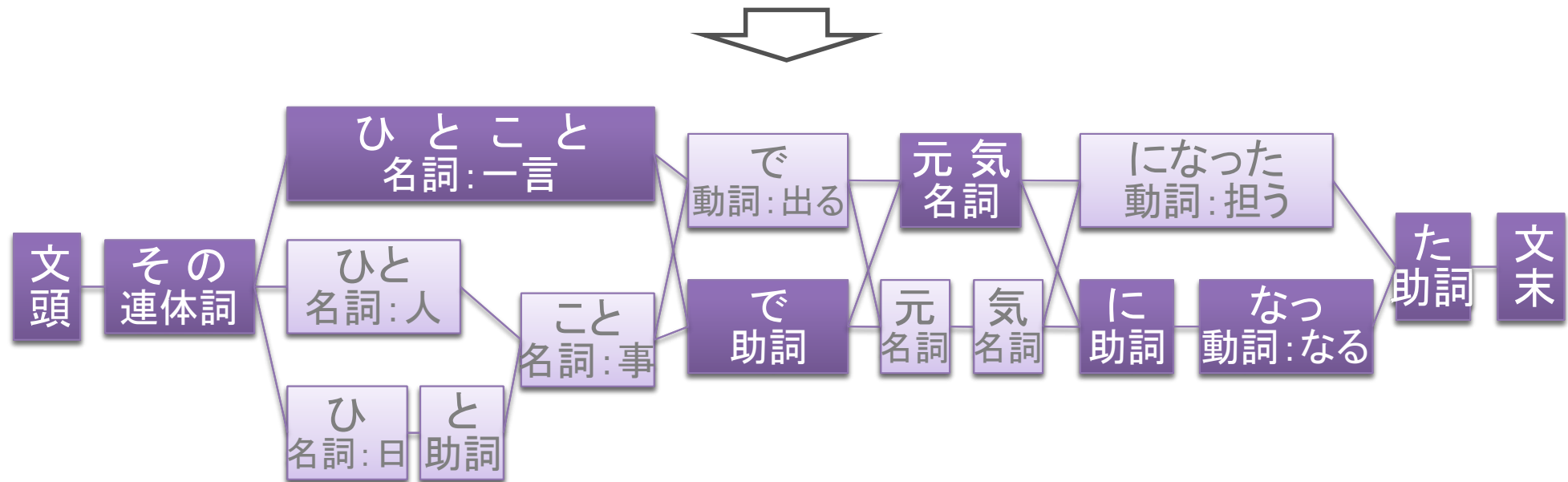
規則性(答えの傾向)

- 名詞の後には助詞が一番続きやすく、つぎに名詞が続きやすい
- 助詞の後には名詞が一番続きやすく、つぎに動詞が続きやすい
- 「が」「を」「に」「で」という文字は助詞になりやすい
- 長い単語が候補にあるなら、短い単語に細切れにするよりも良い

基本戦略①：教師あり学習

規則性(答えの傾向)

- 名詞の後には助詞が一番続きやすく、つぎに名詞が続きやすい
- 助詞の後には名詞が一番続きやすく、つぎに動詞が続きやすい
- 「が」「を」「に」「で」という文字は助詞になりやすい
- 「なる」という動詞は「になう(担う)」という動詞よりも出やすい
- 長い単語が候補にあるなら、短い単語に細切れにするよりも良い



問題！

訓練事例をいくら作っても足りない

単語の組み合わせを考える必要のあるタスク

(名詞) と (名詞) に (動詞)

10万種類

10万種類

1万種類

彼女 と 仙台 に 行きました

盛岡 と 仙台 に 行きました

彼女と 仙台に 行きました

盛岡と 仙台に 行きました

I visited Sendai with her.

I visited Mori. and Sendai.

基本戦略②

大量の生データから得られる特徴を利用

彼女と 仙台に 行きました



「彼女と行く」はよく言う(560万回)

盛岡と 仙台に 行きました



「盛岡と行く」は言わない(23回)

彼女 と 仙台 に 行きました



「彼女や仙台」は言わない(4回)

盛岡 と 仙台 に 行きました



「盛岡や仙台」はよく言う(20万回)

実際の解析例

米航空宇宙局は太陽系外に存在する観測史上最小の惑星を発見した。

* 0 5D 3/4 3.287031

米	名詞,固有名詞,地域,国,*,*,米,ベイ,ベイ,,	B-ORGANIZATION
航空	名詞,一般,*,*,*,航空,コウクウ,コークー,,	I-ORGANIZATION
宇宙	名詞,一般,*,*,*,宇宙,ウチュウ,ウチュー,,	I-ORGANIZATION
局	名詞,接尾,一般,*,*,*,局,キョク,キョク,,	I-ORGANIZATION

は 助詞,係助詞,*,*,*,は,ハ,ワ,, 0

* 1 2D 1/2 1.773937

太陽系	名詞,一般,*,*,*,太陽系,タイヨウケイ,タイヨーケイ,,	0
外	名詞,接尾,一般,*,*,*,外,ガイ,ガイ,,	0
に	助詞,格助詞,一般,*,*,*,に,ニ,ニ,,	0

* 2 3D 1/1 0.000000

存在	名詞,サ変接続,*,*,*,存在,ソンザイ,ソンザイ,,	0
する	動詞,自立,*,*,サ変・スル,基本形,する,スル,スル,,	0

* 3 4D 2/3 0.000000

観測	名詞,サ変接続,*,*,*,観測,カンソク,カンソク,,	0
史上	名詞,一般,*,*,*,史上,シジョウ,シジョー,,	0
最小	名詞,一般,*,*,*,最小,サイショウ,サイショー,,	0
の	助詞,連体化,*,*,*,の,ノ,ノ,,	0

* 4 5D 0/1 0.000000

惑星	名詞,一般,*,*,*,惑星,ワクセイ,ワクセイ,,	0
を	助詞,格助詞,一般,*,*,*,を,ヲ,ヲ,,	0

* 5 -1D 1/2 0.000000

発見	名詞,サ変接続,*,*,*,発見,ハツケン,ハツケン,,	0
した	動詞,自立,*,*,サ変・スル,連用形,する,シ,シ,,	0
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ,,	0
。	記号,句点,*,*,*,。,,。,,。,,	0

EOS

固有表現抽出

項

ガ

項

ニ

述語

(存在する)

述語-項
構造解析

ガ

項

ヲ

述語

(発見する)

係り受け
解析



「行間を読む」

週末に母とやなぎに行ってみました。

うどん屋の名前

うどんを食べるために

私は評判のきつねうどんを頼んだの

私はうどんを食べた

やなぎで

ですが、期待通りの味に大満足。

うどんが美味しいと

きつねうどんの

「行間を読む」 常識に基づく推論

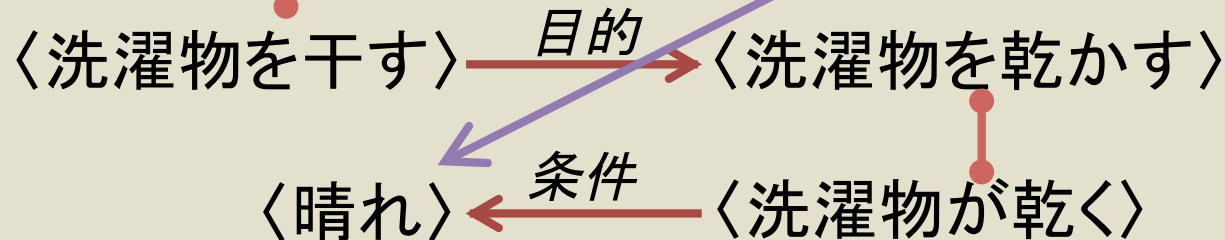


庭に洗濯物を干したとたんに雨が降ってきた

庭に洗濯物を干す

雨が降る

ガッカリ



常識的な
知識

歴史

1990

2000

2010

1980

応用

意味処理

基盤技術

深い言語理解

合理主義

プラン/意図理解

知識に基づく推論

精緻な文法

大問題！

膨大な知識をどうやって集めるか

言葉の理解に必要な知識

- 「洗濯物を干す」目的は「洗濯物を乾かす」こと
- 「洗濯物が乾く」条件は「晴れている」こと
- 「うどん」は、「レストラン(うどん屋)で食べる」
- 「(料理を)注文する」ことを「頼む」とも言う
- 「レストラン」では、「料理を注文」して、しばらくすると料理が「出て」きて、それを「食べる」

歴史

1990

2000

2010

1980

応用

意味処理

基盤技術

深い言語理解

合理主義

プラン/意図理解

知識に基づく推論

精緻な文法

知識獲得の
ボトルネック

経験主義

省略の解析

述語項構造解析

固有表現抽出

統計的形態素/構文解析

彼氏と仙台に行きました

彼氏と 仙台に 行きました

彼氏と仙台に行きました。



彼氏	と	仙台	に	行	ま	し	た
名詞	助詞	名詞	助詞	動詞	助動詞	助動詞	助動詞

実際の解析例

米航空宇宙局は太陽系外に存在する観測史上最小の惑星を発見した。

* 0 5D 3/4 3.287031

米	名詞,固有名称,地域,国,*,*,米,ベイ,ベイ,,	B-ORGANIZATION
航空	名詞,一般,*,*,*,航空,コウクウ,コークー,,	I-ORGANIZATION
宇宙	名詞,一般,*,*,*,宇宙,ウチュウ,ウチュー,,	I-ORGANIZATION
局	名詞,接尾,一般,*,*,*,局,キョク,キョク,,	I-ORGANIZATION

固有表現抽出

は 助詞,係助詞,*,*,*,は,ハ,ワ,, 0

* 1 2D 1/2 1.773937

太陽系	名詞,一般,*,*,*,太陽系,タイヨウケイ,タイヨーケイ,,	0
外	名詞,接尾,一般,*,*,*,外,ガイ,ガイ,,	0
に	助詞,格助詞,一般,*,*,*,に,ニ,ニ,,	0

項

ガ

項

ニ

述語

(存在する)

述語-項
構造解析

ガ

項

ヲ

述語

(発見する)

* 2 3D 1/1 0.000000

存在	名詞,サ変接続,*,*,*,存在,ソンザイ,ソンザイ,,	0
する	動詞,自立,*,*,サ変・スル,基本形,する,スル,スル,,	0

* 3 4D 2/3 0.000000

観測	名詞,サ変接続,*,*,*,観測,カンソク,カンソク,,	0
史上	名詞,一般,*,*,*,史上,シジョウ,シジョー,,	0
最小	名詞,一般,*,*,*,最小,サイショウ,サイショー,,	0
の	助詞,連体化,*,*,*,の,ノ,ノ,,	0

