上位意味クラス推定を用いた語義曖昧性解消

藤田 早苗,♡ **Francis Bond,**♠ 藤野 昭典 ♡
♡ {sanae,a.fujino}@cslab.kecl.ntt.co.jp, ♠ bond@ieee.org
♡ NTT コミュニケーション科学基礎研究所, ♠ 情報通信研究機構

1　はじめに

通常、多くの語は文脈によって異なる複数の語義を持つ。これらの語義から文脈に応じて正しい語義を選択する語義曖昧性解消 (WSD) は、近年、構文解析結果の正解選択 (Fujita et al., 2007) や、機械翻訳 (Chan et al., 2007) の精度向上に有効であることが示されている。

WSD については、多くの先行研究がある。日本語では、教師ありの手法には及ばないものの、教師なしで高い精度を得た、拡張 Lesk を用いる手法 (Baldwin et al., 2008) や、意味情報と構文情報の両方を用いた教師ありの手法 (Tanaka et al., 2007) などが提案されている。しかし、我々は WSD の結果を構文解析にも利用するため、本稿では、構文解析結果を利用しない手法を提案する。

WSD の難しさの原因の一つとして、膨大な量の語義を推定するために十分な学習データの構築が難しいということがあげられる。加えて、クラス数が膨大であれば、多くの機械学習ツールは利用が困難である。そのため、提案手法では WSD を 2 段階に分ける。 1 段階目では、主体, 場所, 具体物, 抽象物などの (上位) 意味クラスを推定する。意味クラスは語義に比べて数がはるかに少ないため、比較的少ない学習データからでも十分な精度を得ることができる。また、上位意味クラスが定まれば、語義が一意に決まる事も多い。 2 段階目では、 1 段階目で推定した意味クラスを用いて語義そのものを推定する。

なお、実験には、檜コーパス (Bond et al., 2006) を用いる。檜コーパスは、辞書 (Lexeed(笠原ら, 2004)) の語義文、例文、新聞 (京大コーパス、以下、 KC) に対するツリーバンク、および、センスバンクから構成される。 また、この辞書には、語義毎に日本語シソーラスである日本語語彙大系 (池原ら, 1997) の意味クラス (意味属性) が付与されている。

2　上位意味クラスの推定

本章では、上位意味クラスの推定方法について述べる。語彙大系は、 2,710 の意味クラスからなり、深さ 0 から 11 までの階層 (レベル) に分けられている。そのうち、レベル 2 から 5 までの意味クラス[1] を用いて実験を行ない、語義曖昧性解消に有効なレベルを調査する。

2.1　訓練／テストデータ

檜センスバンクの内容語は、 Lexeed の語義によってタグ付けされている。しかし、本章では、 (上位) 意味クラスを推定するため、檜センスバンクの語義タグを上位意味クラスへと置換し、訓練／テストデータを作成する。例として、運転手 $_1$ の語義文を文 (1) に示す。文 (1) の下に **cat** で示した行は、各語義タグにリンクされている語彙大系の意味クラス、 **lvl X** で示した行は、レベル X における上位意味クラスを示している。本稿では、語義が複数の意味クラスにリンクされている場合、最初の意味クラスのみを用いている。

表 1に、訓練／テストデータの数を示す。訓練データは、 1 語義平均 5.1 文 (例文) から 17.7 文 (KC) である。しかし、上位意味クラスに集約した場合、レベル 5 でも 1 クラス平均 340.9 文 (例文) から 539.7 文 (KC) となる。このように、上位意味クラスに集約する事により、データスパースネスに強くなる。

Corpus	Set	文数	対象語数	全語数
語義文	Train	67,202	175,709	613,216
	Test	4,942	15,932	54,276
例文	Train	106,528	133,616	432,514
	Test	8,942	12,416	41,019
KC	Train	141,968	211,567	947,298
	Test	5,408	12,581	53,703

