

語義曖昧性解消における統語的依存関係の寄与について

羽鳥 潤* 宮尾 祐介* 辻井 潤一*†

* 東京大学大学院 学際情報学府

* 東京大学大学院 情報理工学系研究科

†National Centre for Text Mining, UK ‡School of Computer Science, Manchester University

1 導入

語義曖昧性解消 (WSD) に対する従来の典型的なアプローチは、個別に学習された分類器を用いて各多義語に対する分類問題を独立して解くものだった。しかし、この手法は語義間の依存関係を全く考慮しておらず、また、訓練データが与えられている語義に対してしか適用できないという問題があった。

これに対して、近年、Ciaramita and Altun (2006) や Mihalcea et al. (2007) は語義の繋がりを考慮し、WSD を単語列に対する系列ラベリング問題として解く方法を提案して一定の成果を収めた。しかし、彼らの利用した語義依存関係は WORDNET の synset 間の、しかも隣り合う語の間にしか働かない単純な依存関係だけであった。そこで本研究では、語義の依存関係をより適切に表現できると考えられる係り受け解析木を用いて、主辞と従属語の間の語義依存関係を利用して WSD を行うモデルを提案する。この手法では、木構造上の条件付き確率場 (T-CRF)(Tang et al., 2006) の利用により、語義の割り当ては文章ごとに最大確率を持つ組み合わせとして出力される。

また、WSD における最大の問題の一つに、十分な訓練データが利用できない事によるデータスパースネスがある。一般に、語義依存関係を取り入れるためには 2 つの語義の組み合わせを素性として定義する必要があるため、素性数の増加によりこの問題はさらに深刻化する。そこで、我々はこれらの素性にいくつかの粗粒度の意味ラベルを組み合わせる事でこの問題に対処した。特に、近年有効性の確認されている supersense(Ciaramita and Altun, 2006) などを利用している。このような粗粒度の意味ラベルを組み合わせる事により、訓練データ中に出現しなかった語 (語義) に対しても語義を推定できる点にも注意されたい。

ところで、WORDNET の synset を基準に WSD を行うモデル (Synset モデル) の構築は、十分な訓練データを利用する事の難しさ、及び語義タグを付与する際の注釈者間一致率の低さ*1 などから困難が大きく、近年では、粗粒度の意味ラベルを利用した WSD*2 でも多くの NLP アプリケーションに十分有用である事が示唆されている (例えば Ide and Veronis (1998))。そこで、我々は supersense を粗粒度の意味ラベルとして素性に組み合わせるだけでなく、それ自体を基準とした WSD を行うモデル (Supersense モデル) も構築し、これらのモデルを比較した。

*1 SENSEVAL-3 English all-words task で 63.4%。

*2 この場合、注釈者間一致率は 90% 程度まで上がる事が確認されている (Navigli et al., 2007)。

2 手法

2.1 条件付き確率場 (CRF)

条件付き確率場 (CRF) は Lafferty et al. (2001) によって導入されたグラフベースの識別モデルであり、NLP の多くの系列ラベリング問題で最高精度を記録している。このモデルでは、与えられた観測列 \mathbf{x} に対するラベル列 \mathbf{y} の条件付き確率は、

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left[\sum_{e \in E, j} \lambda_j f_j(e, \mathbf{y}_e, \mathbf{x}) + \sum_{v \in V, k} \mu_k g_k(v, \mathbf{y}_v, \mathbf{x}) \right] \quad (1)$$

によって与えられる。ここで、 \mathbf{y}_e と \mathbf{y}_v は枝 e と頂点 v に対応するラベル列、 f_j は g_k は素性ベクトル、 λ_j と μ_k は重みベクトルに対応しており、 $Z(\mathbf{x})$ は規格化関数である。木構造上の条件付き確率場 (T-CRF)(Tang et al., 2006) はこれを木構造に対して適用したものであり、係り受け関係に基づく語義依存関係をうまく取り込む事が出来ると考えられる。

2.2 グラフの構築

CRF のグラフを構築する際、語義間の出来るだけ有効な依存関係を利用するために、一部の単語の間引きを行った。まず、Synset モデルでは内容語 (名詞・動詞・形容詞・副詞) 以外の機能語は WSD の対象としないので、係り受け解析木は内容語だけからなる木に構成される。図 1 では、構文解析器の出力木 (左) が、機能語の *the*・*in* を省いた木 (右) に変換される様子が示されている。この際、*in* の主辞 *confidence* と従属語 *bank* は木の中で直接接続され、代わりに *in* はその枝に対する素性として定義される*3。図 2 には線状鎖の場合が示されているが、この場合でもこのプロセスは同様に行われる。また、Supersense モデルの場合には、特に名詞と動詞に絞って WSD を行うため、それ以外の単語は取り除かれた。

