社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＦＩ－ 93 (4)
2009－ＮＬ－189 (4)
2009／1／22

枝分かれ同時確率モデルを用いた対象－属性－属性値関係の抽出

五十嵐力†, 藤本浩司‡, 但馬康宏†, 小谷善行†
† 東京農工大学大学院 工学府
‡ テンソル・コンサルティング株式会社

本稿では、一文内に含まれている単語の三つ組が対象－属性－属性値の 3 項関係になる確率を求める統一的な確率モデルを提案する。従来手法においては対象を定めた上で属性－属性値ペアを求めるなどの 2 段階処理で抽出することが多いが、提案手法においては対象、属性、属性値のすべてを対等に扱う。文中の三つ組が 3 項関係を持つ事例は非常にスパースなものになるが、これを近似した上で枝分かれ同時確率モデルにより推定することでスパースな事例にも対応可能である。評価実験において、枝分かれ同時確率モデルを用いることの有効性を示した。

Automatic Extraction of Object-Attribute-Value Relations from Text Based on Nested Joint Probability Model

Chikara IGARASHI†, Koji FUJIMOTO‡, Yasuhiro TAJIMA†, Yoshiyuki KOTANI†
† Department of Computer and Information Sciences,
Tokyo University of Agriculture and Technology
‡ Tensor Consulting Co. Ltd.

This paper describes an probability model to extract triplets (Object, Attribute, Value) from a sentence. A lot of traditional techniques need two or more processes such as extracting Attribute-Value pairs after deciding an Object. In contrast, this probability model equally treats Object, Attribute, and Value. Three words with the relation in a sentence ralery exists in the case data. Therefore, This probability model estimates a joint probability based on Nested Joint Probability Model. Our experiment showed that the effectiveness of the use of Nested Joint Probability Model to extract triplets.

1．はじめに

Web マイニングにおける研究のひとつに、文書集合から対象語に関する属性と属性値を抽出するというものがある。この(対象, 属性, 属性値) の 3 項関係は、他の対象との比較を容易にしたり、その対象の特徴ベクトルとしての用途が考えられる。3 項関係を文章中から自動的に収集することにより、商品情報に関する文章であればマーケティング、一般文章であれば語彙知識の獲得が期待できる。

類似研究として、意見抽出を目的とした (対象, 属性, 評価値) の 3 項関係を抽出する研究[1][2][3]が多くなされているが、本研究では意

見性があるかどうかに関わらず網羅的に情報を（対象，属性，属性値）の三つ組として抽出することを目的とする。これはすなわち、意見抽出研究のように属性語や属性値語の範囲を限定しないことを意味している。また[2][3]では対象物がトピックとなっている範囲を限定した上で属性－評価の 2 項関係を抽出するという 2 段階の処理となっている。これに見られるように 3 項関係抽出研究の多くが、それぞれの要素を対称に扱っておらず、ヒューリスティクスを用いて 3 項関係にしているといえる。しかし本研究においては、3 項関係の三つ組（対象，属性，属性値）のそれぞれを対称に扱い、この三つ組が 3 項関係を持つかどうかを統一的な確率モデルを用いることで判定する。

2．対象－属性－属性値の 3 項関係

ある「対象」に関するある「属性」が「属性値」である、という関係が存在するときに、「対象」「属性」「属性値」の三つ組に 3 項関係があると定義する。例えば、「ポチの色は白い」という文があれば、対象が「ポチ」、属性が「色」、属性値が「白い」という三つ組がその文から抽出される 3 項関係である。主に一つの対象に対して複数の属性－属性値の組が存在する。

