「常識的」推論規則のコーパスからの自動抽出

鳥澤 健太郎
北陸先端科学技術大学院大学情報科学研究科
Email: torisawa@jaist.ac.jp

1 はじめに

知的な言語処理を行うにあたっては, テキストやユーザー発話の言外の意味を常識にしたがって「推論」することが必要であり, 推論に関する多くの研究がなされて来た. しかしながら, そのような推論で利用されるべき推論規則が満足のいく量, 収集されたことはなく, 知的な言語処理を実現する上で問題となっていた. 本稿では, 我々の持つ常識を表すような推論規則を, コーパスから自動的に抽出する手法を提案する. 本手法で獲得された推論規則には, 例えば以下のようなものがある.

- 「もし, X が本を書くならば, { 普通 or しばしば }X が Z に本を出版する」
- 「もし, X がビールを買うならば, { 普通 or しばしば }X がビールの代金を払う」

以上のように提案手法で獲得される推論規則は自然言語で書かれており, X, Y といった変数を含む. このような規則の一つの問題は, 常に例外が付きまとうことである. 例えば, 本を書いても出版されず, お蔵入りになることもある. 規則中に { 普通 or しばしば } とあるのは, このような例外を許容させるためである. 残念ながら, 現在の手法は「普通」と「しばしば」の差を判別することができないためその両方を出力している.

一般に常識に関する推論規則の妥当性を評価することは難しい. 提案手法に自然言語で書かれた規則を出力させたのは, この評価をより容易にする目的もある. 本研究では複数の被験者 (6名) に自然言語で書かれた規則を読んでもらい, 妥当な規則であるかどうかの判定を行ってもらった. 我々は実験に先だって「普通」「しばしば」「もし A ならば B」といった表現に人工的な定義を与えることはせず, 被験者には各自の持つ解釈や直観を元に, 規則の妥当性を判断してもらった. また, 被験者には「普通」と「しばしば」の2つの副詞の内, それを付加すると規則が妥当でなくなるものを規則から消去するように依頼した. つまり, 評価後の規則には, 「普通」と「しばしば」がそれらを付加したときに規則が妥当であるときに限り, 付加されていることになる.

これまでにも推論規則を自動的に獲得する手法は研究されている. [2, 1] また, 異なる表現であるが, 同じ意味を持つもの同士を結び付ける推論規則, paraphrase に限定された自動収集については [3, 5, 6] がある. また, [8] では, 異なる意味を持つ表現であるが, 同一の常識的なシナリオの中で生じやすい出来事の間に生じる関係 (例えば, 「ビールを飲む」と「ビールに酔う」, 「ビールを買う」) などの間に生じる関係) を自動的に獲得する試みがなされている. 利用している手法という観点からは, 格フレームの意味的類似性を利用するもの [3, 8, 6], 特定の言語表現のパターンから規則の獲得をおこなうもの [2, 1], 単一言語のパラレルコーパスを利用するもの [5] と言ったように分類できる.

本研究では, このうち2種類の手法を組み合わせて使用するアルゴリズムを開発した. まず, 二つの動詞句 (あるいは文) を含む並列句表現をコーパスから抽出する. (より詳しくは, 主として連用中止形, あるいはテ形で二つの動詞句／文が結ばれた並列句表現を抽出する.) ついで, 後述する仮説に基づき, 推論規則を生成する種になれそうな並列句表現だけを取り出し, それを元に推論規則を生成する. 後述の仮説とは, 二つの動詞句の間での名詞の共有可能性に関するものである.

2 推論規則に関する仮説

本手法では, 例えば,

- ビールを飲み, 酔った

といった並列句表現で動詞句ないしは文を二つ含むものから,

- もし, X がビールを飲むならば, { 普通 or しばしば }X がビールに酔う

といった推論規則を獲得する. ここで並列句表現に注目したのは, コーパス中での頻度が大きいからである. 別の可能性として, 乾ら [1] は, 出来事間の論理的関係を表す「ため」などの接続詞に注目し推論規則の抽出を行っているが, 我々の実験では, そのような接続詞を含む表現に比べて, 並列句表現の方が数十倍のオーダーでコーパス中により多く現われていており, より幅広く多数の規則を獲得するという観点から並列句表現に注目することとした.

