

テキスト意味理解の深化と応用

Recent Advances in Text Understanding and its Applications

黒橋 禎夫（京都大学）

Sadao KUROHASHI (Kyoto University)



目次

Table of Contents

1. 深層学習のインパクト

Impact by deep learning

– 意味のベクトル表現

Distributional semantic representation

– End-to-end学習

End-to-end learning

2. テキスト意味解析の深化

Recent Advances in Text Understanding

3. 実課題への応用

Application to Real World Task

– 意見の集約

Opinion summarization

– 行政サービスの対話ボットシステム

Chatbot for administrative service of local government

意味のベクトル表現

Distributional semantic representation

分布仮説: 文脈が似ている単語は類似している

distributional hypothesis: words that are used and occur in the same contexts tend to purport similar meanings

共起語に基づくベクトル表現 (高次元)

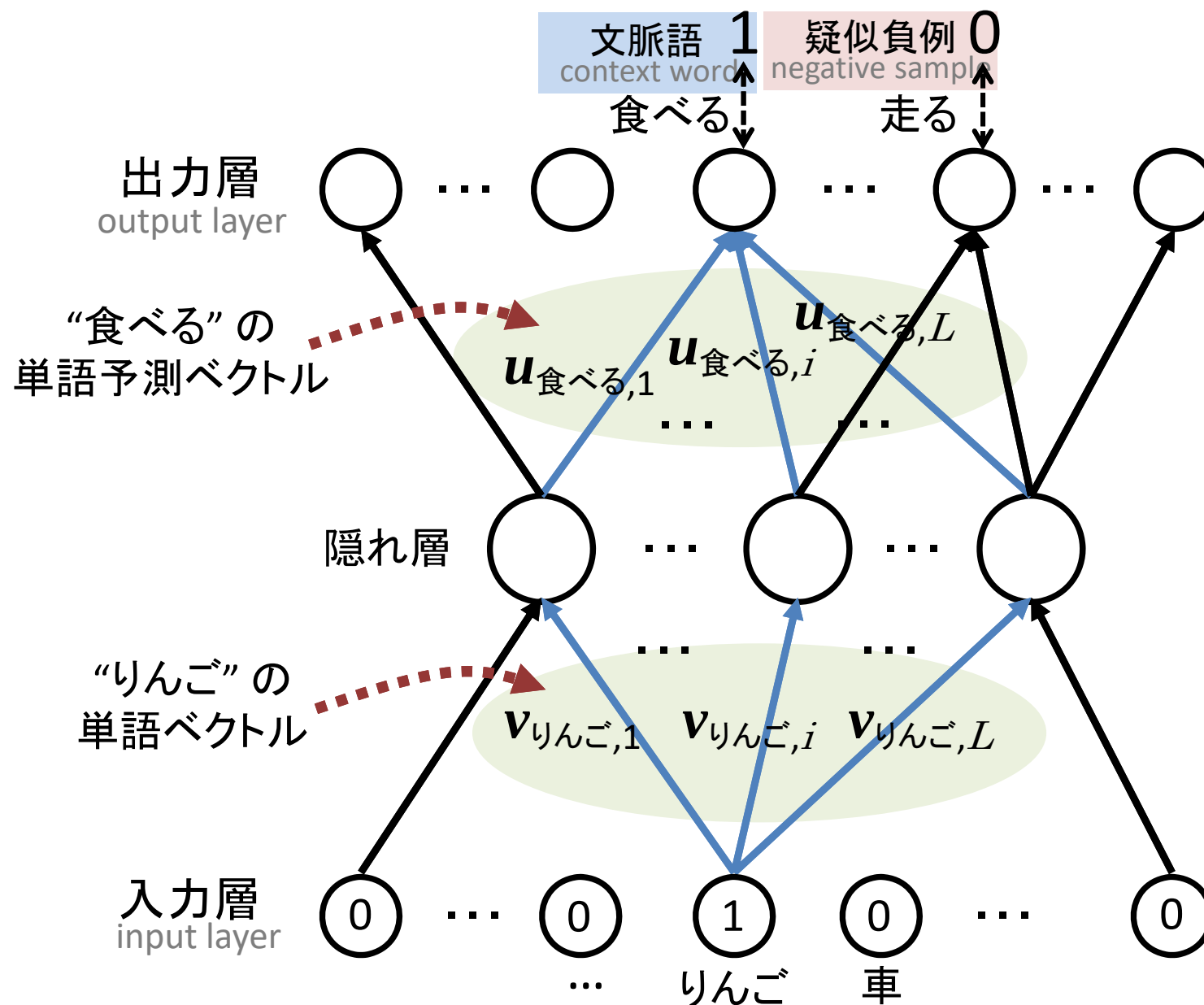
co-occurring words

りんご	(0 0 1 0 0 0 0 1 0 0 0 0)
みかん	(0 0 1 0 0 0 0 0 1 0 0)
車	(0 0 0 0 0 0 1 0 0 0 1 0)
...	

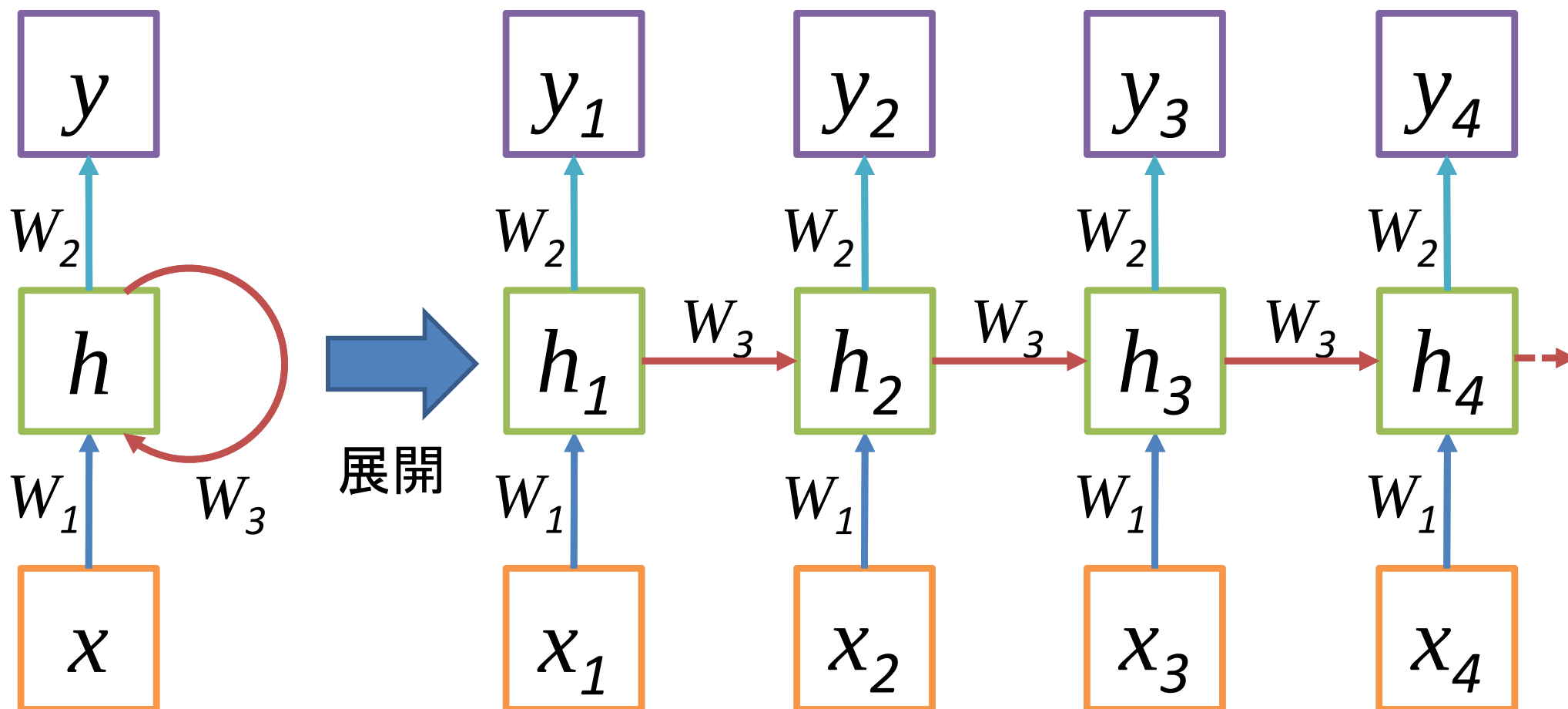
word embeddingのベクトル表現 (低次元)

りんご	(0.82 -1.18 0.74 ...)
みかん	(0.86 -1.23 0.54 ...)
車	(0.15 1.18 -0.34 ...)
...	

Word2Vec [Mikolov+ 2013]



RNN (Recurrent Neural Network)

