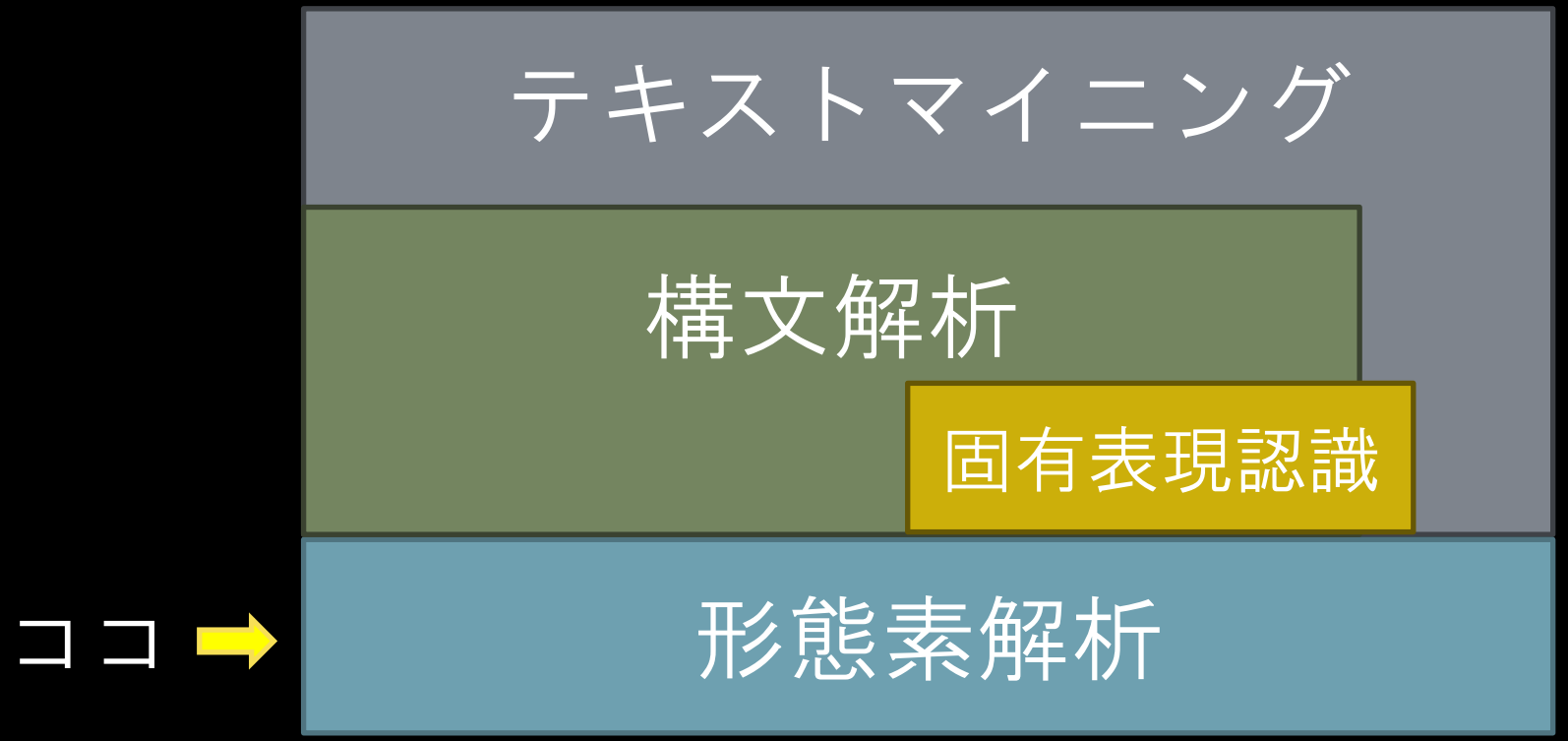
日本語未知語のテキストからの自動獲得

村脇 有吾　黒橋 禎夫
京都大学大学院情報学研究科

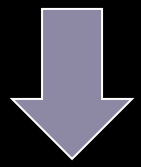
2011年7月7日
第1回テキストマイニングシンポジウム@日本IBM本社

前処理としての形態素解析



日本語の形態素解析

入力: 発表します



発表	名詞–サ変名詞
