語義曖昧性解消における統語的依存関係の利用

羽鳥 潤・宮尾祐介・辻井潤一（東京大学）

導入

- 語義曖昧性解消（WSD）
  - テキスト中に登場する各多義語がどのような語義で用いられているかを判別する問題
  - 機械翻訳・情報検索・情報抽出などに応用される

入力文	The	man	destroys	public	confidence	in	banks
意味ラベル	the	男 人類	破壊する 損なう	公の 公共の	自信 信頼	in	銀行 土手

WordNet

- WordNet（Miller 1995）
  - 英語における最大の計算機可読辞書
  - 語義の定義だけでなく、語義の階層構造が定義されている。
- Supersense
  - WordNetの名詞・動詞が分類されている、一般的な意味カテゴリ
  - 名詞：person, group, ... (26種)
  - 動詞：change, creation, ... (15種)

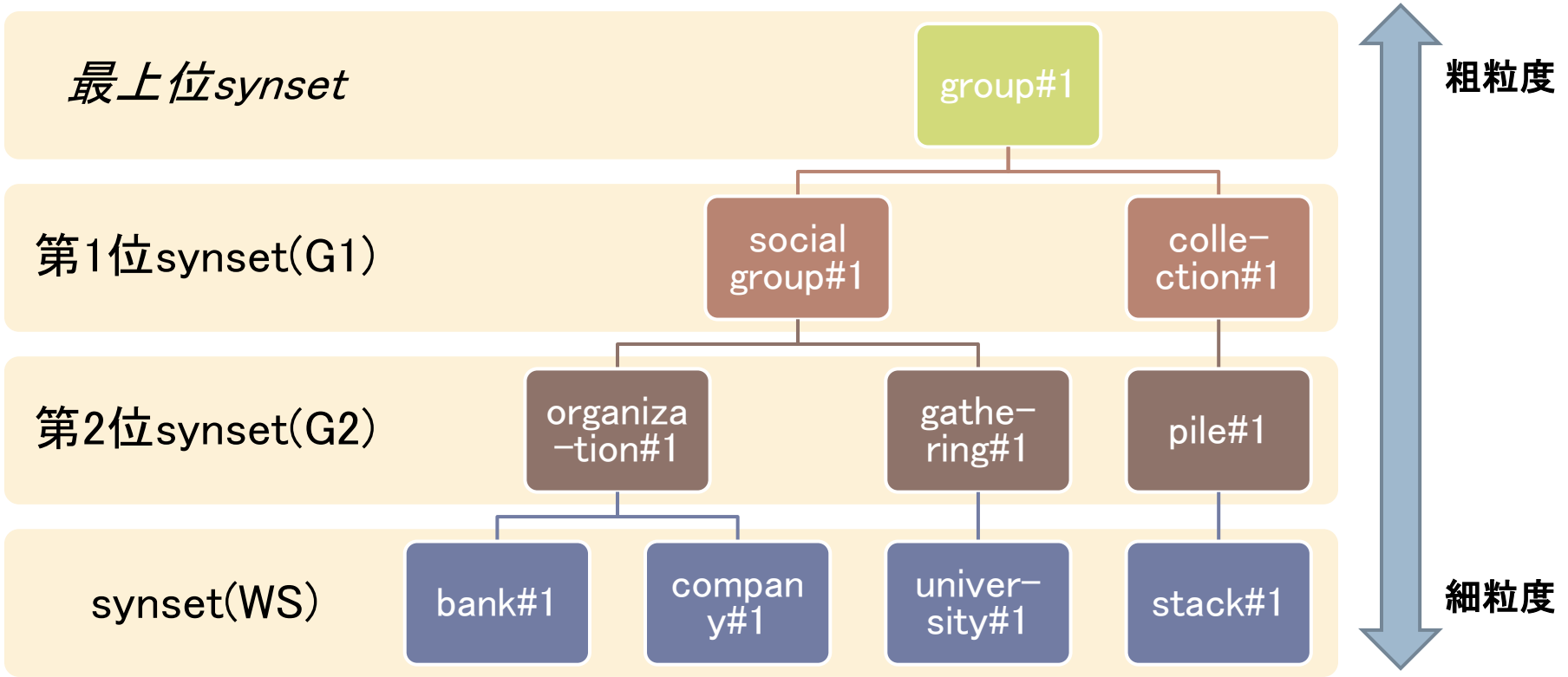
WordNet 2.0 の統計

品詞	単語数	語義数	平均語義数
名詞	114,648	141,690	1.23
動詞	11,306	24,632	2.17
形容詞	21,436	31,015	1.44
副詞	4,669	5,808	1.24
計	152,059	203,145	1.34