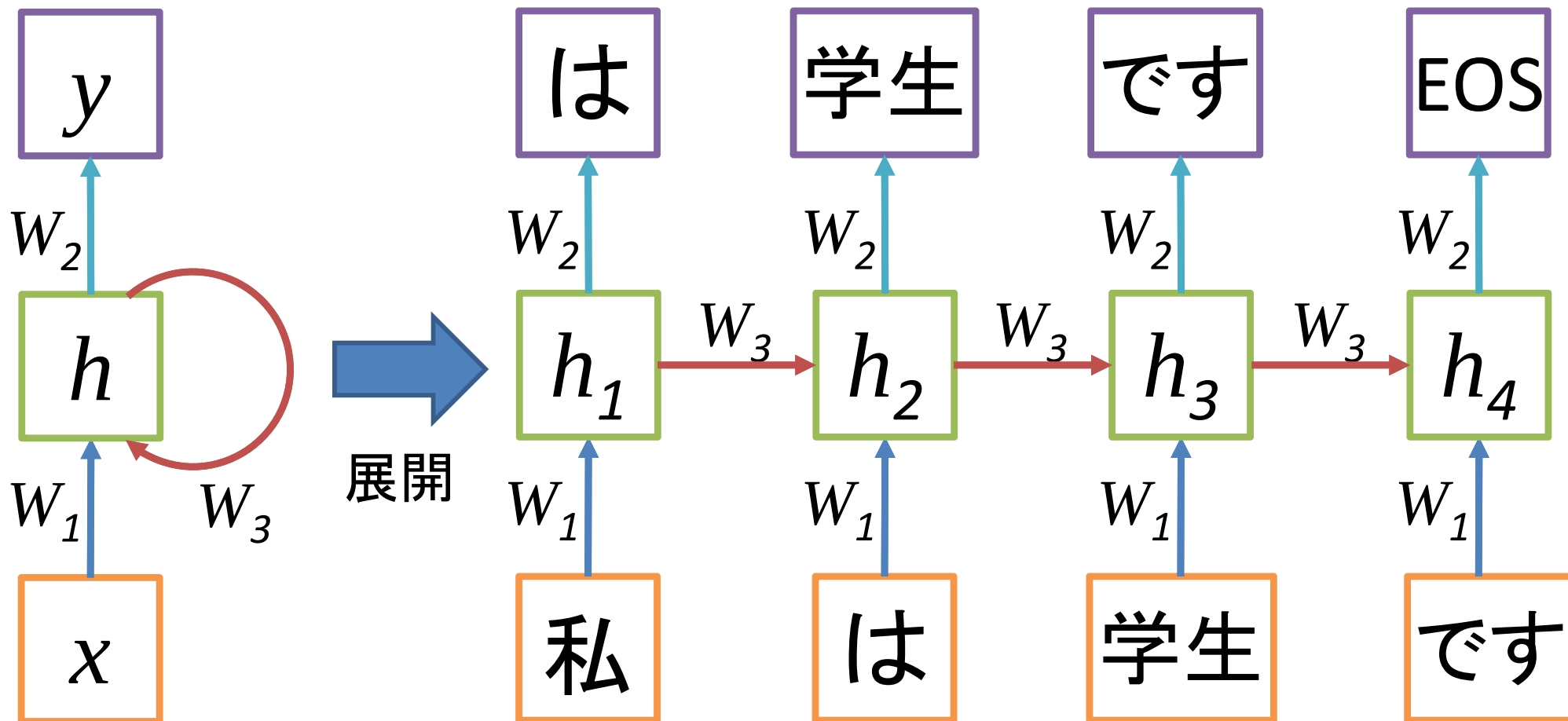


RNN

RNNを時系列に展開したもの

RNN言語モデル

RNN based Language Model

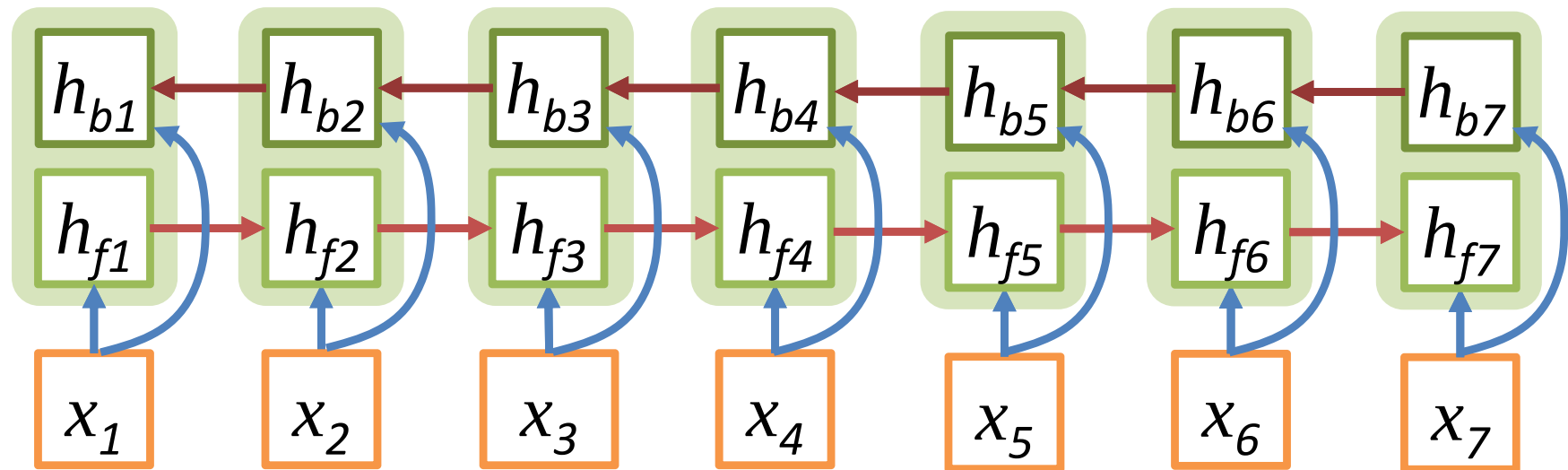


RNN

RNN言語モデル

BiLSTMによる文の意味表現

Semantic Representation of Sentence by BiLSTM

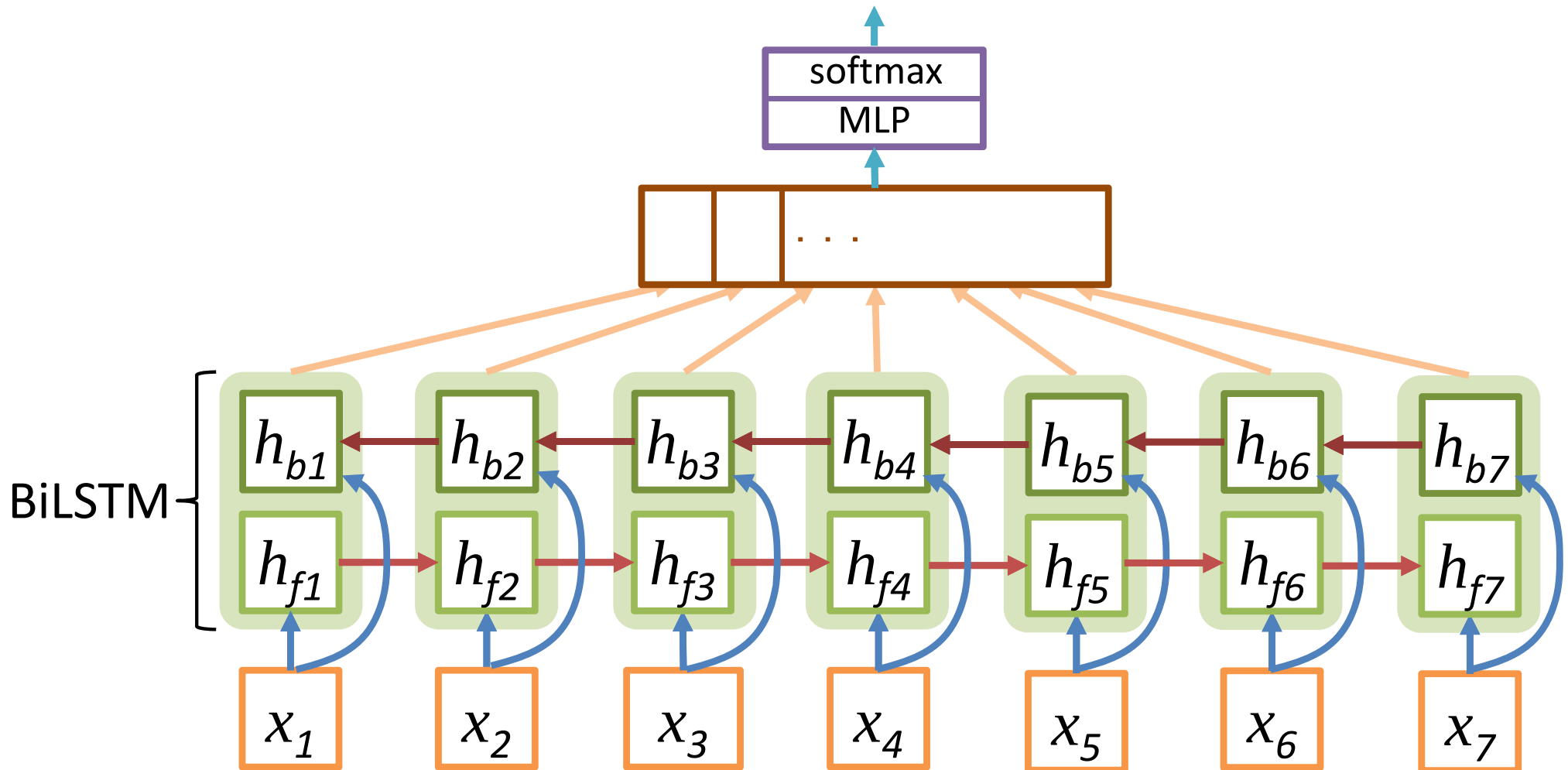


BiLSTMによる文の意味表現

Semantic Representation of Sentence by BiLSTM