し	動詞–サ変動詞
ます	接尾辞–動詞性接尾辞

形態素解析の課題

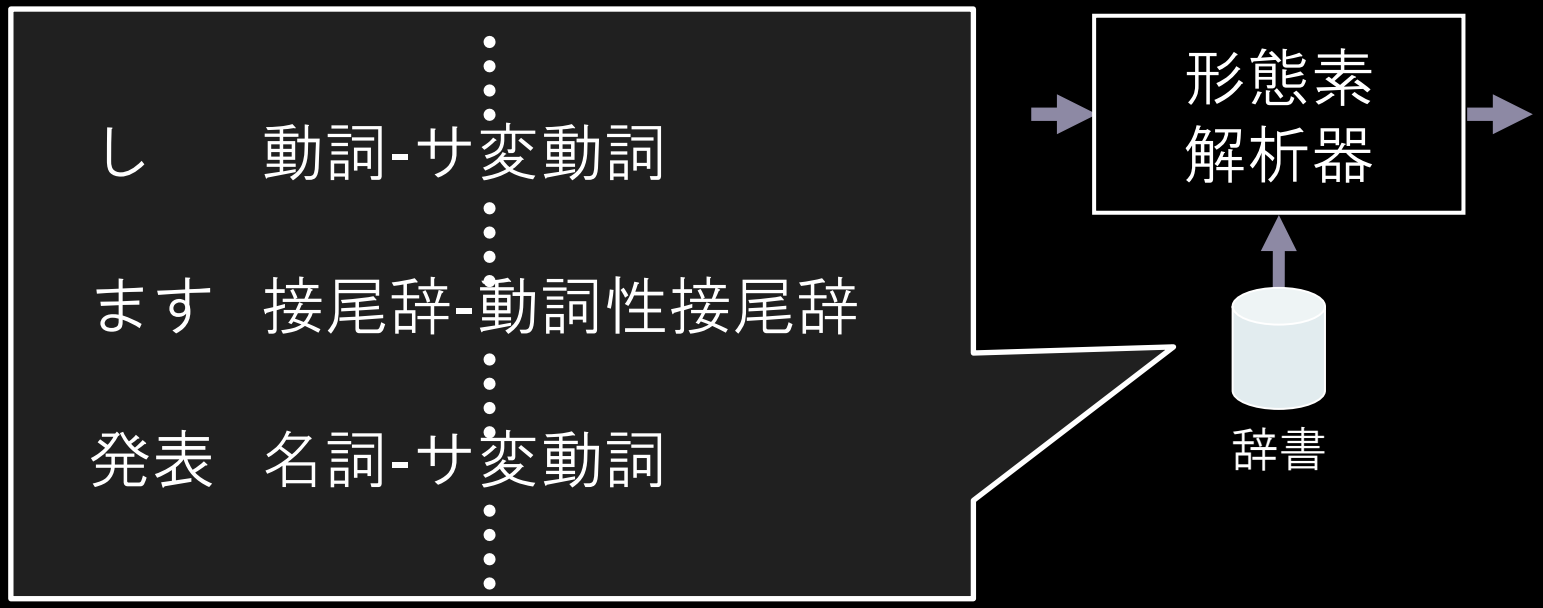
	文法	口語・方言	とっとっと？
一般の語彙	未知語		ググる ぜん動 (蠕動)
特殊な語彙	感動詞		いやっほー
	アスキーアート		(´･ω･`)
規則性を持った語彙・表記	連濁		夫婦げんか
	俗表記		あやιい
	オノマトペ		ほいほい

[笹野+ 2007]

[勝木+ 2011]