bankの第1義～第3義（全10義中）

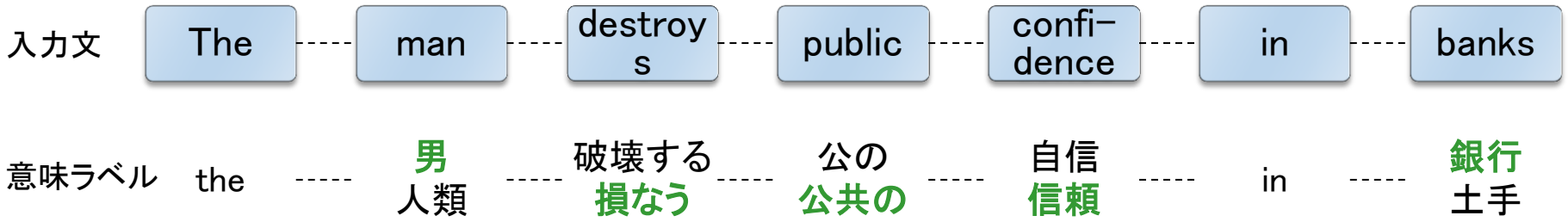
	意味	Synset	Super-sense
1	銀行	banking company#1 ,bank#1, ...	group
2	土手	bank#2	object
3	バンク	bank#3	possession

WordNet ー 語義の階層構造



従来の典型的なアプローチ

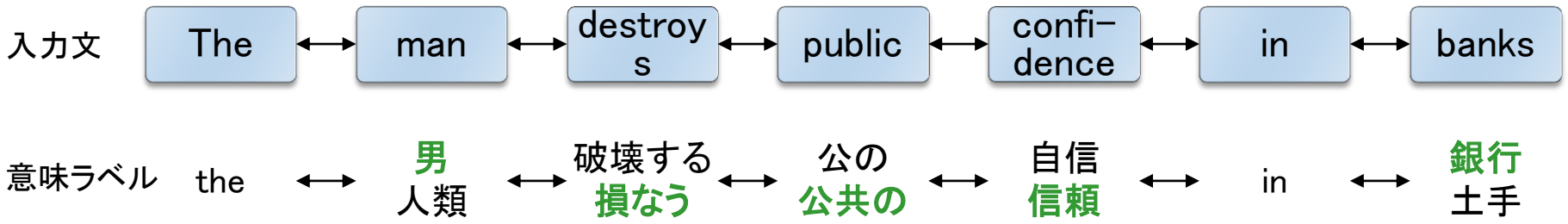
- 文脈情報を特徴量（素性）ベクトルに変換し、機械学習手法（k-NN, SVM, MEなど）を用いて、各語に対する独立した分類問題を解く。



- 素性：*文脈中に出現した単語、対象語の品詞・前後の単語など*
- 問題点
  - 語義依存性を考慮していない
  - 訓練データの与えられている語に対してしか適用できない
  - 深刻なデータスパースネス

系列ラベリングによるアプローチ

- WSDを入力単語列に対する系列ラベリング問題として解く
  - 文章中で隣り合う単語間の語義依存関係を利用できる
  - Ciaramita and Altun (2006)・Mihalcea et al. (2007) など



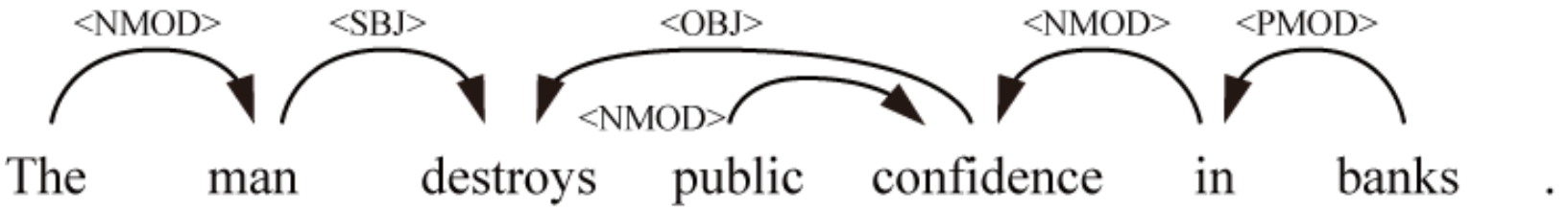
- 既知の問題
  - 長距離の依存関係を拾う事が出来ない
  - 意味的な繋がりのない、不適切な依存関係を拾ってしまう可能性