対象「イチロー」

属性		属性値
職業	：	野球選手
所属球団	：	シアトルマリナーズ
ポジション	：	外野手

図 1　対象「イチロー」に関する 3 項関係

3．文からの 3 項関係抽出

文からの 3 項関係抽出手法に関する類似研究はいくつかあるが、ヒューリスティックに品詞や係り受け関係から三つ組を抽出する方法[1]や、まず対象および属性値を定め、その属性値がどの属性に含まれるかをベイズ推定などにより判定する手法がある。しかしこの場合、あらかじめ抽出規則や属性のリストを定めておかなければならず、限定された対象に関する 3 項関係しか抽出できない。そこで、一文中にある三つの単語を選んだときに、それらが対象、属性、属性値の 3 項関係である確率を求め、その確率によって 3 項関係して抽出するかどうかを判断する確率モデルを考える。ただしこのモデルを扱う上で、3 項関係となる三つ組の単語がすべて文中に含まれていなければならないという前提がある。しかし、例えば対象語がない文であればそれはその文の主題が省略されているためであったり、また属性語が省略された文については前文に記述されているかもしくは属性値語から容易に推定可能である場合が多い。このため本稿では省略された語については言及せず、あくまでも文中に補完されて記述されているものとする。

3．1　三つ組の 3 項関係確率

本稿で使用する記号を次のように定める。

単語：w

文：$S = \{w_1, w_2, \cdots, w_n\}$

3 項関係にある三つ組：$\langle w_O, w_A, w_V \rangle$

ただし w_O：対象語

w_A：属性語

w_V：属性値語

文 S が与えられ、任意の文 S 中の 3 単語 w_i, w_j, w_k を三つ組$\langle w_i, w_j, w_k \rangle$として選んだとする。このとき、$\langle w_i, w_j, w_k \rangle$が文 S における 3 項関係として抽出することが妥当であるかどうかは、事例データにおいてその三つ組が文 S における 3 項関係であるとして抽出された確率から判断することができる。すなわち、3 項関係として抽出されるという事象を PRF とす

ると、三つ組<w_i,w_j,w_k>が文 S から 3 項関係として抽出された確率(3 項関係確率)は

$$P(PRF|<w_i,w_j,w_k>,S)$$

という条件付確率で表せる。また、三つ組<w_i,w_j,w_k>が文 S から 3 項関係として抽出されない確率は

$$P(\overline{PRF}|<w_i,w_j,w_k>,S)$$

となる。このとき、

$$P(PRF|<w_i,w_j,w_k>,S) \vee P(\overline{PRF}|<w_i,w_j,w_k>,S) \quad \cdots(式\ 1)$$

の関係にあれば、三つ組<w_i,w_j,w_k>は文 S から抽出する 3 項関係として妥当であるとみなすことができる。式 1 の右辺と左辺の和は 1 であるため、これは

$$P(PRF|<w_i,w_j,w_k>,S) > 0.5 \quad \cdots(式\ 2)$$

と置き換えられる。したがって、まず 3 項関係確率 P(PRF|<w_i,w_j,w_k>,S)を求めることを考える。この 3 項関係確率の元となる事例データには、文およびその文中の単語による三つ組、そしてその三つ組が 3 項関係として抽出されるか人手により判断して付与したものが記述されているものとする。この事例データから 3 項関係確率を求めるが、抽出文として与えられた文 S が事例データ中の文書集合に含まれている確率が非常に低いため、3 項関係確率は非常にスパースなものとなる。そこで、三つ組間の文法構造が同じ文であれば、三つ組の意味的関係も同一であると仮定する。例えば、「太郎のペットであるポチの色は白い」という文 S_1 に対し、「ポチの色はすごく白い」という文 S_2 があったとする。このとき文 S_1 および S_2 において「ポチの色は白い」という事実が等しく存在し、また 3 単語「ポチ」「色」「白い」の間の文法的関係も同一である。したがって文 S_1 および文 S_2 は、三つ組<ポチ,色,白い>に関して同一の文 S'とみなす。これに伴い、3 項関係確率を

