分布類似度を用いた大規模格フレームの自動構築

濱田 慧 †　笹野 遼平 ‡　柴田 知秀 †　河原 大輔 ††　黒橋 禎夫 †
京都大学 †　東京大学 ‡　情報通信研究機構 ††
{hamada, ryohei, shibata, kuro}@nlp.kuee.kyoto-u.ac.jp, dk@nict.go.jp

1　はじめに

計算機で文章を理解するためには、少なくとも、文章においてどの単語とどの単語がどのような関係をもっているかを明らかにする必要がある。このような単語間の関連性を解析するためには、人間がもっている常識のような幅広い知識が必要となる。そのような知識のうちもっとも基本的なものが「格フレーム」である。格フレームとは、用言とそれに関係する名詞を集めたものであり、例えば「積む」という用言の格フレームとして以下のようなものが考えられる。

{ 従業員、運転手、…} が { 車、トラック、… } に { 荷物、物資 } を 積む

このような格フレームは、構文・格・省略解析のような文章中の要素間の関連性解析から検索、要約、翻訳のような言語処理アプリケーションまで広く応用できると考えられる。格フレームをコーパスから自動構築する手法は河原・黒橋によって提案されている [2, 3]。この手法によって構築された格フレームの例を表 1 に示す。

従来の格フレーム構築では、「荷物を積む」や「経験を積む」といった意味の異なる表現を区別するために、シソーラスに基づく類似度を用いてクラスタリングを行っている。シソーラスは分類語彙表を用いており、類似度は分類項目の木構造における単語間距離に基づいて計算される。しかし、シソーラスから計算される類似度にはいくつか問題点がある。問題の例を表 2 の左側に示す。まず、一般的に似ていない語同士であると考えられるが高い類似度を示す語は、稀にしか用いられないような意味の項目にも語が分類されていることが原因である。例えば、“肝” という語には、[膜・筋・神経・内臓] という項目の他に [心] という項目にも分類されているため、“心理” という語と高い類似度を示す。反対に、一般的に似ていると考えられる語同士の類似度が低いのは適切な項目に分類されていないためである。例えば “データ” と “情報” では “データ” は [本体・代理] という項目に分類され、“情報” は [伝達・報知] という項目に分類されているため、類似度が低くなっている。また、シソーラスにおいて未知語は扱うことができないため、類似度を計算する語のいずれかが未知語であれば類似度は 0 となる。例えば “メモリ” や “ＣＰＵ” などの単語はシソーラスにおいて未知語であるため、これらの類似度は 0 となる。

表 1: 格フレームの例 (積む)

積む:1	
ガ格	人:9、者:6、スタッフ:2、職員:2、…
ヲ格	経験:10088
デ格	現場:22、会社:22、中:19、分野:15…
…	
積む:2	
ガ格	人:4、子供:3、運転手:1、…
ヲ格	荷物:1235、石:521、自転車:189、…
二格	車:969、トラック:260、上:110、…
…	

表 2: 語の類似度の例 (類似度は 0〜1.0)

	シソーラス	分布類似度
似ていないと考えられる語		
肝と心理	0.71	0.02
誤解と警戒	0.71	0.04
似ていると考えられる語		
データと情報	0.14	0.26
企業と会社	0.14	0.20
取り扱えない未知語		
メモリとＣＰＵ	0.00	0.19
紐とロープ	0.00	0.24

このようなシソーラスに基づく類似度の問題によって、格フレームをうまくクラスタリングすることができない場合がある。そこで本研究では語と語の分布により類似度を与える分布類似度をシソーラスの代わりに用いて、格フレームを構築することを提案する。

2　格フレーム構築手法の概要

格フレーム自動構築システムの流れは以下の通りである [2, 3]。

1. コーパスに対して、KNP[4] を用いて構文解析を行い、信頼できる用言・格要素間の関係を取り

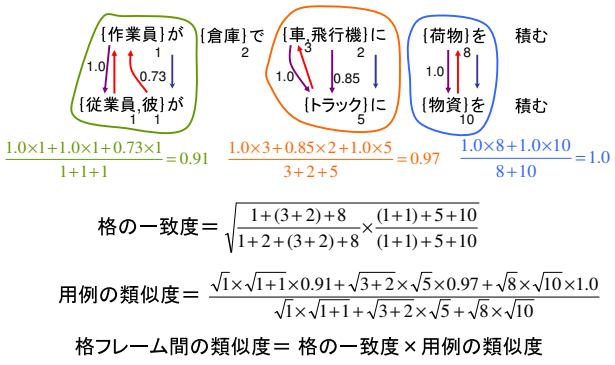

図 1: 格フレーム間の類似度の計算例

出す。

2. 抽出した関係を、用言とその直前の格要素のペアごとにまとめる。ここで作成されたデータを用例パターンと呼ぶ。
3. 頻度の高い上位 N 件の用例パターンに対してクラスタリングを行う。用例パターン同士の類似度を格要素間の類似度をもとに計算し、閾値 (th_{cl}) 以上のものを一つにまとめ、最初の格フレームを得る。本研究では N=50 とした。
4. クラスタリングに用いられなかった用例パターンを、最も類似度の高い格フレームに振り分ける。