2.3 意味ラベル

我々は、WORDNET に含まれている情報を用いて表 1 に示されている 4 つの意味ラベルを利用した。粗粒度のものから順に、synset S_{WS} 、上位 synset $S_{G1} \cdot S_{G2}$ 、及び supersense S_{SS} である。ここで、supersense は WORDNET 内の名詞と動詞が分類されているもっとも一般的な意味クラスであり、名詞に対して *act*, *attribute*, *group*, *person* などを含む 26、動詞に対して *change*, *cognition*, *contact*, *stative* などを含む 15 のクラスが定義されている。これらの意味ラベルのうち、頂点素性に

*3 内容語の間に複数の機能語が存在していた場合にはその組み合わせが素性として定義される。また、係り受け関係のラベルは木の中で一番上にあったものが残される。

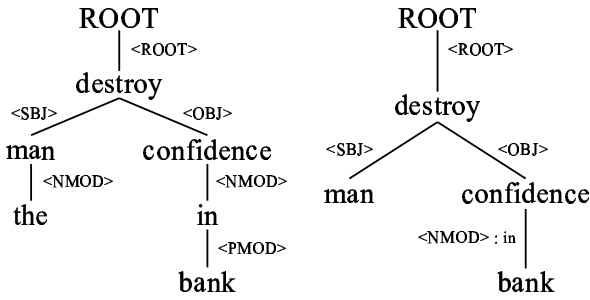


図1 木構造モデルにおける間引きの例

ROOT—the—man—destroy—confidence—in—bank

ROOT——man——destroy——confidence——in——bank

図2 線状鎖モデルにおける間引きの例

表1 “destroy(v)#1” に対する意味ラベルの一覧

S_{WS}	synset	{bank#1, ...}
S_{G2}	第二位 synset	{organization#1, ...}
S_{G1}	第一位 synset	{social group#1}
S_{SS}	supersense	noun.group

は $S_{WS} \cdot S_{SS}$ の2種類を、枝素性には4種類全てのラベルを用いた。ただし、synset以外のラベルは名詞と動詞にのみ用いられ、利用できない意味ラベルは“N/A”に置き換えられる。

2.4 素性

2.4.1 語義頻度

各語義の頻度情報は、訓練データが十分に利用できないWSDでは特に重要である。我々は、SynsetモデルにWORDNET内の語義番号の情報を利用した。WORDNETの語義はその頻度順に並んでいるため、その番号を素性として用いる事で頻度情報を間接的に取り込む事が出来る。

この中で一番目の語義はfirst senseと呼ばれ、WSDにおける特に強いベースラインとして知られている(McCarthy et al., 2004)。Supersenseモデルでは語義番号の代わりにこのfirst senseの情報を素性として使用した。これは、Supersenseモデルでは全語義(全supersense)に対して十分な訓練データが得られるため、一律に語義番号を素性として用いるよりも単語毎の語義頻度情報をうまく取り込めると考えられるからである。

2.4.2 頂点素性

我々は、SynsetモデルにおけるCRFの頂点に対する素性として、多くのWSDモデルで用いられている典型的な文脈素性を利用した。これらの素性は基本的にLee and Ng (2002)の利用したものと同一のもので、対象語・前後の語の語形と品詞、文脈中に登場した語などからなる。但し、統語的な素性としては、主辞の情報だけでなく従属語の情報も利用した。これは、統語的な語義の依存関係がWSDに本当に寄与しているかを明らかにするためで、もし、主辞と従属語の語形・品詞がWSDに十分な情報であれば、語義依存関係の導入によって精度は向上しないと考えられるからである。また、線状鎖モデル

表2 データセットの統計

	#タグ付きの語	#タグ付きの名詞/動詞
brown1/brown2	189,667	135,123
SENSEVAL-2	2,259	1,567
SENSEVAL-3	1,978	1,617

ルにおいてこれに対応する、前後の語の情報は既に素性に含まれている。最後に、Supersenseモデルに対しては有効に働く素性が大きく違う事が知られているため、Ciaramita and Altun (2006)が利用した素性を用いた。