表 1: 訓練 / テストデータ数: ここで対象語とは、 Lexeed の語義でタグ付与された語

2.2　実験: 上位意味クラスの推定

機械学習手法として、 Maximum Entropy Method: **MEM**(Nigam et al., 1999) および、 Conditional Random Fields: **CRF**(Suzuki et al., 2006) を用いて実験を行なっ

[1] 意味クラスは、レベル 2 の場合 ⟨3: 主体⟩ や ⟨533: 具体物⟩ など 9 クラス、レベル 3 の場合 ⟨4: 人⟩ や ⟨706: 無生物⟩ など 30 クラス、レベル 4 の場合 ⟨5: 人間⟩ や ⟨760: 人工物⟩ など 136 クラス、レベル 5 の場合 ⟨6: 人間＜人称＞⟩ や ⟨838: 食料⟩ など 392 クラスに集約される。

(1)	電車 $_1$	や	自動車 $_1$	を	運転 $_1$	する	人 $_4$
cat	⟨988: 乗り物 (本体 (移動 (陸圏)))⟩	-	⟨988: 乗り物 (本体 (移動 (陸圏)))⟩	-	⟨2003: 操縦⟩	-	⟨4: 人⟩
lvl 5	⟨986: 乗り物⟩	-	⟨986: 乗り物⟩	-	⟨1920: 労働⟩	-	⟨4: 人⟩
lvl 4	⟨760: 人工物⟩	-	⟨760: 人工物⟩	-	⟨1560: 行為⟩	-	⟨4: 人⟩
lvl 3	⟨706: 無生物⟩	-	⟨706: 無生物⟩	-	⟨1236: 人間活動⟩	-	⟨4: 人⟩
lvl 2	⟨533: 具体物⟩	-	⟨533: 具体物⟩	-	⟨1235: 事⟩	-	⟨3: 主体⟩

た[2]。なお、本稿では、茶筌による修正済み形態素解析結果を入力として利用する。

以下、利用する素性について述べる。 **CRF**の素性には、 uni-gram, bi-gram, 対象語の前後 2 語の組合せを用いる (表 2)。**MEM**の素性には、対象語自身とその前後の語、対象文中のすべての内容語、および、対象語の前後 3 文字までの文字列を用いる (表 3)。**CRF**と全く同じ素性を利用した実験も行なったが、明らかに精度が下がったため、ここではとりあげない。表 2, 3で、 b_k は k 番目の語の原形、 w_k は表層形、 $p1_k$, $p2_k$, $p3_k$ はそれぞれ、品詞, 品詞細分類 1, 品詞細分類 2 を示す。 Sample は、文 (1) の 5 番目の語 ($i = 5$)「運転」を対象語とした場合の素性の一部である。

Type	Template	Sample
uni-gram	$\langle b_k \rangle$, $\langle w_k \rangle$, $\langle p1_k \rangle$, $\langle p2_k \rangle$, $\langle p3_k \rangle$	⟨自動車⟩, ⟨自動車⟩, ⟨名詞⟩, ⟨名詞 - 一般⟩
組合せ	$\langle b_k, w_k \rangle$, $\langle b_k, p1_k \rangle$, $\langle b_k, p2_k \rangle$, $\langle b_k, p3_k \rangle$, $\langle w_k, p1_k \rangle$, $\langle w_k, p2_k \rangle$, $\langle w_k, p3_k \rangle$, $\langle p1_k, p2_k \rangle$, $\langle p1_k, p3_k \rangle$, $\langle p2_k, p3_k \rangle$	⟨自動車, 自動車⟩, ⟨自動車, 名詞 - 一般⟩, ⟨自動車, 名詞⟩, ⟨自動車, 名詞 - 一般 -*⟩, ⟨名詞, 名詞 - 一般 -*⟩
bi-gram	$\langle b_k, b_{k+1} \rangle$, $\langle w_k, w_{k+1} \rangle$, $\langle p1_k, p1_{k+1} \rangle$, $\langle p2_k, p2_{k+1} \rangle$, $\langle p3_k, p3_{k+1} \rangle$	⟨自動車, を⟩, ⟨自動車, を⟩, ⟨名詞, 助詞⟩, ⟨名詞 - 一般, 助詞 - 格助詞⟩, ⟨名 - 一般 -*, 助 - 格助詞 - 一般⟩