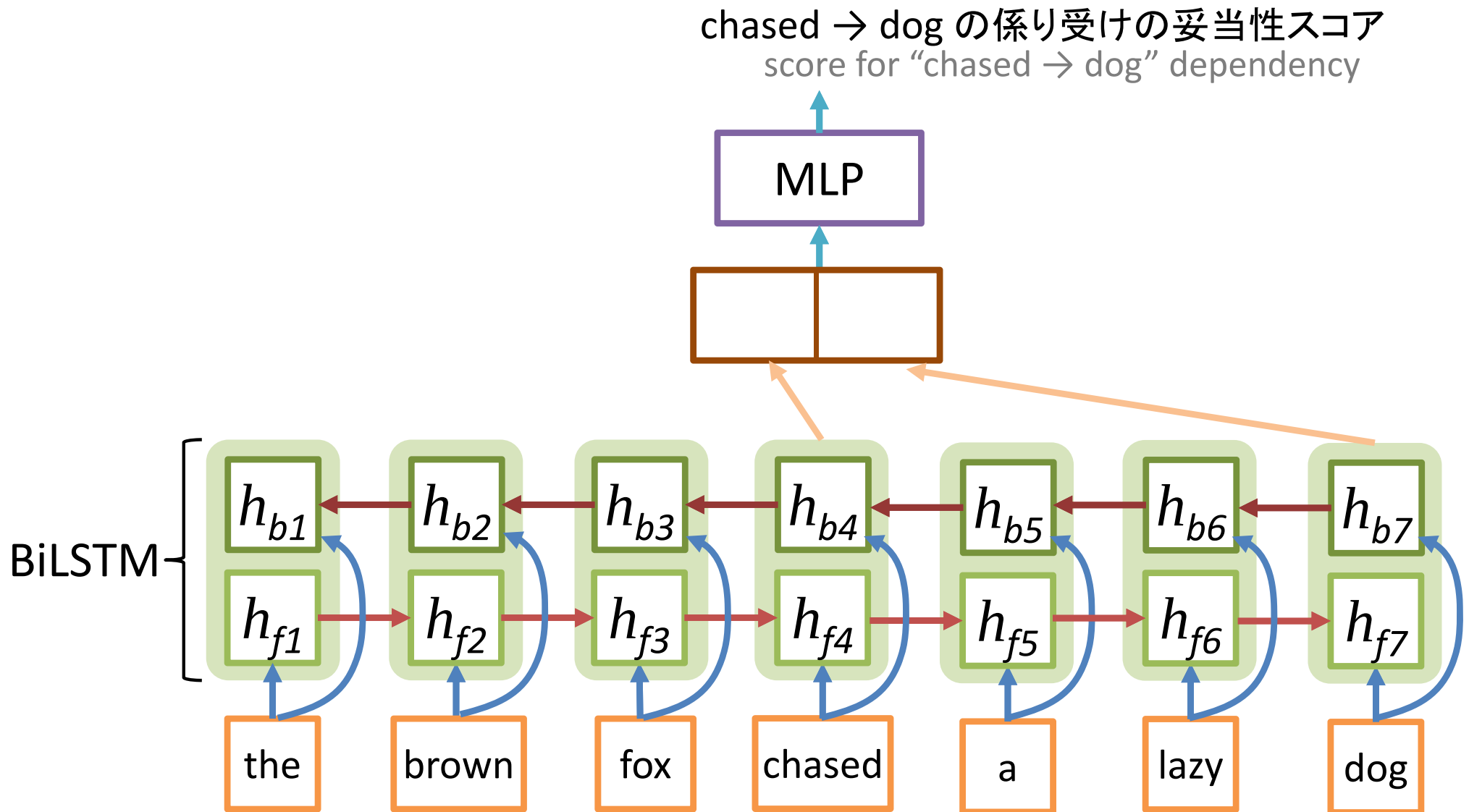
文の解釈 (e.g. 肯定・否定・中立 or 対話行為の種類)

Interpretation of sentence



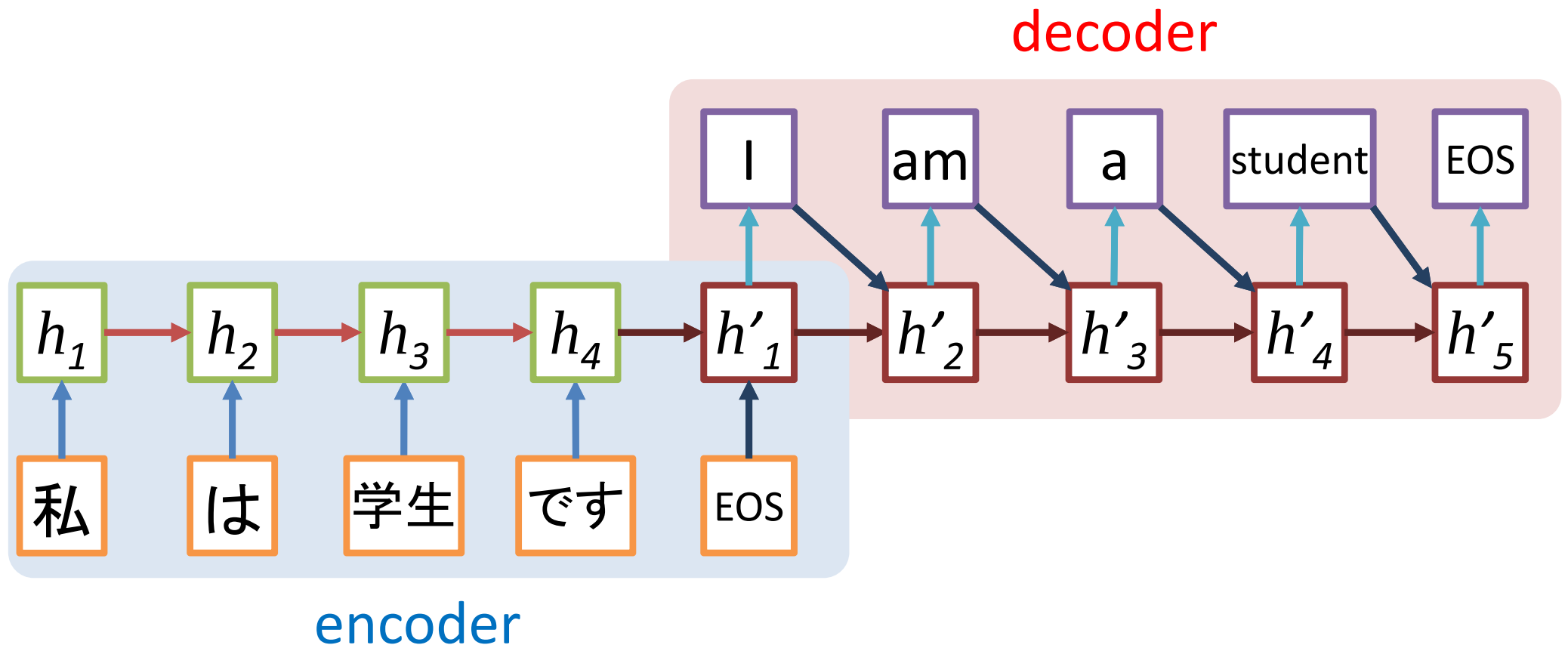
BiLSTMによる構文解析

Parsing by BiLSTM



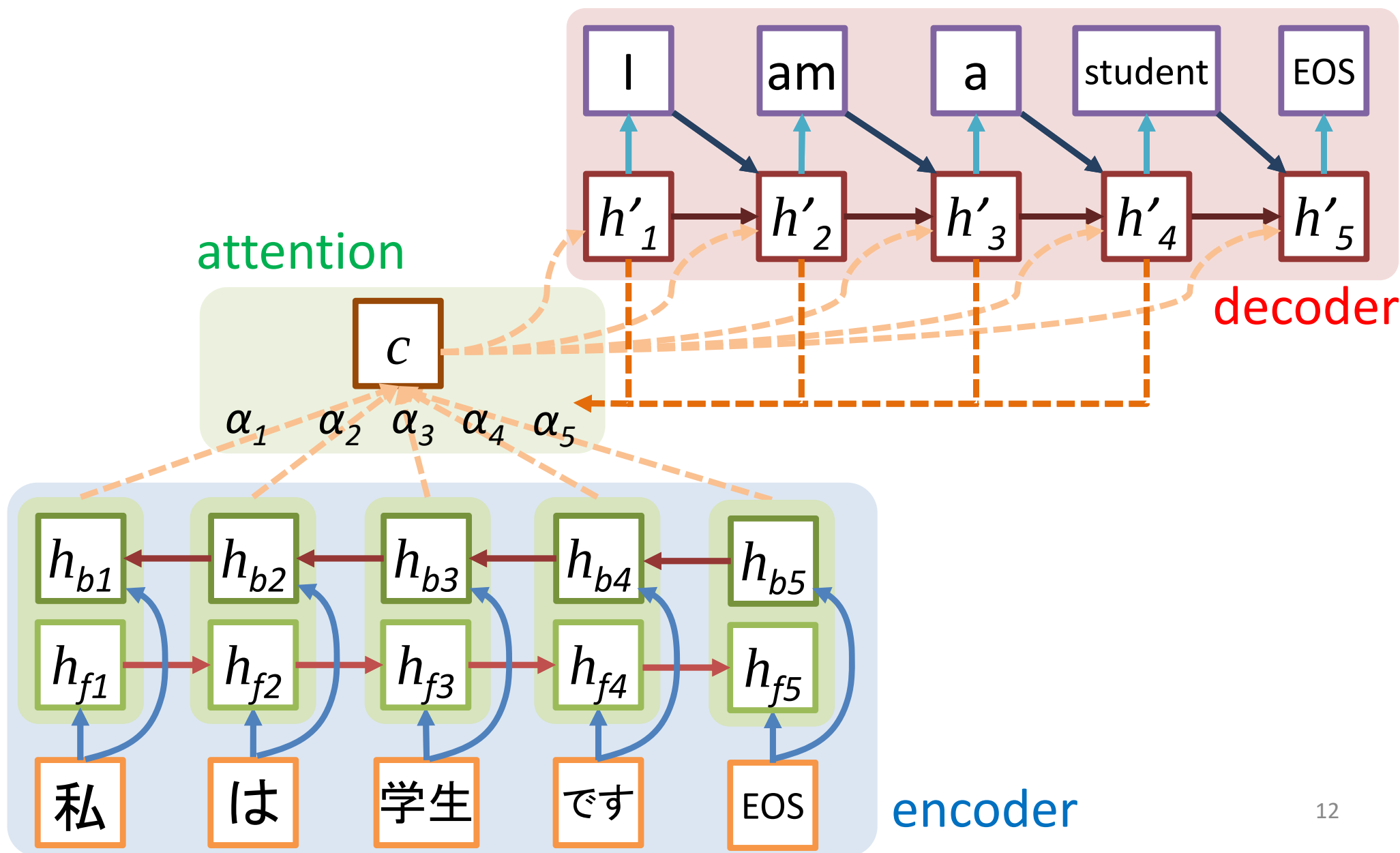
ニューラル機械翻訳モデル

Neural Machine Translation Model [Sutskever+ 2014]