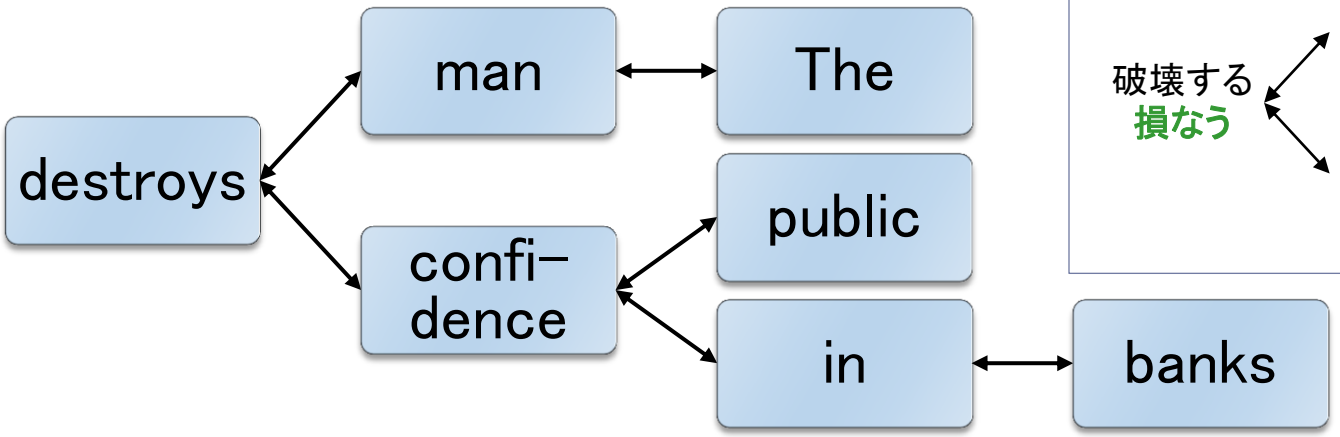
提案手法の特徴

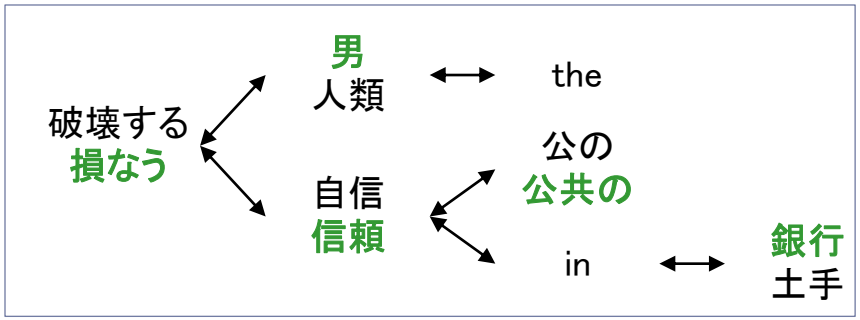
1. 統語的な語義依存関係の考慮

- 係り受け関係にある主辞と従属句の間に語義の依存関係を仮定
- 語義の依存関係をより適切に表現できると考えられる。

構文解析器の出力