* 4 5D 0/1 0.000000

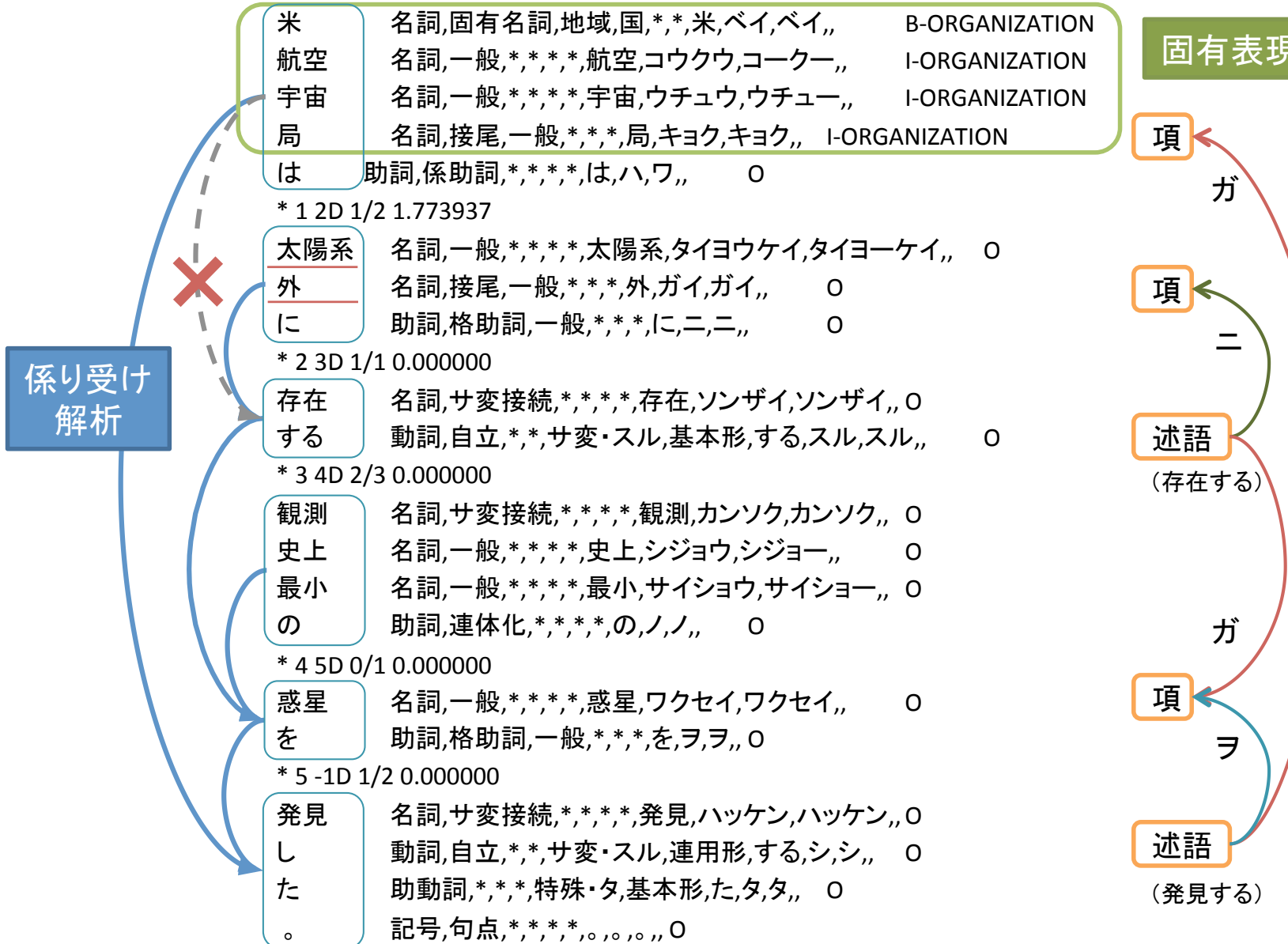
惑星	名詞,一般,*,*,*,惑星,ワクセイ,ワクセイ,,	0
を	助詞,格助詞,一般,*,*,*,を,ヲ,ヲ,,	0

* 5 -1D 1/2 0.000000

発見	名詞,サ変接続,*,*,*,発見,ハツケン,ハツケン,,	0
した	動詞,自立,*,*,サ変・スル,連用形,する,シ,シ,,	0
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ,,	0
。	記号,句点,*,*,*,。,,。,,。,,	0

EOS

係り受け
解析



歴史

1990

2000

2010

1980

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

深い言語理解

合理主義

知識獲得の
ボトルネック

意味処理

プラン/意図理解

知識に基づく推論

省略の解析

精緻な文法

経験主義

述語項構造解析

基盤技術

固有表現抽出

統計的形態素/構文解析

彼氏と仙台に行きました

彼氏と 仙台に 行きました

彼氏と仙台に行きました。



彼氏	と	仙台	に	行	ま	し	た
名詞	助詞	名詞	助詞	動詞	助動詞	助動詞	助動詞

歴史

1990

2000

2010

ビッグデータ

応用

1980

統計的機械翻訳

情報検索 情報抽出・質問応答

知識の獲得

経験主義

深い言語理解

合理主義

知識獲得の
ボトルネック

大規模な
知識獲得

意味処理

プラン/意図理解

知識に基づく推論

精緻な文法

経験主義

述語項構

基盤技術

固有表現抽出

統計的形態素/構文解析

↑
今ここ

単語の意味の知識を自動的に集める



The screenshot shows the Wikipedia page for '国際捕鯨委員会' (International Whaling Commission). The page title is '国際捕鯨委員会' and the subtitle is '提供: フリー百科事典『ウィキペディア (Wikipedia)』'. The main text reads: 'ウィキメディア・カンファレンス・ジャパン 2009で財団・Jay Walsh氏、国会図書館長・長尾真氏が講演。参加・発表・当日スタッフも募集中! 詳しくは公式サイトへ。' There are navigation tabs for '本文', 'ノート', '編集', and '履歴'. The top right corner has links for 'ベータ版を試す' and 'ログインまたはアカウント作成'. The left sidebar shows the Wikipedia logo and the text 'ウィキペディア フリー百科事典'.

国際捕鯨委員会（こくさいほげいいいんかい、International Whaling Commission: **IWC**）は、**国際捕鯨取締条約**に基づき設立された**国際機関**。日本の条約加入は**1951年**^[1]。

- 最近の更新
- おまかせ表示

1 組織・総会

同義語の抽出: 113,401グループ

HIDランプ-高圧放電灯-高輝度放電ランプ

バイオ燃料-エコ燃料-バイオエネルギー

上位下位関係の抽出: 約350万対

イソフラボン-女性ホルモン様成分

キシリトール-甘味添加物

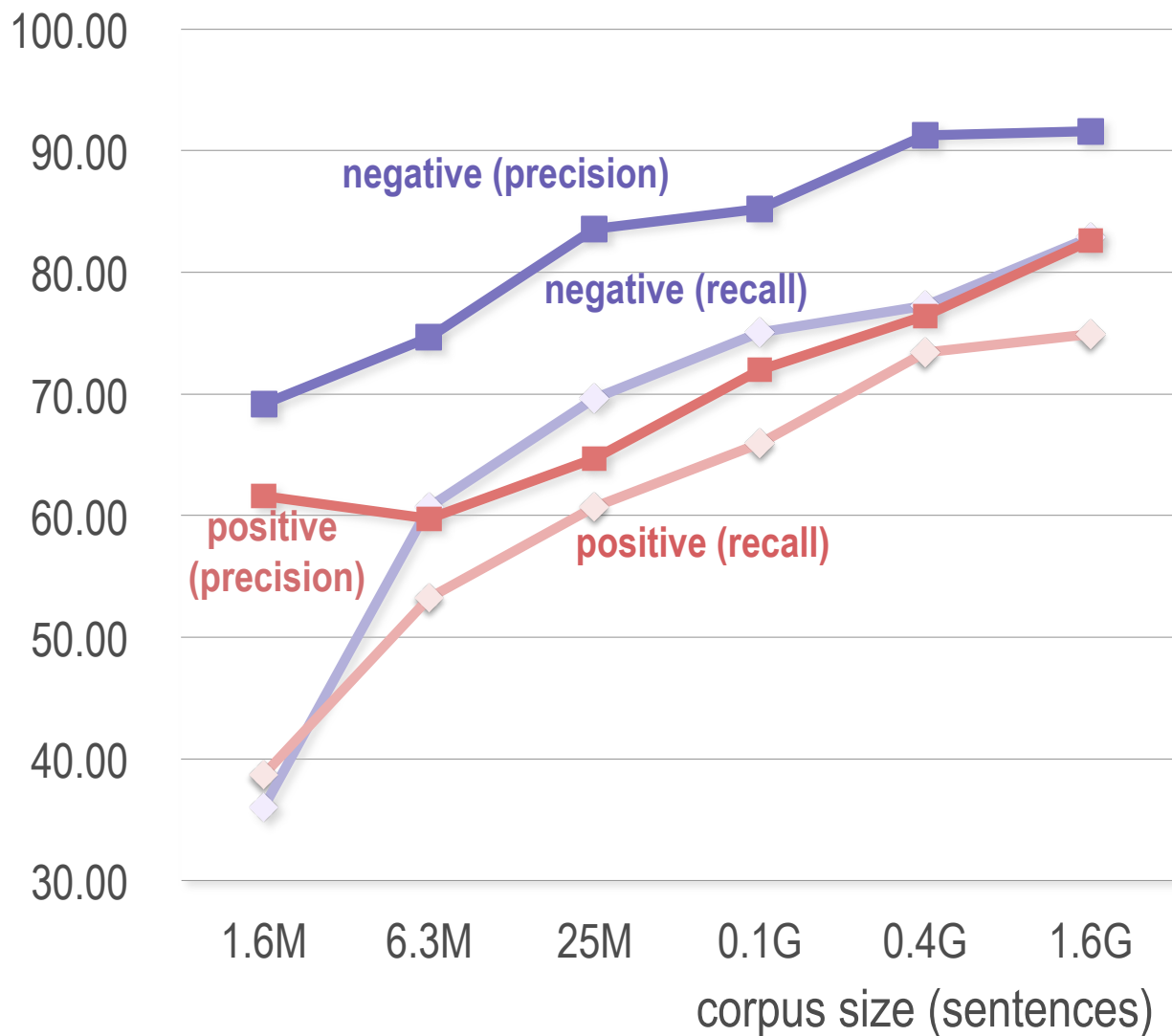
単語の意味の知識を自動的に集める

単語の使われ方から
その単語の意味が推測できる

	を着る	を着せる	に着替える	を着こなす	に包まれる	を買う	を売る	を販売する	を作る	で登場する	を食べる	に行く	使われ方 (共起する動詞)
着物	260	36	7	5	6	24	15	9	18	0	0	0	出現頻度
和服	97	0	11	6	8	2	2	4	0	4	0	0	
着物	20.1	3.7	3.0	2.6	1.7	1.5	1.3	1.2	1.1	.09	.00	.00	割合(%)
和服	28.0	5.0	4.1	3.6	2.2	2.0	1.7	1.6	1.5	1.1	.00	.00	

大規模データの効果

(乾ら@NAIST)



多様なポジ/ネガ表現

(乾ら@NAIST)

ポジティブ

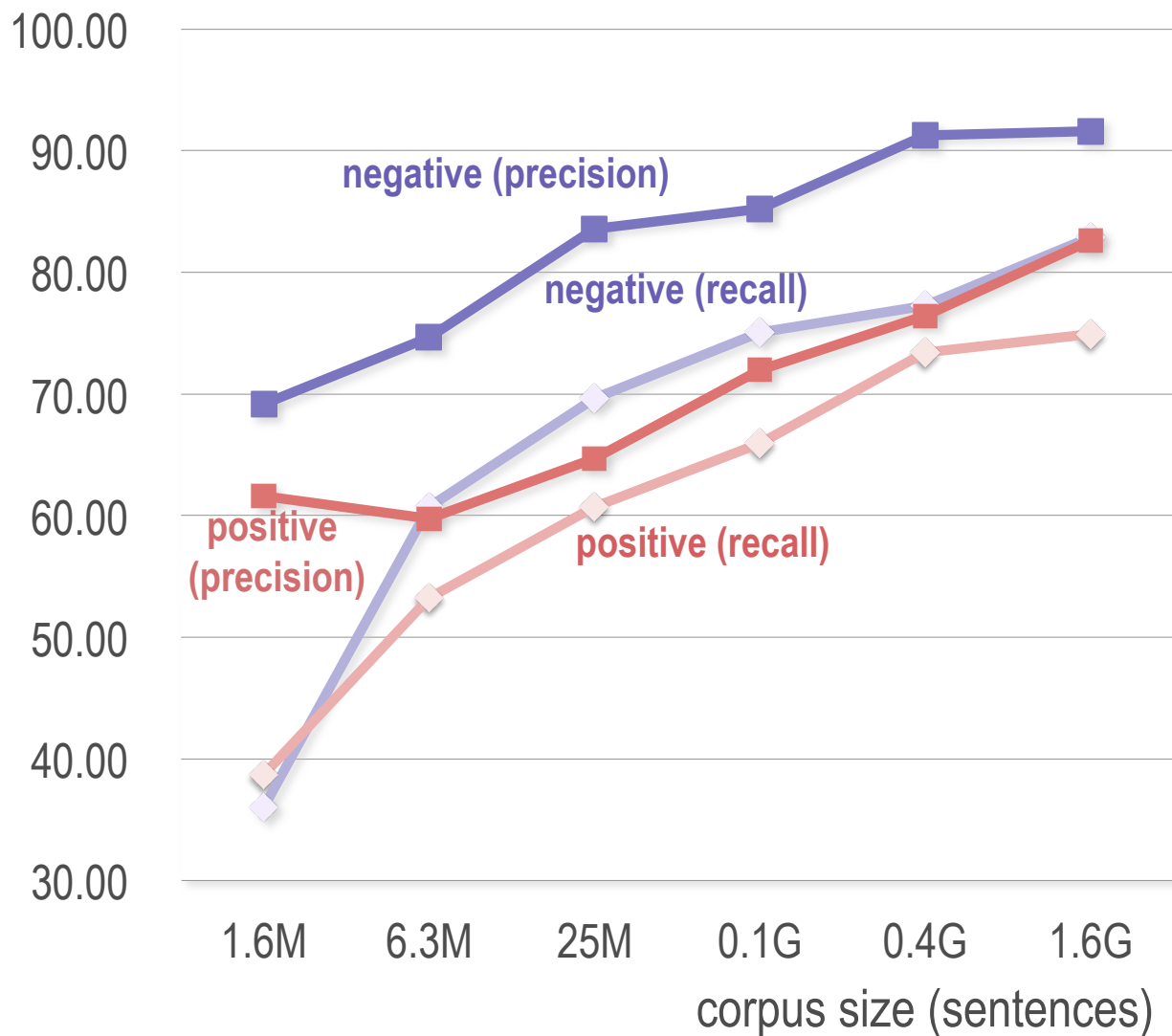
説得力, チームワーク,
モチベーション, 積極性,
開放感, チャレンジ精神,
サービス, 自主性,
達成感, フレンドリー,
レベルアップ, ヒット曲,
スピード感, ビンゴ,
向上心, リゾート気分,
景気回復, 優遇措置,
ポップ, アットホーム

