1B1-05

語の関連に基づく synonymy 集合を用いた概念表現手法の提案

Context Representation Using the Synonymy Set Dependent on the Relativity of Word

関矢 浩史　近藤 健　橋本 誠　高木 友博
Hiroshi Sekiya　Takeshi Kondo　Makoto Hashimoto　Tomohiro Takagi

明治大学 理工学研究科 基礎理工学専攻
Department of Computer Science, Meiji University

Abstract: Word meaning changes dynamically depending on context. We need to specify the context to identify this meaning. However, context varies depending on specificity of the topic and the viewpoint of the writer. In this paper, we propose that a word sequence can be used to identify context. Both contexts identified by word sequences and word sets related to the contexts will be shown concretely. We used 800,000 Reuters news articles, and extracted the word sets using the Confabulation model and five statistical measures as relations. We compared the measures and found that Cogency and Mutual Information were the most effective. We demonstrate the usefulness of the word sequence to identify the context.

1. はじめに

現在，オンラインオフラインを問わず，膨大な量のデジタルドキュメントが存在している．当然，それらの処理（整理や検索等）には，コンピュータの使用が不可欠となっている．このとき，単語の曖昧性は非常に重大な問題のひとつである．多義語はもちろんだが，単語は辞書的な定義以外の意味を持つことがある．例えば，“Hawaii”という単語は「州の名前」であったり「島の名前」であったりするが，「（日本から）Hawaii に行く」といった場合には，感覚的にはこのどちらの意味とも言えず，Hawaii 島を中心としたあるエリアを指しているといったイメージで使われていると考えられる．このように，単語の意味というものは，それが使われる文脈によって動的に変化する．そのため，単語の意味の特定には文脈の特定が必要である．しかし，前述の例でも分かる通り，文脈はトピックの粒度や話者（書き手）の視点などによって様々であり，その種類は膨大である．細かく言えば，少しでも構成する単語や，それらの並びが違うドキュメントは全て別の文脈であるとも言うことが出来るかもしれない．しかし，実際にはその中でも似た文脈というものが存在していると感じられる．よって，それら全てを別々の文脈として扱うというのは好ましくないと考えられる．

そこで本論文では，N-gram モデル[Jelinek, 1990]の考え方に基づいて，対象語の直前の単語列によって文脈が構成されていると仮定した．実験では，その仮定に基づき，連続して出現している単語の列を擬似構文としてコーパス[Reuters]から自動的に抽出した．ここで使用したコーパスは，1996 年から 1997 年のロイターのニュース記事 1 年分，約 80 万文書から成るコレクションである．そして，抽出した各擬似構文によって，それに続くと予測される単語に文脈的な影響が表れている，つまり文脈依存的な synonymy が得られていることを提示することで，この擬似構文が単語の意味を特定するための文脈情報として有効であることを示す．

2. 擬似構文

コンピュータに英語や日本語のような言語を理解させる上で最も難しい問題のひとつが言葉の持つ曖昧性である．一般的に，我々（人間）は，言葉の意味を特定するために文脈情報を用いている．言い換えると，文脈を特定することが出来れば言葉の意味も特定することが出来るということになる．ここで文脈を特定するのに有効であると考えられる情報として構文[Goldberg, 1995]が挙げられるが，抽出のルールを人間が作る必要があり，そのルールも複雑になりがちである．対して，対象語の直前の単語列を文脈と見なすというのは非常に単純である．この仮定は一見すると非常に乱暴に見えるが，同様の仮定に基づく N-gram モデルは多くの統計的な自然言語処理システムで機能している．我々は今回，この文脈を構成しているこれらの単語列を擬似構文と呼んでいる．

我々が今回擬似構文として用いた対象語の直前の 4 単語の列の例を図 1 に図示する．例えば“The Japanese Society for Artificial Intelligence”という 6 単語で構成される単語列があったとする．ここで対象語が“Artificial”だったとすると，擬似構文は“The Japanese Society for”となる．

