対訳コーパスからの動詞の格フレーム獲得における動詞の多義性の類別*

宇津呂 武仁
奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

自然言語コーパスから動詞の格フレームなどの語彙的知識を獲得する際には，一つの動詞の持つ複数の意味を分けて，それぞれの意味ごとに語彙的知識を獲得しなければならない．我々はこの問題に対して，日英対訳コーパスを用い，英語側の情報を参照することにより，日本語動詞の多義性類別が効率よく行なえることを示した [宇津呂 93, 宇津呂 94]．また，日本語コーパスだけを用いる場合には，類義動詞の用例を利用することにより多義性の類別が行なえることを示した [内元 94]．しかし，これらの研究においては，シソーラス中で意味的に近いとみなせる範囲をあらかじめ明示的に指定する必要があったため，特定のシソーラスの影響を受けるおそれがあった．また，個々のシソーラスについて，どの範囲が意味的に近いとみなせるかどうかを経験的に知る必要があった．

これに対して，最近，単語と単語クラスの間の共起性を計るための Association Score と呼ばれる尺度が提案され，シソーラスのような単語クラス体系中の任意の位置における意味的なまとまりを扱うものとして用いられている [Resnik92, Framis94]．そこで，本論文では，対訳コーパスからの動詞の格フレーム獲得において，動詞の多義性を類別する際に，この Association Score の考え方によりシソーラス中の意味的なまとまりを検出し，意味的に近い用例のクラスタリングを行なう手法を述べる．特に，日本語および英語のそれぞれの言語のシソーラスを利用した Association Score を定義し，それぞれの言語のシソーラス中の意味的なまとまりを考慮に入れながら，クラスタリングを行なう．

2 用例の表現形式

多義性の類別に用いるデータとしては，対訳例の依存構造を部分構造同士で対応させたもの [Matsumoto93] を用いる．ここでは，対訳例の依存構造を照合した結果を以下のような素性構造で表現する．

$$\left[\begin{array}{l} pred : v_J \\ sem_E : SEM_E \\ p_1 : \left[\begin{array}{l} pred : n_{J1} \\ sem : SEM_{J1} \end{array}\right] \\ \quad\vdots \qquad\quad \vdots \\ p_m : \left[\begin{array}{l} pred : n_{Jm} \\ sem : SEM_{Jm} \end{array}\right] \end{array}\right]$$

ただし，v_J は日本語動詞を，$p_1,\ldots,p_m$ は日本語の表層格ラベルを，$n_{J1},\ldots,n_{Jm}$ は日本語格要素の名詞を表す．また，SEM_Eは，英語側の述語の英語シソーラス中の意味カテゴリの集合を表す．SEM_{Ji}についても同様に，各要素の名詞の日本語シソーラス中の意味カテゴリの集合を表す．ある語が複数の意味を持つ場合，この語はシソーラス中の複数の (葉の位置の) 意味カテゴリをもつので，ここでは，SEM_E および各 SEM_{Ji} をシソーラス中の意味カテゴリの集合によって表現し，これを**意味ラベル**と呼んでいる．ただし，$c_{E1},\ldots,c_{Ek}$ は英語シソーラス中の意味カテゴリ，$c_{J1},\ldots,c_{Jl}$は日本語シソーラス中の意味カテゴリである．

$$SEM_E = \{c_{E1},\ldots,c_{Ek}\}$$
$$SEM_{Ji} = \{c_{J1},\ldots,c_{Jl}\}$$

3 日英単言語シソーラス

本論文の方法では，日本語シソーラスとして分類語彙表 [国研 93] を，英語シソーラスとして Roget のシソーラス [Roget11] を用いる．分類語彙表は6層の階層構造から構成され，階層構造の葉の部分 (第6層目) に総数約 60,000 語の単語が分類されている．また，その体の類 (名詞シソーラス) の部分には，総数約 45,000 語の名詞が分類されており，根節点の一つ下のレベルは，「抽象的関係」，「人間活動の主体」，「人間活動—精神および行為」，「生産物および用具」，「自然物および自然現象」という5つの意味カテゴリからなる．Roget のシソーラス [Roget11] は7層の階層構造から構成されている．ただし，階層構造の葉の部分は必ずしも第7層目にある