2.4.3 枝素性

枝素性として、グラフ中で直接接続された2単語の語義バイグラムを全ての意味ラベルの組み合わせについて定義した。これらは、各語の間の語義依存関係を表現している。また、単純なバイグラムの他に、それと頂点間に省略された語、及び係り受け関係を表すラベルと組み合わせたものも利用した。したがって、1つの枝に対して定義された素性は、合計で $4^2 \cdot 3 = 48$ 種類ある。

3 実験

3.1 実験条件

評価用のデータセットとしては、コーパスの内の内容語にWORDNET語義タグが付けられている、SEMCORのbrown1/brown2セクション、SENSEVAL-2/3 English all-words taskのデータを利用した^{*4}(表2)。brown1/brown2に含まれる352ファイルは、ドメインが偏らないように5つのセットに分けられ、クロスヴァリデーションに用いられた。各ヴァリデーションでは、そのうちの4セットを訓練データ、残りの1セットを半分に分けてデベロップメント・テストセットとして利用した。但し、SENSEVALデータセットをテストセットとして使用するときは、残りの1セットの全体がデベロップメントセットとして使用された。デベロップメントの際には、 L_2 正規化項の強さを決めるパラメータ σ を調整した。また、データセット中の文の構文解析にはSagae and Tsujii (2007)による係り受け解析器を用いた。

表3は、我々が実験で使用したモデルの一覧である。Synset/Supersenseモデルの別、語義依存関係の有無と構造、語義頻度情報を利用したかに応じて、12種類のモデルが定義されている。Synsetモデルの評価は語義タグの付けられている単語(内容語)全てを対象に行ったが、Supersenseモデルの評価は名詞・動詞に限定して行った^{*5}。ここで、ベースラインはすべての単語に対してそのfirst sense(またはfirst senseのsupersense)を割り当てるモデルである。

3.2 語義依存性の寄与

表4は木構造モデルと語義依存関係を考慮していないモデル(非依存モデル)を比較した結果である。数値はそれぞれSEMCOR・SENSEVAL-2/3データセットをテストセットとして計測された5回の精度の平均である。

^{*4} <http://www.cs.unt.edu/~rada/downloads.html> で配布されているWORDNET 2.0バージョンを利用した。

^{*5} この際、Ciaramita and Altun (2006)が行ったのと同様に、いくつかのnoun.Topsに属する単語を実際の意味に対応したスーパーセンスに置き換え、置き換えた後のタグに基づいて評価を行った。

表 3 実験で使用したモデルの一覧

モデル		Synset	Supersense
木構造	頻度あり	Tree-WS-SR	Tree-SS-FS
	頻度なし	Tree-WS	Tree-SS
線状鎖	頻度あり	Linear-WS-SR	Linear-SS-FS
	頻度なし	Linear-WS	Linear-SS
依存なし	頻度あり	NoDep-WS-SR	NoDep-SS-FS
	頻度なし	NoDep-WS	NoDep-SS
ベースライン		Baseline-WS	Baseline-SS

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%

Supersense モデルでは、すべてのデータセットで語義依存関係の導入による精度の向上を確認し、first sense の有無によって精度大きな差は見られなかった。一方、Synset モデルでは、語義番号を利用した時に Supersense モデルと同様の精度向上が確認されたが、利用しなかった場合は語義依存関係の導入によって一様に精度が低下するという結果となった。これは、Synset モデルでは特にデータスパースネスの問題が起りやすく、語義の頻度情報なしでは十分なスムージング効果が得られないためと考えられる。両モデルとも、語義頻度情報と共に語義依存関係を利用した場合、精度は有意⁶に向上している。

同様に、表 5 は線状鎖モデルと非依存モデルの比較結果を示している。木構造モデルの時より語義依存関係の寄与は若干小さいが、Supersense モデル・Synset モデルとも精度向上を確認している。語義頻度情報がない時に語義依存関係が精度を低下させてしまう現象はより顕著に見られ、特に Synset モデルでは有意に低下した。このことから、線状鎖モデルは木構造モデルよりも不適切な依存関係を学習しやすいと言う事が出来る。

最後に、表 6 に木構造モデルと線状鎖モデルの比較結果を載せた。データセット間のばらつきは大きいですが、統計的に有意である差に着目すれば、木構造モデルの方がより高精度である事がはっきりと分かる。