$$P(PRF|<w_i,w_j,w_k>,S) \approx P(PRF|<w_i,w_j,w_k>,S')$$

と近似する。ここで、S'は w_i,w_j,w_k と、その間の文法構造からなる文である。

条件付確率の定義から、この 3 項関係確率は

$$P(PRF|<w_i,w_j,w_k>,S') = \frac{P(<w_i,w_j,w_k>,S',PRF)}{P(<w_i,w_j,w_k>,S')} = \frac{F(<w_i,w_j,w_k>,S',PRF)}{F(<w_i,w_j,w_k>,S')} = \frac{F(<w_i,w_j,w_k>,S',PRF)}{F(<w_i,w_j,w_k>,S',PRF)+F(<w_i,w_j,w_k>,S',\overline{PRF})}$$

…式 3

で求めることができる。すなわち、文構造 S'をもつ文から三つ組<w_i,w_j,w_k>を 3 項関係としてタグ付けした頻度、およびタグ付けしていない頻度を計算すればよい。

3．2　3 項関係確率の分解

式 3 において、F(<w_i,w_j,w_k>,S',PRF)は、事例データ中で 3 項関係をもつとタグ付けされた文 S'中の三つ組<w_i,w_j,w_k>の出現頻度であるが、この出現頻度も非常にスパースなものとなる。そこで、さらに三つ組中の 2 単語間の関係が他の 1 単語とは独立であると仮定する。すなわち、三つ組すべてが与えられてそれが 3 項関係をもつと判断するのではなく、三つ組中の 2 単語が与えられれば 3 項関係のうちの 2 項関係であると判断できると考える。例えば、「ポチの色は白い」という文において、「ポチの色は＊」(＊は任意単語)という部分的な関係だけ与えられても、対象が「ポチ」、属性が「色」であると判断できる。このような関係がどの 2 単語間にも存在すると考えられる。ここで、3 事象の同時生起をベン図により表すと図 2 のようになる。

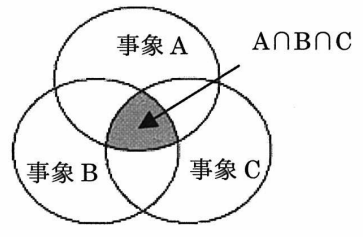

図2　3事象の積事象

$A \cap B \cap C$

$= A \cup B \cup C + A \cap B + A \cap C + B \cap C - A - B - C$

であるから、この A,B,C をそれぞれ $<w_i,*,*>,<*,w_j,*>,<*,*,w_k>$ とすれば(*は任意単語であることを表す)、

$$F(<w_i,w_j,w_k>,S',PRF) = F(S',PRF) - F(<\overline{w_i},\overline{w_j},\overline{w_k}>,S',PRF) + F(<w_i,w_j,*>,S',PRF) + F(<w_i,*,w_k>,S',PRF) + F(<*,w_j,w_k>,S',PRF) + F(<w_i,*,*>,S',PRF) + F(<*,w_j,*>,S',PRF) + F(<*,*,w_k>,S',PRF)$$

…式4

となる。

3．3　枝分かれ同時確率モデル

式4において $F(<w_i,w_j,*,>,S',PRF)$は、文 S' 中の2単語 w_i,w_j をそれぞれ3項関係における対象、属性であるとタグ付けされた頻度を表している。この頻度が w_k の語に依存しないと仮定すると、S'における w_i,w_j,w_k の間の文法構造を d_{ijk} として、

$$F(<w_i,w_j,*>,S') \approx F(w_i,w_j,d_{ijk})$$

と近似される。すなわち、対象が w_i、属性が w_j であり、それらおよび w_k の間の文法的関係が d_{ijk} である文が生起する頻度となる。本手法ではこの頻度 $F(w_i,w_j,d_{ijk})$ を3事象の枝分かれ同時確率モデル[4]によって求める。