ネガティブ

ガン, リバウンド,
いじめ, ストーカー,
リストラ, 不正アクセス,
文字化け, DV,
シワ, 肌荒れ,
エイズ, スパムメール,
ニキビ, 生活習慣病,
クーデター, メモリ不足,
夏バテ, オイルショック,
生ゴミ, ドタキャン

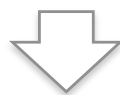
大規模データの効果

(乾ら@NAIST)



常識もビッグデータから集める

天気が良いので、洗濯物がよく乾く
最近、晴れが続いているので、洗濯物がよく乾く
日が当たらないので、洗濯物が乾くのもままならず
お天気は晴れマークが多かったから、洗濯物が乾く
湿度はなかったから、洗濯物が乾く
今日は、空気が乾燥していて、洗濯物がよく乾く
梅雨空が続いたため、洗濯ものが乾かず



天気が良い(晴れる) → 洗濯物がよく乾く

歴史

1990

2000

2010

1980

応用

意味処理

基盤技術

深い言語理解

合理主義

プラン/意図理解

知識に基づく推論

精緻な文法

情報検索

統計的機械翻訳

情報抽出・質問応答

経験主義

知識獲得の
ボトルネック

経験主義

固有表現抽出

統計的形態素/構文解析

ビッグデータ

知識の獲得

知の編集

推論

大規模な
知識獲得

頑健・高度な
言語処理

↑
今ここ

本日の講演内容

我々は今どこにいるか

- ビッグデータ出現のインパクト

知の編集へ

- 情報信憑性判断の支援を例に

その先の言語理解研究へ

- 経験主義と合理主義の間で

知の編集へ

情報信憑性判断の支援を例に



震災直後に流れたデマ



@jusminrom

ジャスミン・ロメロ@高3

【被爆予防方法】 イソジンを水に薄めて大量に飲んで下さい。特に子供に。自然ヨウ素で甲状腺を満たし、放射線ヨウ素の吸収を抑えます。 #jishin

3月13日 Keitai Webから

デマを否定する情報も



サニー

@ZxCG



フォローする

放射能対策でヨウ素剤の代わりにイソジン3滴という全くのデマが流れているので止めてください。イソジンだと14.3cc飲む必要が有りますが、こんなに飲んだら胃腸がただれま
す。 <http://bit.ly/h55YA1>

支援要請の過剰拡散も



13番

@13_ban



フォローする

《緊急：拡散希望》 当院ではICUのみ自家発電中です。燃料のA重油も朝7時でなくなり、人工呼吸器も止まっています。A重油の入手先をご教示ください。

0222482131 (仙台市太白区
広南病院用度課電気室)