またもう一つ, 並列句表現に注目した理由としては, 我々人間にとって自明な推論規則は, 「ため」「ならば」のような論理的接続詞をつかっては表現されにくいという観察がある. 例えば, 「もし X が研修を受けるならば, 普通, X が研修を修了する」という規則が妥当で

あるとする．このような規則がどのような表現から生成できるかを考えると，まずは，以下にあるような，規則を直接的に表した表現や論理的関係を陽に記述した表現が候補となろう．

- もし研修を受けるならば，研修を修了する．
- 研修を受けたら，普通，研修を修了する．

しかしながら，このような自明な推論を論理的関係を陽に記述しつつ表現することはあまりない．一方で，以下のような並列句表現は，論理的関係を陽には表現していないが，頻繁に使われやすく，この点からも並列句からの推論規則の自動獲得が望ましい．

- 研修を受けて，修了した．

もちろん，任意の並列句から推論規則を生成できるわけではない．並列句が単に二つのイベントが比較的独立に生じたことを報告している場合には，そのような並列句から推論規則を生成すると常識に反する推論規則が得られてしまう．例えば，並列句「ビールを飲み，車を運転した．」は，交通事故に関する新聞記事でよくみられるが，これから生成された推論規則「もしXがビールを飲むならば，Xが車を運転する」は常識に反している．我々の手法では，このような並列句を次の仮説に基づき排除する．この仮説は，動詞句間の名詞の共有可能性に関するものである．

- 仮説：もし，二つの出来事 A, B をそれぞれ表す表現 e_A, e_B が同一の並列句表現に現われているときに，e_A に現われている名詞句が e_B にも現われている，あるいは現われやすい時に，e_A と e_B の間には「もし e_A ならば，e_B」という関係が成立しやすい．また，そうでなければ e_A と e_B の間には「もし e_A ならば，e_B」という関係が成立しにくい．

仮に e_A =「ビールを飲む」，e_B =「酔う」と仮定すると，名詞句「ビール」は e_B の中に現われやすい．別のいい方をすると，「ビールに酔う」という表現は頻繁に見られる．上の仮説はある並列句がこのように名詞（この場合はビール）を共有しやすいときに，その並列句からは推論規則を生成できると言っていることになる．

逆に e_A =「ビールを飲む」，e_B =「運転する」と仮定すると，名詞句「ビール」は e_B の中に現われにくい．つまり，「ビールで運転する」といったような表現はあまり見られない．上の仮説は，このような並列句から推論規則を生成するのは不適切であると主張している．

この仮説は次のような観察に基づいてたてられた．

- 観察：もし，二つの出来事 A, B をそれぞれ表す表現 e_A, e_B があり，「もし e_A ならば，e_B」という関係が成立しているならば，二つの出来事 A と B には共通の参加者があり，その参加者を通じて出来事 A, B 間には相互作用がある．

先の「ビールを飲む」の例でいけば，「飲む」と「酔う」という二つの出来事には共通の参加者「ビール」があり，それを介して相互作用が起きているとみなすことができる．一方「飲む」と「運転する」の間では「ビール」が共有されていないがゆえにそのような相互作用が生じていない．よって，両者が比較的独立に生起していると考えられる．ここで問題になるのは，「共通の参加者」とはなにかということである．実は，「飲む」と「運転する」の間でも行為者としての主語が共有されている．従って，単に，e_A, e_B に同時に現われている，あるいは現われやすい名詞句が参照しているものをすべて「共通参加者」とよぶのであれば，「飲む」と「運転する」にも共通参加者があることになってしまう．本研究では，二つの出来事に共通の行為者，すなわち主語が存在したとしても出来事の間に論理的関係が生じる可能性は比較的低いという観察に基づき，e_A 中で「目的語」となっている語で，e_B にも現われやすいものだけを「共通参加者」であるとみなした．この仮定に基づけば，「飲む」と「運転する」には「共通参加者」がないことになる．

3 自動獲得アルゴリズム

以下に推論規則獲得の具体的なアルゴリズムを与える．まず，アルゴリズムは並列句構造をコーパスから抽出することから始まる．[1] ここで，全ての動詞，全ての名詞，全ての助詞をそれぞれ，V, N, Rel で表記することとする．抽出された並列句は3つ組 $\langle C, Args_1, Args_2 \rangle$ として蓄積される．ここで，$C \subseteq V \times V \times R$ であり，R は実数すべてを含む集合であるとする．直観的には，$\langle v_1, v_2, f \rangle \in C$ であれば，動詞 v_1 を主辞とする動詞句 e_1 と，動詞 v_2 を主辞とする動詞句 e_2 が同じ並列句にコーパス中で f 回現われていることを表す．ただし，並列句中で現われる動詞句の順番としては，e_1, e_2 の順でなければいけない．また，今回の実験では構文解析エラーなどの悪影響を避けるため，5回以上共起している動詞対だけを C に含めている．