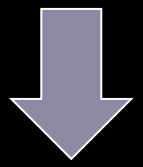
形態素解析における辞書引き

入力: 発表します



未知語による解析誤り

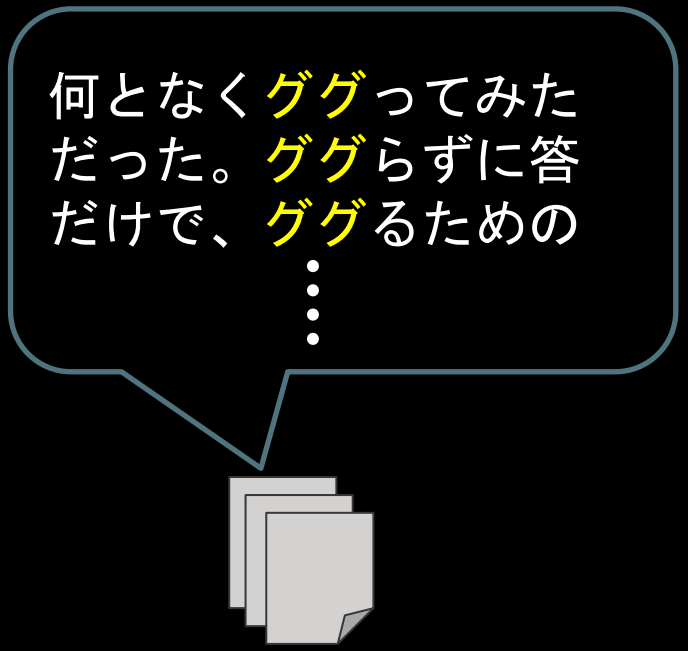
入力: ググったら



ググ	未定義語
ったら	助詞-副助詞

解決策: テキストからの自動獲得

- 知っておきたい未知語は解析したいテキストに何度か出現するもの
- 生テキストから未知語を獲得し、辞書に自動追加



自動獲得向けのタスク整理

問題		自動獲得のサブタスク		例
形態素の単位		複合名詞の分割		フェルミエネルギー ⇒ フェルミ + エネルギー
品詞	形態レベル	未知語検出	未知語同定 (未知語候補の列挙と選択)	ググる:ラ行動詞 さっぽろ:名詞-未分類
	意味レベル	名詞の意味分類		さっぽろ:名詞-地名
応用処理に用いる付加情報		異表記の認識		店舗= 店舗 キメる = 決める ??
		ドメイン分類 [Hashimoto+ 2008]		ベートーヴェン:文化・芸術