枝分かれ同時確率モデルは、枝分かれ構造をもつ事象の同時確率を求めるモデルであり、より上位の水準における事象の同時確率から計算することで事例の少ないデータにも対応できるという特徴がある。例えば図3のように単語の品詞構造も枝分かれ構造である。

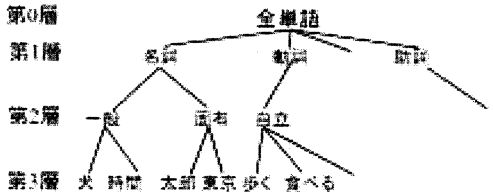

図3　枝分かれ事象の例：単語の品詞構造

枝分かれ事象 A,B,C の i 層の水準を A_i,B_i,C_i とする。水準 $A_iB_jC_k$ に対し、その共通の同時水準 $A_{i-1}B_{j-1}C_{k-1}$ が生起したときの条件付交互作用を

$$I_{A_iB_jC_k \mid A_{i-1}B_{j-1}C_{k-1}}$$

と表す。ここで、3事象 A,B,C に対する交互作用は

$$I_{ABC} = \frac{P(ABC)P(\overline{A}\overline{B}C)P(\overline{A}B\overline{C})P(A\overline{B}\overline{C})}{P(\overline{A}BC)P(A\overline{B}C)P(AB\overline{C})P(\overline{A}\overline{B}\overline{C})}$$

として定義される。すなわち、$I_{ABC}=1$ のとき A,B,C は独立して生起する。これは交互作用がないといえる。また $I_{ABC}>0$ のとき A,B,C は共起しやすく、$I_{ABC}<0$ のとき A,B,C は共起しにくい。$I_{ABC}=0$ であるとき、A,B,C は排他的関係にある。

枝分かれ同時確率モデルでは、条件付交互作用 $I_{A_iB_jC_k \mid A_{i-1}B_{j-1}C_{k-1}}$ が1であるとみなせるとき、上位水準から $A_iB_jC_k$ の同時確率 $P(A_iB_jC_k)$を計算するというものである。

$I_{A_iB_jC_k \mid A_{i-1}B_{j-1}C_{k-1}}=1$ を展開することで、$P(A_iB_jC_k)$に関する方程式

$$Z_3P(A_iB_jC_k)^3+Z_2P(A_iB_jC_k)^2+Z_1P(A_iB_jC_k)+Z_0=0$$

ただし

$Z_3=P(A_{i-1}B_{j-1}C_{k-1})$

$Z_2=E_1E_2+E_1E_3+E_2E_3 - O_0O_1-O_0O_2-O_0O_3-O_1O_2-O_1O_3-O_2O_3$

$Z_1=E_1E_2E_3+O_0O_1O_2+O_0O_1O_3+O_0O_2O_3+O_1O_2O_3$

$Z_0=-O_0O_1O_2O_3$

$E_1=P(A_{i-1}B_{j-1}C_k)-P(A_iB_{j-1}C_k)-P(A_{i-1}B_jC_k)$

$E_2=P(A_{i-1}B_jC_{k-1})-P(A_{i-1}B_jC_k)-P(A_iB_jC_{k-1})$

$E_3=P(A_iB_{j-1}C_{k-1})-P(A_iB_{j-1}C_k)-P(A_iB_jC_{k-1})$

$O_0=P(A_{i-1}B_jC_k)$

$O_1 = P(A_iB_{j-1}C_k)$
$O_2 = P(A_iB_jC_{k-1})$
$O_3 = P(A_{i-1}B_{j-1}C_{k-1})$
$-P(A_iB_{j-1}C_{k-1}) - P(A_{i-1}B_jC_{k-1}) - P(A_{i-1}B_{j-1}C_k)$
$+ P(A_{i-1}B_jC_k) + P(A_iB_{j-1}C_k) + P(A_iB_jC_{k-1})$

を得る。すなわち、上位水準の生起確率から方程式の係数を求め、方程式を解くことによって$P(A_iB_jC_k)$が求まる。この際、上位水準の生起確率に対してもより上位の水準の確率から再帰的に求めることができる。

