格フレームを選択する三手法の比較

内山 将夫　　板橋 秀一
筑波大学

1 はじめに

単文(入力文)の格フレームを決めることは，機械翻訳における訳語選択など，自然言語処理の様々な分野で必要とされる．この選択に，実例と入力文との類似性を利用する手法が提案されている[1]．このような手法においては類似性の定義が重要な問題である．従来の研究では，グラフの上における距離が用いられることが多かった．それに対して，本稿では，入力文を生成する格フレームを確率最大という観点から選択する三手法について比較する．比較する手法は，(1)選択制限＋概念体系，(2)事例＋概念体系，(3)事例＋共起関係，である．まず，2節において各手法を定式化し，3節で実験の結果を述べる．

2 各手法の定式化

入力文s[1]の，確率最大の格フレームfとは次式を満すものである．

$$f = \arg\max_F \Pr(F|s) = \arg\max_F \Pr(F)\Pr(s|F). \quad (1)$$

$\Pr(s)$は最大化に無関係なので除いた．$\Pr(S,F)$は格フレームFである単文Sが生起する確率である．その他の確率は$\Pr(S,F)$から導出される．

格フレームFはn個の格スロットの集合である．格スロット$\mathcal{S}_i$は，それを充足する概念についての条件M_iと，格フレームにおける役割r_iの対$\langle M_i, r_i\rangle$である．入力文sの形態素のうちで，$\mathcal{S}_i$を充足する形態素をw_iとする．たとえば，$F = \{\mathcal{S}_1, \mathcal{S}_2 | \mathcal{S}_1 = \langle \mathrm{ANI}, \mathrm{agent}\rangle, \mathcal{S}_2 = \langle \mathrm{WALK}, \mathrm{act}\rangle\}$であり，$s$ = “太郎が歩く”，であるとすると，w_1 = “太郎”，w_2 = “歩く”，である．簡単のため，入力文を構成する形態素により，すべての格スロットは過不足なく充足されると仮定する．

[1]不特定の値が代入されている確率変数を英語のアルファベットの大文字，特定の値が代入されている確率変数を英語のアルファベットの小文字で表す．

入力文を，格スロットと対応づけられた形態素の集合により表す．形態素には一般に複数の概念が対応するので，入力文は複数の概念集合に対応する．w_iの任意の概念をC_iで表すと，入力文の概念集合$C(s)$は$\{C_1, ..., C_n\}$で表される．添字は対応する格スロットを示す．入力文が表す概念集合は，(1)式のF,sを置き換えた，次式を満すfを決めることにより決まる．

$$f = \arg_F \max_{F,C(s)} \Pr(\mathcal{S}_1, ..., \mathcal{S}_n)\Pr(C_1, ..., C_n|\mathcal{S}_1, ..., \mathcal{S}_n). \quad (2)$$

単純化のため以下の3式を仮定すると，(3)となる．

$$\begin{aligned}\Pr(\mathcal{S}_1, ..., \mathcal{S}_n) &= \Pr(\mathcal{S}_1)...\Pr(\mathcal{S}_n),\\ \Pr(C_i|\mathcal{S}_1, ..., \mathcal{S}_n) &= \Pr(C_i|\mathcal{S}_i),\\ \Pr(C_i, C_j|\mathcal{S}_1, ..., \mathcal{S}_n) &= \Pr(C_i|\mathcal{S}_1, ..., \mathcal{S}_n)\Pr(C_j|\mathcal{S}_1, ..., \mathcal{S}_n).\end{aligned}$$

$$f = \arg_F \max_{F,C(s)} \prod_{i=1}^{n} \Pr(\mathcal{S}_i)\Pr(C_i|\mathcal{S}_i). \quad (3)$$

$\Pr(\mathcal{S}_i)$は，(格フレームを捨象した)格スロット$\mathcal{S}_i$の生起確率であり，$\Pr(C_i|\mathcal{S}_i)$は，格スロット$\mathcal{S}_i$に概念C_iが生起する確率である．

さらに，概念の生起確率が格によらないと仮定すると，以下の2式が成り立ち，$\Pr(r_i)$は最大化に関係ないので除くと，(4)式となる．

$$\begin{aligned}\Pr(\mathcal{S}_i) &= \Pr(M_i, r_i) = \Pr(M_i)\Pr(r_i),\\ \Pr(C_i|\mathcal{S}_i) &= \Pr(C_i|M_i, r_i) = \Pr(C_i|M_i),\end{aligned}$$

$$f = \arg_F \max_{F,C(s)} \prod_{i=1}^{n} \Pr(M_i)\Pr(C_i|M_i). \quad (4)$$

2.1 選択制限＋概念体系

木構造である概念体系(シソーラス)が与えられているときには，M_iとしてシソーラス上のノードを指定できる．C_iは典型的にはシソーラスの葉に位置するが，中間ノードでもかまわない．