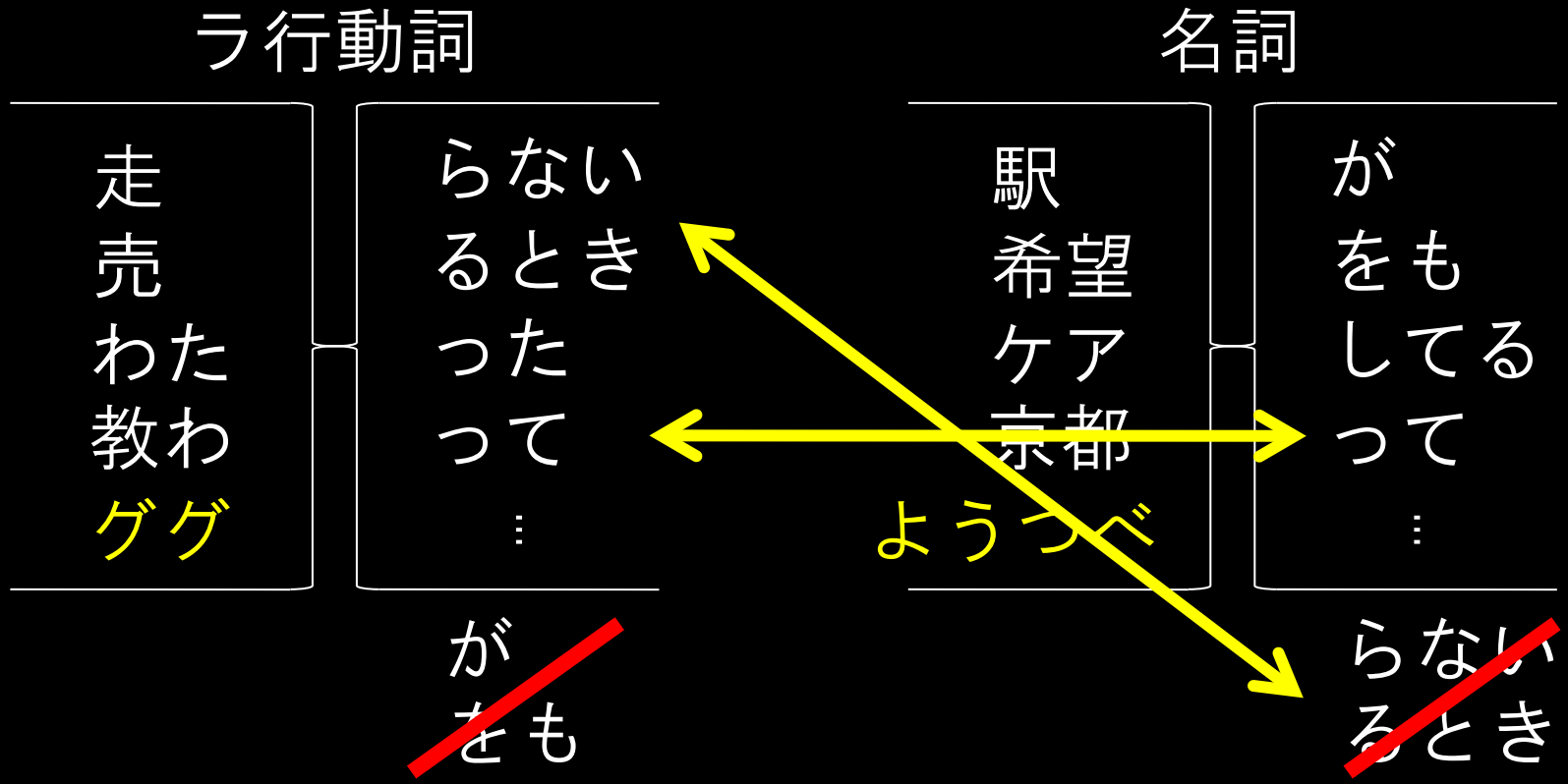
自動獲得の基本戦略

- 手がかりとしてメタ知識が必要
  - 形態素はテキスト中でどう振る舞うか
- メタ知識を自動構築
  - 文法・基本語彙を人手で整備済みであることを利用

メタ知識の例: 形態変化の規則性

	既知語		未知語	
ラ行動詞	走	–らない	ググ	–らない
	走	–るとき	ググ	–るとき