#Sendai #earthquake

#eqjp #jishin

← 返信 ↻ リツイート ★ お気に入りに登録

273

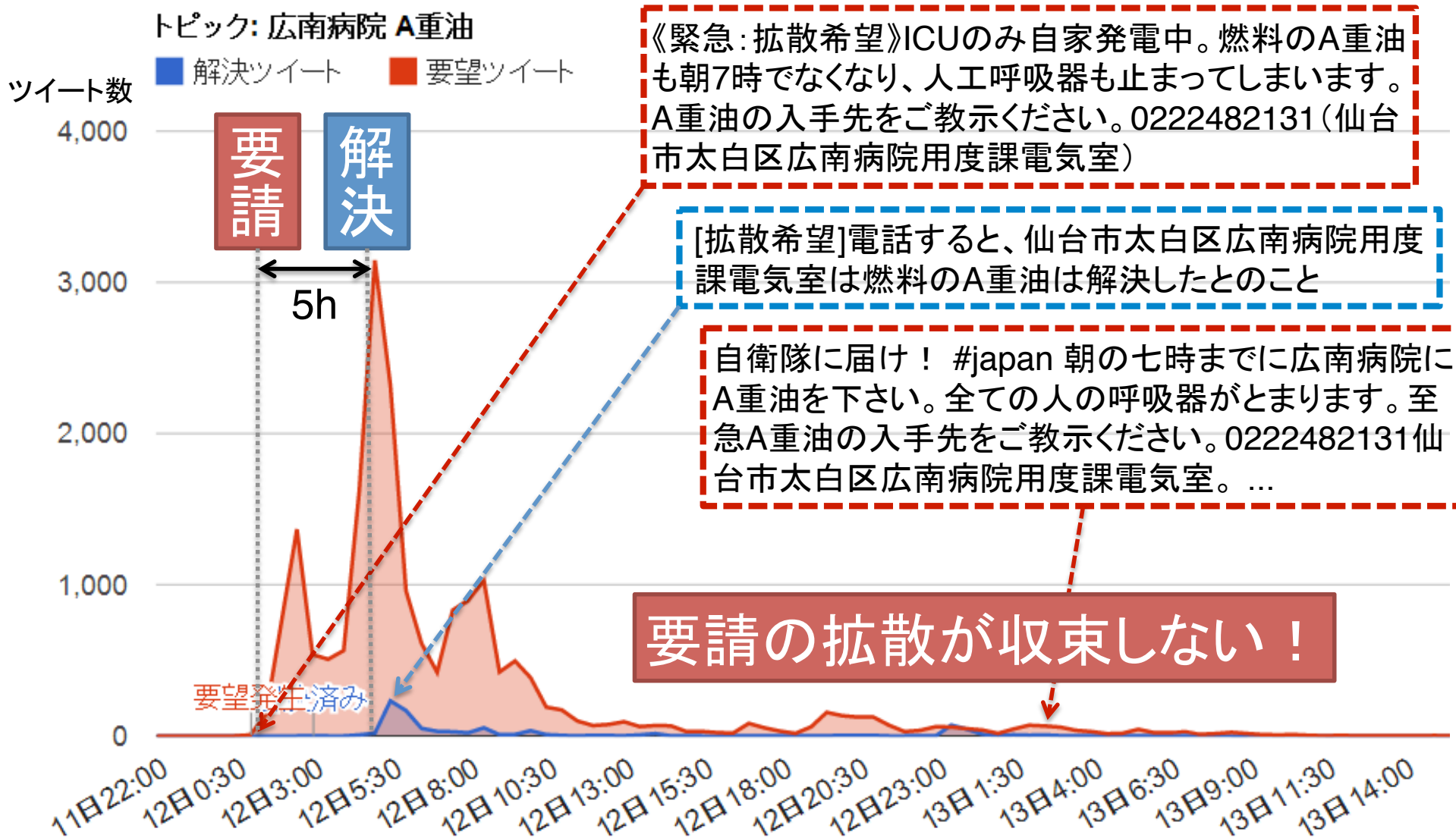
件のリツイート

2

FAVORITES

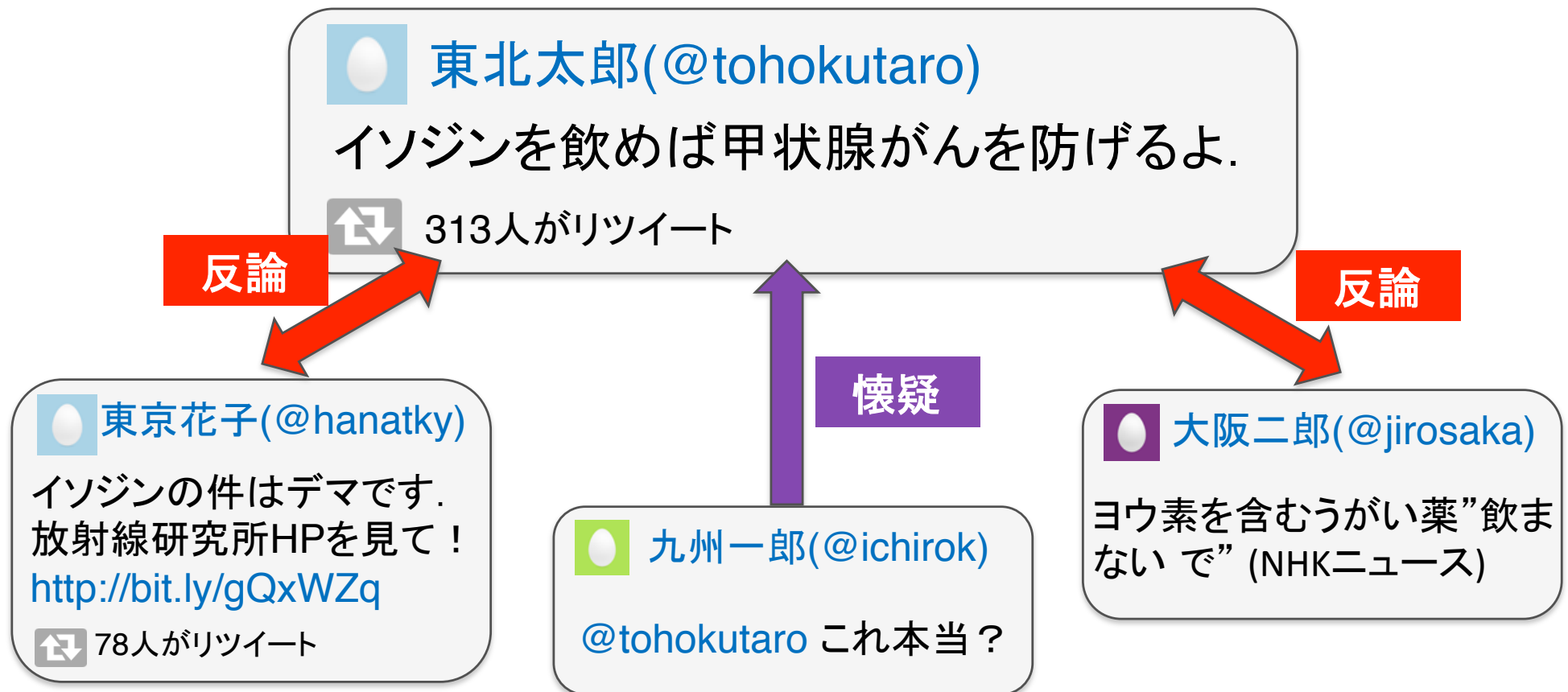


支援要請の過剰拡散も



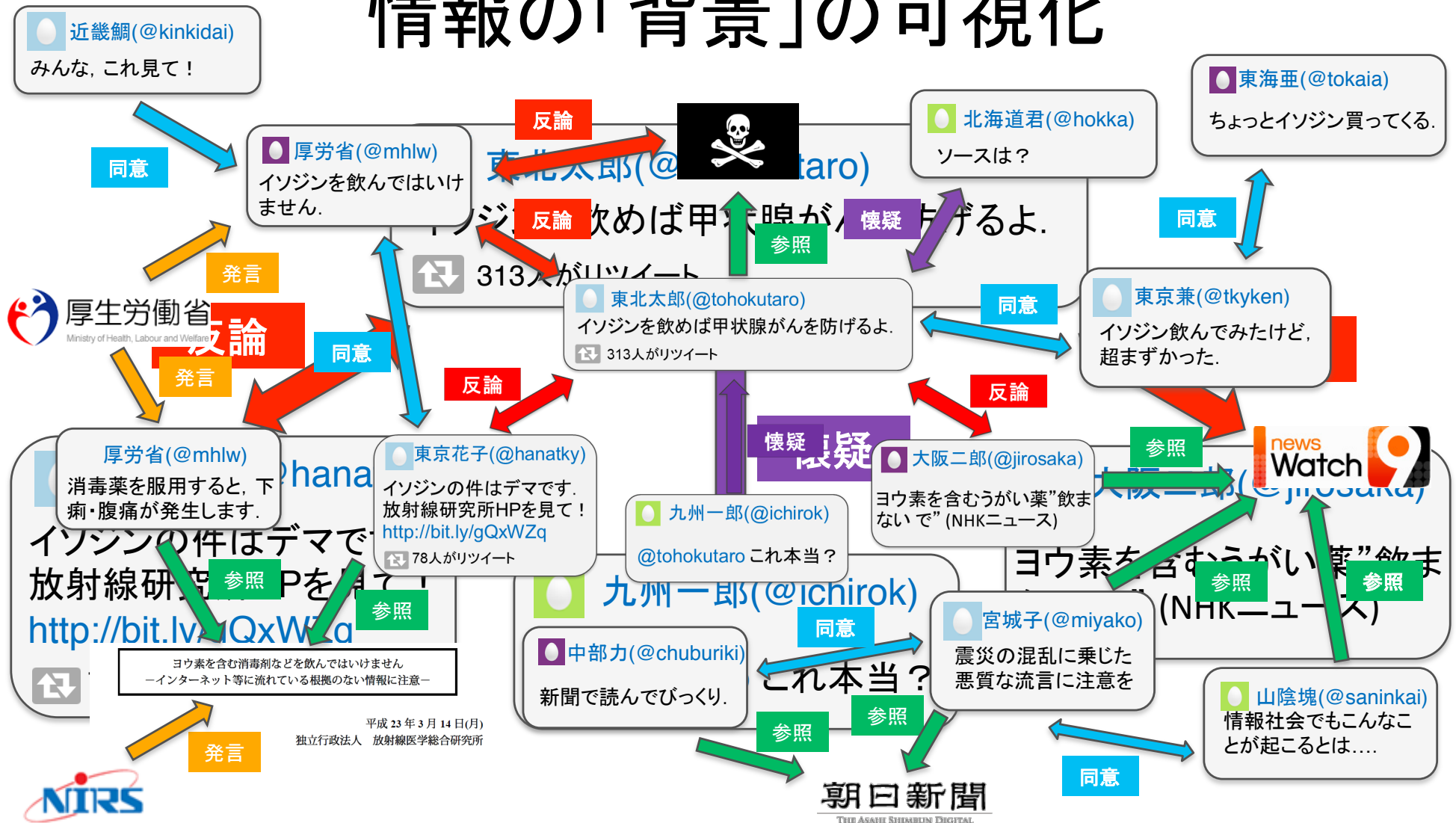
情報の信頼性をどう担保するか

情報の「背景」の可視化



情報の信頼性をどう担保するか

情報の「背景」の可視化



言論マップ



質問

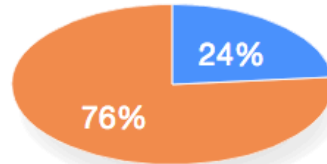
イソジンで被曝を防げる?

肯定

否定

39

★根拠 21



12

★根拠 1

肯定

被曝を軽減するにはイソジンがいい

否定

イソジンは被曝対策には使えません

平時の一般のトピックでも

質問：

フェミダンプラスはコラーゲンとヒアルロン酸の働きで、
同意 (459文) 噛むだけでお肌つやつや！若返りと美肌を提供します log1 log2

ただ、研究職の私から言わせてもらうと、コラーゲンは体に入っただけからといって、肌を潤いをもたらすわけではありません log1 log2

• コラーゲンは健康な肌ならきれいな網目構造になっており、まるでベッドのスプリングのように表皮を支えて弾力のある肌を保っています log1 log2

• お肌にハリがなくなってきたり、たるんでくるのかな、と思うときにコラーゲンやヒアルロン酸を摂取するのは効果的 log1 log2

フェミダンプラスはコラーゲンとヒアルロン酸の働きで、噛むだけでお肌つやつや！若返りと美肌を提供します log1 log2

• ただしコラーゲンは肌以外にも骨や血管・内臓にも使われるので不規則な生活で体内がぼろぼろだと先に体内の改善に使われるので不規則な生活だといくら一生懸命に摂取しても効果が期待できません log1 log2

ニッピコラーゲンといったコラーゲンですが、肌から浸透しないので、出来てしまったしわに対しての効果は期待できません log1 log2

• 美肌づくりに欠かせないコラーゲンですが、実はコラーゲンは分子が大きいので、お肌に入りにくい！という難点があります log1 log2

根拠 (10文)

完了

同意・対立・根拠の認識

事実性(モダリティ)解析 + 構造的アライメント + 関係パターン認識

Collagen has beautiful skin effects.
コラーゲンには美肌効果がある

+ 存在
+ existence

- 存在
- existence

コラーゲンをサプリメントで飲んでも、
Even when taken as a supplement,

途中でアミノ酸に分解されてしまうので、
(collagen) ends up broken down into amino acids so,

(Φが)肌をきれいにする効果は期待できません。
we cannot expect (collagen) to be effective at making skin pretty.

事実性の推定

そのイベントは起こったのか起こっていないのか

流出しなかった（流出しなかった）

流出したではないか（流出した）

流出しただけでない（流出した）

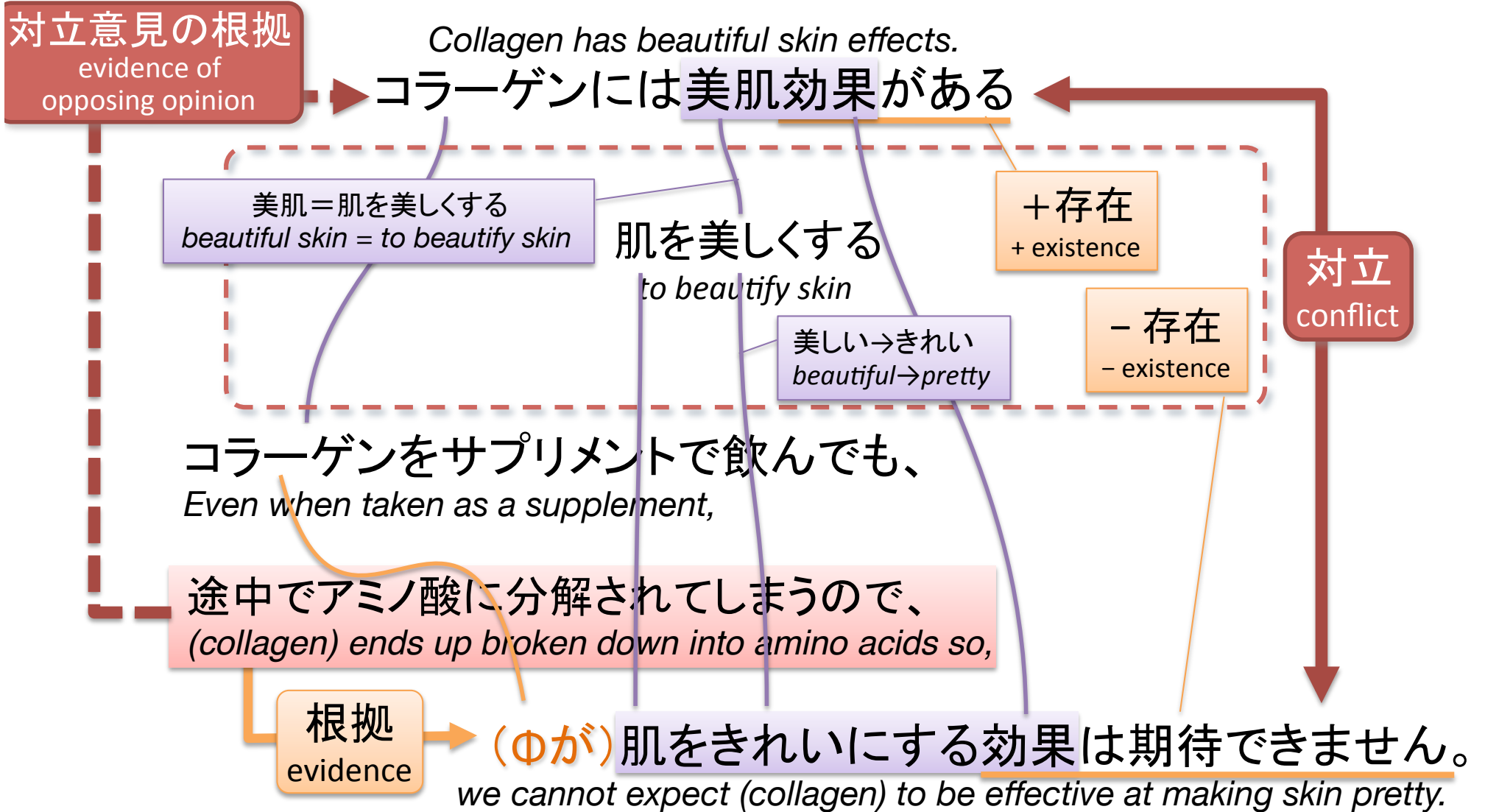
流出を防いだ（流出しなかった）

流出を抑えた（流出した，でも少し）

流出を止めたい（流出した，今も流出）

同意・対立・根拠の認識

事実性(モダリティ)解析 + 構造的アライメント + 関係パターン認識



言論マップで利用している語彙知識

- 日本語WordNet (Bond+ 2009) : 56,741 synsets, 92,241 語
 - 過剰: 冗多, 過多, 冗長, 余った, 余計, 余分
 - 効果: 効目, 実効, 効, 利き目, 効き目, 利き, 効力, 効き, 機能, 効能
 - 病気 ⇔ 具合い, 丈夫, 健康, フィットネス, 健全, 多幸
- 事象間関係知識 (Matsuyoshi+ 2008) : 約140万対
 - <誰か>が<何か>を摂取する – <誰か>が<何か>を取る (上位語)
 - <何か>が普通だ – <何か>が異常だ (反義)
- Wikipediaから抽出した同義語: 113,401 synsets
 - HIDランプ – 高圧放電灯 – 高輝度放電ランプ
 - バイオ燃料 – エコ燃料 – バイオエネルギー
- Wikipediaから抽出した上位下位関係 (Sumida+ 2008) : 約350万対
 - イソフラボン – 女性ホルモン様成分
 - キシリトール – 甘味添加物
- 文脈類似度に基づく動詞間含意関係 (ALAGIN) : 約170万対
 - 引き起こす – もたらす, 予防する – 防げる, ケアする – 防止する

うがいはインフルエンザの予防に効果的だ

賛成意見の根拠

(33文)

反対意見の根拠

(6文)

あと、現在の豚インフルエンザの特徴は、日本でも冬に流行する、インフルエンザA型に似ていて、湿気・熱に弱いので、マスク・手洗い・うがいでかなりの予防はできるとのこと log1 log2

ウイルスは、粘膜に接着後、
わめて短時間で粘膜内に進入
効果は低いと考えられてい
防にはなりません log1 log2
耳より健康教室
でもほぼ落ち着き始めたので
B型の流行もまだ懸念される

入るということですから、
やうがいも有効でも、イン
ど効果はないともいわれて
漢方110腰痛・関節痛・

インフルエンザは、咳やく
ぶきによってヒトからヒトに
の着用、うがい・手洗いは
が、それにも増して、イン
接種が大切です log1 log2
健康長寿ネットーインフル

またインフルエンザは、イ
た人の咳やくしゃみに含まれ
ことによって感染するので、

よくうがい励行！と言われますが、インフルエンザウイルスは咽頭（のど）、眼、また鼻粘膜に付着すると約20分で細胞中に取り込まれてしまうので、外出後のうがいは予防には無効であり、外出時にマスクをすることをお勧めします log1 log2

インフルエンザウイルスは、粘膜に接着後、20～30分というきわめて短時間で粘膜内に進入しますので、うがいの効果は低いと考えられていますが、ある程度の予防にはなりません log1 log2

賛否によるネット情報の自動分析は世界初

言語処理学会 最優秀発表賞(2010)

言語処理学会 若手奨励賞(2010)

CICLing Best Paper Award(2011)

日経新聞技術トレンド調査 総合3位(2011)

情報処理学会 山下記念研究賞(2012)

2011年(平成23年)5月30日(月曜日)