また，$Args_1$ は $V \times V \rightarrow 2^{N \times Rel \times R}$ となるような関数であり，並列句中で動詞に係っている名詞とその頻度を表す．仮に $\langle v_1, v_2, f \rangle \in C$ であり，$\langle n, p, f_n \rangle \in Args_1(v_1, v_2)$ であったとすると，これは動詞 v_1, v_2 を含む並列句内で，v_1 に名詞 n が助詞 p を介して f_n 回係っていることを示す．$Args_2$ も $Args_1$ と同様に，並列句中で動詞に係っている名詞とその頻度を表しているが，$Args_1(v_1, v_2)$ が v_1 に係っている名詞を示すのに対して，$Args_2(v_1, v_2)$ は v_2 に係っている名詞を示す．

仮に以下の文「ビールを買い，代金を払った．」が5回コーパス中で観測され，この文以外に「買う」と「払う」が共起している文が無いとする．このときには，$\langle$買う，払う，5$\rangle \in C$, $\langle$ビール，を，5$\rangle \in Args_1$(買う，払う)，

[1] 実際には「ので」などの論理的接続詞を含む文も利用しているが，我々の実験ではその数が並列句の2.5%程度であり，以下では並列句に説明を限定する．

- $Score(v,n,v_c) = MAX\{\{Score_v(v,n,v_c)\} \cup \{Score_{arg}(v,n,arg) | \langle arg, p_{arg}, r_{arg}\rangle \in Args_2(v,v_c) \wedge r_{arg} > \theta_{arg}\}\}$
- $Score_v(v,n,v_c) = \sum_{\alpha \in Class}\{P_V(v_c|\alpha)P(\alpha|n)(P(\alpha|\langle v, \mathtt{wo}\rangle) + P_C(\alpha|\langle v,v_c\rangle))\}$
  ただし, wo は助詞「を」を表す.
- $P_V(v_c|\alpha) = \sum_{p \in Rel} P(\langle v_c, p\rangle|\alpha)$ ただし $\alpha \in Class$
- $P_C(\alpha|\langle v,v_c\rangle) = \frac{1}{Z_{v,v_c}} \sum_{\langle w,\mathtt{wo},r\rangle \in Args_1(v,v_c)}\{P(\alpha|w) \cdot r\}$ ただし $\alpha \in Class.$ and $Z_{v,v_c} = \sum_{\alpha \in Class}\sum_{\langle w,\mathtt{wo},r\rangle \in Args_1(v,v_c)}\{P(\alpha|w)\cdot r\}$
- $Score_{arg}(v,n,arg) = \sum_{\alpha \in Class}\{P(\alpha|n)\}\{P(\langle arg, no\rangle|\alpha)\}\{P(\alpha|\langle v, \mathtt{wo}\rangle) + P_C(\alpha|\langle v,v_c\rangle)\}$

図 1: 並列句判別のためのスコア関数

$\langle$代金, を, 5$\rangle \in Args_2$(買う, 払う) となる.

以上をもとに, 適切な推論規則を生成できる並列句を弁別するスコア関数を定義した. 具体的な関数は, 図 1 にある $Score(v,n,v_c)$ であるが, v, v_c は並列句に現われる動詞, n は動詞 v の目的語となる名詞である. また, 並列句中では, v は v_c よりも前に現われるとする. 例えば、並列句「ビールを飲み, 酔った」に関するスコアの値は, $Score$(飲む, ビール, 酔う) で表されることになる. また, 結果として生成される推論規則では v は前提部の動詞, v_c は結論部の動詞となる.

図 1 には $Class$ という集合が現われているが, これは EM 法による単語クラスタリング [4, 7] で用いられる集合で, その各要素は単語の意味クラスの「名前」として機能している. また, 名詞 n と $Class$ の要素 α に対して, $P(n|\alpha)$ や $P(\alpha|n)$ といった確率も現われているが, これは EM 法によって推定される確率で, $P(\alpha|n)$ は名詞 n が意味クラス α の要素として使用される確率を表している. [2]同様に, 動詞 v と助詞 p, $Class$ の要素 α に対して, $P(\alpha|\langle v,p\rangle)$ という確率も使われているが, これは助詞 p を介して動詞 v に係る名詞が単語クラス α に属する確率を表している.