$\Pr(M_i)$ = M_iの支配下の葉概念の生起確率の和 (5)

と定義する．次に，$\varphi_{M_i}(C_i)$ を，ノード M_i が葉 C_i を支配しているときには 1 そうでないときには 0 をとる関数として定義すると，(6) 式が成り立ち，(4) 式は (7) 式になる．

$$\Pr(C_i|M_i) = \frac{\Pr(C_i, M_i)}{\Pr(M_i)} = \frac{\Pr(C_i)}{\Pr(M_i)} \times \varphi_{M_i}(C_i). \quad (6)$$

$$f = \arg_F \max_{F,C(s)} \prod_{i=1}^{n} \Pr(C_i) \times \varphi_{M_i}(C_i). \quad (7)$$

2.2 事例

(1) 式を，(3) 式と同様な条件付き独立性を仮定して，(8) 式のように変形する．$\Pr(s, s'|F)$ は，格フレーム F である文が二つあったとして，それが入力文 s と事例 s' である確率である．また，$C(s') = \{C'_1, ..., C'_n\}$ において，C'_i は事例 s' の単語 w'_i の任意の概念である．添字は対応する格スロットを示す．

$$\begin{aligned} f &= \operatorname*{argmax}_F \sum_{s'} \Pr(F) \Pr(s, s'|F) \\ &= \arg_F \max_{F,C(s)} \sum_{C(s')} \prod_{i=1}^{n} \Pr(M_i) \Pr(C_i, C'_i|M_i) \\ &= \arg_F \max_{F,C(s)} \sum_{C(s')} \prod_{i=1}^{n} \Pr(M_i) \Pr(C_i|M_i) \Pr(C'_i|M_i). \end{aligned} \quad (8)$$

これを計算するのは大変なので次式を計算する．

$$f = \arg_F \max_{F,C(s),C(s')} \prod_{i=1}^{n} \Pr(M_i) \Pr(C_i|M_i) \Pr(C'_i|M_i) \quad (9)$$

選択制限＋概念体系のときとは異なり，M_i は与えられていない．M_i は C_i と C'_i との関係から (9) 式を満すようなものが選ばれる．

2.2.1 事例＋概念体系

概念体系としてシソーラスが与えられていて，C_i と C'_i とが固定されているときには，(9) 式を満す M_i は C_i と C'_i との共有ノードのうち最小の高さのものとなる．確率は，(5) 式と (6) 式を利用して計算される．

2.2.2 事例＋共起関係

概念集合 $\mathcal{C}$ と概念 c とが与えられたとき，c と同一の役割 r で同一の格フレームに使用されたことがあるかどうかで，$\mathcal{C}$ を $\mathcal{C}(c)^+$ と $\mathcal{C}(c)^-$ に分割する．

$$\begin{aligned} \mathcal{C}(c)^+ &= \{c' | \exists f (r(c, f) \wedge r(c', f))\} \\ \mathcal{C}(c)^- &= \mathcal{C} - \mathcal{C}(c)^+ \end{aligned}$$

$r(c, f)$ は，概念 c が格フレーム f において役割 r で使用されたことを示す．

概念 C_i と C'_i が与えられたとき，$\Pr(M_i)$ と $\Pr(C_i|M_i)$ と $\Pr(C'_i|M_i)$ とを以下のように定義する．

$$\Pr(M_i) = \sum_{c \in (\mathcal{C}(C_i)^+ \cup \mathcal{C}(C'_i)^+)} \Pr(c), \quad (10)$$

$$\Pr(C_i|M_i) = \Pr(C_i)/\Pr(M_i), \quad (11)$$

$$\Pr(C'_i|M_i) = \Pr(C'_i)/\Pr(M_i). \quad (12)$$

これらを利用して (9) 式が計算される．

2.2.3 事例を利用した二つの手法の共通性と選択制限との違いと

選択制限により格フレームを選ぶ方法では，格スロットを充足する概念の条件は所与であった．しかし，事例による方法では，入力文と事例との関係から条件が決まる．事例による手法では，(11) 式などから，(9) 式が (13) 式のようになる．

$$f = \arg_F \max_{F,C(s),C(s')} \prod_{i=1}^{n} \Pr(C_i) \Pr(C'_i|M_i). \quad (13)$$