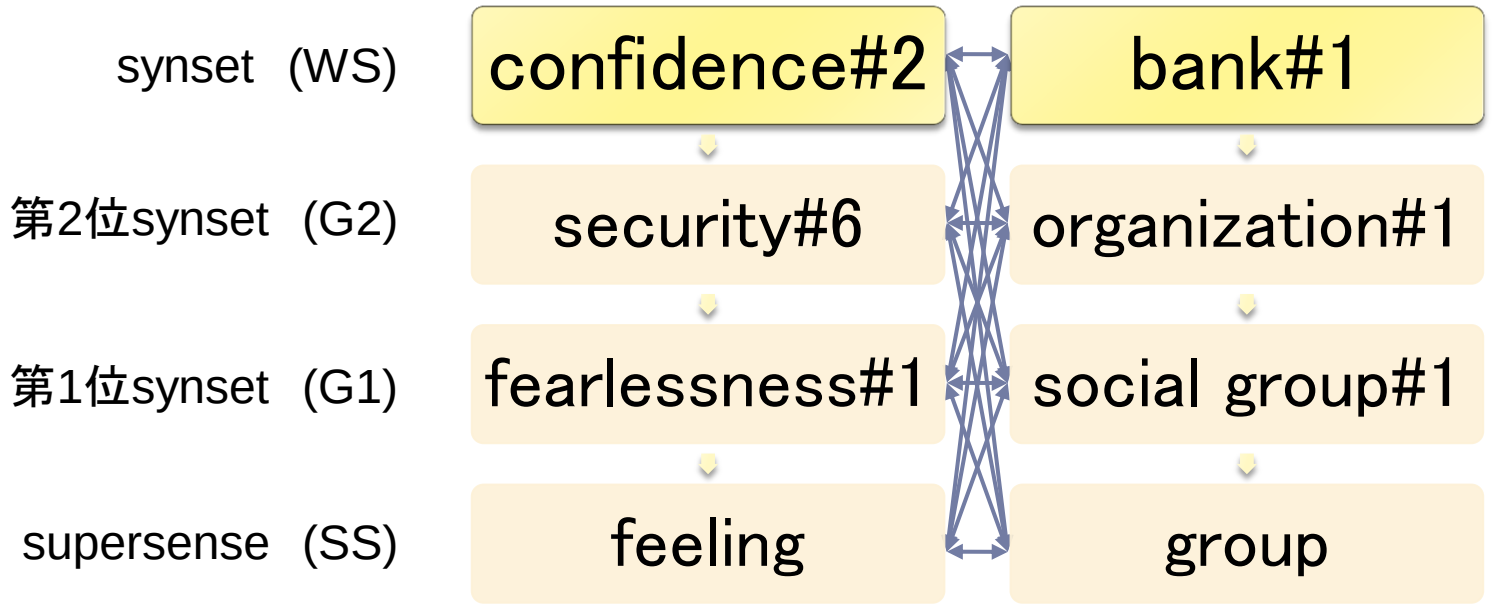




提案手法の特徴

2. 細／粗粒度意味ラベルの同時使用

- 精度を維持したまま、データスパースネス問題を軽減できる
- 訓練データが与えられていない語に対しても語義を推定できる



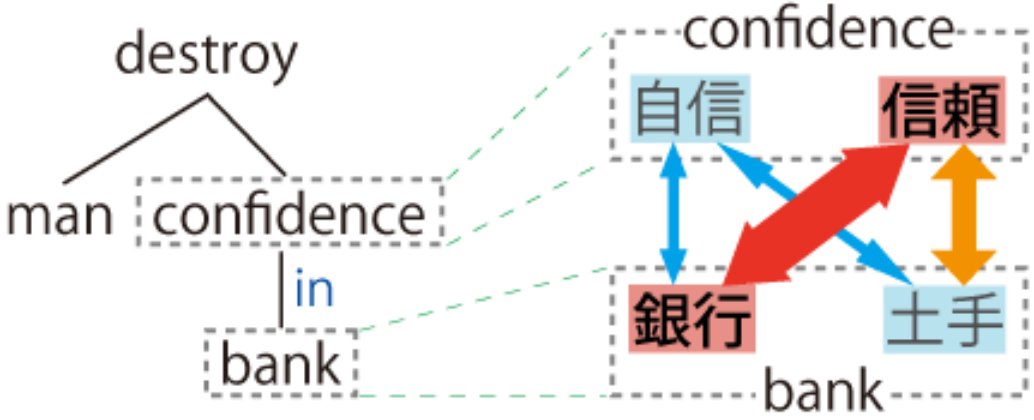
木構造上の条件付き確率場(T-CRF)

- 条件付き確率場(CRF)