スコアの定義の概要は次のように述べることができる. $Score(v,n,v_c)$ は, $Score_v(v,n,v_c)$ と $Score_{arg}(v,n,arg)$ の二種類のスコアの最大値である. 前者 ($Score_v(v,n,v_c)$) は名詞 n が v_c になんらかの助詞 p を介して直接係る場合を扱っており, 後者 $Score_{arg}(v,n,arg)$ は v_c と n の間に別の名詞 arg が介在する場合を扱っている.

以下では, $Score_v(v,n,v_c)$ に説明を限定する. 基本的に, $Score_v(v,n,v_c)$ は, 与えられた名詞 n が属するのとと同じ単語クラス (α) の単語の v への係りやすさ $P(\alpha|\langle v, \mathtt{wo}\rangle) + P_C(\alpha|\langle v,v_c\rangle)$ と, v_c への係りやすさ ($P_V(v_c|\alpha)$) の積であるとみなすことができる. つまり, v と v_c が名詞 n を共有する, そのしやすさを表しており, 全節で述べた仮説を実現したものであると考えることができる. また, 注意すべき点は, 単純に動詞と名詞の間での係り受けの起きやすさだけを判定しているのではなく, 項 $P_C(\alpha|\langle v,v_c\rangle)$ によって, v, v_c を含む並列句の中での, 係り受けのしやすさをも考慮にいれていることである. これにより, 例えば, 「ビールを飲み, 酔う」や, 「青酸カリを飲み, 死ぬ」といった並列句が多く現われるコーパスを用いると, $Score_v$(飲む, ビール, 酔う) > $Score_v$(飲む, ビール, 死ぬ) であるのに対して, $Score_v$(飲む, 青酸カリ, 酔う) < $Score_v$(飲む, 青酸カリ, 死ぬ) となり, 「もし, X がビールを飲むならば, X がビールに酔う」「もし, X が青酸カリを飲むならば, X が青酸カリで死ぬ」といった規則は生成されやすいが, 「もし, X がビールを飲むならば, X がビールで死ぬ」「もし, X が青酸カリを飲むならば, X が青酸カリに酔う」といった規則は生成されにくいといった望ましい性質が実現される.

実際に推論規則を生成するにあたっては, 以上のスコアの計算だけでなく, X, Y, Z などの変数を規則に挿入するなどの処理が必要となる. 特に, 並列句から生成される推論規則では, その前提部と結論部で主語が共有されることが多く, それを表す変数の共有を規則中に挿入することが望ましい. (例「もし, X が青酸カリを飲むならば, X が青酸カリで死ぬ」の変数 X) また, 前提部に現われた名詞が帰結部に現われることも多く, それらも規則に反映すべきである. (例「もし, X が青酸カリを飲むならば, X が青酸カリで死ぬ」の「青酸カリ」) 本研究では, そのような処理をヒューリスティックなルールによって行っているが, その説明は割愛する.

4 実験

提案手法によりコーパスから推論規則を獲得し, その獲得された規則の妥当性を 6 名の被験者にチェックしてもらった. 並列句の抽出に使ったコーパスは, 新聞３３年分である. また, 推論規則の生成にあたっては, 並列句の抽出につかった新聞とは別の新聞から動詞とその目的語を２００組抽出し, それらを前提部とするような推論規則を生成した. 生成された推論規則は, 前述のスコア関数がある閾値を越えたもので, ２１７個あった.

さらに, 同一の前提部に対して, 以下に挙げたもう一つのスコアで, 提案手法で生成したとの同数の推論規則

[2]本稿にある実験では, 単語クラスの数すなわち, $|Class|$ として 2500 を設定し, 新聞３３年分の構文解析結果から, 37638 個の単語の分類を行い, その結果を使用している.

- もしXが記者会見を開くならば, { 普通 | しばしば }X が記者会見で正式に Y を表明する.
- もしXが自民党を批判するならば, { 普通 | しばしば }X が自民党の対案を語る.
- もしXが監督を務めるならば, { 普通 | しばしば }X が監督の手腕を振るう.
- もしXが服を作るならば, { 普通 | しばしば }X が服を Z に着せる.
- もしXが服を作るならば, { 普通 | しばしば } 服が売れる.