目的鮮明な産学連携 評価

2011年度技術トレンド調査

技術トレンド調査総合順位1~10位

順位	研究開発者 (発表日)	成果の概要	評価点
1	東京大学、コニカミノルタエムシーナド (2)	軟骨映すX線装置開発、	11.14
2	日立製作所 (2010年12)	NEC、東北大学、奈良先端科学技術大学院大学、横浜国立大学 (1月31日)	ネットの情報、賛否を分析・表示、信頼性判断に有用
3	NEC、東 奈良先端科学 学院大学、	雨大 水産総合研空	天然ウナギの卵発見、世

技術振興機
教授の成
装置は、
1位の

X
始

「技術トレン
日本経済新聞社は、主な技術調

誤情報の自動収集

よく反論されている命題をマイニング

市原市のコスモ石油千葉製油所LPGタンクの爆発により、千葉県、近隣圏に在住の方に有害物質が雨などと一緒に飛散するという虚偽のチェーンメールが送られています。

【リツイートで拡散希望】 コスモ石油の爆発で有害物質の雨が降る件はデマ。広げてしまった方はツイート削除の上、訂正を

コスモ石油の爆発による有害物質流出のおそれの件、コスモ石油から正式にアナウンス出てます。ツイッターで出回ってるような事はないそうです。HPは繋がりにくいのでスクリーンショットです。

誤情報の自動収集

よく反論されている命題をマイニング

トピック	スコア	命題
田尻智さん	18.7	海外で、ポケモンの生みの親の田尻智さんが亡くなった
コスモ石油千葉製油所	11.8	@市原市のコスモ石油千葉製油所LPGタンクの爆発により、千葉県、近隣圏に在住の方に有害物質が雨などと一緒に飛散する
女性暴行	10.4	@「阪神大震災の際には女性暴行が増えた」
有毒物質	8.6	コンビナート火災に関し『有毒物質が発生し、雨に混じって降ってくるので肌をさらさないように』
尾田栄一郎先生	5.3	尾田栄一郎先生が15億円を寄付
円資産	5.2	池田経済産業副大臣は、「日本企業が円資産を買い戻している
空中投下	3.0	『物資の空中投下が日本で認められていない』
サーバルーム	1.9	サーバルーム、ラックで腹部を怪我した
イソジン	1.6	被爆予防にイソジンを飲むと良い
茨城県知事	1.2	茨城県知事災害派遣要請出していない
食塩	1.0	「食塩にはヨウ素が含まれており被曝の防止に効果がある」

誤情報の「一生」を可視化

誤情報の拡散と訂正の拡散を時系列分析

トピック「コスモ石油」

■ デマ訂正ツイート ■ デマツイート

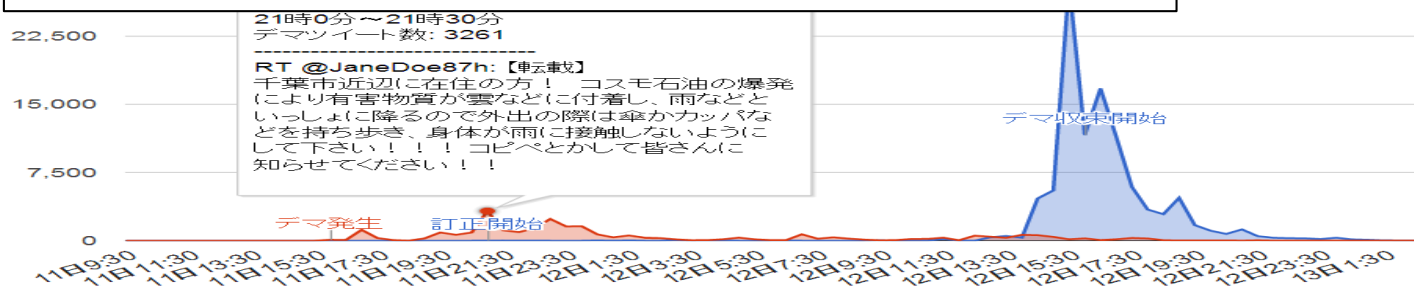


以下のボタンをクリックすると拡大・縮小します

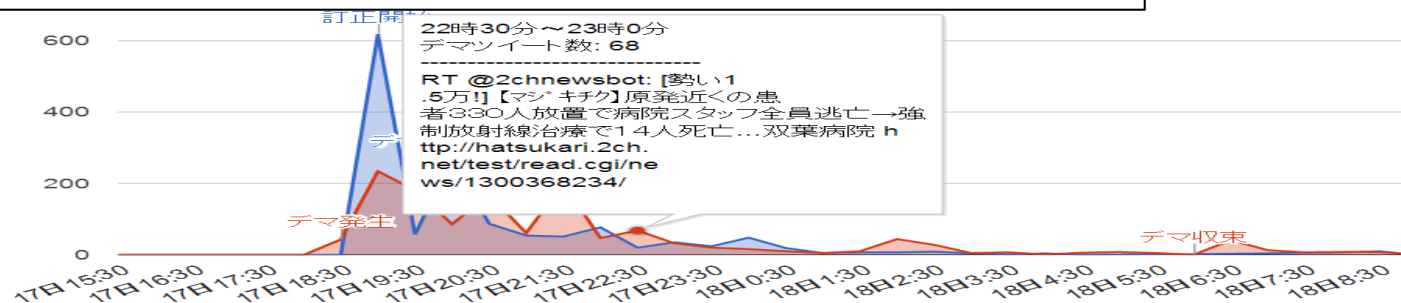
[11日9時-13日3時](#) [13日3時-14日21時](#) [14日21時-16日15時](#) [16日15時-18日9時](#) [全体に戻す](#)

様々なパターンがあることも明らかに

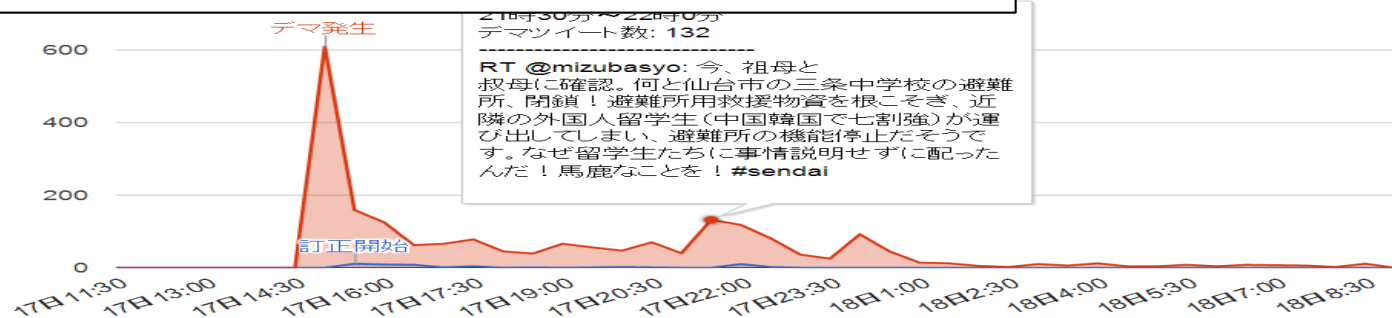
コスモ石油: 5時間で訂正, 24時間で収束



双葉病院: 30分で訂正, 11.5時間で収束



三条中学校: 30分で訂正, 収束せず



返信・引用による論述関係認識

@*** 925RTs 2011-03-11 23:26:35
 コスモ石油の爆発により有害物質が雲などに付着し、雨などと一緒に降るので外出の際は傘かカッパなどを持ち歩き、身体が雨に接触しないようにしてください!!! コピペとかして皆さんに知らせてください。

@▽▽▽ 442RTs 2011-03-11 19:58:50
 フジテレビで言ってましたな RT@***:【拡散希望】千葉市近辺に在住の方! コスモ石油の爆発により有害物質が...雨に接触しないようにしてください!!!

検出された
 <反論>関係 ——
——
 <類似>する
 ツイート ——

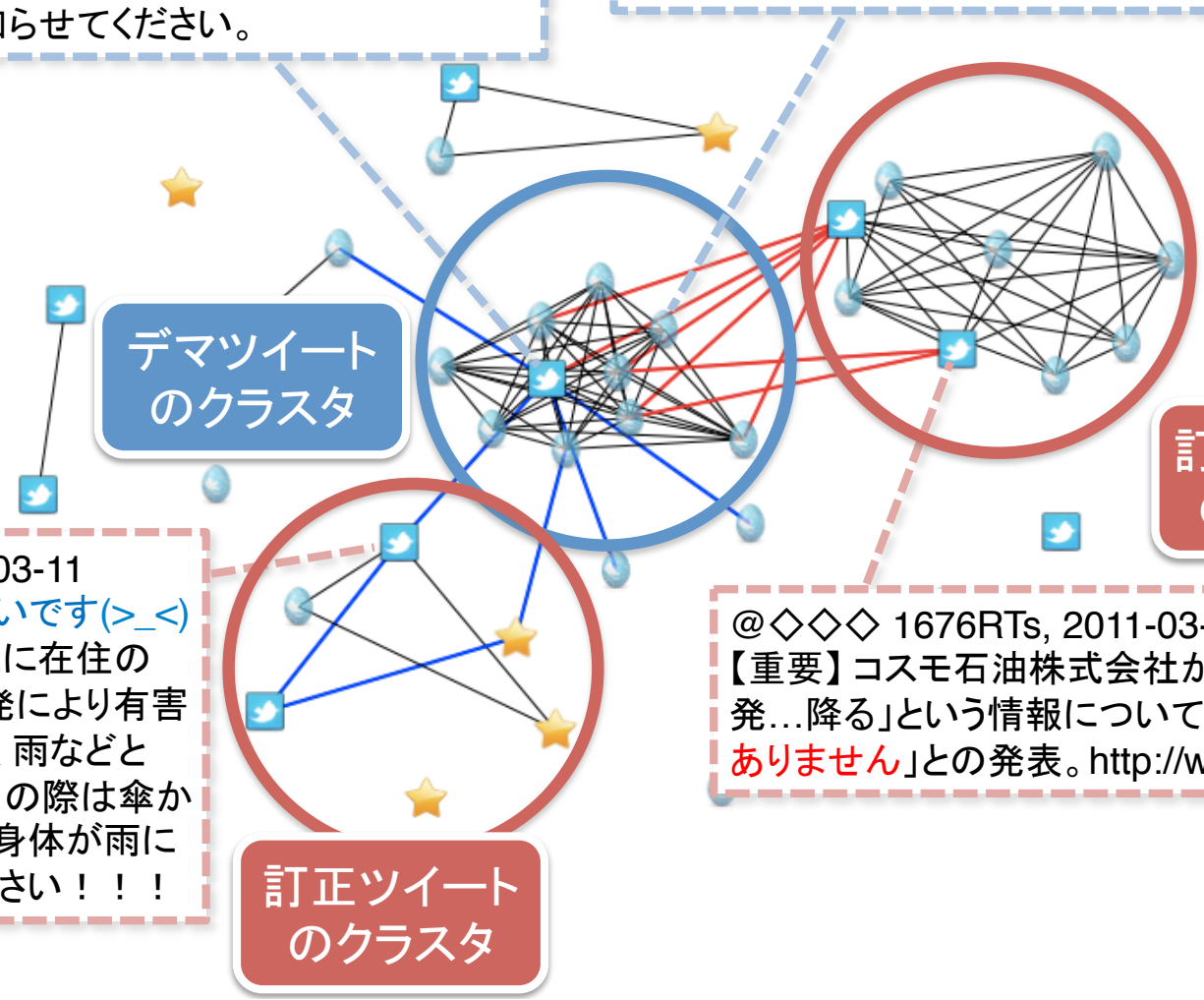
デマツイートの
 クラスタ

訂正ツイートの
 クラスタ

@☆☆☆ 0RTs, 2011-03-11
 23:49:24 ガセネタらしいです(>_<)
 RT @***: 千葉市近辺に在住の方! コスモ石油の爆発により有害物質が雲などに付着し、雨などと一緒に降るので外出の際は傘かカッパなどを持ち歩き、身体が雨に接触しないようにして下さい!!!