表 2: CRF で利用した素性: ここで、 i 番目の語が対象語とすると、 uni-gram と組合せでは、 $k = i-2, ..., i+2$、 bi-gram では、 $k = i-2, ..., i+1$。

Type	Template	Sample
内容語	$\langle b_l \rangle$	⟨電車⟩, ⟨自動車⟩, ⟨人⟩
uni-gram	$\langle b_j \rangle$, $\langle w_j \rangle$, $\langle p1_j \rangle$, $\langle p3_j \rangle$	⟨運転⟩, ⟨運転⟩, ⟨名詞⟩, ⟨名詞 - サ変接続 -*⟩
文字列	$\langle cb1_i \rangle$, $\langle cb2_i \rangle$, $\langle cb3_i \rangle$ $\langle ca1_i \rangle$, $\langle ca2_i \rangle$, $\langle ca3_i \rangle$	⟨を⟩, ⟨車を⟩, ⟨動車を⟩ ⟨す⟩, ⟨する⟩, ⟨する人⟩

表 3: MEM で利用した素性: ここで、 i 番目の語が対象語とすると、 uni-gram では、 $j = i-1, ..., i+1$。

2.3 結果と議論: 上位意味クラス推定

上位意味クラスの推定結果を表 4に示す。ベースライン (BL) は、訓練データ中の最頻の意味クラスを選択した場合の精度である。本手法では、各レベルまでの全てのクラスが選択可能になっているため、語によっては有り得ないクラスが選択される場合がある。例えば、「ドライバー」という語は、レベル 2 では、⟨3: 主体⟩か ⟨533: 具体物⟩ しか取り得ない。しかし、本手法では、⟨388: 場所⟩ などのクラスも選択可能である。そこで、このようなエラーを修正するため、有り得ないクラスが選択された場合、**CRF**では、可能なクラスの中で最頻のクラスへと変更する。また、**MEM**では、可能なクラスの中で最も確率の高いクラスへと変更する[3]。表 4において、「修正前」で示した精度は、推定結果そのままの精度であり、「修正後」で示した精度は、有り得ないクラスを修正した場合の精度である。

表 4の修正前の結果から、**CRF**はより深いレベルでの精度が比較的高いことがわかる。しかし、**CRF**は**MEM**より多くの時間とメモリを必要とする。そこで、いくつかの値 (* を付与した数値) は、p2 を用いないで得た。但し、 p2 を用いない場合、精度は 0.1-0.2 % 程度悪くなる[4]。

表 4の修正前の結果では、いくつかの条件で、**MEM**の方が、**CRF**より高い精度を出している。しかし、修正方法は**MEM**の方が有利であるにも関わらず、修正後は、**CRF**ベースの精度の方が全て高くなっている[5]。 修正後は**CRF**の方が全て高くなった理由として、**MEM**は最初から頻度に重点がおかれた学習方法であるのに対し、**CRF**は比較的系列としての妥当性に重点がおかれた学習方法であるため、頻度による修正は**CRF**に対してより効果的であるという理由が考えられる。

表 4から、修正後の精度は、修正前に比べて、いずれも上昇している。そのため、次章の語義曖昧性解消では、修正後の結果を利用する。

3 語義曖昧性解消

本章では、前章で獲得した上位意味クラスを用いた語義曖昧性解消 (WSD) について述べる。まず、 WSD のための素性について述べる。 WSD のコンテストである SENSEVAL-2 日本語辞書タスクにおいて、最も高い精度を得た村田ら (2003) のシステム (以下、**MRT**) をほぼ再実装し、我々のシステムと比較する。我々が再実装したシステム (以下、**CRL**) と、**MRT**の違いについては以下に述べる。