格フレーム間の類似度は、共通の格に含まれるそれぞれの用例ごとにもっとも似ている相手格中の用例をみつけ、その類似度の平均をその格間の類似度とし、そのそれぞれの格の類似度の重み付け平均を求め、それに格の一致度「対応づけられた格の用例数/全格用例数」をかけることによって求められる (図 1)。

3 分布類似度による格フレーム構築

3.1 分布類似度の計算

分布類似度とは「語の出現分布の似ている語は意味も似ている」という考え方に基づいて計算される語と語の類似度である [1, 5]。分布類似度による類似度計算の概要は以下の通りである。

1. まず、ある語 w と曖昧性のない係り受け関係にある語とその間の関係をペアとして抽出する。このペアを共起要素と呼ぶ。例えば「医者に診せる」という文からは、「診せる:ニ」が医者の共起要素として抽出される。抽出した共起要素のうち、語 w と共起要素 v の自己相互情報量 (Pointwise Mutual Information, PMI) が正となるもののみを以下で用いる。

$$PMI(w,v) = \log \frac{P(w,v)}{P(w)P(v)} \quad (1)$$

表 3: 共起要素の例 (医者)

共起要素	自己相互情報量
不養生:ノ	12.801
診せる:ニ	12.199
完治:カラ	12.142
診る:ニ	10.934
解剖:ガ	10.203

表 4: 類義語の例 (医者)

類義語	Jaccard 係数
医師	0.226
医	0.219
教師	0.196
上司	0.190
親父	0.171

「医者」の共起要素の例を表 3 に示す。

2. 語 w_1,w_2 の分布類似度 sim_{dis} を各々の共起要素の重複率と定義し、Jaccard 係数によって計算する。

$$sim_{dis}(w_1,w_2) = Jaccard(w_1,w_2) \quad (2)$$
$$= \frac{|T(w_1) \bigcap T(w_2)|}{|T(w_1) \bigcup T(w_2)|} \quad (3)$$

ただし、$T(w)$ は共起要素の集合をあらわす。このように計算して得られた「医者」の類義語の例を表 4 に示す。

このようにして得られる分布類似度を用いると、シソーラスを用いた類似度計算で問題となっていた単語間の類似度は表 2 の右側のようになる。本研究では分布類似度が 0.18 以上のものは似ている語と考える。

上記の分布類似度の計算の結果より、シソーラスでの類似度計算における問題点を改善できると考えられる。

3.2 分布類似度を用いた格フレーム構築

分布類似度計算を用いた格フレーム構築では、クラスタリングの閾値 th_{cl} を 0.18 と設定した。これは格フレーム間の類似度のオーダーが用例間の類似度とほぼ同じと考えられるからである。

シソーラスを用いた語の類似度計算は、分類コードを利用して高速に計算することができる。一方、分布類似度を用いる場合、2 語の共起要素の重複率を計算する必要があり、格要素に現れる 100 万語近い語同士の類似度をあらかじめ計算しておくことは非現実的であり、また、格フレーム構築時にその場で分布類似度を計算すると、計算機クラスタを利用しても大規模格フレームを構築するのに非現実的な時間がかかってしまう。

そこで、頻度の高い語同士 (具体的には頻度上位 3 万語) の類似度はあらかじめ計算しておき、それ以外

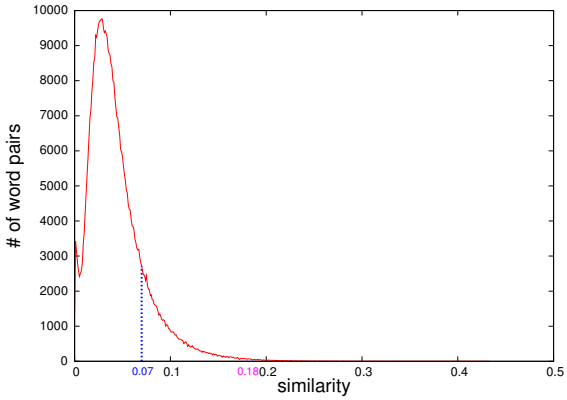

図 2: 分布類似度の頻度分布

表 5: 構築した格フレームの統計情報

用言数		64661
(内訳)	動詞	42323
	形容詞	2789
	名詞+判定詞	19299
用言あたりの平均格フレーム数		7.0
格フレームあたりの格の平均数		4.3
格あたりの平均用例数		134.2
格あたりの平均異なり用例数		9.9