図 4 に枝分かれ同時確率モデルを用いて$P(A_iB_jC_k)$を求めるアルゴリズムを示す。

1. 3事象 A_i,B_j,C_k を入力
2. A_i,B_j,C_k のうち二つが最上位水準であれば、A_i,B_j,C_k の観測同時確率を返す
3. A_i,B_j,C_k のうち一つが最上位水準であれば、2次の枝分かれ同時確率を計算して返す
4. A_i,B_j,C_k のいずれも最上位水準でなければ
   4-1. A_i,B_j,C_k の周辺確率から、係数 Z_3,Z_2,Z_1,Z_0 を計算
   4-2. 3次方程式 $Z_3P^3+ Z_2P^2 +Z_1P+Z_0=0$ の解 P を $A_iB_jC_k$ の予測同時確率とする
   4-3. $A_iB_jC_k$ の観測同時確率に対して、交互作用の有無を検定する
      4-3-1.交互作用があれば、観測同時確率を返す
      4-3-2.交互作用がなければ、予測同時確率を返す

図4　枝分かれ同時確率モデルによる3事象の同時確率計算アルゴリズム

3．4　2項関係確率の枝分かれ事象

3 項関係確率を求めるための二つ組の生起頻度 $F(w_i,w_j,d_{ijk})$を枝分かれ同時確率モデルにより求める上で、各事象の枝分かれ構造を決める必要がある。本手法では、w_i,w_j,w_k の事象構造を形態素解析器 ChaSen[5]における品詞階層構造とした。また d_{ijk} の事象構造を、1層目に係り受け解析器 CaboCha[6]による w_i,w_j,w_k の文節の係り受けパターン、2層目に w_i,w_j,w_k それぞれの文節の句の分類とする。文節の係り受けパターンは、$w_i \rightarrow w_j$、$w_j \rightarrow w_k$、$w_k \rightarrow w_i$ のそれぞれの文節間において、直接係る、間接的に係るなど 12 パターンを用意した。また文節の句は、例えば w_i であれば w_i を含む句において w_i 以降の文字列としている。

例として、「私の名前はリーです」という文における三つ組<私, 名前, リー>に対して、それぞれの事象階層は

w_i　：名詞－代名詞－一般－私
w_j　：名詞－代名詞－名前
w_k　：名詞－固有名詞－人名－姓－リー
d_{ijk}　：a+a+d　－　の+は+です。

のようになる。

4．評価実験

4．1　収集事例

本手法の評価に用いた事例データは、抽出する 3 項関係のタグ付けができるだけ主観的にならないよう、文意がはっきりしているものとして外国人向けの日本語学習教材[7]から 301 文を抜粋した。この収集文に対して、まず人手により省略語の補完を行い、次に CaboCha による形態素解析、係り受け解析を行う。そして形態素解析結果の特定品詞語(名詞, 動詞-自立, 形容詞-自立, 副詞-一般, 未知語)を用いて三つ組を全通り作成する。この三つ組に対して人手で 3 項関係があるかどうかのタグ付けを行った。これにより、すべての三つ組の数が 11,766 件、うち 3 項関係ありが 284 件、3 項関係なしが 11,482 件であった。

4．2　実験方法

評価実験のために、式 4 における各頻度を枝分かれ同時確率モデルにより計算し、式 3 の判定式が 0.5 より大きければ 3 項関係があると判定するシステムを作成した。[4]によれば枝分かれ同時確率計算における観測頻度の交互作用の有無の判定に χ^2 検定を用いているが、χ^2 検定は期待頻度が小さい場合に誤差が出るという問題がある。このため本システムにおいては G 検定を用いている。