[2]実際は、 Support Vector Machine (SVM, (Chang and Lin, 2001)), でも実験したが、**MEM**より精度が低く、時間も非常にかかったためここでは取り上げない

[3]**MEM**では、全クラスの確率が簡単に獲得できたため。

[4]* を付与した条件以外で比較した場合。

[5]実際には、**MEM**の結果を、 CRF と同様に、可能な最頻の意味クラスに修正する方法も試したが、精度はやや下がった。

Cor-	語義文					例文					KC				
pus		修正前		修正後			修正前		修正後			修正前		修正後	
Lvl	BL	**CRF**	**MEM**	**CRF**	**MEM**	BL	**CRF**	**MEM**	**CRF**	**MEM**	BL	**CRF**	**MEM**	**CRF**	**MEM**
2	91.3	**96.0**	95.4	**96.3**	95.7	87.4	88.7	**89.4**	**92.0**	91.8	90.0	93.3	**95.3**	**96.3**	95.8
3	83.5	**92.0**	90.8	**92.5**	91.4	80.1	84.0	**84.3**	**87.6**	87.4	83.0	89.8*	**91.8**	**93.4***	92.8
4	79.2	**90.6**	89.3	**91.2**	90.2	76.7	**82.0**	80.8	**85.7**	84.9	80.0	88.2*	**89.4**	**91.9***	90.8
5	70.1	**85.9***	85.1	**86.7***	86.6	67.7	**77.9***	75.4	**81.9***	81.0	70.6		86.6		88.8

表 4: 上位意味クラス推定結果 (**CRF**/**MEM**): 但し、 * を付与した数値は、 p2 を素性として利用していない。

MRTは、表 3の素性以外に、次の (a)-(c) の情報を利用して素性を作成している。 (a) KNP による構文解析結果、 (b) 図書館などで書類の分類に用いられる国際十進分類法 (UDC) のコード、 (c) 日本語のシソーラスである分類語彙表 (国立国語研究所, 2004) の分類番号。

上記 (a)-(c) のうち、本実験では (a) と (b) は利用しない。 (a) を利用しない理由は、 WSD の結果を構文解析に利用するため、構文解析を WSD の前処理としては行なわないためである。 (b) を利用しない理由は、 UDC コードが檜コーパスには付与されていないためである。

ここで、 (c) の分類語彙表は日本語約 96,000 語が収録されており、深さ 5 の木構造になっている。最初のレベルでは 4 クラスにわけられ、レベル 3 で 95 クラス、レベル 5 で 895 クラスに分けられる。語彙大系も分類語彙表も、共に日本語のシソーラスであるが、語彙大系が主に一般名詞を分類するために作られたのに対し、分類語彙表は、機能語を含む全ての語を分類対象としている。 (c) について、**MRT**は、レベル 3 と 5 のクラスを両方利用している。しかし、彼らは字面にマッチした最初のクラスを利用しており、本稿のようにより適切なクラスを推定するようなことはしていない。語彙大系と分類語彙表から得られるタイプと粒度は異なり、特に分類語彙表には機能語が含まれることから、異なる効果が得られると考えられるため、我々も (c) の素性は利用する。

また、**MRT**は、 JUMAN/RWC の形態素解析結果を両方利用しているが、本稿では茶筌による形態素解析結果のみ利用している。

つまり、**CRL** 用には分類語彙表から獲得した素性を表 3に追加し、我々のシステムには、分類語彙表および推定した上位意味クラスによる素性を追加する。この時、推定したレベルより上位レベルの意味クラスも利用する。例えば、前章でレベル 3 の意味クラスを推定した場合、レベル 2 の意味クラスも素性として追加している。

実験には、 SENSEVAL-2 での対象語 (名詞、動詞各 50 語) を用いた。但し、 Lexeed にない 2 語、および、訓練 / テストデータのいずれかに出現しなかった語を除いている。実際の対象語数を表 5に示す。