は格フレーム構築時に動的に計算する。図 2 は分布類似度の頻度分布を表しており、閾値 (th_{dis}) よりも高いものだけを保持し、その他は 0 とする。ここで、閾値周辺で類似度が不連続にならないように、類似度を補正する。つまり類似度は以下の式で与えられる。

$$sim_{dis}(w_1, w_2)\prime = \begin{cases} \frac{sim_{dis}(w_1,w_2)-th_{dis}}{1-th_{dis}} & (sim_{dis}(w_1, w_2) \geq th_{dis}) \\ 0 & (sim_{dis}(w_1, w_2) < th_{dis}) \end{cases} \quad (4)$$

具体的には、$th_{dis} = 0.07$ とし、この時、格フレームクラスタリングの閾値 th_{cl} は $\frac{0.18-0.07}{1-0.07} = 0.118$ に補正される。このように計算することによって、200CPU を使うと格フレーム構築時間が 4 日となり、格フレーム構築が現実的な時間で終了するようになった。

4 実験・考察

4.1 実験

分布類似度を用いて、Web コーパス 5 億文より格フレームを構築した。構築した格フレームの統計情報を表 5 に示す。また、構築した具体的な格フレームの例を表 6 に示す。

「積む」の例では、シソーラスではマージされなかった “修業” “研鑽” “練習” などの用例が同じ格フレームにマージされた。また、どこにも振り分けられなかった “ノウハウ” が「積む:1」の [ヲ格] に振り分けられた。

表 6: 分布類似度での格フレームの例 (積む)

積む:1	
ガ格	人:9、者:6、スタッフ:2、職員:2、子供:2…
ヲ格	経験:10092、体験:431、ノウハウ:17、…
デ格	現場:22、会社:22、中:19、分野:16、企業、…
…	
積む:2	
ガ格	者:6、選手:4、僧:3、人:2、私:1、少女:1、…
ヲ格	修業:2116、研鑽:1647、練習:1516、訓練:946、…
デ格	基:118、下:60、元:38、店:31、大学:18、…
…	
積む:3	
ガ格	人:3、運転手:1、人間:1、業者:1、娘:1、…
ヲ格	荷物:1243、自転車:189、道具:125、機材:113、…
二格	車:977、トラック:261、船:142、バイク:20、…
…	

表 7: シソーラスを用いて構築した格フレームと分布類似度を用いて構築した格フレームの比較評価

	th	dis	比較
味わう	△	○	○
著しい	△	△	△
裏切る	×	○	○
襲う	△	△	△
恐れる	×	△	○
折る	×	△	○
飾る	△	○	○
固まる	○	○	△
寄与する	△	○	○
心がける	△	△	△

	th	dis	比較
積極的だ	×	△	○
散る	○	△	×
積む	△	○	○
解く	×	○	○
逃げる	×	△	○
冷やす	△	○	○
復元する	△	△	△
防ぐ	△	△	△
守る	△	△	△
躍進する	×	△	○

次に、構築した格フレームのうち、以下の 20 用言について人手で評価した。

「味わう」「著しい」「裏切る」「襲う」「恐れる」「折る」「飾る」「固まる」「寄与する」「心がける」「積極的だ」「散る」「積む」「解く」「逃げる」「冷やす」「復元する」「防ぐ」「守る」「躍進する」

まずは、シソーラスに基づいた類似度を用いて構築した格フレームと分布類似度を用いて構築した格フレームをそれぞれ○、△、×の 3 段階で評価した。そしてシソーラスに基づいた類似度を用いて構築した格フレームと分布類似度を用いて構築した格フレームを比較し、分布類似度を用いた方が用法ごとにうまくクラスタリングされていると判断できるものは○、どちらともいえないものには△、そうでないものは×という 3 段階で評価した。それぞれの評価結果を表 7 に示す。20 用言のうち、シソーラスに基づいた類似度を用いた場合よりも分布類似度を用いた方がよいと判断されたのは 12 用言であった。