3.3 Synset モデル vs Supersense モデル

我々は supersense を基準として Synset モデルと Supersense モデルの精度比較を行った (表 7)。その結果、語

⁶ 有意性検定は McNemar's Test による。*** が 0.001, ** が 0.01, * が 0.05 の有意水準に対応している。

表 5 語義依存性の寄与 (線状鎖モデル)

	SEM	SE2	SE3
Linear-SS-FS	83.69%	78.44%	80.19%
NoDep-SS-FS	83.39%	78.24%	79.89%
Diff.	+0.29%***	+0.20%	+0.30%*
Linear-SS	79.12%	77.21%	78.22%
NoDep-SS	79.11%	77.09%	78.24%
Diff.	+0.01%	+0.12%	-0.21%
Baseline-SS	83.02%	76.26%	78.48%
Linear-WS-SR	77.38%	67.89%	66.24%
NoDep-WS-SR	77.16%	67.87%	66.02%
Diff.	+0.21%**	+0.02%	+0.22%
Linear-WS	72.87%	67.68%	65.82%
NoDep-WS	73.27%	68.48%	65.88%
Diff.	-0.41%***	-0.81%**	-0.05%
Baseline-WS	75.06%	65.38%	63.40%

表 6 木構造モデルと線状鎖モデルの比較

	SEM	SE2	SE3
Tree-SS-FS	83.60%	78.93%	80.24%
Linear-SS-FS	83.69%	78.44%	80.19%
Diff.	-0.08%	+0.48%***	+0.05%
Tree-SS	79.15%	77.78%	78.96%
Linear-SS	79.12%	77.21%	78.22%
Diff.	+0.03%	+0.58%*	+0.74%*
Tree-WS-SR	77.46%	68.51%	66.32%
Linear-WS-SR	77.38%	67.89%	66.24%
Diff.	+0.08%	+0.62%**	+0.08%
Tree-WS	73.19%	68.38%	65.50%
Linear-WS	72.87%	67.68%	65.82%
Diff.	+0.32%***	+0.71%*	-0.32%

表 7 Synset モデル・Supersense モデルの精度比較 (supersense 基準, 有意性検定は行っていない)

	SEM	SE2	SE3
Tree-WS-FS	84.34%	80.87%	79.62%
Tree-SS-SR	83.60%	80.24%	78.93%
Diff.	+0.74%	+0.63%	+0.69%
Tree-WS	80.65%	80.21%	79.27%
Tree-SS	79.15%	78.96%	77.78%
Diff.	+1.50%	+1.25%	+1.49%
Baseline-SS	83.02%	76.26%	78.48%

義頻度情報ありのモデルで 0.63–0.74%, 語義頻度情報なしのモデルで 1.25–1.50%, Synset モデルが Supersense モデルよりも高い精度を示した。この事は、supersense の粒度が多くの NLP タスクに十分であるといえども、それ自体の粒度が粗すぎて十分 WSD に必要な情報を表現できていない可能性がある事を示している。語義頻度情報による精度の底上げがない場合、この差が非常に大きくなっている事からもこの事は裏付けられる。

3.4 他のシステムとの比較

表 8 に、今回実験に使用した Synset モデルで最高の精度を記録した Tree-WS-SR と、他の WSD モデルとの比較結果を示す。この比較実験では、訓練データとして brown1/brown2 の全体を利用し、SenseEval-2 all-words task データをデベロップメントセットとして使

表 8 SENSEVAL-3 English all-words task データに基づく他のシステムとの比較結果

システム	精度
PNNL (Tratz et al. 2007)	67.0%
Simil-Prime (Kohomban et al. 2005)	66.1%
Our model (Tree-WS-SR)	65.4%
Gambl (Decadt et al. 2004)	65.2%
SenseLearner (Mihalcea et al. 2004)	64.6%
Baseline (Baseline-WS)	62.2%

表 9 最も強く寄与した語義依存関係の種類

正解例		不正解例	
種類	観測数	種類	観測数
動詞-名詞	26	動詞-名詞	22
名詞-名詞	10	動詞-動詞	6
名詞-形容詞	4	名詞-名詞	5
動詞-動詞	3	名詞-修飾語	4
名詞-修飾語	3	動詞-副詞	3
動詞-修飾語	3	動詞-修飾語	2
その他	5	その他	4