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left\{ \sum_{v \in \mathcal{C}_1} \sum_{k} \lambda_k g_k(v, \mathbf{y}_v, \mathbf{x}) + \sum_{u,v \in \mathcal{C}_2} \sum_{j} \lambda_j h_j(u, v, \mathbf{y}_u, \mathbf{y}_v, \mathbf{x}) \right\}$$

x: 入力列, y: 出力ラベル列
u,v: 節点
g: 節点素性, h: 枝素性
λ: 各素性の重み
Z: 分配関数

- 語義依存性の強さは、枝素性に対する重みとして学習される。



"The man destroys confidence in banks."の解析例。
矢印の太さは語義の依存関係の強さを示す。

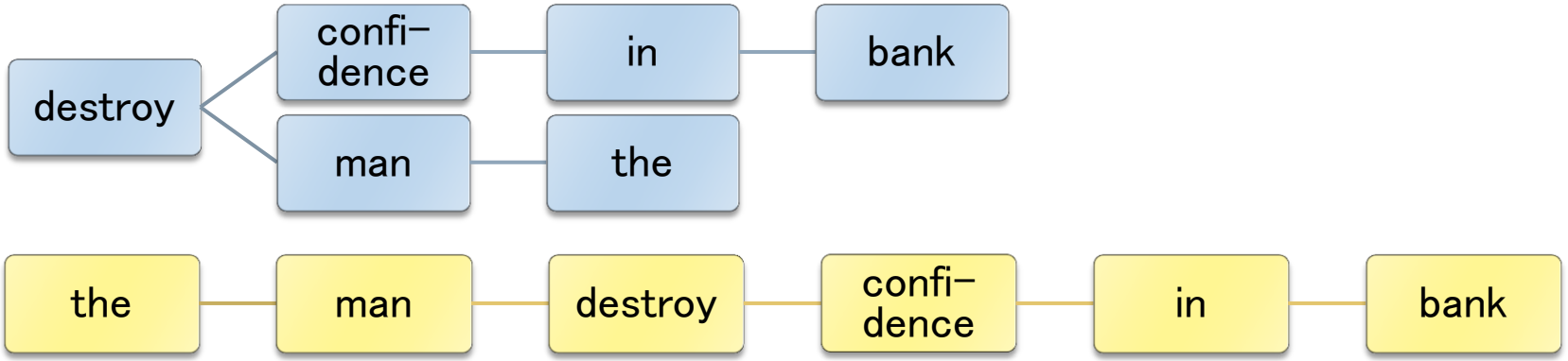
本研究の位置付け

- 語義依存関係の利用
  - Mihalcea and Faruque (2004) で線形の依存関係は利用されているが、統語的な語義依存関係を利用した点、語義依存関係の寄与を陽に示した点は初。

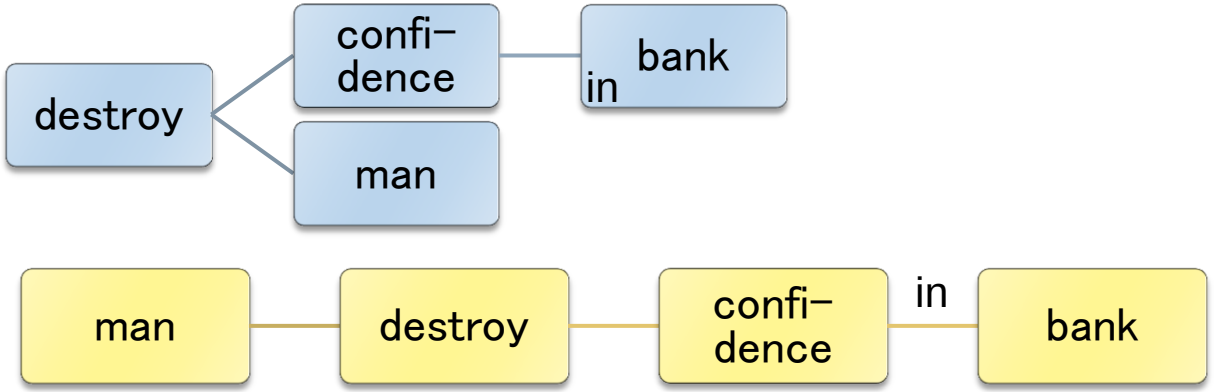
- 粗粒度意味ラベルの使用
  - Mihalcea et al. (2007)などの例があるが、細／粗粒度ラベルを同時に確率モデルに組み込んだモデルは初。
    - CRFの利用により、任意の数のラベルを同時に組み込む事が出来る。