	走	–った	ググ	–った
	走	–って	ググ	–って
名詞	駅	–が	ようつべ	–が
	駅	–を	ようつべ	–を
	駅	–なら	ようつべ	–なら
	駅	–って	ようつべ	–って

形態変化の規則性は制約



形態論的制約を用いた未知語同定

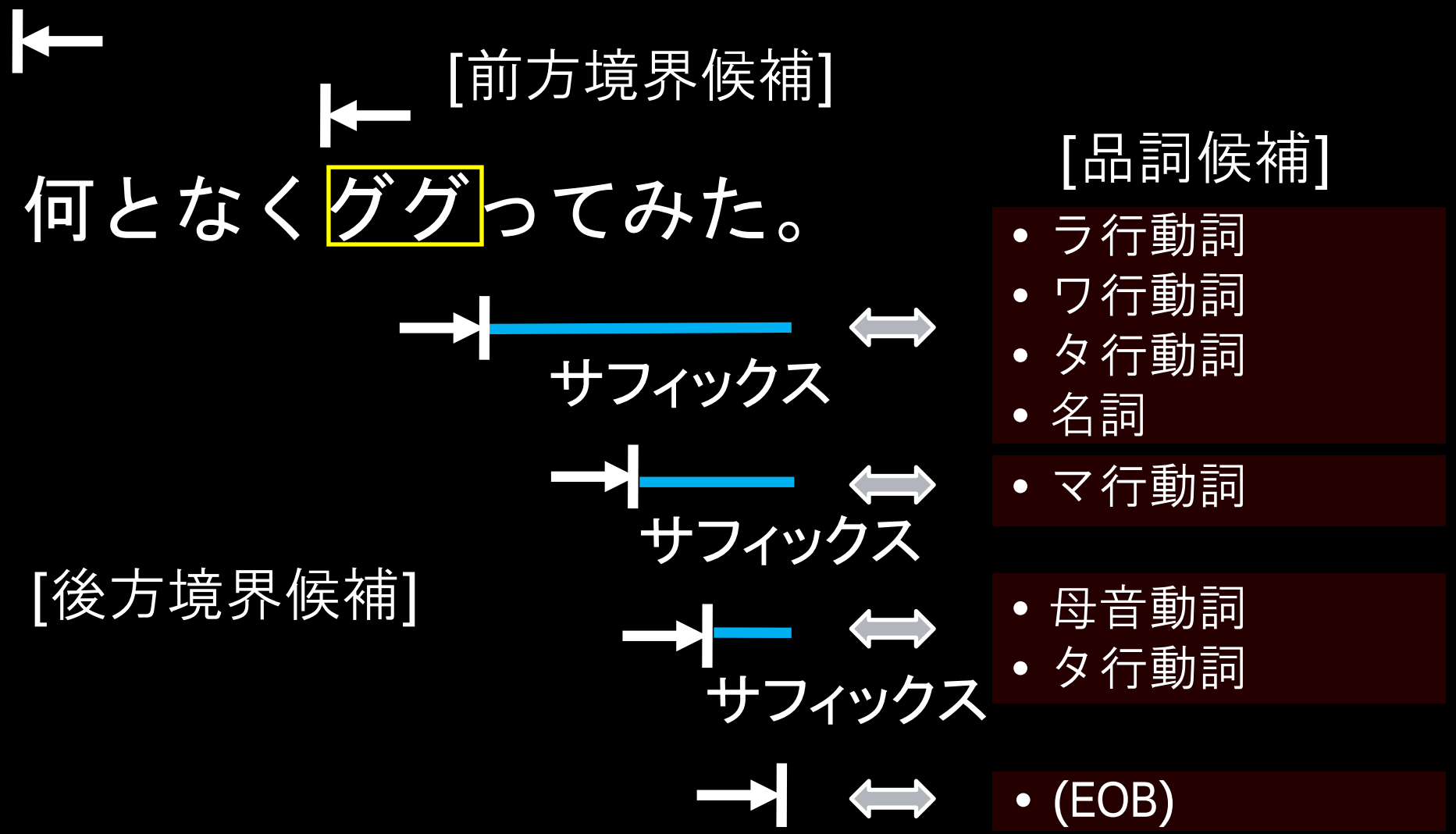
何となくググってみた ➡
- ラ行動詞,
- ワ行動詞,
- タ行動詞 or
- 名詞