*Classifying Multiple Senses in Verbal Case Frame Acquisition from Bilingual Corpora

とは限らない．葉の部分に約100,000語の単語が分類されている．品詞分類よりも意味分類を優先させており，異なる品詞間でも意味的類似度が適切に表現されている．根節点の一つ下のレベルは，Abstract Relations, Space, Matter, Intellect, Volition, Affections という6つの意味カテゴリからなる．

本論文では，シソーラスを，各節点が意味カテゴリである木構造として扱う．また，$\preceq$を，各節点の意味カテゴリの上位・下位関係とし，意味カテゴリ c_1 が c_2 よりも下位の場合に，$c_1 \preceq c_2$ と表す．特に，意味ラベル $SEM=\{c_1,\ldots,c_n\}$ については，その要素となっているどの意味カテゴリよりも下位であるとする．すなわち，

$$\forall c \in SEM, \quad SEM \preceq c$$

であるとする．これにより，意味ラベル内の意味カテゴリの曖昧性については，あらゆる可能性を考慮して取り扱う．

4 動詞の多義性の類別

4.1 Association Score の考え方

従来から，単語間の共起性を計る尺度としては，単語間の相互情報量を用いたもの [Church90] がよく使われている．これに対して，Association Score は，単語間の共起性を単語と単語クラスの共起性に拡張した尺度として用いられている [Resnik92, Framis94]. [Resnik92, Framis94] では，特に，動詞と名詞クラスとの間の共起性を計る尺度について述べている．

いま，コーパス中の全動詞の集合を $\mathcal{V}$，全名詞の集合を $\mathcal{N}$とすると，ある動詞 $v(\in \mathcal{V})$ および名詞クラス $c(\subseteq \mathcal{N})$ の共起確率 $\Pr(v,c)$ は以下のようになる．

$$\Pr(v,c) = \frac{\sum_{n\in c} \text{共起頻度}\,(v,n)}{\sum_{v'\in\mathcal{V}}\sum_{n'\in\mathcal{N}} \text{共起頻度}\,(v',n')}$$

これを用いると，動詞 vと名詞クラス c の共起性を表す Association Score $A(v,c)$ は，vが c と共起する条件付き確率と vと c の相互情報量の積として，以下のように定義される．

$$A(v,c) = \Pr(c \mid v)\log\frac{\Pr(v,c)}{\Pr(v)\Pr(c)}$$

ここで，第一項の条件付き確率が vと c の共起の一般性を表し，第二項の相互情報量が vと c の共起の強さを表す．[Resnik92] では，実際にこの Association Score が高かった動詞・名詞クラス対の例として，*call-someone, climb-stair, cook-meal* などを挙げている．

4.2 二言語の情報を利用した Association Score

Association Score の考え方の特徴は，共起確率と相互情報量の積で共起性を計ることによって単語クラスを含む組の間の共起性を適切に計ることができる点にあり，この考え方は広く利用可能な一般的なものであるといえる．一方，本論文は，二言語の情報で構成された用例中の英語述語の意味カテゴリおよび日本語格要素の意味カテゴリを用いて，意味的に近い用例のクラスタリングを行ない，日本語動詞の多義性を類別することを目的としている．そこで，ここでは，日本語動詞 v_Jについて，ある格 p の格要素としてどのような意味カテゴリの名詞をとる時に，どのような意味カテゴリの英語述語に翻訳されやすいかという共起性を Association Score によって表現することを考える[1]．