モデル ― 木構造モデル／線状鎖モデル

- 統語的な語義依存関係の有効性を調べるために、
  木構造モデル・線状鎖モデルの両モデルを構築して比較する。

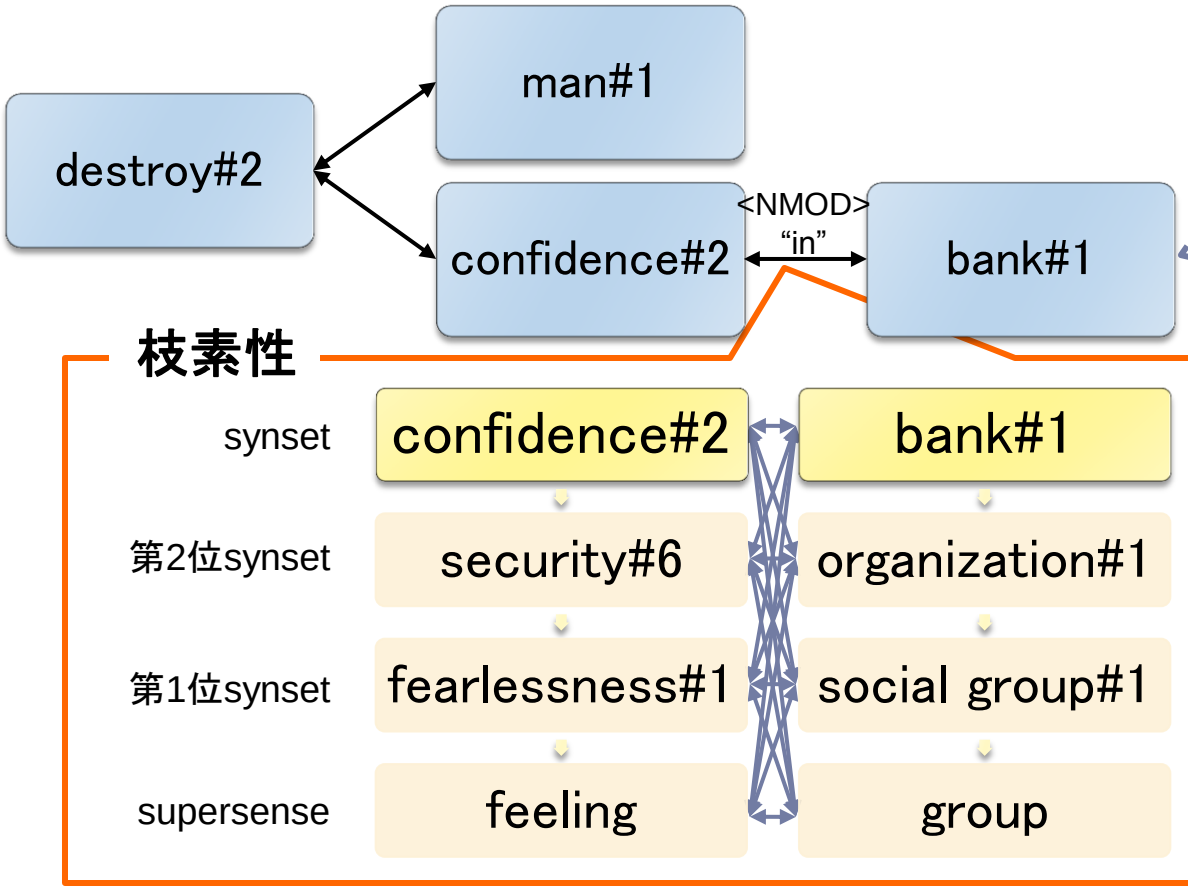


- 但し、事前に対象語以外を除去して間引きを行っておく。



モデル ー 素性

- 節点素性はLee et al. (2004)・Ciaramita et al. (2006) に基づく
- 枝素性として語義依存素性を導入



節点の素性
- 単語の語形・品詞
- 前後の文脈中に出現する語
- 前後の単語の原形・品詞
- 親子の単語の語形・品詞
- 語義頻度

3種類の枝素性

confidence#2—bank#1
confidence#2—in—bank#1
confidence#2—NMOD—bank#1
（合計 $3 \times 4^2 = 48$ 種類）

実験条件

- データセット
  - SEM: SemCor (Miller et al., 1993)
  - SE2/3: SENSEVAL-2/3 English all-words task data (Snyder et al., 2004)
- 実験条件
  - SEMを5つのセットに分割し、5回交差検定を行う
  - development セットを用いて、CRF のL2正規化項をチューニング
- 有意性検定：McNemar's test による(有意水準0.05)

	# タグ付 語	# タグ付名詞/動詞
SEM	189,667	135,123
SE2	2,259	1,567
SE3	1,978	1,617

- WordNetのsynsetレベルでWSDを行うSynsetモデルの他に、supersenseレベルのWSDを行うSupersenseモデルを構築して精度を比較した。