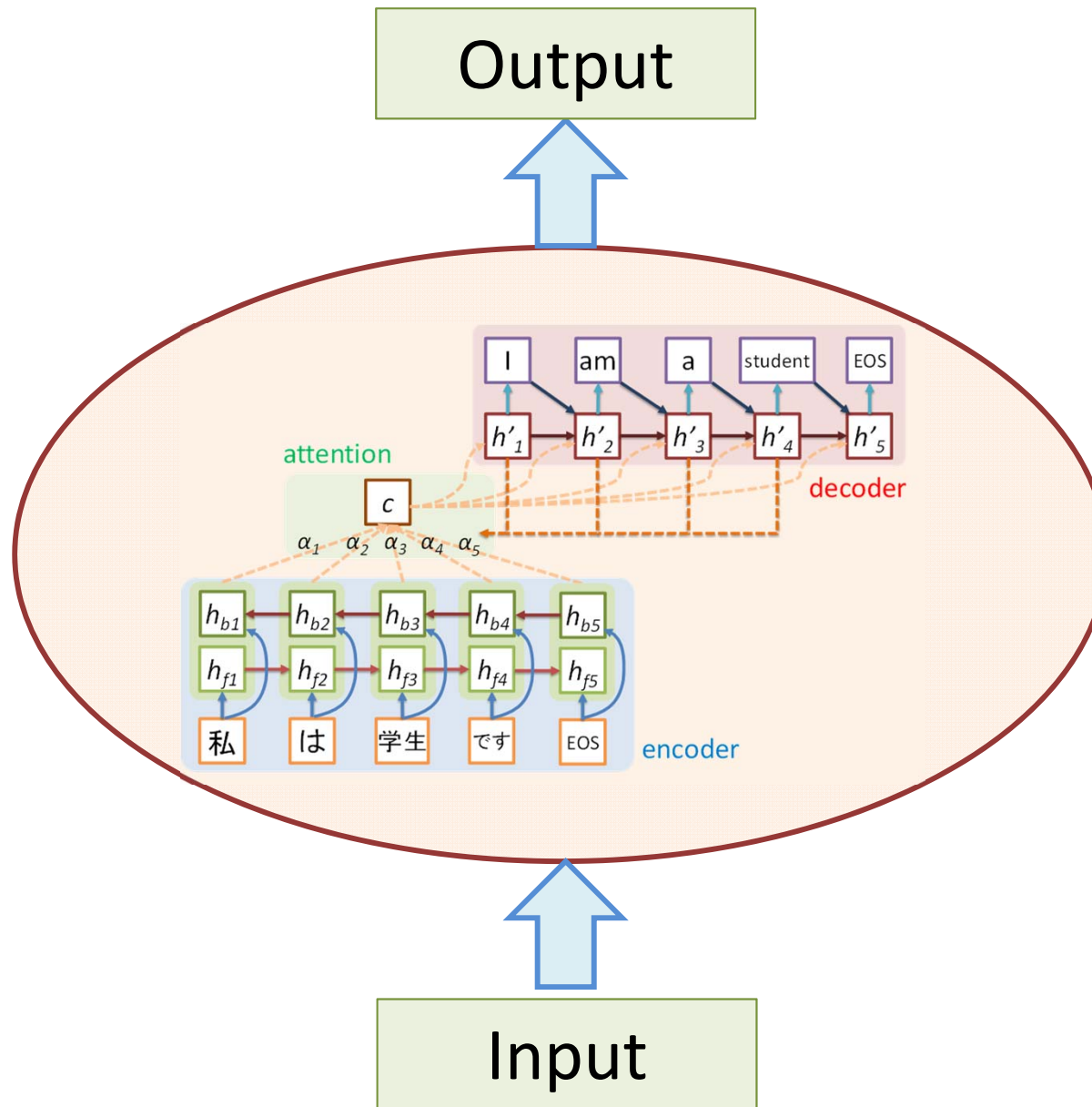
ニューラル機械翻訳モデル

Neural Machine Translation Model [Bahdanau+ 2015]



End-to-End 学習

End-to-End Learning



End-to-End 学習

End-to-End Learning

機械翻訳

Machine Translation

A battle to ask Prime Minister Shinzo Abe's five-year administration started. He emphasized the achievements of ...

文書要約

Summarization

安倍首相、政権5年の実績強調

DNN

安倍晋三首相の5年間の政権運営を問う戦いが始まった。経済政策の実績を強調し、...

雑談対話

Chatbot

早く寝たほうがいいですね

DNN

今日は仕事でヘトヘトだ！

End-to-End 学習

End-to-End Learning

天気予報コメント生成^{*1}

Weather Forecast Generation

今日は雲が広がり、ニワカ雨の可能性
があります。折りたたみ傘が
あると安心。

DNN

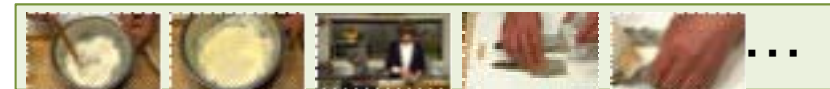
気温					気圧					湿度				
28.3	28.2	27.9	...	...	1017	1016	1015	...	...	...	...	...	...	...
28.2	28.1	...	...	...	1015	1015	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

ビデオクリップ説明文生成^{*2}

Video Description Generation

A woman is cooking in the kitchen.

DNN



*1 村上, 笹野, 高村, 奥村: 数値予報マップからの天気予報コメントの自動生成, 言語処理学会 第23回年次大会 (2017.3)

*2 Laokulrat+: Generating Video Description using Sequence-to-sequence Model with Temporal Attention, COLING2016

Deep Learning

Pros

- 意味表現

semantic representation

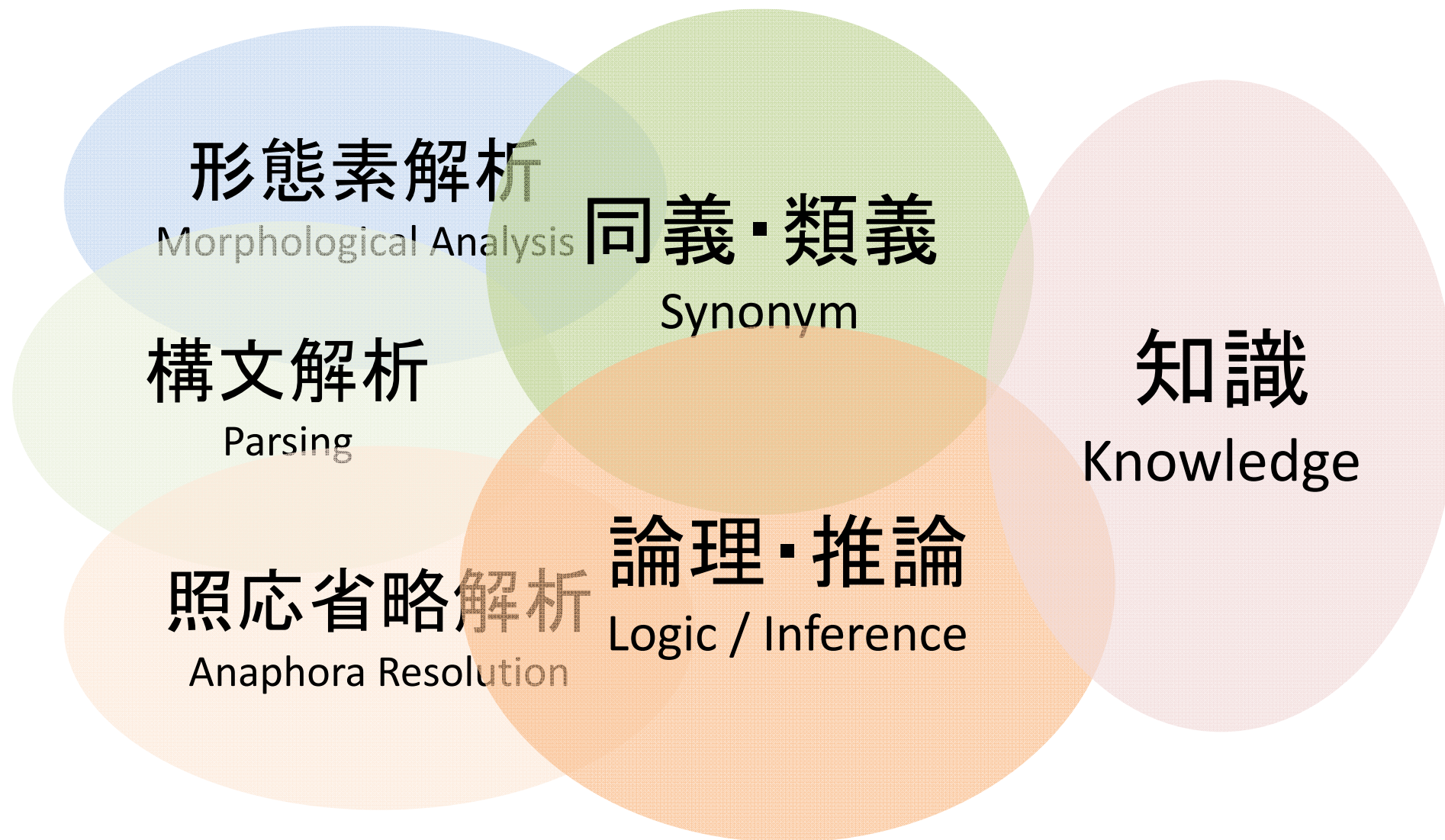
- End-to-end learning

Cons

- Data hungry

NLPのアプローチ

Approach



Establishment of Knowledge-Intensive Structural Natural Language Processing and Construction of Knowledge Infrastructure (2013.10—2019.3)

Research Director Sadao KUROHASHI (Kyoto University)

Main Research Collaborator Kentaro INUI
(Tohoku U)



Daisuke BEKKI
(Ochanomizu U)

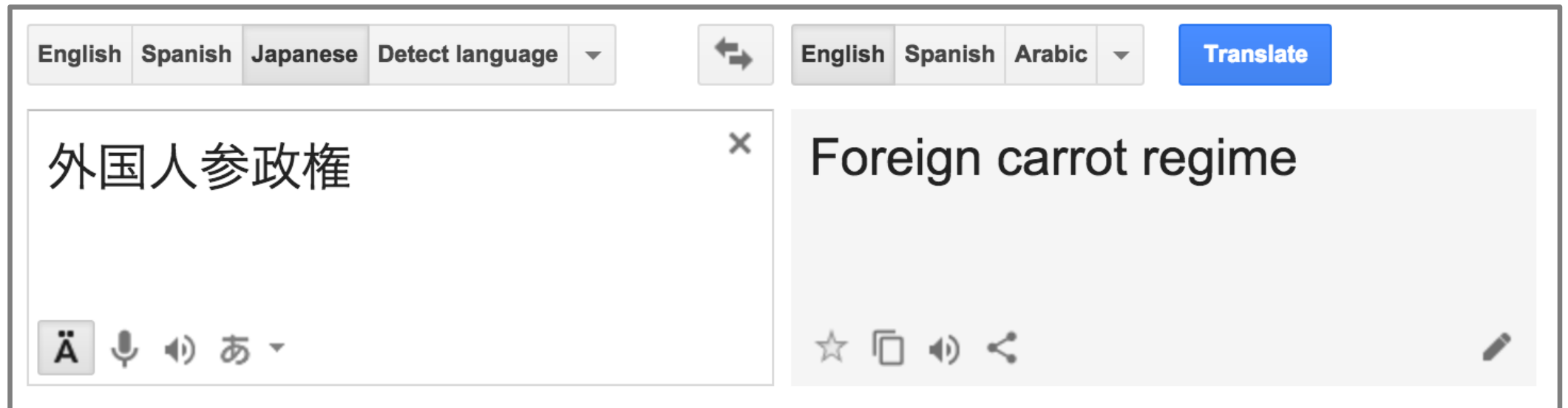


RNN言語モデルに基づく形態素解析

Morphological Analysis using RNN Language Model

[Morita+ EMNLP2015]

外国人参政権



<https://translate.google.co.jp/>

RNN言語モデルに基づく形態素解析

Morphological Analysis using RNN Language Model

[Morita+ EMNLP2015]

外国人参政権



The screenshot shows the Google Translate web interface. On the left, the source language is set to Japanese (日本語) and the text input field contains '外国人参政権'. On the right, the target language is set to English (英語) and the translated text is 'Foreign baclick administration'. The interface includes language selection buttons, a 'Translate' button, and various utility icons like a star, a document icon, a speaker icon, and a share icon.