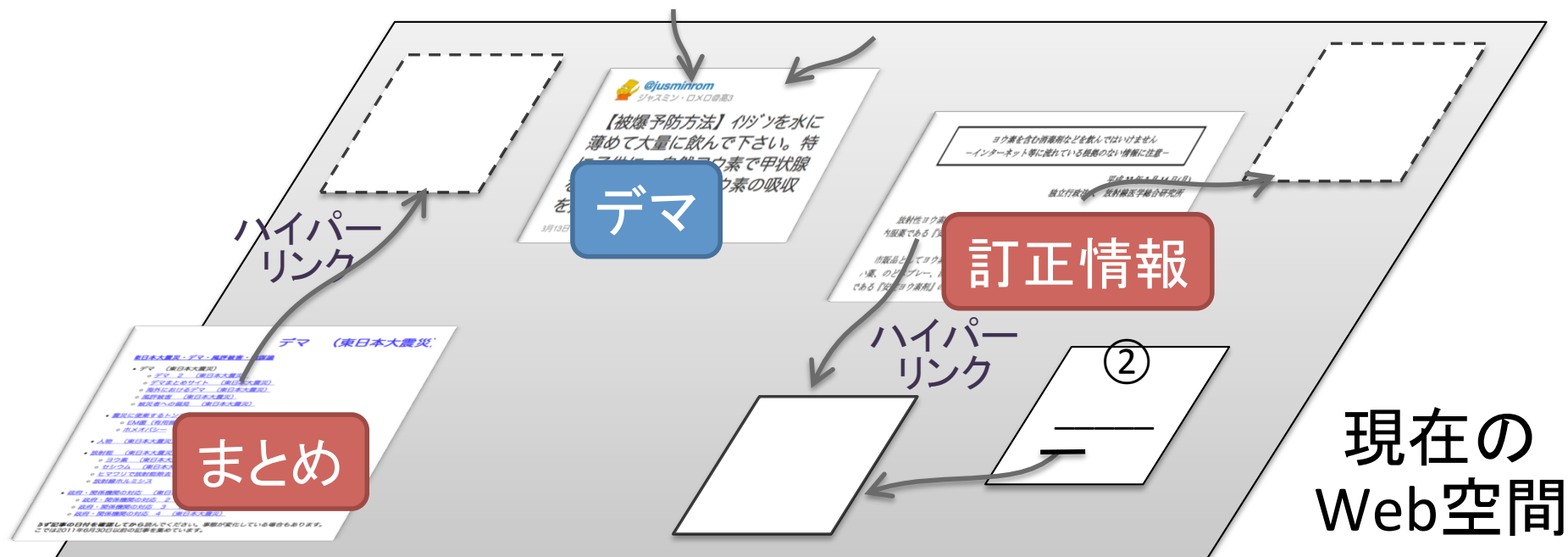
訂正ツイートの
 クラスタ

@◇◇◇ 1676RTs, 2011-03-12 14:46:29
 【重要】コスモ石油株式会社から「コスモ石油の爆発...降る」という情報について、「**このような事実はありません**」との発表。[http://www.cosmo-oil.jp/...](http://www.cosmo-oil.jp/)



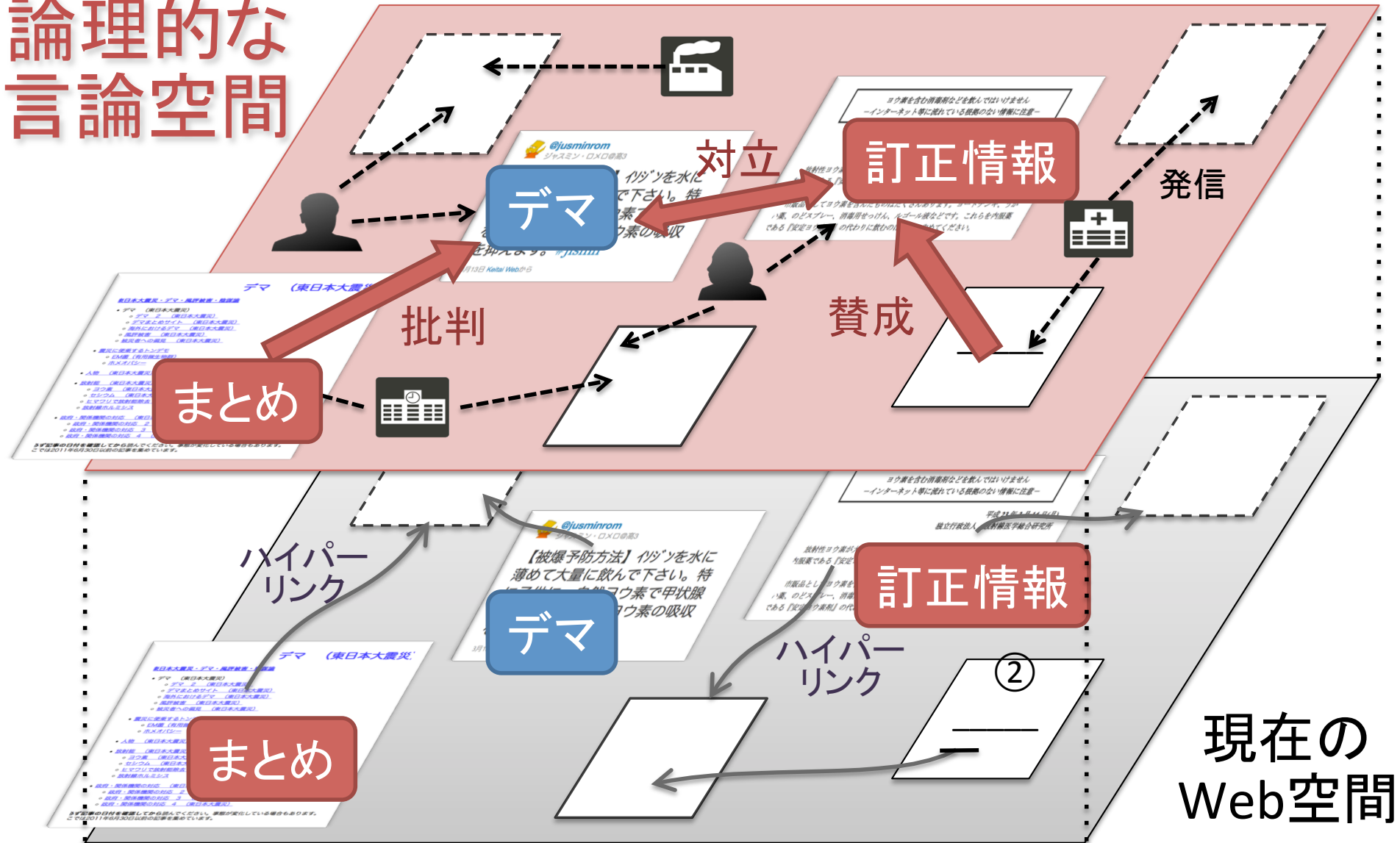
ウェブの情報の流れを変える

誤情報を見ている人に
訂正情報が伝わらない！



ウェブの情報の流れを変える

論理的な
言論空間



現在の
Web空間

言論マップ2G

命題ネット

(controversialな命題, 根拠, 因果関係, etc.)

ゲーム脳は恐ろしい

塩素は健康を害す

イソジンで
放射能を解
毒

温暖化によっ
て海面が上

コラーゲンは肌に良い

コラーゲンはシミには
効果がない

ビタミンCで
風邪を予防
できる

コラーゲンは肌から浸透しない

コラーゲンの上手な摂り方
コラーゲンドリンク

コラーゲンの補給で、モチモチとした弾力のあるお肌を取り戻すことができるので、積極的にコラーゲンを摂りましょう。

なお、コラーゲンの再合成を助けるビタミンCが加えられていることも大事です。

コラーゲン化粧品とコラーゲンサプリメントについて
OKWave
がんばるう日本!
Abille | OKGuide | OKetter | ARIGATO | OKMusic | OKStars

コラーゲンは肌から浸透しないので、出来てしまったシミに対する効果は期待できません

Web上の言明

質問(1)

クターシラボドモホルンリンクルやニビコラーゲンといった、コラーゲン成分の効能をうたった化粧品がありますよね？
最近そのひとつを使い出しているのですが、最近コラーゲンサプリメントにも興味を持っている調べてるとコラーゲンはヒフから塗っても浸透しないと書いてあり「じゃあコラーゲン配合の商品はイカサマ？」と不安になりました。事実はどうなんでしょうか？効果ゼロなのでしょうか？

閉ってます

誤情報の自動収集

よく反論されている命題をマイニング

誤情報モニタリングポータルサイト

現在話題の誤情報 (注目度順)

- | | | |
|---|--|------|
| 1 | 比例は候補者名を書けば大丈夫 | 919点 |
| 2 | ワタミ以外の自民候補者名を書けばいい | 391点 |
| 3 | ネガキャン工作人員の間で喧伝されている
演説会場でのカンパ募集が公選法違反 | 266点 |
| 4 | ゆーなっち健全 | 207点 |
| 5 | 因みに近親相姦で奇形児が産まれる | 113点 |
| 6 | X党の比例区候補Aを落選させるためには、他党に投票する
よりX党の他の候補者に投票する方が効果的..... | 101点 |
| 7 | これをアインシュタインは双子のパラドックスと言った | 100点 |
| 8 | 前にエートスという聞いた事のない名前を調べてたら、いき
なりbuvery | 94点 |
| 9 | 外山一派は権力からカネもらって活動してる | 87点 |

Web上の情報伝達を変える

Mozilla Firefox

Search Bookmarks and History

Google

(Untitled)



深海魚油
ソフトカプセル
Deep sea Fish Oil
Soft Capsules

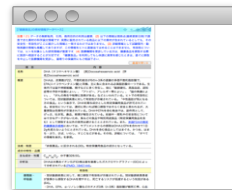
〇〇社によって製造された深海魚油のソフトカプセルです。...

主な成分である ω-3脂肪酸は、血液のより良い循環を促進し、高コレステロールの改善、血栓の防止、高血圧の改善、心臓の健康維持をサポートします。また、双極性障害、ADHDの改善に有効です。

Find: [] Next Previous Highlight all Match case

「裏を取る」作業を支援

「ω-3脂肪酸はADHDの改善に有効です」の対立ページ



健康食品データベース DHA..

[独立行政法人 国立栄養研究所]

小児のADHDは血中DHAレベルの低さと関連があるが、DHAは小児の注意欠陥多動性障害(ADHD)に対して、経口摂取で効果がないことが示唆されている。DHAを1日あたり345mg摂取しても改善はみられなかった...



魚油のオメガ3脂肪酸、頭を良くする効果ない=英研究

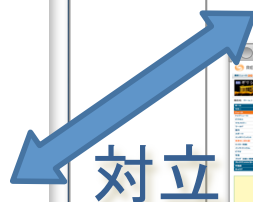
[ロイター] 英国で行われた研究によると、魚油に含まれるオメガ3脂肪酸は、心臓の健康のためには良いが、脳の機能を高める効果はないという。ロンドン大学の研究者らが専門誌「アメリカン・ジャーナル・オブ・クリニカル...

「ω-3脂肪酸はADHDの改善に有効です」の根拠ページ



生活習慣をサポートする『オメガ3脂肪酸（後編）』

[サブスマートUSA] ... ADHD、統合失調症、アルツハイマー疾患のような精神・認知に関する疾患を患う人は、血中のオメガ3脂肪酸レベルが低い人が多いことから、このような精神・認知に関する疾患に対し、オメガ3脂肪酸は有効性が期待できます。...



対立



根拠

本日の講演内容

我々は今どこにいるか

- ビッグデータ出現のインパクト

知の編集へ

- 情報信憑性判断の支援を例に

その先の言語理解研究へ

- 経験主義と合理主義の間で



その先の言語理解研究へ

「行間を読む」言語処理への挑戦

談話処理(「行間を読む」解析)

多様な問題の複合

週末に母とやなぎに行ってみました。

固有表現認識

うどん屋の名前

談話関係解析

うどんを食べるために

私は評判のきつねうどんを頼んだの

??????

私はうどんを食べた

多義性解消

注文した

ですが、期待通りの味に大満足。

??????

うどんが美味しいと

照応解析

きつねうどんの

談話処理の難しさ

- 複合的問題
 - さまざまな部分問題が相互に依存
 - 語義、省略・照応、談話関係、意図、...
 - どの問題から解くかを予め指定するのが困難
(単純な分類器学習に還元するのが難しい)
- 知識の組み合わせ
 - かなり大規模な知識ベースが利用可能に
 - いつどの知識を組み合わせるべきかを予め指定するのが困難
(分類器の素性の設計が難しい)

大規模・多種類の知識を使いこなす
ための仕組みを研究する段階に

Ed shouted at Tim because he was angry.

X is angry → X shouts

Ed shouted at Tim because he crashed the car.

X crashes Y's car → Y shouts at X ?

仮説推論 (abduction)

観測に対する最良の仮説(説明)を求める推論

仮説



スプリンクラー消し忘れ?