結果 ー 語義依存関係の寄与

- 木構造・線状鎖モデルとも、語義依存関係の導入により精度が有意に向上。

語義依存関係による精度向上

モデル	木構造			線状鎖		
	SEM	SE2	SE3	SEM	SE2	SE3
Supersenseモデル（依存）	83.60%	78.93%	80.24%	83.69%	78.44%	80.19%
Supersenseモデル（非依存）	83.39%	78.24%	79.89%	83.39%	78.24%	79.89%
差	+0.21%	+0.69%	+0.35%	+0.29%	+0.20%	+0.30%
ベースライン	83.02%	76.26%	78.48%	83.02%	76.26%	78.48%
Synsetモデル（依存）	77.46%	68.51%	66.32%	77.38%	67.89%	66.24%
Synsetモデル（非依存）	77.16%	67.87%	66.02%	77.16%	67.87%	66.02%
差	+0.29%	+0.64%	+0.30%	+0.21%	+0.02%	+0.22%
ベースライン	75.06%	65.38%	63.40%	75.06%	65.38%	63.40%

赤字：有意な精度向上($p<0.05$)，**黒字**：有意でない差

結果 ― 細／粗粒度意味ラベルの寄与

- 細／粗粒度の意味ラベルのそれぞれが、WSDの精度向上に寄与している。

粗粒度意味ラベルの利用による精度向上

[Synsetモデル]	語義頻度あり		
	SEM	SE2	SE3
粗粒度ラベルあり	77.46%	68.51%	66.32%
粗粒度ラベルなし	77.40%	68.39%	66.06%
差	**+0.04%**	**+0.11%**	**+0.26%**
語義依存なし	77.16%	67.87%	66.02%

- Synsetモデルは Supersenseモデルよりも大きく優れている。

Synset／Supersenseモデルの精度比較
（木構造、supersenseを対象とした時の精度）

頻度	モデル	SEM	SE2	SE3
あり	Synset	84.34%	80.87%	79.62%
	Supersense	83.60%	80.24%	78.93%
	差	+0.74%	+0.63%	+0.69%
ベースライン		83.02%	76.26%	78.48%

結果 ー 木構造モデルと各モデルの比較

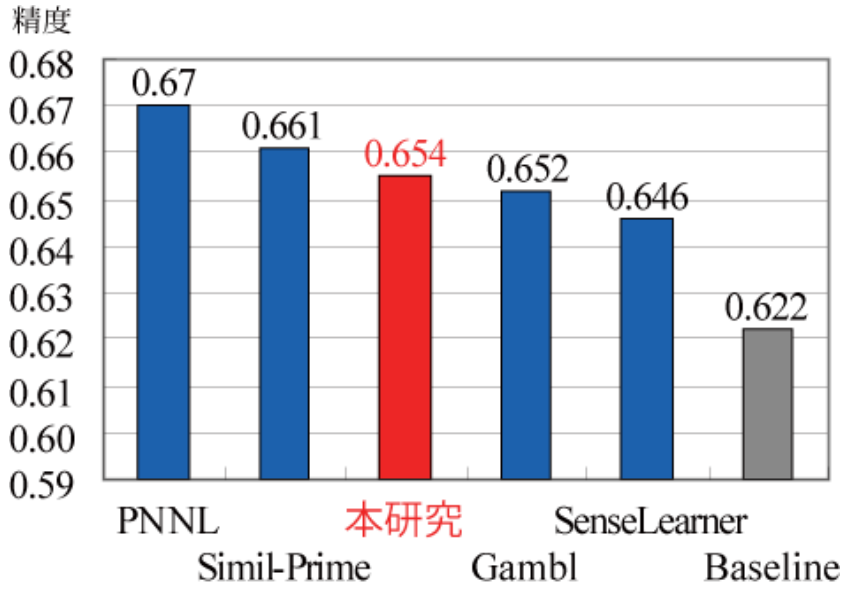
- 木構造モデルは、線状鎖モデルより有意に高い精度を示した。
- （参考）木構造synsetモデルは、世界最高水準にあるシステムと比較して同程度の精度を記録。