我々は、**MRT**と同様、語と品詞の組合せ毎のモデルを作成した。また、**MRT**は、 SVM とナイーブベイズの両方を組み合わせて利用しているが、本実験では SVM(Chang and Lin, 2001) のみを利用している。また、**MRT**では、多項式カーネルを利用しているが、我々の実験では線形カーネルの方が精度が高くなったため、線形カーネルを利用した。

Corpus	名詞		動詞		合計	
No.	Wd	Pol	Wd	Pol	Wd	Pol
語義文	44	6.4	46	9.6	90	8.1
例文	41	6.6	46	9.4	87	8.1
KC	49	6.3	49	10.4	98	8.4

表 5: WSD の対象語数:Wd は対象語数、 Pol は平均多義数

3.1 結果と議論：語義曖昧性解消

表 6 に WSD の結果を示す。ベースライン (BL) は、訓練データの中での最頻語義を選択した場合の精度である。また BL2 は、前章で推定した上位意味クラスを満たす最頻語義を選択した場合の精度である。

表 6において、**SCRF**は**CRF**、**SMEM**は**MEM**によって推定 / 修正した上位意味クラスを用いたシステムである。全ての結果はベースライン (BL) より有意に改良されている。語義文を除き、**SCRF** の結果が最もよい。

表 6から、上位意味クラスを満たす最頻語義を選択した場合 (BL2) でも、高い精度で語義を推定することができた。一般に、階層が深くなると、上位意味クラスの推定自体の精度は下がるにもかかわらず、より深いレベルの意味クラスを用いる方が、 WSD の精度自体は向上している。

4 議論と今後の課題

本稿では、 WSD における上位意味クラス推定の有効性を示した。しかし、上位意味クラスが曖昧性の削減に効果がない場合も存在する。例えば「アイロン」は、 ”布地のしわをのばすのに使う鉄製の重いこて。” と、 “髪の毛をちぢらせる鉄製のこて。” の二つの意味を持つ。アノテイターはこれらの語義を完全に区別できるが、両方とも、〈915: 家庭用具〉と〈969: 電力機器〉にリンクされており、意味クラスから語義を絞ることはできない。しかし、ほとんどの多義語に対しては、本手法は有効である。

Corpus	Lvl	語義文 名詞	語義文 動詞	語義文 平均	例文 名詞	例文 動詞	例文 平均	KC 名詞	KC 動詞	KC 平均
BL		74.5	56.8	63.8	63.7	56.2	58.3	69.2	62.1	66.1
CRL		81.1	65.3	71.5	79.5	68.5	71.6	80.9	67.0	74.7
BL2	2	76.8	59.9	66.5	66.9	58.8	61.0	69.9	63.4	67.0
(**CRF**	3	80.8	60.6	68.5	69.1	60.5	62.8	75.0	65.4	70.7
修正後	4	80.9	61.6	69.2	71.0	61.3	64.0	76.7	**68.0**	72.8
利用)	5	**83.4**	**67.4**	**73.7**	76.3	65.2	68.3			
BL2	2	77.0	58.5	65.8	65.0	58.0	59.9	70.5	62.4	66.9
(**MEM**	3	81.1	60.3	68.5	69.1	60.5	62.9	74.2	63.3	69.3
修正後	4	81.3	61.3	69.1	69.7	61.6	63.9	75.4	64.3	70.4
利用)	5	82.6	66.6	72.8	72.6	63.1	65.7	77.2	67.5	72.9
SCRF	2	81.3	65.6	71.8	79.5	68.3	71.4	81.3	67.0	74.9
	3	81.5	66.1	72.2	79.5	68.5	71.6	**81.5**	67.0	**75.1**
	4	81.6	66.3	72.3	79.5	68.8	71.7	81.3	67.0	74.9
	5	81.7	67.2	72.9	**80.1**	**69.2**	**72.3**			
SMEM	2	81.5	65.3	71.7	78.5	68.3	71.1	79.9	66.9	74.1
	3	81.3	65.2	71.6	78.5	68.3	71.1	79.8	66.9	74.1
	4	81.7	65.2	71.7	78.9	68.3	71.2	79.8	66.6	73.9
	5	81.6	65.5	71.8	79.2	67.9	71.0	79.7	66.7	73.9