まず，対訳コーパス中から得られる日本語動詞 v_J の用例の集合を $Eg(v_J)$ とする．また，$Eg(v_J)$ の要素のうちで，英語述語の意味カテゴリ SEM_Eが英語シソーラス中の意味カテゴリ c_Eを満たす，すなわち $SEM_E \preceq c_E$となるものの集合を $Eg(c_E|v_J)$，日本語動詞 v_Jの格 p の格要素の名詞の意味カテゴリ SEM_J が日本語シソーラス中の意味カテゴリ c_Jを満たす，すなわち $SEM_J \preceq c_J$となるものの集合を $Eg(c_J|v_J,p)$，$Eg(c_E|v_J)$ と $Eg(c_J|v_J,p)$ の交わり集合を $Eg(c_E,c_J|v_J,p)$ とする．すると，$Eg(v_J)$ 中での $Eg(c_E \mid v_J)$，$Eg(c_J|v_J,p)$，$Eg(c_E,c_J|v_J,p)$ の要素の割合を表す条件付き確率 $\Pr(c_E,c_J|v_J,p)$, $\Pr(c_E|v_J)$, $\Pr(c_J|v_J,p)$ は，以下のように定義される．

$$\Pr(c_E \mid v_J) = \frac{|Eg(c_E \mid v_J)|}{|Eg(v_J)|}$$

$$\Pr(c_J \mid v_J,p) = \frac{|Eg(c_J \mid v_J,p)|}{|Eg(v_J)|}$$

$$\Pr(c_E,c_J \mid v_J,p) = \frac{|Eg(c_E,c_J \mid v_J,p)|}{|Eg(v_J)|}$$

さらに，これらの条件付き確率を用いることにより，日本語動詞 v_Jのある格 p のもとでの，英語シソーラス中の意味カテゴリ c_Eと日本語シソーラス中の意味カテゴリ c_Jの Association Score $A(c_E,c_J|v_J,p)$ が以下のように定義される．

$$A(c_E,c_J \mid v_J,p) = \Pr(c_E,c_J \mid v_J,p)\log\frac{\Pr(c_E,c_J \mid v_J,p)}{\Pr(c_E \mid v_J)\Pr(c_J \mid v_J,p)}$$

[1]ただし，本論文では，v_Jのどの用例も必ず格 p を持ち，格 p の格要素の名詞が v_Jの多義性類別の有力な手がかりとなると仮定している．

4.3 Association Score を用いたクラスタリング

日本語動詞 v_J の用例の集合 $Eg(v_J)$ において，v_J の多義性を類別しながら $Eg(v_J)$ の要素をクラスタリングしていく際には，得られたクラスターが適切であるかどうかの尺度が必要となる．本論文では，前節で定義した Association Score をこの尺度として用い，Association Score ができるだけ大きくなるようにクラスターを構成していく．クラスタリングは，次の二つの手順からなる．

4.3.1 Association Score 最大の上位節点の探索

$Eg(v_J)$ 中のある用例 e の英語述語の意味ラベルが SEM_E，格 p の格要素の名詞の意味ラベルが SEM_J であるとする．ここでは，まず，各用例について，

> $SEM_E \preceq c_E$，$SEM_J \preceq c_J$ を満たすシソーラス中の上位節点 c_E，c_J の組のうちで，$A(c_E, c_J | v_J, p)$ の値が最大となるもの

を求める．これにより，各用例に対して，意味的なまとまりがもっとも大きい上位概念の組 c_E, c_J が求まることになる．また，意味カテゴリの条件 c_E, c_J を満たす用例の集合 $Eg(c_E, c_J | v_J, p)$ を構成することにより，各用例が含まれるべきクラスターも同時に求めることができる．各用例は，どれか一つのクラスターにのみ含まれることが保証されるので，全用例の集合 $Eg(v_J)$ を，Association Score が大きく，しかも互いに素なクラスターに分割できたことになる．