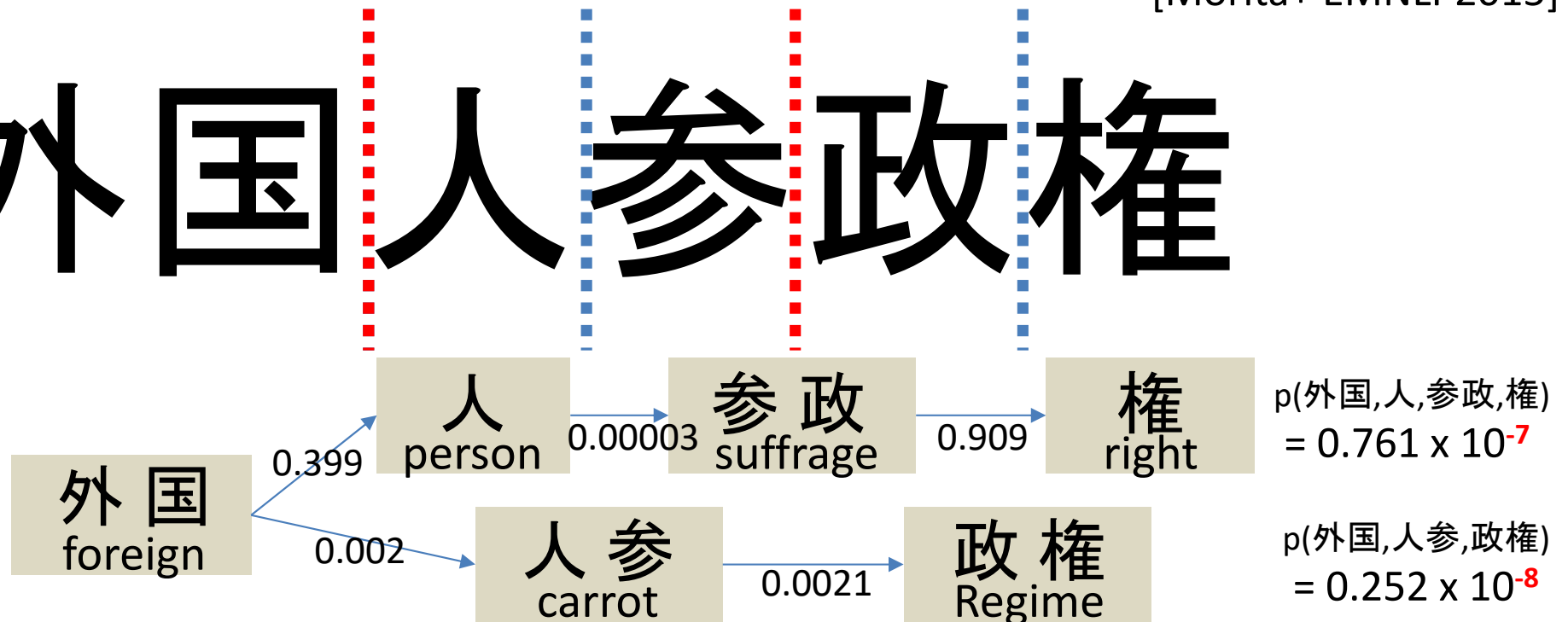
<https://translate.google.co.jp/>

RNN言語モデルに基づく形態素解析

Morphological Analysis using RNN Language Model

[Morita+ EMNLP2015]

外国人参政権

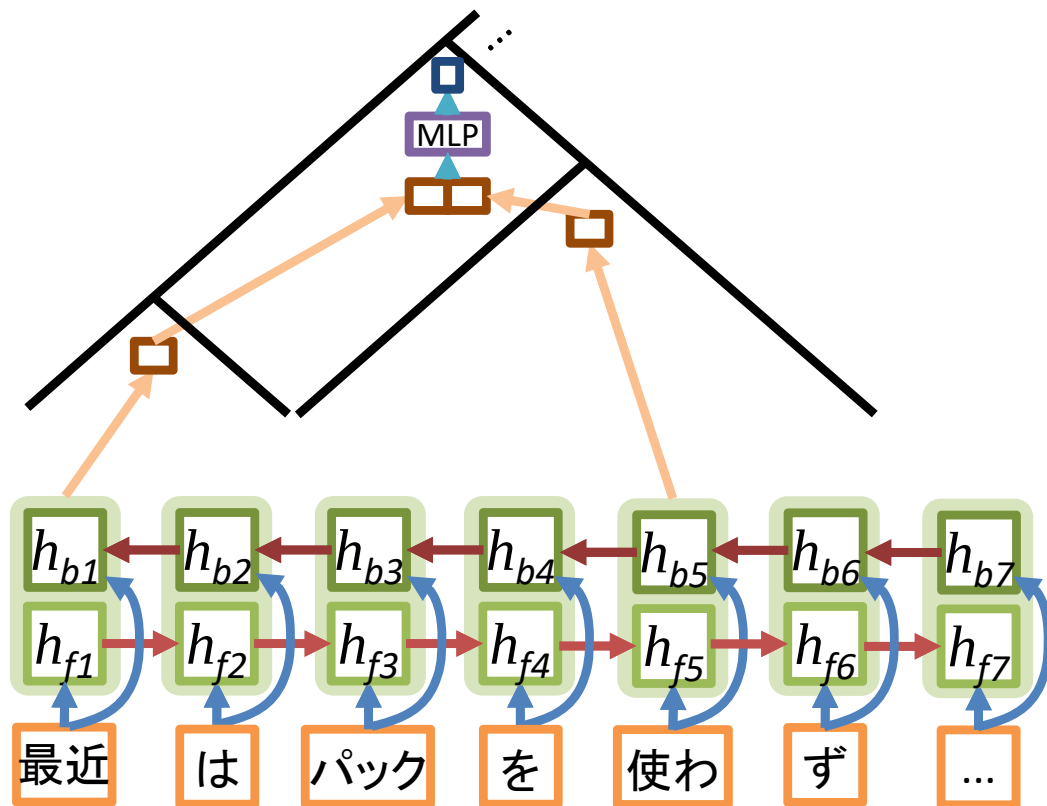


	Segmentation	Seg+POS	Speed (sent/sec)
JUMAN	98.80	97.75	8,800
JUMAN++v1	99.37	98.95	16
JUMAN++v2	99.52	99.21	4,800
Mecab	99.32	98.85	53,000

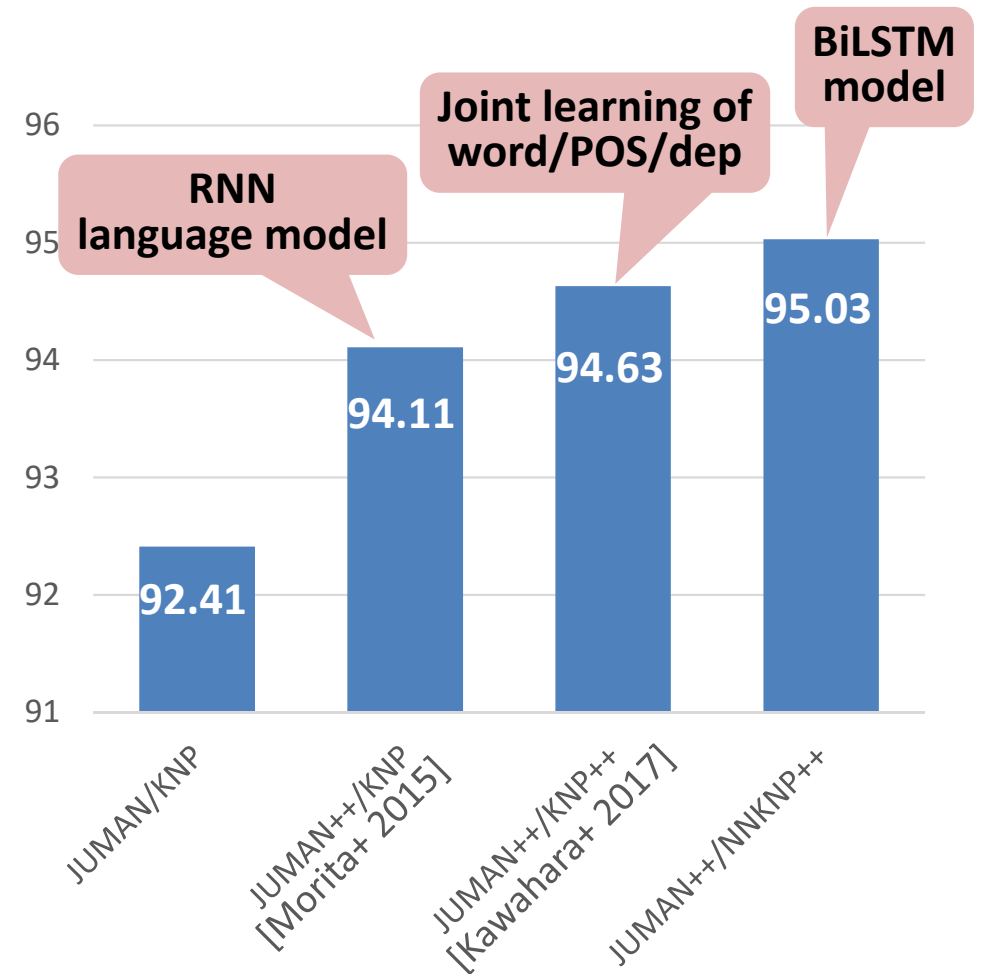
(京大テキストコーパスで評価)