（木構造モデルの精度）ー（線状鎖モデルの精度）

モデル	SEM	SE2	SE3
Synset/語義頻度あり	+0.08%	+0.62%	+0.08%
Synset/語義頻度なし	+0.32%	+0.71%	−0.32%
Supersense/語義頻度あり	−0.08%	+0.48%	+0.05%
Supersense/語義頻度なし	+0.03%	+0.58%	+0.74%

システム	精度
PNNL (Tratz et al. 2007)	67.0%
Simil-Prime (Kohomban et al. 2005)	66.1%
本研究 (木構造synsetモデル)	65.4%
Gambl (Decadt et al. 2004)	65.2%
SenseLearner (Mihalcea et al. 2004)	64.6%
Baseline (最頻の語義)	62.2%

Senseval-3 データセットに基づく比較結果



正解例

- *He also bought … assorted nails, levels and T squares and plumb lines and … that he had no idea how to use or what they were for.*
  - SS:noun.artifact–SS:noun.artifact, α=1.497, λ=0.4035
  - SS:noun.artifact–(COORD)–SS:noun.artifact, α=1.367, λ=0.3126

1	noun.attribute	degree#1, grade#8, level#1	a position on a scale of intensity or amount or quality
2	noun.state	grade#2, level#2, tier#1	a relative position or degree of value in a graded group
3	noun.state	degree#2, level#3, stage#2, point#4	a specific identifiable position in a continum or series or especially in a process
4	noun.attribute	level#4	height above ground
5	noun.artifact	level#5, spirit level#1	indicator that establishes the horizontal when a bubble is centered in a tube of liquid

1	noun.body	nail#1	horny plate covering and protecting part of the dorsal surface of the digits
2	noun.artifact	nail#2	a thin pointed piece of metal that is hammered into materials as a fastener
3	noun.quantity	nail#3	a former unit of length for cloth equal to 1/16 of a yard

不正解例

- *Philadelphia permitted him to seek a better connection after he had refused to reconsider his decision to end his career as a player.*
  - WS:career#2−(NMOD)−SS:noun.person, α=1.071, λ=0.06859

career の全2義

1	noun.act	career#1, calling#1, vocation#1	the particular occupation for which you are trained
2	noun.act	career#2, life history#2	the general progression of your working or professional life

結論／今後の課題

- 語義依存関係は語義曖昧性解消問題に寄与している。
- 係り受け構造は線状鎖構造よりも語義の依存関係を適切に表現している。
- 細粒度・粗粒度の意味ラベルの両方が効果的に働いている。

- 今後の課題
  - 準教師あり学習によるアプローチ
  - ドメイン情報の利用

訂正

誤

表 4　語義依存性の寄与（木構造モデル）

	SEM	SE2	SE3
Tree-SS-FS	83.60%	78.93%	80.24%
NoDep-SS-FS	83.39%	78.24%	79.89%
Diff.	+0.21%**	+0.69%***	+0.35%*
Tree-SS	79.15%	77.78%	78.93%
NoDep-SS	79.11%	77.09%	78.24%
Diff.	+0.04%	+0.69%*	+0.53%
Baseline-SS	83.02%	76.26%	78.48%
Tree-WS-SR	77.46%	68.51%	66.32%
NoDep-WS-SR	77.16%	67.87%	66.02%
Diff.	+0.29%***	+0.64%***	+0.30%*
Tree-WS	73.19%	68.38%	65.50%
NoDep-WS	73.27%	68.48%	65.88%
Diff.	-0.09%	-0.10%	-0.38%
Baseline-WS	75.06%	76.26%	78.48%

正

表 4　語義依存性の寄与（木構造モデル）

	SEM	SE2	SE3
Tree-SS-FS	83.60%	78.93%	80.24%
NoDep-SS-FS	83.39%	78.24%	79.89%
Diff.	+0.21%**	+0.69%***	+0.35%*
Tree-SS	79.15%	77.78%	78.93%
NoDep-SS	79.11%	77.09%	78.24%
Diff.	+0.04%	+0.69%*	+0.53%
Baseline-SS	83.02%	76.26%	78.48%
Tree-WS-SR	77.46%	68.51%	66.32%
NoDep-WS-SR	77.16%	67.87%	66.02%
Diff.	+0.29%***	+0.64%***	+0.30%*
Tree-WS	73.19%	68.38%	65.50%
NoDep-WS	73.27%	68.48%	65.88%
Diff.	-0.09%	-0.10%	-0.38%
Baseline-WS	75.06%	65.38%	63.40%

ありがとうございました。