夜のうちに雨が降った?

観測



自宅の庭が濡れている



隣りの庭も濡れている

「行間を読む」仮説推論

背景知識

「何かを食べる」には「飲食店に行って、それを注文」すればよい

$\text{食べる}(x, z) \rightarrow \text{飲食店}(w) \wedge \text{行く}(x, w) \wedge \text{注文する}(x, z, w)$

$\forall x, z \text{ 食べる}(x, z) \rightarrow \exists w \text{ 飲食店}(w) \wedge \text{行く}(x, w) \wedge \text{注文する}(x, z, w)$

「(店で)注文する」ことを「頼む」とも言う

$\text{注文する}(x, z, w) \rightarrow \text{頼む}(x, z, w)$

「ある食べ物が評判」なら「それを食べたい」と思う

$\text{評判}(z) \rightarrow \text{食べる}(x, z)$

評判(z)

私はきつねうどんを食べた

食べる(x, z)

評判だから食べに行った

仮説

飲食店(w)

食べる目的
で行った

注文する(x, z, w)

やなぎ = 飲食店

y=w

「頼む」=
やなぎで注文する

w=v

観測

私(x) \wedge やなぎ(y) \wedge 行く(x, y) \wedge きつねうどん(z) \wedge 頼む(x, z, v) \wedge 評判(z)

母とやなぎに行った。 評判のきつねうどんを頼んだ。

「行間を読む」仮説推論

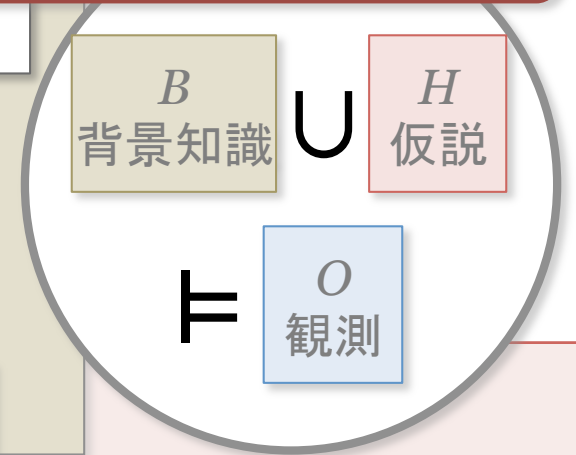
背景知識

「何かを食べる」に **棄却可能** (観測が増えると解釈が変化する)

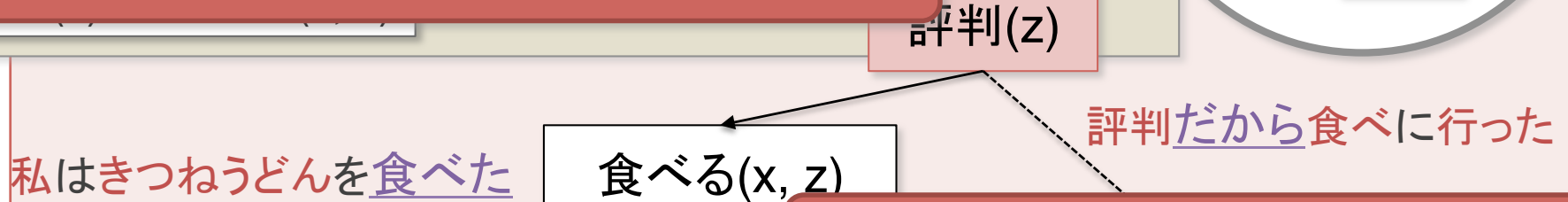
食べる(x, z) → 飲食店(w) ∧ 行く(x, w) ∧ 注文する(x, z, w)

「(店で)注文する」ことを「頼む」とも言う

注文する(x, z, w) → 頼む(x, z, w)



処理の手順を捨象した宣言的知識



仮説

飲食店(w)

食べる目的で行った

注文する(x, z, w)

やなぎ = 飲食店

y=w

「頼む」 = やなぎで注文する

w=v

観測

私(x) ∧ やなぎ(y) ∧ 行く(x, y) ∧ きつねうどん(z) ∧ 頼む(x, z, v) ∧ 評判(z)

母とやなぎに行った。 評判のきつねうどんを頼んだ。

東北大の取り組み

背景知識

「何かを食べる」には「飲食店に行って、それを注文」すればよい

食べる(x, z) → 飲食店(w) ∧ 行く(x, w) ∧ 注文する(x, z, w)

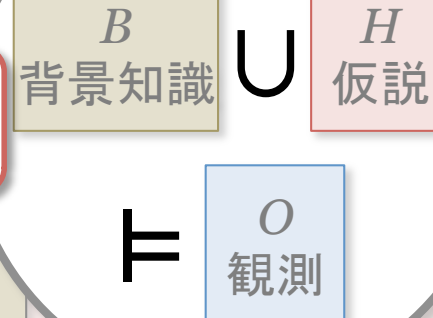
「(店で)注文する」ことを「頼む」とも言う

注文する(x, z, w) → 頼む(x, z, w)

「ある食べ物が評判」なら「それを食べたい」と思う

評判(z) → 食べる(x, z)

③ 言語処理への応用:
NLP若手の会 奨励賞 (2011)



① 世界最速の推論エンジン

7分(US) → 3秒(東北大)

情報処理学会 山下記念賞 (2012)
言語処理学会 年次大会優秀賞 (2012)

評判(z)

評判だから食べに行った

食べる(x, z)

きつねうどんを食べた

② 仮説の評価関数の学習(世界初)

NLP若手の会 奨励賞 (2012)

自然言語処理研究会 学生奨励賞 (2012)
IBIS 2012 Honorable Mention (2012)

仮説

飲食店(w)

食べる目的
で行った

注文す
「頼む」

やなぎ = 飲食店

y=w

やなぎで注文する

w=v

観測

私(x) ∧ やなぎ(y) ∧ 行く(x, y) ∧ きつねうどん(z) ∧ 頼む(x, z, v) ∧ 評判(z)

母とやなぎに行った。 評判のきつねうどんを頼んだ。

仮説選択の高速化

- 仮説推論は重たい処理
 - 重み付き最大充足割当問題 (W-MaxSAT) の一種
→ NP-hard
- これまでは命題論理版の推論を高速化
 - 整数線形計画問題 (ILP) としての定式化 [Santos 94], 非線形計画法と整数線形計画法の併用 [Matsuo & Ishizuka 98], etc.
 - 問題: 一階述語論理 (FOL) の推論を命題論理上で行うのは極めて非効率的
- アイデア: FOL 上の推論を基礎化せずに直接行う
 - Lifted inference [Poole 03, Braz+ 05, etc.]
 - 起源: 定理証明の Resolution [Robinson 65]
 - Lifted inference を整数線形計画法 (ILP) により定式化

整数線形計画法による推論の高速化

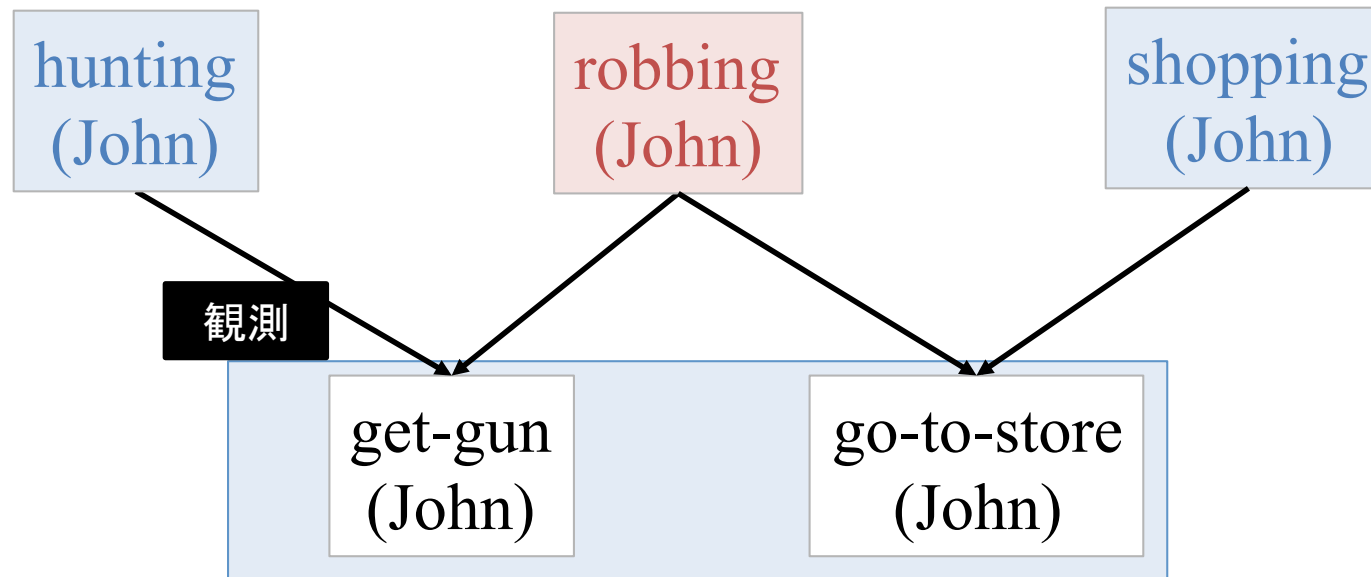
背景知識

hunting(x) \Rightarrow get-gun(x)

robbing(x) \Rightarrow get-gun(x)

shopping(x) \Rightarrow go-to-store(x)

robbing(x) \Rightarrow go-to-store(x)



John got a gun and went into the store.

整数線形計画法による推論の高速化

背景知識

$$S(x) \Rightarrow P(x)$$

$$S(x) \Rightarrow Q(x)$$

$$h_{x=y}$$

$$u_{S(x),S(y)}=1 \Rightarrow h_{x=y}=1$$

$$u_{S(x),S(y)} \leq h_{x=y}$$

目的関数

$$\arg \max_{h_{P(x)}, h_{Q(y)}, \dots, u_{S(x),S(y)}} \text{score}(H) = \mathbf{w}^T \cdot [h_{P(x)}, h_{Q(y)}, \dots, u_{S(x),S(y)}]$$

$$= \mathbf{w}^T \cdot \Phi(h_{P(x)}, h_{Q(y)}, \dots, u_{S(x),S(y)})$$

$$h_{S(x)}=1 \Rightarrow h_{P(x)}=1$$

$$h_{S(x)} \leq h_{P(x)}$$

$$h_{P(x)}=1$$

観測

$P(x)$

$$h_{P(x)}$$

$$u_{S(x),S(y)} \in \{0, 1\}$$

$Q(y)$

$$h_{Q(y)} \in \{0, 1\}$$

$$u_{S(x),S(y)}=1 \Rightarrow h_{S(x)}=h_{S(y)}=1$$

$$u_{S(x),S(y)} \leq (h_{S(x)} + h_{S(y)})/2$$

推論効率の評価実験

- データセット
 - 背景知識
 - WordNet 3.0 [Fellbaum 98]
 - FrameNet 1.5 [Ruppenhofer+ 10]
→ 約37万の推論規則
 - 観測 (入力)
 - RTE-2 開発セット
 - RTE: 2つのテキストの対が含意関係にあるかを判別
 - 1,600 文章, 平均 20.3 リテラル
- ツール
 - 論理式への変換: Boxer [Bos 08]
 - ILP ソルバー: Gurobi Optimizer 4.6

評価実験の結果

後ろ向き推論の深さを変えたときの推論時間

深さ	潜在仮説集合の平均サイズ	2分以内に最適解と判断された問題	平均推論時間
1	89	99.8%	0.33 秒
2	479	98.4%	1.07 秒
3	1,171	95.8%	2.63 秒

比較参考値

- Mini-TACITUS [Mulkar-Mehta+ 07]: **30分以上**
- Markov Logic Networks [Richardson & Domingos 06] による実装 [Blythe+ 11]: **7分**

評価関数の学習

学習の目標

$$\mathbf{w}^T \cdot \Phi(H') + \Delta < \mathbf{w}^T \cdot \Phi(H)$$

出力された仮説 (H')

$e = h$ (**Ed** = **he**) ... 誤

$crash(e, u_2)$

$irritated(e)$

正解の仮説 (H)