だった。ググらずに答 ➡
- ラ行動詞

だけで、ググるための ➡
- ラ行動詞, or
- 母音動詞

⋮

参考: 候補列挙の一般化


[前方境界候補]
何となくググってみた。
[品詞候補]
サフィックス
サフィックス
サフィックス
[後方境界候補]
• ラ行動詞
• ワ行動詞
• タ行動詞
• 名詞
• マ行動詞
• 母音動詞
• タ行動詞
• (EOB)

獲得実験

高精度

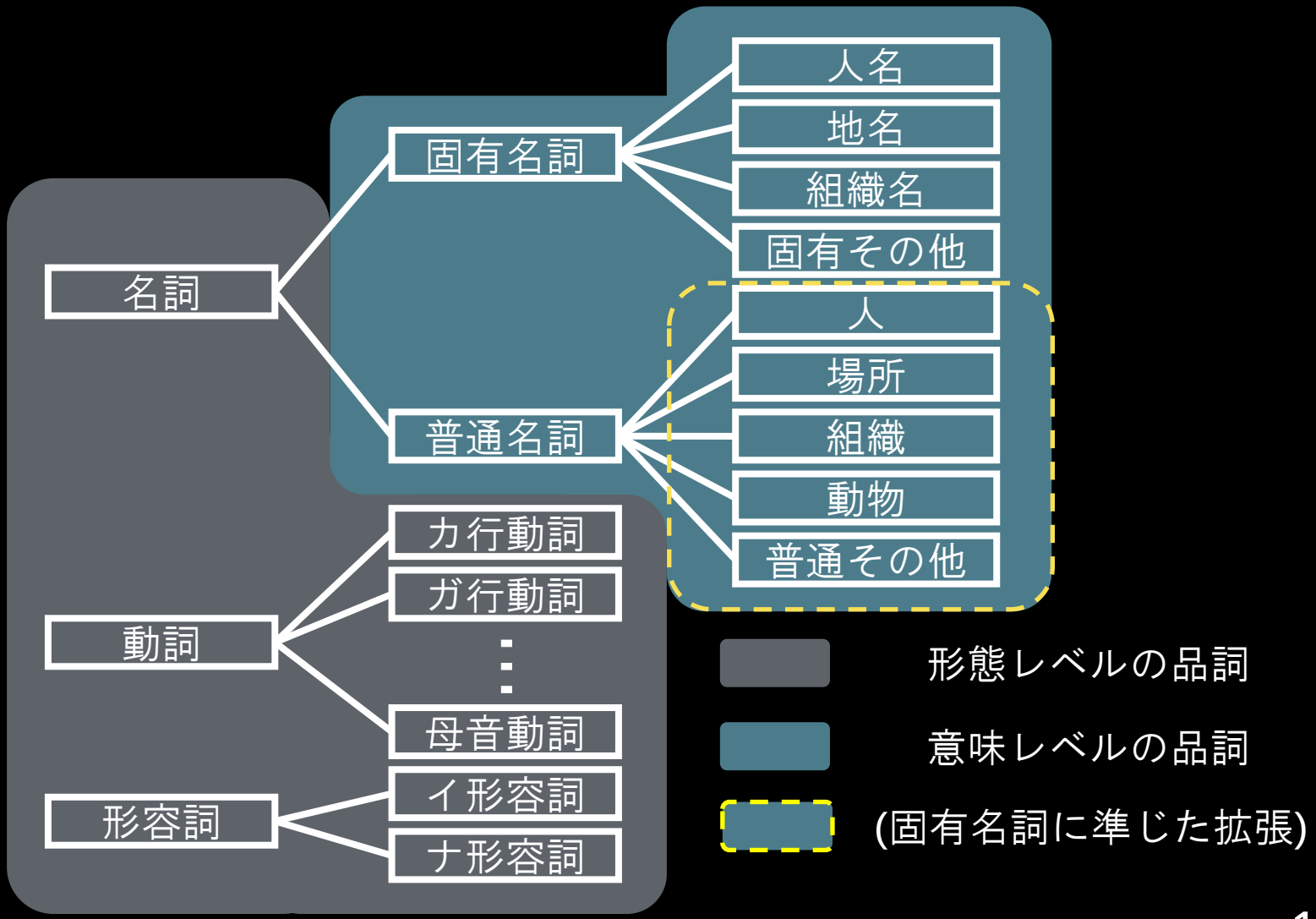
少数用例で同定できる

文書セット	精度	獲得時に比較した用例数	獲得未知語の例
捕鯨問題	187 / 190 (98.4%)	5	モラトリアム 混獲
赤ちゃんポスト	74 / 75 (98.5%)	4	円ブリオ基金 助産師
JASRAC	460 / 470 (97.9%)	5	ぱくる シャ乱Q
ツンデレ	215 / 221 (97.3%)	5	アキバ ドジっ娘

自動獲得向けのタスク整理

問題		自動獲得のサブタスク		例
形態素の単位		複合名詞の分割		フェルミエネルギー ⇒ フェルミ + エネルギー
品詞	形態レベル	未知語検出	未知語同定 (未知語候補の列挙と選択)	ググる:ラ行動詞 さっぽろ:名詞-未分類
	意味レベル	名詞の意味分類		さっぽろ:名詞-地名
応用処理に用いる付加情報		異表記の認識		店舗= 店舗 キメる = 決める ??
		ドメイン分類 [Hashimoto+ 2008]		ベートーヴェン:文化・芸術

品詞の2段階獲得


名詞
固有名詞
人名
地名
組織名
固有その他
普通名詞
人
場所
組織
動物
普通その他
動詞
カ行動詞
ガ行動詞
母音動詞
形容詞
イ形容詞
ナ形容詞
形態レベルの品詞
意味レベルの品詞
(固有名詞に準じた拡張)