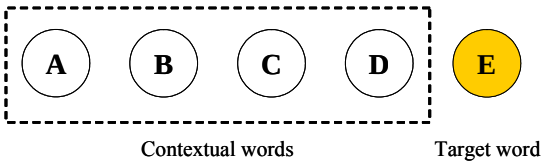

図 1　単語列と対象語の位置関係

3. 単語予測モデル概要

3.1 Confabulation モデル

一般的な N-gram モデルでは，対象語の直前の単語列をひとつのグループとして扱う．例えば，“$abcde$”という 5 単語の列が与えられた場合，5 番目の単語の予測には最初の 4 単語のセットを用いる．

$$n-gram(abcd,e)=P(e|abcd) \qquad (1)$$

一方，Hecht-Nielsen[Hecht-Nielsen, 2004a] [Hecht-Nielsen, 2004b]は，同様の単語予測実験において，対象語の直前の単語列の要素それぞれを別々に取り扱う手法を提案し（図 2），これを Confabulation モデルと呼んでいる．

連絡先：関矢 浩史，明治大学 理工学研究科 基礎理工学専攻，214-0034 川崎市 多摩区 東三田 1-1-1，Tel: 044-934-7483，Fax: 044-934-7912，sekiya@cs.meiji.ac.jp

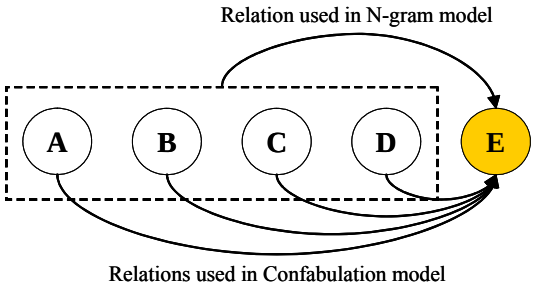

図 2 N-gram と Confabulation モデルの比較

Confabulation モデルでは単語列の要素それぞれが独立に対象語と関連を持っていると仮定している．例えば，Forward Probability がその関連度として用いられた場合，対象語は最初の 4 単語それぞれとの条件付き確率の積によって予測される．

$$conf(abcd,e) = P(e|a)P(e|b)P(e|c)P(e|d) \quad (2)$$

この違いにより，N-gram モデルを用いて予測可能な単語が実際にコーパス内に存在する 5 単語の列に制限されてしまうのに対し，Confabulation モデルは予測可能な単語にそのような制限はかからない．

人間は知っている単語であれば，それを新たな文脈の中でも妥当な使い方をすることができる．そのような柔軟で強力な言語表現を実現できる可能性を考慮し，我々は本実験において，Confabulation モデルを採用することとした．

3.2 尺度

我々は，Confabulation モデルの関連度として，比較のために“Forward Probability”，“Cogency”，“Mutual Information”，“Jaccard Coefficient”，“Chi Square”という 5 つの統計学的な尺度を用いて実験を行った．

例えば，“xy”という単語列が与えられ，x を文脈語，y を対象語としたとき，Forward Probability は文脈語が生起したときにどの程度対象語が生起するかという確率として表現される．

$$forward(x,y) = P(y|x) \quad (3)$$

ここで，“$x_1x_2...x_i...x_ny$”という単語列が与えられ，x を文脈語群，y を対象語としたときには，Forward Probability を Confabulation モデルの考え方に適用し，以下のような式を用いた．

$$forward.conf = \prod_{i=1}^{n} forward(x_i,y) \quad (4)$$

Backward Probability は Forward Probability とは反対に，対象語が生起したときにどの程度文脈語が生起するかという確率として表現される．この確率は，ニューロンの生理実験に基づいており，Hecht-Nielsen はこの尺度を Cogency と呼んでいる．