$t = h$ (**Tim** = **he**) ... 正

$angry-with(e, t)$

=

$angry-with(u_3, h)$

教師が与えなくてもよい

$do-bad-thing(u_1, h)$

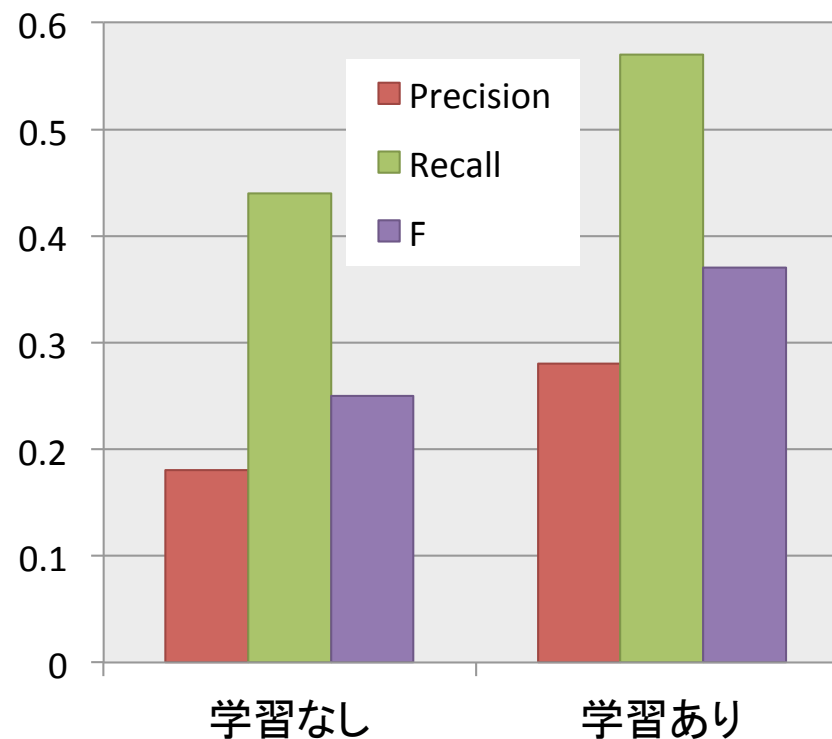
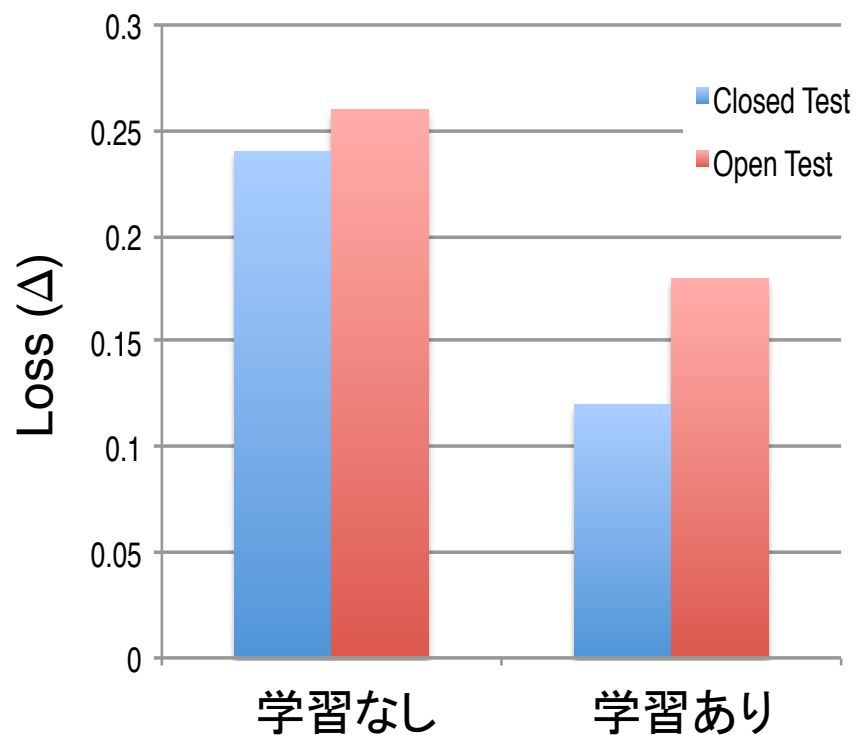
観測

$Ed(e) \wedge shout-at(e, t) \wedge Tim(t) \wedge male(h) \wedge crash(h, c) \wedge car(c)$

Ed shouted at **Tim** because **he** crashed the car.

学習の効果

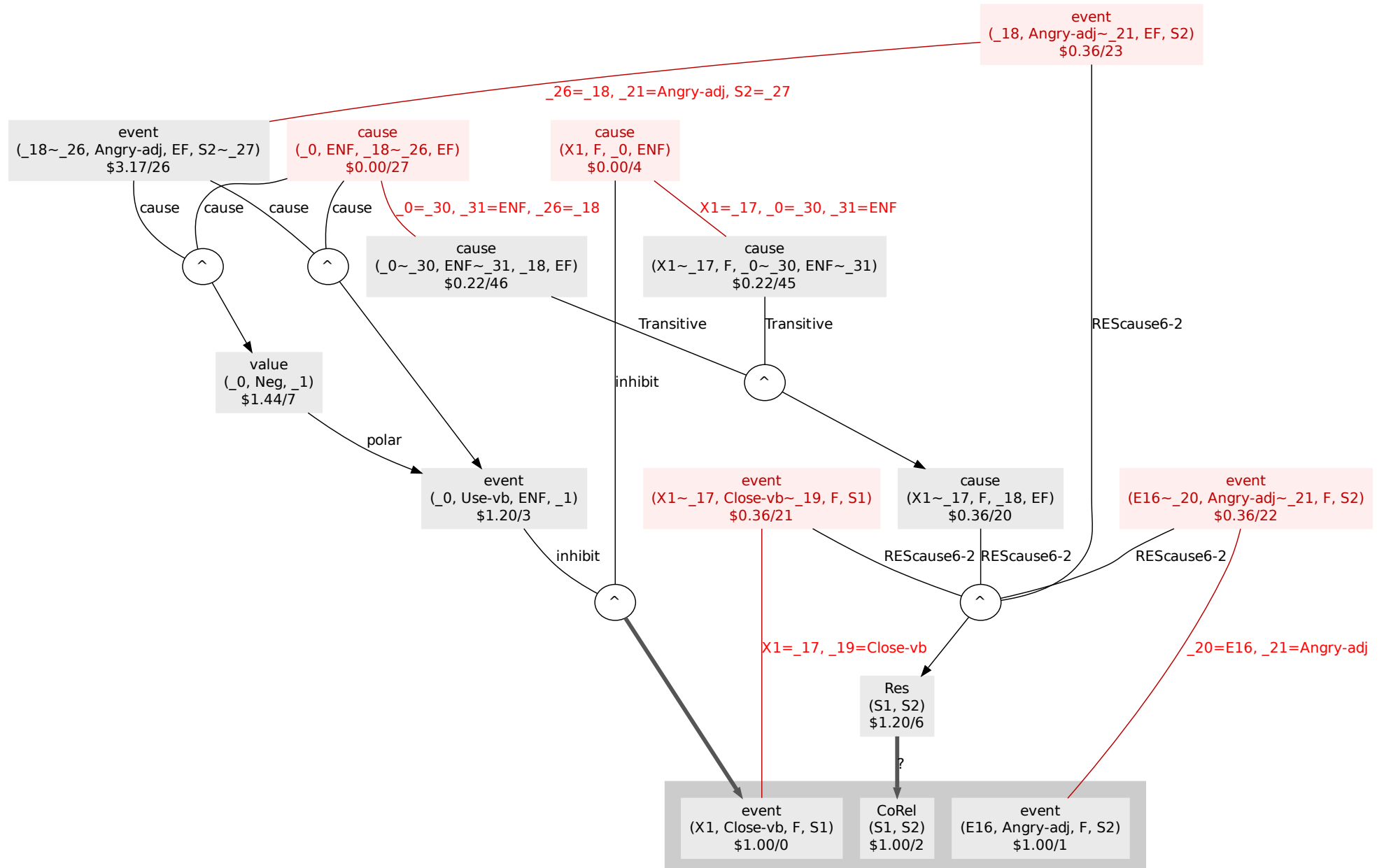
プラン認識課題 [Ng & Mooney 92]



(S1: A lot of traffic once used Folsom Dam Road.)

S2: Right now, the road is closed.

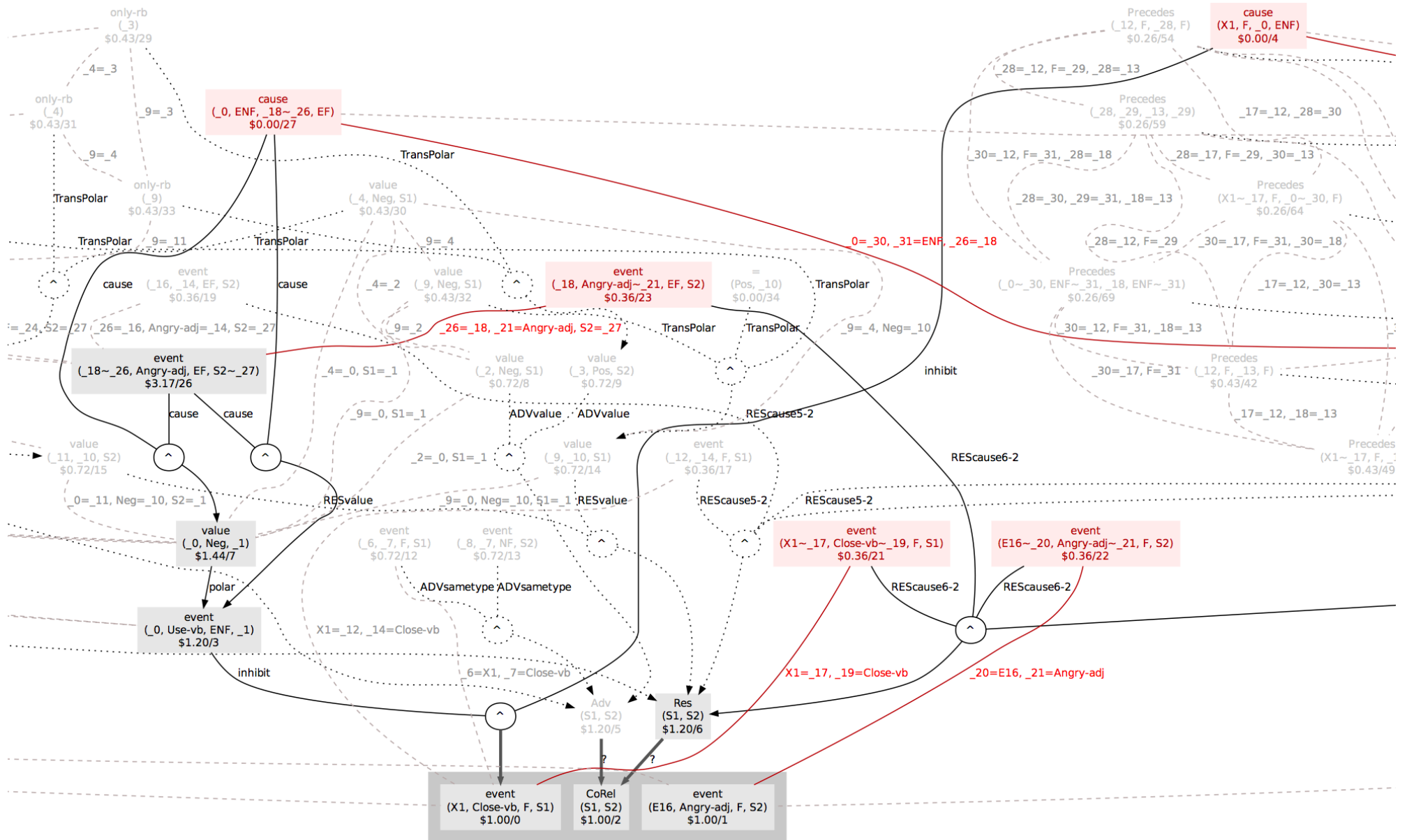
S3: Most of the people who used the road every day are angry.



(S1: A lot of traffic once used Folsom Dam Road.)

S2: Right now, the road is closed.

S3: Most of the people who used the road every day are angry.



まとめ

ゲーテンベルク以来の情報爆発

地球規模のコミュニケーションの支援へ

鍵は自然言語処理による情報・知識編集

言語処理技術自身もビッグデータで飛躍

「行間を読む」言語処理は知識×推論

言語処理から人工知能にブレークスルーを