(7) 式では，C_i が M_i を充足するかしないかが $\varphi_{M_i}(C_i)$ により示されていた．事例の場合には，充足するかしないかの程度が $\Pr(C'_i|M_i)$ により示される．M_i は C_i と C'_i とにより決まる．

M_i は，もし，C_i と C'_i とが当該の格スロットを共に充足するならば，それら以外に，どれくらいの概念がその格スロットを充足するかを示す．シソーラスを利用した場合には，M_i は，C_i と C'_i との共有ノードのうちで最小の高さのものとなる．この支配下にある概念は C_i や C'_i と同様に当該の格スロットを充足する．共起関係を利用した場合も同様であり，$\mathcal{C}(C_i)^+ \cup \mathcal{C}(C'_i)^+$ に含まれる概念は，C_i や C'_i と同様に当該の格スロットを充足する．どちらの手法においても，C_i と C'_i との関係から，概念集合を二つの同値類に分割する．一つは，C_i と C'_i とを含み，当該の格スロットを C_i と C'_i と同様に充足できる概念の集合である．他方は，その補集合である．

3 格フレームの選択実験

EDR日本語単語辞書，EDR日本語共起辞書，EDR概念体系辞書，EDR日本語動詞共起パターン副辞書[3]を用いて，単文の格フレームを選択する実験を行なった．

単語辞書は，入力文の概念集合を得るために使用した．共起辞書は，名詞概念と動詞概念との共起関係と頻度を得るために使用した．概念体系辞書は，(5)式の計算に用いた．ただし，概念体系辞書は多重継承を許すにもかかわらず，中間節点の生起確率を支配下の節点の生起確率の和としたため，実際よりも大きい値が与えられたことになる．共起パターン辞書からは，入力文の格フレームの候補を得た．これらには選択制限として格スロットを充足する概念の上位概念が記載されている．

3.1 格フレームの選択法

概念の確率は，日本語共起辞書に記載されている頻度に比例したものを与えたが，共起辞書にない概念についても，頻度0.1を与えた．条件付き確率については，名詞概念については，(6)式や(11)式で計算したが，動詞概念については，M_l と C_i とが一致するときに1，そうでないときには0を与えた．よって，格フレーム一つに対して最大で一つの動詞概念が対応する．また，格フレームと名詞概念との共起関係は，当該の格フレームの動詞概念と名詞概念との共起関係と同等であるとした．

2節では，入力文と格フレームとにおいて，格スロットと形態素との対応がとれていることを仮定したが，実際には，それは成り立たない．入力文における格の数を l，格フレーム F における格の数を $m(F)$，対応のとれた格の数を $n(F)$ とし，(14)式と(15)式により格フレームを選択した．選択制限の場合には，

$$f = \arg_F \max_{F,C(s)} \prod_{i=1}^{n(F)} \Pr(C_i) \times \varphi_{M_i}(C_i) \times \alpha^{l+m(F)-2n(F)}. \quad (14)$$

事例を利用した手法の場合には，

$$f = \arg_F \max_{F,C(s)} \prod_{i=1}^{n(F)} \Pr(C_i) \Pr(C'_i|M_i) \times (\alpha\beta)^{l+m(F)-2n(F)} \quad (15)$$

である．α や β は適当にえらんだ小さい値であり，$\alpha = 0.01/$(全概念の延べ数)，$\beta = 1/$(全概念の延べ数)とした．α は概念の確率の最低値，β は概念の条件付き確率の最低値に相当する．これらは対応のとれていない格スロットに対するペナルティである．

(15)式では(9)式と異なり，それぞれの格を満す名詞概念 C'_i の組み合わせについて事例 s' による制約がない．共起辞書における二項関係のみを使ったため，格を満す名詞概念間の組合せが失われたためである．

3.2 実験用の単文

実験用の単文は，IPALから，(1)格フレームが共起パターン辞書に登録されていて，(2)共起名詞概念が一つ以上ある．という2条件を満すもの70文を無作為抽出した．次に，それらに対して，係助詞を適当な格助詞に直し，人名を彼／彼女に変更した．また，名詞句からは主辞のみを抜き出し，動詞を基本形(終止形)に変更し，形態素に分離した．たとえば，「部屋は花の香りに満ちている」ならば「部屋-が　香り-に　満ちる」が処理の対象となる．