品詞の2段階獲得

さ + っ + ぽ + ろ

… 南北線のさっぽろ駅のすぐそばに …
… 大通とさっぽろで通り越す。 …
[BOS] さっぽろを元気にする路面電車 …

さっぽろ:名詞-未分類

… 南北線のさっぽろ駅のすぐそばに …
… 大通とさっぽろを通り越す。 …
[BOS] さっぽろを元気にする路面電車 …

手がかり：語彙的選好

- Xを通り越す ⇒ 地名 or 場所？
- Xを遣わす ⇒ 人名 or 人？
- Xが多い ⇒ 普通名詞？
- どのX ⇒ 普通名詞？
- 2人のX ⇒ 人？

既知語のテキスト中での振る舞いから獲得し、獲得語 (未知語) に適用

固有名詞と普通名詞の区別がうまくいかない

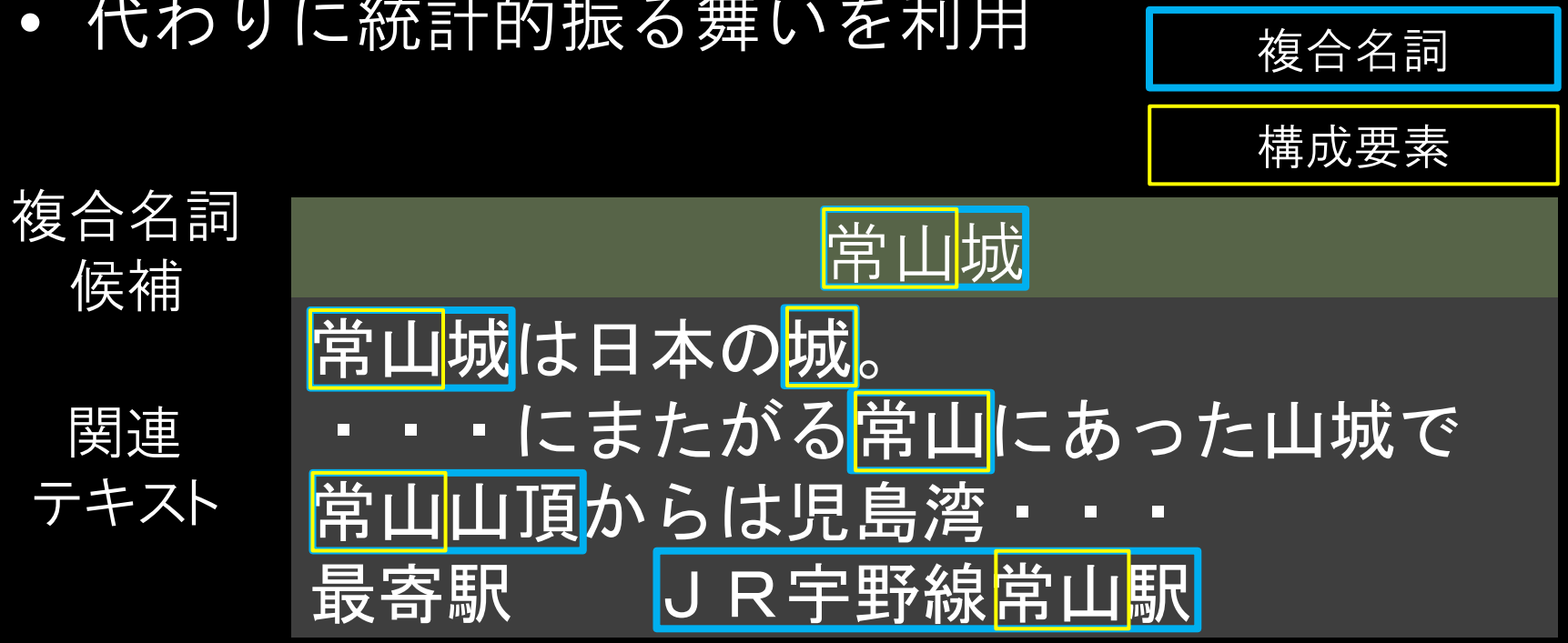
- 人名/人をその他から区別する手がかりは豊富:
  - Xが言う, Xに聞く, Xと暮らす
- 普通名詞を固有名詞から区別する手がかりは豊富:
  - 数値表現: 多くのX, <NUM>人のX, Xが増える
  - 属性表現: Xになる, Xを兼ねる
- 人名を人から区別する
  - 相沢X, X浩志
    - 架空の人名やオンライン・ハンドルには効かない
  - Xさん, X大統領
  - 人名専用の用言はない?

自動獲得向けのタスク整理

問題		自動獲得のサブタスク		例
形態素の単位		複合名詞の分割		フェルミエネルギー ⇒ フェルミ + エネルギー
品詞	形態レベル	未知語検出	未知語同定 (未知語候補の列挙と選択)	ググる:ラ行動詞 さっぽろ:名詞-未分類
	意味レベル	名詞の意味分類		さっぽろ:名詞-地名
応用処理に用いる付加情報		異表記の認識		店舗= 店舗 キメる = 決める ??
		ドメイン分類 [Hashimoto+ 2008]		ベートーヴェン:文化・芸術

頻度統計に基づく複合名詞分割

- 構成要素間に形態的マーカが存在しない
  - 常山 -Φ 城
- 代わりに統計的振る舞いを利用



まとめ

- テキストから未知語を自動獲得
- 人手による辞書登録を前提とした設計となっている。自動化向けにタスクを整理
- 獲得に必要な知識は、人手で登録した語彙を用いて自動構築し、未知語に適用