$$cogency(x,y) = P(x|y) \quad (5)$$

ここで，文脈語が複数ある場合には，Cogency を Confabulation モデルの考え方に適用し，以下のような式を用いた．

$$cogency.conf = \prod_{i=1}^{n} cogency(x_i,y) \quad (6)$$

Mutual Information は偶然に共起すると期待される確率に対して実際どの程度共起したかという比の形で表現される．

$$mutual(x,y) = \log\frac{P(x\cdot y)}{P(x)P(y)} \quad (7)$$

ここで，文脈語が複数ある場合には，Mutual Information を Confabulation モデルの考え方に適用し，以下のような式を定義した．

$$mutual.conf = \sum_{i=1}^{n} mutual(x_i,y) \quad (8)$$

Jaccard Coefficient は文脈語か対象語のどちらかが生起した確率に対して，共起確率がどの程度だったかという比の形で表現される．

$$jaccard(x,y) = \frac{P(x\cap y)}{P(x\cup y)} \quad (9)$$

ここで，文脈語が複数ある場合には，Jaccard Coefficient を Confabulation モデルの考え方に適用し，以下のような式を定義した．

$$jaccard.conf = \frac{\sum_{i=1}^{n} P(x_i\cap y)}{P(\bigcup_{i=1}^{n} x_i\cup y)} \quad (10)$$

Chi Square は文脈語と対象語それぞれの生起の間に統計的に有意な関連があるかどうかという尺度として表現される．具体的には，x と y の共起の回数を $C(x\cdot y)$ と表したときに，$C(x\cdot y)=A$，$C(x\cdot\bar{y})=B$，$C(\bar{x}\cdot y)=C$，$C(x\cdot\bar{y})=D$ と置くと，

$$chi(x,y) = \frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (11)$$

のように書くことができる．ここで，N は全ての組み合わせの総和を表している．

ここで，文脈語が複数ある場合には，Chi Square を Confabulation モデルの考え方に適用し，以下のような式を定義した．

$$chi.conf = \sum_{i=1}^{n} chi(x_i,y) \quad (12)$$

4. 検証･考察

4.1 検証方法

実際に，擬似構文が言葉の意味を特定する文脈情報として有効であることを示すために，それぞれの擬似構文によって，後に続くと予測される単語の集合が文脈的に制限を受けている

ことを示す. これらの擬似構文はロイターのニュース記事のコーパスから自動的に抽出されたものである. 具体的には 5 単語の幅をもつウインドウを用いてコーパス全体から単語列を抽出した. また, 簡単化のために, 本実験ではコーパスにおいて生起確率の高い上位 3 万語を対象とした. つまり, コーパスから抽出した 5 単語の列はすべてこの上位 3 万語によって構成されているということである. これらの単語列を用いて, 我々は Confabulation モデルによる対象語の予測実験を行った. また, その結果を用いて Confabulation モデルと前述した 5 つの統計的な尺度の相性についても比較を行った.

4.2 得られた Synonymy と擬似構文の有効性

以下は, コーパスから抽出された擬似構文と, それぞれの擬似構文に続くと予測された単語集, つまり文脈的な synonymy である(表 1, 2, 3, 4). また, これらは全て Confabulation モデルに Cogency をその関連度に用いて得られたものである.

表 1 “government of prime minister”に対する synonymy

"government of prime minister"	
Related word	Relationship value
Meles	0.005868408
Gediminas	0.004884470
Necmettin	0.002793888
Abdellatif	0.002729859
Rafik	0.002481266
Andris	0.002072932
Zhan	0.001895817
Romano	0.001258259
Thorbjoern	0.001004164
Bashkim	0.000987456
Janez	0.000791000
Meraj	0.000789541
Vaclav	0.000654863
Branko	0.000635066
Benjamin	0.000618136
Kengo	0.000487491
Benazir	0.000452870
Alain	0.000445664
Inder	0.000442266
Aleksander	0.000355988
Mesut	0.000263984
Lionel	0.000245173
Chavalit	0.000223583
Zulfikar	0.000219513
Tiit	0.000209493
Nawaz	0.000189236
Sheikh	0.000176216
Ryutaro	0.000108052
Paavo	0.000102548

表 2 “by Italian prime minister”に対する synonymy

"by Italian prime minister"	
Related word	Relationship value
Romano	0.002757365
Giulio	0.000255617
Silvio	0.000059596