4.3.2 意味的に近い複数のクラスターの統合

前節の処理の結果得られるクラスターは，日英シソーラス中の単一の意味カテゴリの組で表現されているため，意味的に近い用例だけが含まれる可能性は高いものの，動詞 v_J の意味の区別の数から比べると，全体としてクラスター数が多くなってしまう傾向がある[2]．そこで，ここでは，前節の処理で得られたクラスターのうちで，本来，動詞 v_J の意味としては同じであると考えられるものを検出してこれらのクラスターを統合することを行なう．また，意味的に近いかどうかを判定する際には，統合することによって Association Score が増加するかどうかを調べることとする．つまり，日英シソーラス中の単一の意味カテゴリの組で表現するよりも，複数の意味カテゴリを用いて表現した一つのクラスターとみなした方が Association Score が大きくなる場合に，クラスターの統合を行ない一つのクラスターとみなす．

具体的には，まず，二つのクラスター $Eg(c_{E1}, c_{J1} | v_J, p)$，$Eg(c_{E2}, c_{J2} | v_J, p)$ について，これらが統合可能かどうかを調べるために，英語側の意味カテゴリ c_{E1}，c_{E2} あるいは日本語側の意味カテゴリ c_{J1}，c_{J2} がシソーラス中で重なりを持つかどうかを見る．この条件は，以下の式によって調べる．

$$Eg(c_{E1} | v_J) \cap Eg(c_{E2} | v_J) \neq \emptyset \text{ または}$$
$$Eg(c_{J1} | v_J, p) \cap Eg(c_{J2} | v_J, p) \neq \emptyset$$

これは，英語側にも日本語側にも重なりのないクラスターを統合することを避けるための条件である．

また，二つのクラスター $Eg(c_{E1}, c_{J1} \mid v_J, p)$，$Eg(c_{E2}, c_{J2} | v_J, p)$ を統合したクラスターは，意味カテゴリの選言を用いて

$$Eg(c_{E1} \vee c_{E2}, c_{J1} \vee c_{J2} \mid v_J, p)$$

と表現される．このクラスターに対する Association Score も，定義に従い同様に計算される．

次に，統合可能であると判定されたクラスターの組に対して，あらゆる統合のパターンについて各クラスターの Association Score を計算し，全クラスターについて Assocation Score の平均をとった値が最大となる統合のパターンを求める．すなわち，全用例の集合 $Eg(v_J)$ の最終的な分割結果が $Eg(C_{E1}, C_{J1} | v_J, p), \ldots, Eg(C_{Et}, C_{Jt} | v_J, p)$ となったとして (ただし，各 C_{Ei}, C_{Ji} は，シソーラス中の意味カテゴリの任意個の選言である)，Association Score の平均値

$$\frac{\sum_{i=1}^{t} A(c_{Ei}, c_{Ji} | v_J, p)}{t}$$

を最大とするような統合を行なう．

4.4 例

講談社学術文庫の和英辞典 [清水 79] 中の対訳例から動詞「買う」を含む用例を集め，Association Score を用いたクラスタリングを行なって多義性を類別した結果を表 1に示す．ただし，日本語の格としては,「を」格に注目した.「歓心を買う」および「努力を評価する」という意味の用例が,「反感などを招く」という意味を表すクラスターに混在している点と,「評価する」という意味を表すクラスターが二つに分かれてしまっている点以外は，概ね適切な類別が行なえている．これら

[2] 例えば，[春野 94] では，Association Score を応用して，複数の格の情報を考慮した格フレームの有効度を定義し，日本語動詞の格フレーム獲得を行なっているが，人手で記述した格フレーム辞書中の格フレーム数の約 5 倍の数の格フレームが獲得されており，同様の傾向が見られている．