BiLSTMに基づく日本語構文解析

BiLSTM based Japanese Dependency Parsing



Kyoto University Text Corpus (40K sen)



ベクトル表現を用いた構造的意味合成

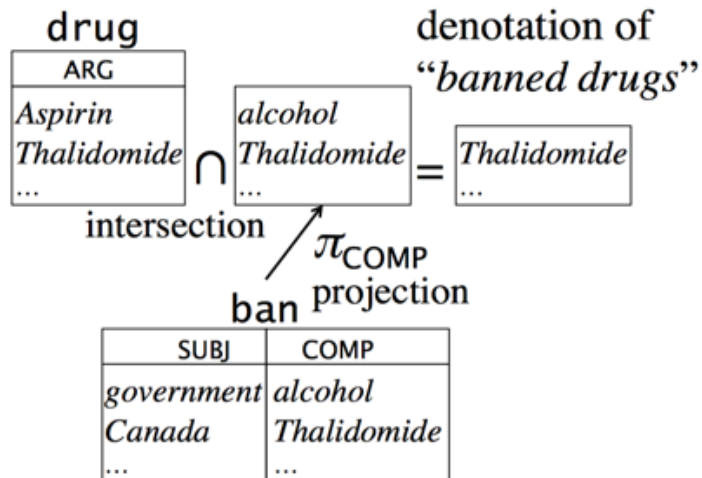
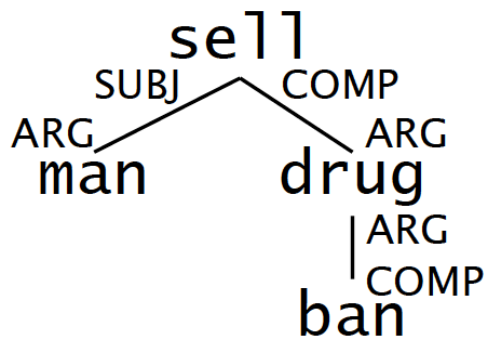
Vector-based DCS

[Tian+ ACL 2016]

Dependency-based Compositional Semantics

Composes meanings from denotations of words
(Liang+ 2013)

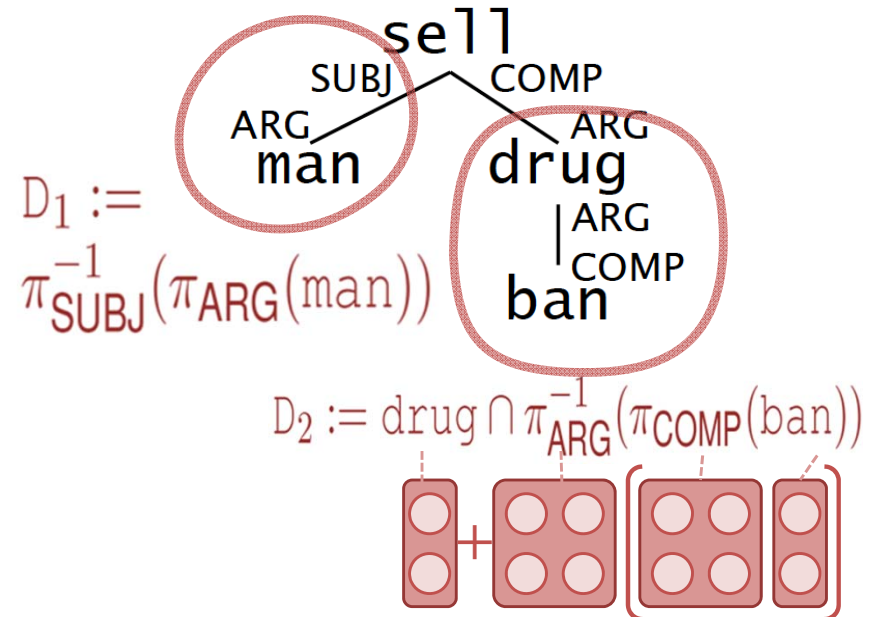
A man sells banned drugs.



Distributed representations of DCS

$$D_3 := \text{sell} \cap D_1 \cap \pi_{\text{COMP}}^{-1}(\pi_{\text{ARG}}(D_2))$$

A man sells banned drugs.



ベクトル表現を用いた構造的意味合成

Vector-based DCS

[Tian+ ACL 2016]

DCS-vec

bag-of-words

“those who arrested”



nearest neighbors

0.91 police

0.88 noble

0.87 ma

0.86 su

“those who were arrested”



nearest neighbors

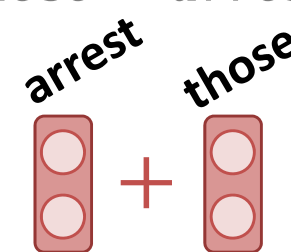
0.99 prisoner

0.98 noble

0.97 policeman

0.96 chen

“those” “arrested”



nearest neighbors

0.75 those

0.75 arrested

0.75 jailed

0.73 detained

Achieved SOTA w/o tuning in various benchmark tests

phrase similarity estimation

	AN	NN	VO
vecDCS	0.51	0.49	0.41
-no matrix	0.52	0.46	0.42
-no inverse	0.47	0.43	0.38
vecUD	0.44	0.46	0.41
GloVe	0.41	0.47	0.41
Grefenstette and Sadrzadeh (2011)	-	-	-
Blacoe and Lapata (2012):RAE	0.31	0.30	0.28
Grefenstette (2013a)	-	-	-
Paperno et al. (2014)	-	-	-
Hashimoto et al. (2014):Wadd _{nl}	0.48	0.40	0.39
Kartsaklis and Sadrzadeh (2014)	-	-	-

relation classification

vecDCS	81.2
-no matrix	69.2
-no inverse	79.7
vecUD	69.2
GloVe	74.1
Socher et al. (2012)	79.1
+3 features	82.4
dos Santos et al. (2015)	84.1
Xu et al. (2015)	85.6

sentence completion

vecDCS	50
-no matrix	60
-no inverse	46
vecUD	31
N-gram (Various)	39-41
Zweig et al. (2012)	52
Mnih and Teh (2012)	55
Gubbins and Vlachos (2013)	50
Mikolov et al. (2013a)	55

論理推論と意味類似度計算の統合

Combining Logical Inference and Textual Similarity

- 含意関係認識 (RTE): T ならば H といえるか？

Recognizing Textual Entailment (RTE): Does T entail H ?

T: *There was no man sawing logs in the yard.*

$\neg \exists x \exists e (\text{man}(x) \ \& \ \text{saw}(e) \ \& \ \text{past}(e) \ \& \ \text{subj}(e, x) \ \& \ \text{obj}(e, \text{logs}) \ \& \ \text{in}(e, \text{yard}))$



深い意味解析＋論理推論

Deep Parsing + Logical Inference

H: *Each worker wasn't cutting wood in the garden.*

$\forall x(\text{worker}(x) \rightarrow \neg \exists e (\text{cut}(e) \ \& \ \text{past}(e) \ \& \ \text{subj}(e, x) \ \& \ \text{obj}(e, \text{wood}) \ \& \ \text{in}(e, \text{garden})))$

- 文の意味類似度計算: T は H と意味的に類似しているか？

Semantic Textual Similarity (STS): How similar T and H are?

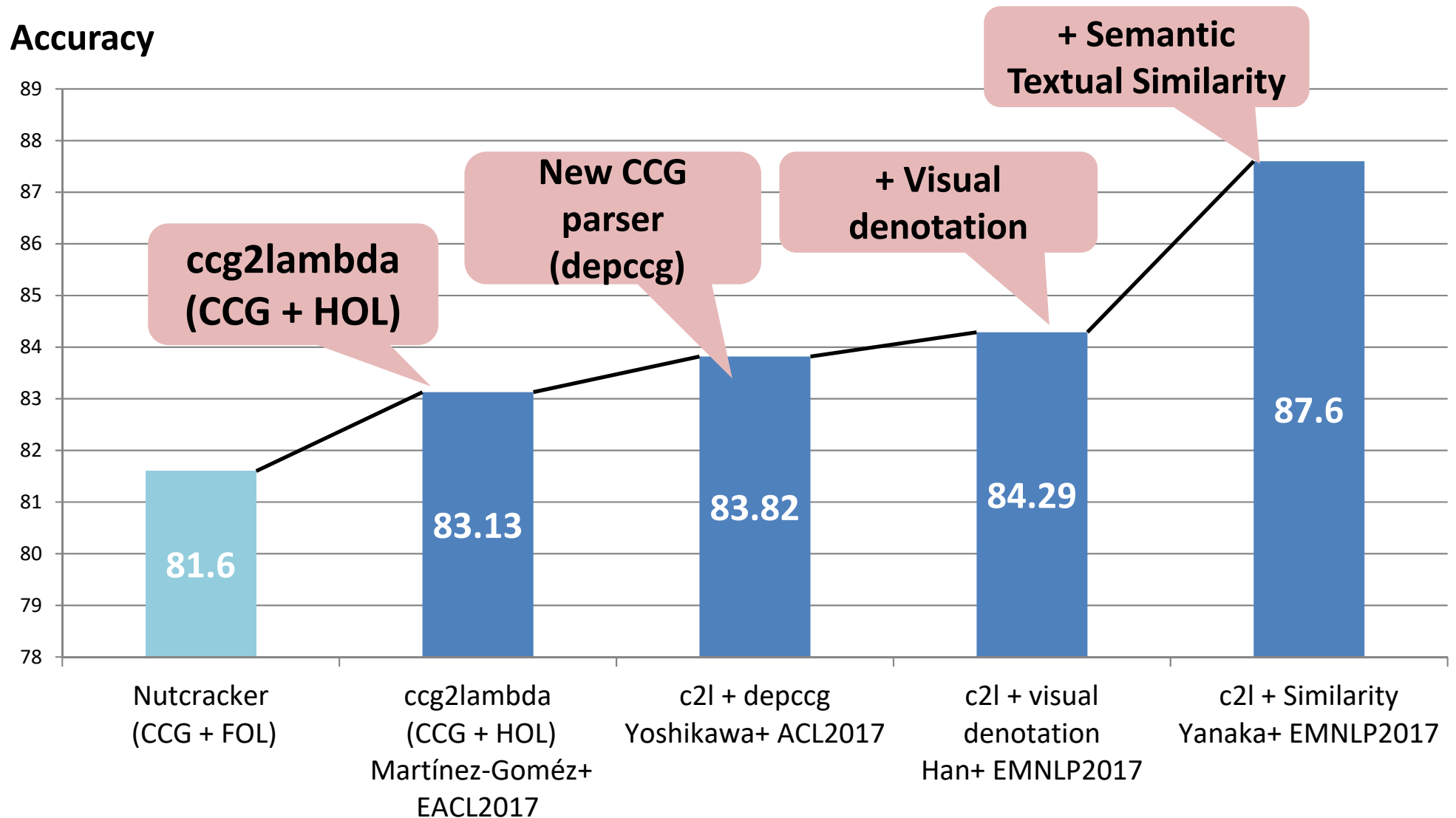
T: *There was no man sawing logs in the yard*