表 6: 語義曖昧性解消結果 (SVM)

また、上位意味クラス推定の結果のみを構文解析などに利用する事もできる。実際、 Fujita et al. (2007) によると、構文解析の正解選択においてレベル 2 の意味クラスが最も精度向上に寄与している。

今後の課題としては、英語等の他の言語でも同様の効果を得られるかどうか実験を行ないたい。また、本稿では、パッケージ化された CRF の学習ツールを用いたが、形態素解析ツールである mecab(Kudo et al., 2004) での実装と同様に、可能な字面と上位意味クラスの組合せを辞書として登録し、組み合わせを制限することで、精度と学習速度の向上をはかりたい。

5 おわりに

本稿では、上位意味クラス推定を用いた語義曖昧性解消 (WSD) 方法を提案した。本手法では、まず上位意味クラスを推定してから、その推定結果を用いて WSD を行なう。上位意味クラスの推定では、 CRF と MEM を用いた実験を行ない、共に高い精度を得た。また、 WSD でも、 SENSEVAL-2 で最も高い精度を出した方法より高い精度を得る事ができた。これにより、提案手法である上位意味クラス推定を用いた WSD は効果的であるといえる。

謝辞

本研究で利用した CRF の学習ツールは、鈴木潤氏に御提供いただいた物 (Suzuki et al., 2006) です。この場を借りてお礼申し上げます。

参考文献

Timothy Baldwin, Su Nam Kim, Francis Bond, Sanae Fujita, David Martinez, and Takaaki Tanaka. 2008. Mrd-based word sense disambiguation: Further extending lesk. In *The Third International Joint Conference on Natural Language Processing (IJCNLP-2008)*.

Francis Bond, Sanae Fujita, and Takaaki Tanaka. 2006. The Hinoki syntactic and semantic treebank of Japanese. *Language Resources and Evaluation*, 40(3–4):253–261.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40.

Chih-Chung Chang and Chih-Jen Lin. 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Sanae Fujita, Francis Bond, Stephan Oepen, and Takaaki Tanaka. 2007. Exploiting semantic information for hpsg parse selection. In *ACL 2007 Workshop on Deep Linguistic Processing*, pages 25–32.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 230–237.

Kamal Nigam, John Lafferty, and Andrew McCallum. 1999. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67.

Jun Suzuki, Erik McDermott, and Hideki Isozaki. 2006. Training conditional random fields with multivariate evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 217–224.

Takaaki Tanaka, Francis Bond, Timothy Baldwin, Sanae Fujita, and Chikara Hashimoto. 2007. Word sense disambiguation incorporating lexical and structural semantic information. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 477–485.

池原 悟, 宮崎 雅弘, 白井 諭, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳史, 林 良彦. 1997. 日本語語彙大系. 岩波書店.

国立国語研究所. 2004. 分類語彙表 *CD-ROM (*増補改訂版版*)*. 大日本図書.

村田 真樹, 内山 将夫, 内元 清貴, 馬 青, 井佐原 均. 2003. 技術資料 SENSEVAL-2J 辞書タスクでの CRL の取り組み - 日本語単語多義性解消における種々の機械学習手法と素性 の比較. 自然言語処理学会論文誌, 10(3):115–134.

笠原 要, 佐藤 浩史, Francis Bond, 田中 貴秋, 藤田 早苗, 金杉 友子, 天野昭成. 2004. 「基本語意味データベース:lexeed」の構築. *2004-NLC-159*, pages 75–82.