図 2: 生成された推論規則の例

被験者	A	B	C	D	E	F
「普通」						
提案手法 (%)	45	50	43	31	39	48
ベースライン (%)	23	34	29	19	24	31
「しばしば」						
提案手法 (%)	79	67	80	59	64	67
ベースライン (%)	56	55	64	39	44	49

表 1: 各被験者毎の推論規則の妥当性

を生成し, それをベースラインとした.

$$Base(v,n,v_c) = f(v,v_c)\cdot\sum_{\alpha\in Class}\{P(\alpha|n)(P(\alpha|\langle v,\mathtt{wo}\rangle)+P_C(\alpha|\langle v,v_c\rangle))\}$$

このスコアは, 前節で与えたスコア中の動詞 v と v_c が v の目的語 n を共有する度合を単純に v と v_c の共起頻度 $f(v,v_c)$ で置き換えたものと考えることができる. また, v と n の間の係りやすさも考慮に入っている. $(P(\alpha|\langle v,\mathtt{wo}\rangle)+P_C(\alpha|\langle v,v_c\rangle))$ つまり, このスコアは, より高頻度で共起し, なおかつ n が v に係りやすいような動詞対からは適切な推論規則が生成できるという仮説に基づいたスコアである.

表 1 に各被験者によるチェックの結果を示す. 「普通」, 「しばしば」とあるのは, それらを規則に付加したときに妥当である規則の割合をそれぞれしめす. すべての項目, すべての被験者で, 提案手法がベースラインよりも 10%から 20%程度良い結果を出している. また, 被験者間で判断にばらつきがあるが, 表 2 に判断の一致の度合を示す. 被験者数とあるのは, 一致して妥当であると判断した被験者の数であり, 例えば, 6名中5名の被験者が推論規則の半分以上に対して, 「しばしば」を付加した場合に妥当であると判断している. この値は必ずしも高くは無いが, 最初の試みとしては妥当な値であると考えている. 最後に図 2 に実験で生成された推論規則で妥当と思われるものの例を挙げる.

5 まとめ

我々の持つ常識を表現できる推論規則をコーパスから自動的に獲得することを試みた. 自動獲得にあたっては, 二つの動詞をふくむ並列句表現から推論規則を生成したが, 動詞の間で名詞を共有しやすい並列句表現からは妥当な推論規則が生成されやすいという仮説をたて, 実験でその有効性を示した. また, 実験の結果得られた推論規則には確かに我々の持つ常識を表現する推論規則が含まれていた.

被験者数	6	5	4
「普通」			
提案手法	10. 2%	25. 1%	34. 4%
ベースライン	6. 0%	11. 6%	17. 2%
「しばしば」			
提案手法	34. 9%	50. 7%	66. 0%
ベースライン	19. 1%	31. 6%	43. 7%

表 2: 妥当性判定の一致の度合

参考文献

[1] Koji Inui, Kentaro Inui, and Yuji Matsumoto. 接続助詞「ため」を含む複文から因果関係知識を獲得する. In *IPSJ SIG Notes*, volume NL-150-25, 2002. in Japanese.

[2] Christopher S.G. Khoo, Syin Chan, and Yun Niu. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting of the Association for the Computational Linguistics*, pages 336–343, 2000.

[3] Dekang Lin and Patrick Pantel. Discovery of inference rules for question answering. *Journal of Natural Language Engineering*, 7(4):343–360, 2001.

[4] Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of 37th Annual Meeting of the ACL*, pages 104–111, 1999.

[5] Yusuke Shinyama, Satoshi Sekine, and Kiyoshi Sudo. Automatic paraphrase acquisition from news articles. In *Proceedings of Human Language Technology (HLT2002)*, 2002.

[6] Kentaro Torisawa. A nearly unsupervised learning method for automatic paraphrasing of Japanese noun phrases. In *Proceedings of Workshop on Automatic Paraphrasing: Theories and Applications*, pages 63–72, Tokyo, Japan, 2001.

[7] Kentaro Torisawa. An unsupervised method for canonicalization of Japanese postpositions. In *Proceedings of 6th Natural Language Processing Pacific Rim Symposium*, Tokyo, Japan, 2001.

[8] Kentaro Torisawa. An unsupervised learning method for associative relationships between verb phrases. In *Proceedings of COLING 2002*, 2002.