表 1: 「を」格に注目した動詞「買う」の用例のクラスタリング

クラスター No.		英語カテゴリ (レベル, 例)	日本語カテゴリ (レベル, 例)	統合前 用例数	統合前 Ass.Sc.	統合後 用例数	統合後 Ass.Sc.
1	1	*buy*(葉レベル)	車 (葉レベル)	4	0.024	67	0.227
	2		140001(6 レベル, 品)	2	0.149		
	3		13220(5 レベル, 絵)	3	0.019		
	4		14590(5 レベル, 本)	4	0.149		
	5		15111(5 レベル, 宝石)	1	0.012		
	6		131(3 レベル, 書物)	3	0.048		
	7		15(2 レベル, 株)	2	0.036		
	8		11(2 レベル, 土地)	3	0.054		
	9	Purchase(葉−1 レベル,*buy,pay*)	14(2 レベル, 生産物)	34	0.149		
	10,11	*treat oneself to*	外車/指輪 (葉レベル)	1,1	0.070		
	12	*purchase*(葉レベル)	地所 (葉レベル)	1	0.083		
	13	*bring*(葉レベル)	土産 (葉レベル)	1	0.062		
	14	*get*(葉レベル)	おもちゃ(葉レベル)	1	0.070		
2	15	*incur*(葉レベル)	130(3 レベル, 恨み, 反感)	5	0.185	6	0.406
	16	Motive/Excitation (葉−1 レベル,*arouse,rouse*)	反感 (葉レベル)	3	0.169		
	17	*disgust*(葉レベル)	ひんしゅく (葉レベル)	1	0.083		
	18	*appreciate*(葉レベル)	努力 (葉レベル)	1	0.083		
	19	*win*(葉レベル)	敏心 (葉レベル)	1	0.083		
3	20	*use*(葉レベル)	手腕 (葉レベル)	1	0.083	1	0.083
4	21	*get an opinion of*(葉レベル)	男 (葉レベル)	1	0.083	1	0.083
計				75	—	75	—

の問題に関しては，それぞれの意味の用例が少数しかないことが原因であると思われる．

5 おわりに

対訳コーパスからの動詞の格フレーム獲得において，動詞の多義性を類別する方法として，二言語のシソーラスを利用した Association Score を用いて意味的に近い用例のクラスタリングを行なう手法を述べた．本論文では，あらかじめ決められたただ一つの格にのみ注目する方法を述べた．今後は，あらゆる格を考慮し，しかも動詞の多義性類別の有力な手がかりとなる格にのみ注目するように，本論文の Association Score を拡張することを考えている．

参考文献

[Church90] Church, K. W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*, Vol. 16, No. 1, pp. 22–29 (1990).

[Framis94] Framis, F. R.: An Experiment on Learning Appropriate Selectional Restrictions from a Parsed Corpus, *Proceedings of the 15th COLING*, pp. 769–774 (1994).

[春野 94] 春野雅彦：最小汎化を用いたコーパスからの動詞格フレーム学習, 日本ソフトウェア科学会第 11 回全国大会論文集, pp. 409–412, 日本ソフトウェア科学会 (1994).

[国研 93] 国立国語研究所: 分類語彙表, 秀英出版 (1964,1993).

[Matsumoto93] Matsumoto, Y., Ishimoto, H. and Utsuro, T.: Structural Matching of Bilingual Texts, *Proceedings of the 31th Annual Meeting of ACL*, pp. 23 – 30 (1993).

[Resnik92] Resnik, P.: WordNet and Distributional Analysis: A Class-based Approach to Lexical Descovery, *Proceedings of the AAAI-92 Workshop on Statistically-Based Natural Language Programming Techniques*, pp. 48–56 (1992).

[Roget11] Roget, S. R.: *Roget's Thesaurus*, Crowell Co. (1911).

[清水 79] 清水護, 成田成寿（編）：和英辞典, 講談社学術文庫 (1979).

[内元 94] 内元清貴, 宇津呂武仁, 長尾眞：動詞の語彙的知識獲得における類義語の用例を用いた多義性の類別, 情報処理学会研究報告, Vol. 94, No. 47 (94–NL–101), pp. 105–112 (1994).

[宇津呂 93] 宇津呂武仁, 松本裕治, 長尾眞：二言語対訳コーパスからの動詞の格フレーム獲得, 情報処理学会論文誌, Vol. 34, No. 5, pp. 913–924 (1993).

[宇津呂 94] 宇津呂武仁：コーパスからの動詞の語彙的知識獲得における多義性の類別について,「自然言語処理における学習」シンポジウム論文集, pp. 174–181, 電子情報通信学会・日本ソフトウェア科学会 (1994).