H: *Each worker wasn't cutting wood in the garden*

シソーラス＋内容語のベクトル表現

Thesaurus + Vector representations for words

RTE Experiments (SICK dataset)



実課題への応用

Application of NLP to Real World Tasks

Complaint Purchase Center

(不満買取センター)

Complaints about **Hamburger Shops**

ハンバーガーはおいしいけど、

hamburger tasty

ソースが多すぎて、口の周りが必ず汚れる。

sauce much mouth mouth always dirty

ソース多すぎて食べにくい。

sauce too much hard to eat

テリヤキソースとマヨネーズが多すぎて

teriyaki sauce mayo too much

食べにくいのが超不満!

hard to eat very discontent

ソースが垂れて最後に残り過ぎてもったいない。

sauce drool last left too much wasteful

日曜日の昼間は喫煙席は、いらないと思います。

Sunday daytime smoking area unnecessary

店内タバコ臭い。喫煙席デカすぎ。

shop tobacco stink smoking area too big



Complaint Purchase Center

(不満買取センター)

Insight Tech

Complaints about Hamburger Shops

ハンバーガーはおいしいけど、

hamburger tasty

ソースが多すぎて、口の周りが必ず汚れる。

sauce much mouth mouth always dirty

ソース多すぎて食べにくい。

sauce too much hard to eat

テリヤキソースとマヨネーズが多すぎて

teriyaki sauce mayo too much

食べにくいのが超不満!

hard to eat very discontent

ソースが垂れて最後に残り過ぎてもったいない。

sauce drool last left too much wasteful

日曜日の昼間は喫煙席は、いらんと思います。

Sunday daytime smoking area unnecessary

店内タバコ臭い。喫煙席デカすぎ。

shop tobacco stink smoking area too big



でてきたり
出てしまい
く-ガ-出てしまい

はみ出して
はみ出て
はみ出て来て
はみ出る

少ない
少なく
少な過ぎ
増やしてほしい

たっぶり
たっぶりな

残り過ぎて
残る
残ってしまう

多すぎる
多い
多すぎて

こぼれます
こぼれやすい
こぼれやすくて
飛び散り
く-ガ-こぼれて
こぼれて
こぼれる

因果関係グラフ Causality Graph

query: 手間がかかる depth: 1 edge: 20

筋が多く脂肪がとり除きにくいので
 ? ?
 使うと
 use
 リサイクルに出さなきゃいけないので
 この木綿豆腐を買ってます。
 レンジで温めなきゃいけないので
 where ?
 タネがあるので、
 there are seeds

手間がかかる ≠ 手間が省ける
 require a lot of trouble save the trouble

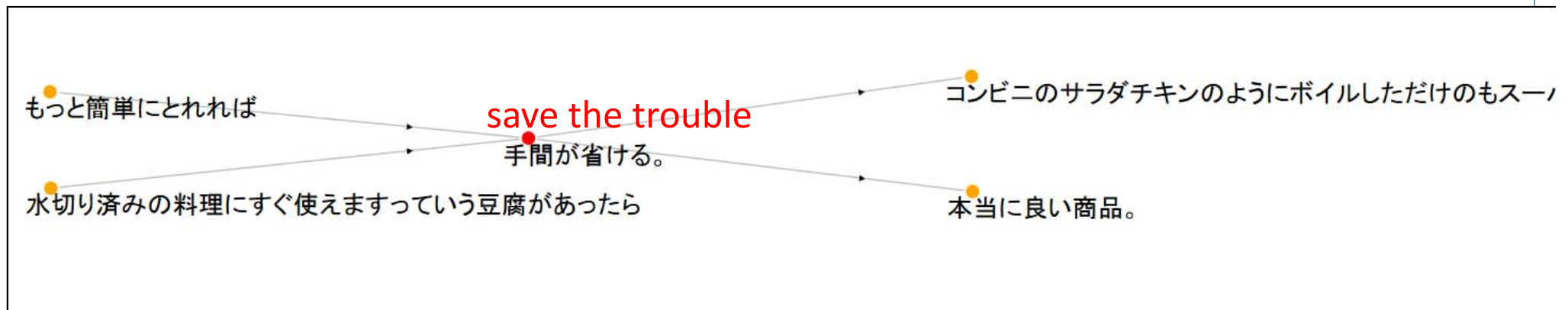
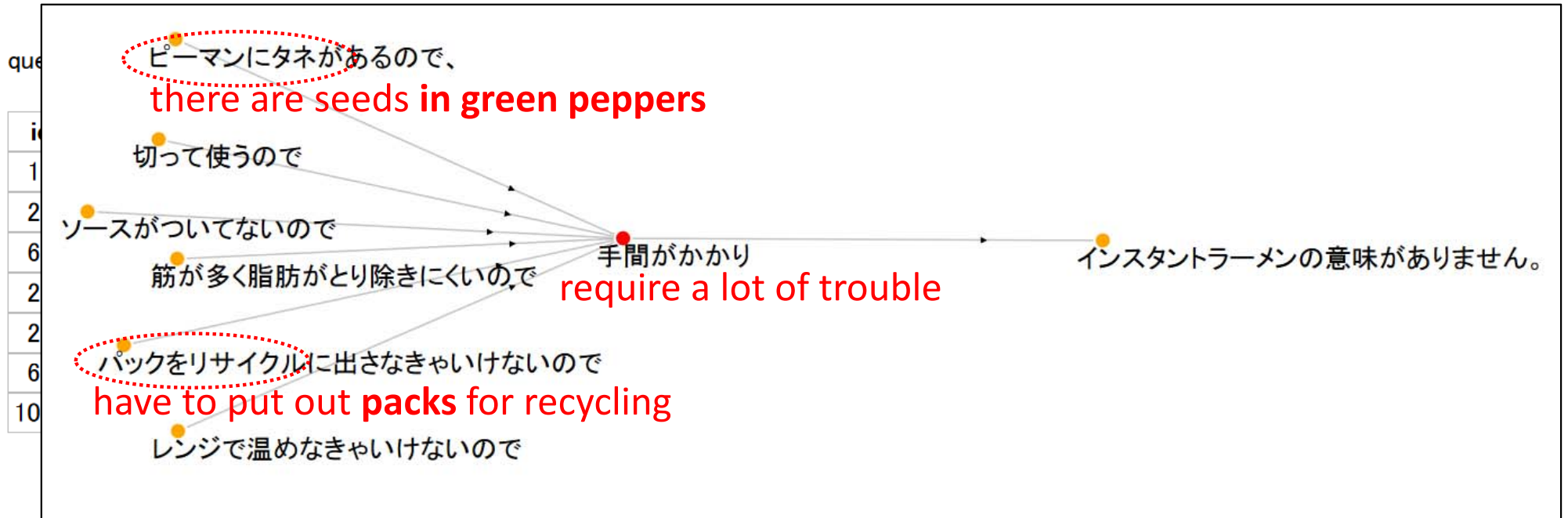
[cid:34]
 手間がかかり [3]
 手間が省ける。 [1]
 料理の際に手間がかかる。 [1]
 キヌアも五穀も茹でる手間がはぶけて [1]
 調理に手間がかかる。 [1]
 調理する手間がはぶけるので、 [1]

手間がかかり

因果関係グラフ Causality Graph

w/ 頑健な同義・反義判定、節の類似度計算、省略解析

handling synonym/acronym, phrasal similarity, zero anaphora



行政サービスの対話ボットシステム

Chatbot for Administrative Service

(Collaborating with **LINE** Corporation)

行政ホームページ
FAQs on Local Gov Sites

対話フローチャート生成
Dialogue Flowchart

粗大ごみの出し方について知りたい。
No.13341

✓ 回答

粗大ごみの収集は有料です。
金属を含むものは30cm以上、金属を含まないものは50cm以上が粗大ごみです。

粗大ごみの出し方は、次の通りです。

- 1 収集申込みをする
- 2 粗大ごみ処理券を購入する
- 3 収集日に粗大ごみ処理券を貼って出す

1 収集申込みをする

出す粗大ごみの一番長いところを測ってからお申し込みください。
お申し込みの際に、処理手数料、収集日、受付番号を案内いたします。
また、収集場所の確認を行います。

お住まいの地域別収集日の3日前（12月31日～1月3日、土、日曜日を除く）までにお申し込みください。申込みが多い場合は、次回の収集日になる場合がありますので、お早めにお申し込みください。

2 粗大ごみ処理券を購入する

粗大ごみ処理手数料は長さなどにより、200円、500円、1,000円の3区分があります。
申込みの際に確認した金額分の「川崎市粗大ごみ処理券」を購入してください。
「粗大ごみ処理手数料の減免の申請（美化活動や生活保護世帯など）」は生活環境事業所に直接お問い合わせください。
（管轄の生活環境事業所は下記の関連ホームページをご参照ください。）