表 3 “is former prime minister”に対する synonymy

"is former prime minister"	
Related word	Relationship value
Giulio	0.000511235
Petre	0.000478109
Tiit	0.000235680
Zulfikar	0.000186992

まず表 1 を見ると, “government of prime minister”という擬似構文に対して, 世界の首相の名前の集合が抽出できていることが分かる. 表 2 からは, “by Italian prime minister”という擬似構文に対して, イタリアの首相の名前の集合が抽出されている. ここで, “Romano (Prodi)”は表 1, 表 2 に共通して見ることができる. さらに表 3 では, “is former prime minister”という擬似構文に対して, 元首相の名前の集合が得られている. また, 表 2 と表 3 は共に, イタリアの元首相である“Giulio (Andreotti)”の名前を含んでいることが分かる.

表 4 “and said prime minister”に対する synonymy

"and said prime minister"	
Related word	Relationship value
Ryutaro	0.001058507
Vo	0.000756075
Romano	0.000661768
Pavlo	0.000546340
Gyula	0.000523615
Paavo	0.000448649
Viktor	0.000349859
Benjamin	0.000342884
Nawaz	0.000323693
Costas	0.000299586
Yitzhak	0.000271878
Meraj	0.000263180
Begum	0.000220172
Mirko	0.000214548
Vaclav	0.000188366
Bashkim	0.000187661
Kamal	0.000180351
Petre	0.000179291
Inder	0.000135792
Roumen	0.000122250
Wlodzimierz	0.000120161
Necmettin	0.000115626
Alain	0.000113479

ここで表 4 に示した擬似構文“and said prime minister”は, 実際にはコーパス中に 5 回しか出現していない. しかし, 表から分かるように 23 種類の単語の集合がこの擬似構文から得られている. 言い換えると, 少なくとも 18 単語はこの擬似構文とは実際には共起していなかったということになる. さらに, この表に含まれている単語は全て実在する首相の名前であった.

これらの結果は, 擬似構文が様々な粒度のトピックや視点を表現することができることを示しているものと考えられる. また, 得られた synonymy ひとつひとつがある文脈依存的な概念を表しているものと捉えることができる. さらに, 表 4 からは confabulation モデルがコーパス内では実際には共起していな

い組み合わせであっても，共起する可能性が十分に高い，つまり関連が十分にあると予想されるものであれば抽出できるということを示していると考えることができる．

4.3 尺度の比較

ここでは，前述した 5 つの統計的な尺度について，Confabulation モデルに用いる関連度としての有効性を比較していく．

まず，全ての尺度の結果に共通して言えることは，後に続くと予想された対象語の集合が実際にコーパス中に生起した単語列に縛られてはいないということである．つまり，この意味においてはどの尺度も Confabulation モデルに適用可能であると考えられる．

次に，各尺度についての比較であるが，その観点として，擬似構文に対して関連するとして得られた単語群のコーパス中における頻度の傾向と，各擬似構文との組み合わせが妥当なものであるかどうかの 2 点を主に比較した．得られた単語群の頻度については，全体的に分散していた方がよいと考えられるが，本論文では，文脈の特定性を重視しているため，その中でも高頻度語よりも低頻度語の方が重要であると考えている．

まず Forward Probability を用いた場合には，得られた単語群はほとんど全てが最も高頻度でコーパス中に現れた単語（例："the", "a", "to"等）によって構成されていることが分かった．また，擬似構文との組み合わせの多くは文法的に妥当とは言えないものが数多く含まれていた．

一方，Cogency と Mutual Information を用いた場合には，得られた単語群はそのような高頻度語をそれほど含んでおらず，擬似構文との組み合わせについてもほぼ妥当なものであった．つまり，Cogency と Mutual Information は Forward Probability とは逆の性質を持っていると言うことができる．但し Cogency については，本実験には採用していない最も頻度が低い語（例：綴り間違い，数字を含む単語等）に対して非常に高い値を与えてしまう傾向があることには注意しなければならない．