本手法の比較対象として、クローズドデータテストにおいて、枝分かれ同時確率モデル使用

の有無、および式4による三つ組の分解の有無を比較した。また leave-one-out クロスバリデーション法によるオープンデータテストにより枝分かれ同時確率モデル使用の有無による比較を行った。

4．3　実験結果

3 項関係抽出実験のクローズドデータテスト結果を表1、オープンデータテストの結果を表2に示す。

表1　3項関係抽出精度(クローズド)

手法	適合率	再現率	F 値
枝分かれ同時確率なし 3項関係分解なし	0.936	0.978	0.956
枝分かれ同時確率なし 3項関係分解あり	0.973	1.000	0.986
枝分かれ同時確率あり 3項関係分解あり	0.973	1.000	0.986

表2　3項関係抽出精度(オープン)

手法	適合率	再現率	F 値
枝分かれ同時確率なし	0.851	0.161	0.272
枝分かれ同時確率あり	0.271	0.507	0.353

また、枝分かれ同時確率使用による各2項関係抽出精度を表3に示す。

表3　2項関係抽出精度(オープン)

抽出ペア	適合率	再現率	F 値
w_i - w_j	0.496	0.590	0.539
w_j - w_k	0.711	0.798	0.752
w_k - w_i	0.459	0.718	0.560

5．評価実験に対する考察

表1の結果から、枝分かれ同時確率モデルや3 項関係の分解の有無に関わらずクローズドデータテストでは高い精度を得られることがわかった。これは事例さえあれば3項関係を正しく抽出できるということである。したがって、文 S を三つ組の係り受け関係だけの構造 S'に簡略化しても、3項関係が変わらないという仮定が裏付けられたと考えられる。

表2の結果から、オープンデータテストにおいては枝分かれ同時確率モデルを用いることで用いない場合よりも F 値が向上した。オープンデータテストにおいてはテスト用の三つ組が事例にほとんど存在しないため、枝分かれ同時確率モデルなしのモデルでは $F(<w_i,w_j,w_k>,S',PRF)$, $F(<w_i,w_j,w_k>,S',\overline{PRF})$ が両方とも0になってしまい判定できない。これを推定することで判定可能になる枝分かれ同時確率モデルの有効性が示されたと考えられる。

6．まとめ

本稿では、文中から3項関係をもつ三つ組を抽出する統一的な確率モデルを提案した。また枝分かれ同時確率モデルにより少数データにも対応できるシステムを作成し、提案する確率モデルにおいて枝分かれ同時確率モデルとの併用が有用であることを示した。

三つ組の階層構造をさらに考慮し、品詞構造以外にも概念構造などを取り入れることで精度の向上が見込めると考えられる。

参考文献

[1] 鈴木泰裕, 高村大也, 奥村学, Weblog を対象とした評価表現抽出, 人工知能学会 SIG-SW&ONT-A401-02, 2004.

[2] 土田正明, 水口弘紀, 久寿居大, 対象－属性－評価の3項関係同定による評判情報抽出, 言語処理学会第 13 回年次大会論文集, pp. 412-415, 2007.

[3] 飯田龍, 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一, 意見抽出を目的とした機械学習による属性-評価値対同定, 情報処理学会研究報告-言語処理 NL165-4, pp. 21-28, 2005.

[4] 藤本浩司, 乾伸雄, 小谷善行, 枝分かれ構造をもつ同時確率モデルによる形態素解析, 情報処理学会論文誌 vol. 39, No. 7pp. 2101-2111, 1998.

[5] 松本裕治, 北内啓, 平野善隆, 松田寛, 高岡一馬, 浅原正幸, 形態素解析システム『茶筌』version 2.2.9 使用説明書, 奈良先端科学技術大学院大学, 2002.

[6] 工藤拓, 松本裕治, Support Vector Machine を用いた Chunk 同定, 情報処理学会 自然言語処理研究会研究報告 NL140, pp9-16, 2000.

[7] 寺内久仁子, にほんご1・2・3　上, アルク, 1996