3.3 実験結果

70文について格フレームの平均個数は3.9であった．ただし，共起名詞概念が一つもない格フレームは計数から除いた．除かない場合には6.3である．各々の格スロットあたりの共起概念の延べ数について，平均値は14，中央値は2であった(共起概念が0のものを除くと 平均値 = 21，中央値 = 5)．

選択制限＋概念体系により格フレームを選択した結果を表1に示す．選択制限の結果が正解を含む場合と含まない場合とに分けて記述した．「格フレーム数」には格フレームの平均個数，「選択フレーム数」には選択された格フレームの平均個数を載せてある．選択制限に合格した場合には $25/32 = 0.78$ という高い割合で正解となるが，カバー率 $(32/70 = 0.46)$ は低い．そのため正解率も低い $(25/70 = 0.36)$．

事例＋概念体系による正解率は $42/70 = 0.60$ であり，事例＋共起関係による正解率は $35/70 = 0.50$ である．これらの手法を符号検定により片側検定で比較すると，事例＋概念体系は，選択制限＋概念体系と事

	格フレーム数	選択フレーム数	正解数	不正解数	合計	正解率
正解を含む	3.8	2.1	25	7	32	
正解を含まない	4.0	0.7	0	38	38	
全体	3.9	1.3	25	45	70	36 %

表 1: 選択制限＋概念体系

例＋共起関係に比べて，それぞれ有意水準 0.01，0.05 で有意に高く，正しい格フレームを選択する．また共起関係＋事例は選択制限＋概念体系と比べて 0.06 の有意水準で，正しく格フレームを選択する．

正解率が低い原因は，格スロットと概念との共起が少ないことにあると考えられる．中央値が 2 ということは，半数以上の格スロットは 2 個のみの概念としか共起していないということである．これは事例を利用した解析における問題の一つである [4]．

		選択＋概念		事例＋共起	
		正解	不正解	正解	不正解
事例＋概念体系	正解	14	28	32	10
	不正解	11	17	3	25

表 2: 事例＋概念体系と他の手法との関係

事例＋概念体系で正解／不正解であった単文について，他の二つの手法ではどうなのかを表2に示す．表から，事例＋概念体系と選択制限＋概念体系とは補完関係にあることがわかる．どちらかの手法で正解となる割合は $(14+28+11)/70 = 0.76$ である．一方，事例＋概念体系と事例＋共起関係とでは，前者が後者を包含するといえる．前者で不正解であったもので後者で正解であったものは 3 例のみである．

選択制限は，擬人的な表現など事前に予想がつきにくいものに対して有効ではないようである．たとえば，「ダンプカーが砂利を運ぶ」や「小鳥がねぐらに戻る」や「プロセスが走る」のような文は選択制限では取り扱いが難しくなる．このような表現には事例のほうが適している．一方，「彼が不平を並べる」のように比較的適用範囲が狭いものには選択制限が適しているようである．

4 おわりに

単文の格フレームを選択する手法として，(1) 選択制限＋概念体系，(2) 事例＋概念体系，(3) 事例＋共起関係，の三つを比較した．これらの共通点は，確率最大の格フレームを選択する点である．相異点は，格スロットの充足条件が所与であるかないか，概念集合の分割様式が静的か動的かである．

実験によると，事例＋概念体系が他の二つよりも有意に優れていた．事例＋概念体系と選択制限＋概念体系では，両者が補完関係にある．事例＋概念体系と事例＋共起関係では，前者が後者を包含する．

事例＋概念体系が事例＋共起関係を包含するということは，両者の性質が似ていることを示している．よって，共起関係の利用法として，上述の方法よりも精密な手法を用いれば，事例＋共起関係による方法が事例＋概念体系に匹敵する可能性がある．

参考文献

[1] 黒橋禎夫，長尾真：“格フレーム選択における意味マーカと例文の有効性について”自然言語処理研究会 91-11，情報処理学会，(1992)．

[2] 計算機用日本語基本動詞辞書 IPAL (Basic Verbs) 説明書，情報処理振興事業協会技術センター (1987)．

[3] 日本電子化辞書研究所：“ＥＤＲ電子化辞書マニュアル”，http://www.iijnet.or.jp/edr (1995)．

[4] 浦本直彦：“用例に基づく多義性解消における学習のための一手法”，言語処理学会第 1 回年次大会発表論文集 pp.65-68，(1995)．