4.2 考察

分布類似度を用いることによって改善された点を以下に示す。

- 似ていると考えられる語であるのにシソーラスでマージされていなかったものが、分布類似度ではマージされた

  シソーラス:{ 修業 } を積む、{ 研鑽 } を積む
  　　{ 練習 } を積む
  分布類似度:{ 修業、研鑽、練習…} を積む
- 頻度の少ない用例の振り分けで、新たに適切な語 (未知語を含む) が振り分けられた

  シソーラス:{ 病、症、…} を防ぐ
  分布類似度:{ 病、症、癌、エイズ…} を防ぐ
  (未知語の例)
  シソーラス:{ パソコン、コンピューター }
  　　が固まる
  分布類似度:{ パソコン、コンピューター、マッ
  　　ク、Windows、…} が固まる

分布類似度を用いることによって悪くなった点を以下に示す。

- 似ていないと考えられる語で、シソーラスでマージされなかったたものが、分布類似度ではマージされた。

  シソーラス:{ 桜、花、…} が散る
  　　{ 魚 } が散る
  分布類似度:{ 花、桜、魚、…} が散る

“花” と “魚” の共起要素は、「買う:ヲ」、「選ぶ:ヲ」、「開く: ヲ」、「育てる:ヲ」などであり、このような共起要素のため “花” と “魚” が高い類似度を示す。今回は、分布類似度を計算する際に Jaccard 係数を用いたが、他の係数を利用することも検討する予定である。

その他には以下のような例が見られた。

- 分布類似度によって新たに用例が振り分けられたが、適切でないものも目立つ

  シソーラス:{ 食事、食 } を心がける
  分布類似度:{ 食事、食、散歩、塩、給食 }
  　　を心がける

“食事” と “給食” は類似度 0.12 とある程度の類似度を示しており、うまく振り分けられた。“散歩” は「運動を心がける」とまとまるとよいと思われるが、“散歩” と “運動” の類似度は 0.05 で、“散歩” と “食事” の類似度 0.11 となっているため、“散歩” は「食事を心がける」の格フレームに振り分けられる。“塩” と “食事” の類似度は 0.06 であるが、他に振り分けられる格フレームがなかったため、「食事を心がける」の格フレームに振り分けられた。

- マージされて欲しい語が、シソーラスでも分布類似度でもまとまらなかった

  　　{ 住居 } を復元する
  　　{ 建物 } を復元する

“住居” と “建物” はシソーラスにおいてそれぞれ [住居]、[家屋・建物] という意味項目に分類されているため、類似度は高くない。分布類似度でも類似度は 0.12 であった。“住居” しか持たない共起要素に [帰る:へ]、[押しかける:ニ] などが、“建物” しか持たない共起要素に [そびえる:ガ]、[踏み入れる:ニ] などがあった。

- 分かれて欲しい語が、シソーラスでも分布類似度でも分かれなかった

  　　{ 痛み、波、…} が { 私 } を襲う

“痛み” と “波” はシソーラスを用いた類似度で 0.28、分布類似度で 0.12 であるが、共通の直前格 [私:ヲ] を持っているためマージされる。

分布類似度を用いることにより、未知語を取り扱えるようになり、類似度が 0 をとる語のペアが格段に減ったため、ほとんどの用例がどこかの格フレームに振り分けられるようになった。したがって、よりまとまっていて規模の大きい格フレームを構築することができた。したがって分布類似度を用いた格フレーム構築は、シソーラスによる類似度を用いた場合よりも有効であると考えられる。

5 おわりに

本稿では、分布類似度を用いた格フレームの自動構築について述べた。今回の実験により、人手のリソースを用いることなく、既存の格フレームと同程度あるいはそれ以上の精度の格フレームを構築できることがわかった。今後としては、シソーラスと分布類似度の長所をうまく組み合わせて格フレームを構築する予定である。

参考文献

[1] J.R.Firth. *Studies in Linguistic Analysis, chapter A synopsis of linguistic theory.* oxford, 1957.

[2] 河原大輔, 黒橋禎夫. 用言と直前の格要素の組を単位とする格フレームの自動構築. 自然言語処理, Vol. 9, No. 1, pp. 3–19, 2002.

[3] 河原大輔, 黒橋禎夫. 格フレーム辞書の漸次的自動構築. 自然言語処理, Vol. 12, No. 2, pp. 109–131, 2005.

[4] 黒橋禎夫, 長尾眞. 並列構造の検出に基づく長い日本語文の構文解析. 自然言語処理, Vol. 1, No. 1, pp. 35–57, 1994.

[5] 相澤彰子. Web コーパスを用いた語の類似度計算に関する考察. 情報処理, Vol. 2007, No. 67, pp. 45–52, 2007.