用した。精度の数値は、SENSEVAL-3 で提供されている scorer2 で測定した。我々のモデルは、現在のところ最高精度を記録している PNNL などのシステムに若干劣る結果となったが、SENSEVAL-3 の上位 2 システムである Gambl・SenseLearner より高い精度を記録した。他のシステムがより多くの訓練データを使用している事、訓練器の組み合わせなどによる機械学習的なチューニングを多く行っている事などを考慮すると、我々のシステムは比較的シンプルなモデルと少ない訓練データでこの精度を記録している事から、これらのシステムに十分匹敵する性能を持っていると考えられる。

4 分析

語義依存関係の寄与の分析のため、Tree-WS-SR と NoDep-WS-SR の出力を比較し、前者が正解だが後者が不正解だった例（正解例）、前者が負正解だが後者が正解だった例（不正解例）を合わせて 100 個取り出して、54 個の正解例と 46 個の不正解例を得、各例に対して働いていた枝素性の中で最大のスコアを持つものを素性の代表として取り出した。表 9 は、正解例と不正解例のそれぞれについてこれらの素性の種類を集計したものである。正解例と不正解例の両方で最もよく働いていたのは動詞-名詞関係であり、これは全体の 48% を占めている。その他の多くの素性は比較的頻度が少ないが、名詞-名詞関係だけは正解例の中で 19% と比較的大きな割合を占めている。特に、正解例の中で見つかった 10 例の名詞-名詞関係のうち 4 例は名詞-of-名詞関係であったが、このような関係は不正解例中に 1 件も見つからなかった事から、正解例に特に強く寄与していると考えられる。

また、表 10 ではこれらの依存関係がどの意味ラベルの組み合わせによるものかを集計している。ここで見てとれるのは、正解例の方が不正解例よりも粗粒度の意味ラベルの寄与を受けているという傾向である。特に、正解例の 37% は supersense-supersense 関係によっており、supersense と synset 間の関係や supersense と第一位 synset 間の関係などの粗粒度意味ラベル同士の関係も効果的に働いている事がわかる。一方、正解例の中

表 10 正解例と不正解例に対して寄与した語義依存関係の意味ラベルの組み合わせ（行が主辞、列が従属語のラベルに対応）

正解例				不正解例				
	WS	G2	G1	SS	WS	G2	G1	SS
WS	2			5	4			4
G2	2	1	1		1	1		
G1	1			5	2			4
SS	7		8	19	7	2	5	14

で synset 同士の関係の寄与が比較的少ないのは、訓練データのサイズが小さいためにデータスパースネスがあり、これらの素性が効果的に働きづらかった事を示している。

5 結論

本稿では、語義の統語的な依存関係に着目し、文中の全内容語の語義曖昧性を同時に解消する手法を提案した。実験では、語義依存関係が精度の向上に寄与している事、粗粒度と細粒度の意味ラベルの組み合わせが効果的に働く事、及び、統語的な語義依存関係が線状鎖上の関係よりも文のセマンティクスをよく捉えている事の三点が確認された。但し、Synset モデルでは語義依存関係の導入によりデータスパースネスが起こりやすく、語義の頻度情報との適切な組み合わせが必要である事が分かった。一方で、Supersense モデルは語義頻度情報なしでもロバストに動作したが、精度の面からは Synset モデルに劣っており、より細粒度の語義の寄与を考慮した方が高精度のモデルを構築できる事が分かった。

参考文献

- Ciaramita, M. and Y. Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*.
- Ide, N. and J. Veronis. 1998. Word sense disambiguation: The state of the art. *Comp. Linguistics*, 24:1-40.
- Lafferty, J., A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*.
- Lee, Y. K. and H. T. Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proc. of EMNLP*.
- McCarthy, D., R. Koeling, J. Weeds, and J. Carroll. 2004. Finding predominant senses in untagged text. In *Proc. of ACL*.
- Mihalcea, R., A. Csomai, and M. Ciaramita. 2007. UNT-Yahoo: SuperSenseLearner: Combining SenseLearner with SuperSense and other coarse semantic features. In *Proc. of SemEval-2007*.
- Navigli, R., K. Litkowski, and O. Hargraves. 2007. Semeval-2007 task 07: Coarse-grained english all-words task. In *Proc. of SemEval-2007*.
- Sagae, K. and J. Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proc. of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*.
- Tang, J., M. Hong, J. Li, and B. Liang. 2006. Tree-structured conditional random fields for semantic annotation. In *Proc. of the 5th Int'l Semantic Web Conf.*