購入先
市内のコンビニエンスストア、市内の郵便局（ゆうちょ銀行）

粗大ごみの出し方について知りたい。

「粗大ごみの出し方について知りたい。」についてお答えします

粗大ごみの収集は有料です。
金属を含むものは30cm以上、金属を含まないものは50cm以上が粗大ごみです。
どれが聞きたいですか

- 1 収集申込みをする
- 2 粗大ごみ処理券を購入する
- 3 収集日に粗大ごみ処理シールを貼って出す

編集 追加 削除

1 収集申込みをする

1 収集申込みをするについてお答えします
出す粗大ごみの一番長いところを測ってからお申し込みください。
お住まいの地域別収集日の3日前（12月31日～1月3日、土、日曜日を除く）までにお申し込みください。
申込みが多い場合は、次回の収集日になる場合がありますので、お早めにお申し込みください。

編集 追加 削除

ユーザー質問とのマッチング
Flexible Match with User's Query

同義語: 住所 = 所在地,
Synonym ゴミを出す = ゴミを捨てる

固有名詞: 防災無線がよく聞こえ
Named Entity ません → 防災テレホン
サービスについて知りたい

意図推定: 運動不足を解消したい
Intension → プールを探しています

2 粗大ごみ処理券を購入する

2 粗大ごみ処理券を購入するについてお答えします
粗大ごみ処理手数料は長さなどにより、200円、500円、1,000円の3区分があります。
申込みの際に確認した金額分の「川崎市粗大ごみ処理券」を購入してください。
「粗大ごみ処理手数料の減免の申請（美化活動や生活保護世帯など）」は生活環境事業所に直接お問い合わせください。

編集 追加 削除

行政サービスの対話ボットシステム

Chatbot for Administrative Service

(Collaborating with **LINE** Corporation)

対話フローチャート生成
Dialogue Flowchart

Chatbot による対話例
Chatbot System



粗大ごみの出し方について知りたい。

「粗大ごみの出し方について知りたい。」についてお答えします
粗大ごみの収集は有料です。
金属を含むものは30cm以上、金属を含まないものは50cm以上が粗大ごみです。
どれが聞きたいですか

- 1 収集申込みをする
- 2 粗大ごみ処理券を購入する
- 3 収集日に粗大ごみ処理シールを貼って出す

編集 追加 削除

1 収集申込みをする

1 収集申込みをするについてお答えします
出す粗大ごみの一番長いところを測ってからお申し込みください。
お住まいの地域別収集日の3日前（12月31日～1月3日、土、日曜日を除く）までにお申し込みください。
申込みが多い場合は、次回の収集日になる場合がありますので、お早めにお申し込みください。

編集 追加 削除

2 粗大ごみ処理券を購入する

2 粗大ごみ処理券を購入するについてお答えします
粗大ごみ処理手数料は長さなどにより、200円、500円、1,000円の3区分があります。
申込みの際に確認した金額分の「川崎市粗大ごみ処理券」を購入してください。
「粗大ごみ処理手数料の減免の申請（美化活動や生活保護世帯など）」は生活環境事業所に直接お問い合わせください。

編集 追加 削除

既読 午後 9:35 粗大ゴミを出したい

「粗大ごみの出し方について知りたい。」についてお答えします
粗大ごみの収集は有料です。
金属を含むものは30cm以上、金属を含まないものは50cm以上が粗大ごみです。
粗大ごみの出し方は、次の通りです。
1 収集申込みをする
2 粗大ごみ処理券を購入する
3 収集日に粗大ごみ処理シールを貼って出す
どれが聞きたいですか

午後 9:35

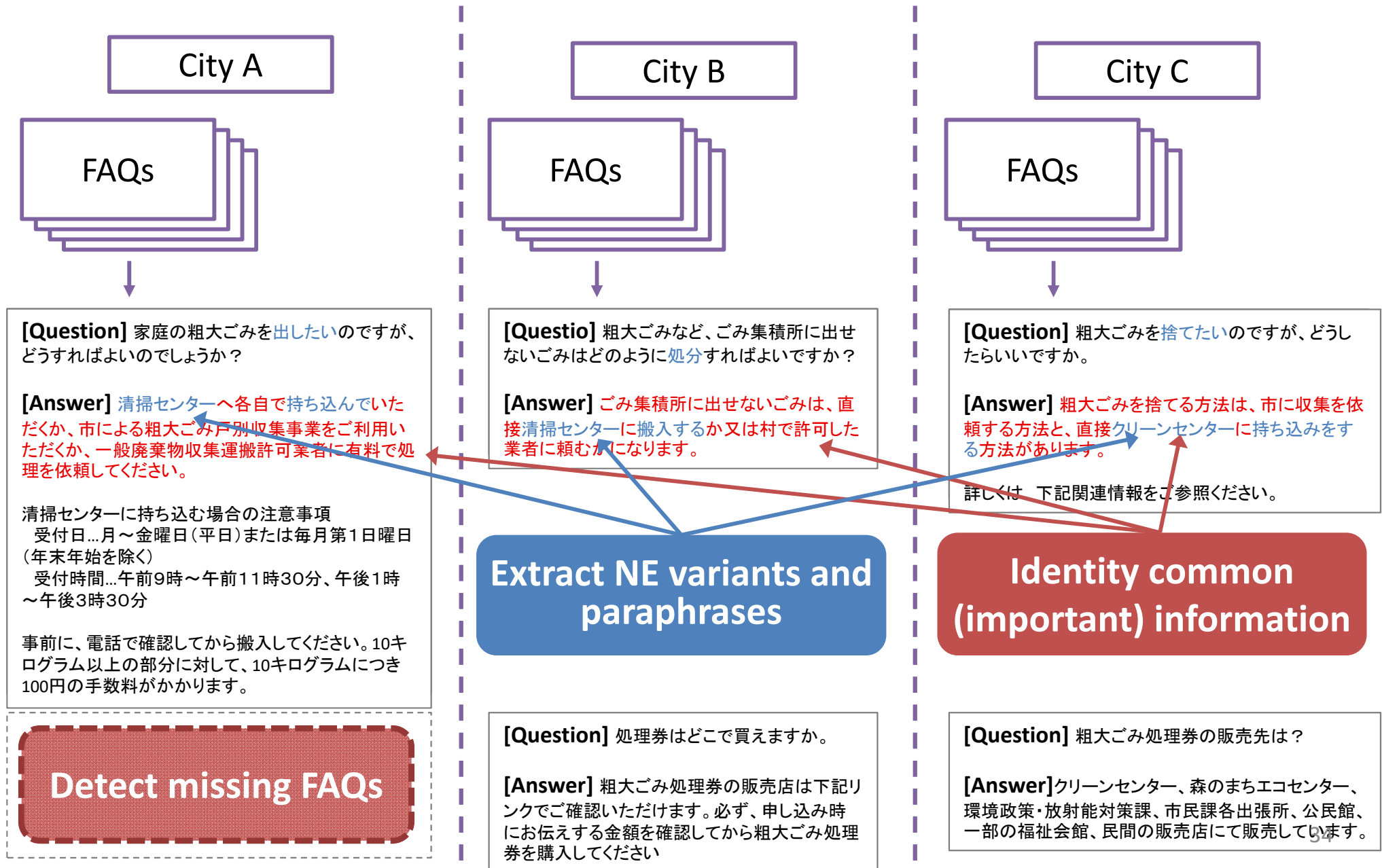
既読 午後 9:36 収集申込み

「1 収集申込みをする」についてお答えします
出す粗大ごみの一番長いところを測ってからお申し込みください。
お申し込みの際に、処理手数料、収集日、受付番号を案内いたします。
また、収集場所の確認を行います。
お住まいの地域別収集日の3日前までにお申し込みください。
申込みが多い場合は、次回の収集日になる場合がありますので、お早めにお申し込みください。

33% 午後 9:36

FAQ mining (Inui G)

Comparative analysis of FAQs on local gov sites



RTE for Question Answering

(Bekki G)

行政テキスト Local gov. texts

金属を含むものは30cm以上、金属
まないものは50cm以上が粗大ごみ

Any piece of garbage not containing metal
over 50cm in size is categorized as oversized

知識 CCGによる統語・意味解析
意味表現へのマッピング
Semantic parsing, Mapping to Formulas
プラステックの引き出しは粗大ごみではない。
Plastic does not contain metal.

質問文 Question sentence

40センチぐらいのプラスチックの引き出し
は粗大ごみですか？

Is a plastic drawer of about 40cm in size
categorized as oversized garbage?

ccg2lambda

<https://github.com/mynlp/ccg2lambda>

高階論理の意味表現 (HOL representation)

(SR1) $\text{contain}(x, \text{metal}) \ \& \ (\text{size}(x) \geq 30\text{cm}) \rightarrow \text{oversized_g}(x)$

(SR2) $\text{contain}(x, \text{metal}) \ \& \ (\text{size}(x) < 30\text{cm}) \rightarrow \neg \text{oversized_g}(x)$

(SR3) $\neg \text{contain}(x, \text{metal}) \ \& \ (\text{size}(x) \geq 50\text{cm}) \rightarrow \text{oversized_g}(x)$

(SR4) $\neg \text{contain}(x, \text{metal}) \ \& \ (\text{size}(x) < 50\text{cm}) \rightarrow \neg \text{oversized_g}(x)$

(KB) $\text{plastic}(x) \rightarrow \neg \text{contain}(x, \text{metal})$

論理推論＋
文意味類似度計算
Logical Inference + Textual Similarity

はい (yes): 粗大ごみである

(Ans-Pos) $\text{plastic}(x) \ \& \ \text{drawer}(x) \ \& \ (\text{size}(x) \approx 40\text{cm}) \rightarrow \text{oversized_g}(x)$

いいえ (no): 粗大ごみではない

(Ans-Neg) $\text{plastic}(x) \ \& \ \text{drawer}(x) \ \& \ (\text{size}(x) \approx 40\text{cm}) \rightarrow \neg \text{oversized_g}(x)$

Answer : **NO**

まとめ

Summary

- **分散意味表現**によるテキスト解析の深化

Distributional semantic representation improves text analysis

- 形態素・構文解析

Morphological and dependency analysis

- 構文を考慮した句類似度計算

Syntax-sensitive phrase similarity estimation

- 論理推論と意味類似度計算の統合

Combination of logical inference and textual similarity

- **実課題への応用**

Application to Real World Task

- 意見の集約

Opinion summarization

- 行政サービスの対話ボットシステム

Chatbot for administrative service of local gov