Jaccard Coefficient を用いた場合には，得られた単語群は高頻度語も低頻度語も含んでおり，擬似構文との組み合わせについてもほぼ妥当なものであった．しかし，高頻度語ほど高い値を与えられる傾向が見られた．また，擬似構文ごとに得られた単語群との関連度の値が大きく分散しており，どの単語を採用するかについて，一貫した判断基準の決定が非常に難しいものであった．

Chi Square を用いた場合には，得られた単語群は低頻度語をほとんど含んでいなかった．また，擬似構文との組み合わせについても妥当とは言えないものが多く見られた．Chi Square は高頻度語に高い値を与える傾向にあり，この点で，Forward probability に似た性質を持っていると言うことができる．本来，Chi Square は正と負の 2 種類の相関を測ることができるが，本実験においては，正の相関のみを用い，負の相関については用いていない．また，特定の状況では Chi Square の値は必ずしも信頼できないという報告もある[Cochran, 1954]．これらのことが Chi Square を用いた結果に悪影響を及ぼしている可能性があると考えられる．

5. まとめ

実験によって，擬似構文が局所的な文脈情報として有効であるということ，さらにそれらによって得られた synonymy がある文脈依存的な概念を表していることを示した．また，Confabulation モデルについてもコーパス中に実際に生起した単語列に縛られることなく，妥当だと思われる単語を抽出することができることを示した．

6. 今後の展望・課題

本論文において，我々は擬似構文を局所的な文脈として利用したが，大域的な文脈情報というものもあるはずだと考えている．そこで，大域的な文脈情報を抽出する手法についての検討を行っている．

本実験では，N-gram モデルと同様の仮定，つまり対象語の直前の単語列のみを文脈語と見なした．しかし，対象語の直後にある単語列についても同様に局所的な文脈情報を保持していると考えられる．また，その文脈情報は直前の単語列が保持するものとは特徴が異なるものであると期待している．そこで，それらの局所的な文脈情報としての有効性について，さらなる実験を行う予定である．

我々は本実験において，対象語の直前の 4 単語を用いたが，文脈語は多すぎても少なすぎても好ましくないと考えている．よって，文脈を表現するための適切な単語数というものについて検討が必要である．

局所的な文脈情報を用いて単語の意味を捉えることで，我々は今後，文脈に依存した動的な概念表現の実現に向けて取り組んでいく予定である．また，本手法を情報検索システムの精度改善のために利用することについても検討中である．

参考文献

[Jelinek, 1990] F. Jelinek: “Self-organized language modeling for speech recognition,” Readings in Speech Recognition, pp. 450-506, Morgan Kaufmann Publishers, 1990.

[Goldberg, 1995] A. E. Goldberg: “Constructions: A Construction Grammar Approach to Argument Structure,” University of Chicago Press, 1995.

[Hecht-Nielsen, 2004a] R. Hecht-Nielsen: “A Theory of Cerebral Cortex,” UCSD Institute for Neural Computation Technical report #0401, 2004.

[Hecht-Nielsen, 2004b] R. Hecht-Nielsen: “A Theory of Cerebral Cortex,” UCSD Institute for Neural Computation Technical report #0404, 2004.

[Reuters] Reuters Corpus @ NIST or Reuters Corpus
http://trec.nist.gov/data/reuters/reuters.html
http://about.reuters.com/researchandstandards/corpus/

[Callan, 1994] J. P. Callan, W. B. Croft and J. Broglio: “TREC and TIPSTER Experiments with INQUERY,” Information Processing and Management, 31(3), pp.327-343, pp. 327-343, 1994.

[Xu, 2000] J. Xu and W. B. Croft: “Improving the effectiveness of information retrieval with local context analysis,” ACM Transactions on Information Systems (TOIS), Volume 18, Issue 1, pp. 79-112, 2000.

[Gauch, 1997] S. Gauch and J. Wang: “A Corpus Analysis Approach for Automatic Query Expansion,” Proceedings of the Sixth International Conference on Information and Knowledge Management (CIKM'97), pp. 278-284, 1997.

[Cochran, 1954] W. G. Cochran: “Some methods for strengthening the common $\chi 2$ tests,” Biometrics, 10, pp. 417-451, 1954.