2F2-4

カテゴリ情報を考慮したWikipediaからの含意関係の抽出

Extracting Entailment Rules using Categorical Information from Wikipedia

田中　翔平[*1]　岡崎　直観[*2]　石塚　満[*1]
Shohei Tanaka　Naoaki Okazaki　Mitsuru Ishizuka

[*1]東京大学大学院情報理工学研究科
Graduate School of Information Science and Technology, University of Tokyo

[*2]東北大学大学院情報科学研究科
Graduate School of Information Sciences, Tohoku University

Textual entailment recognition and extraction play an important role in semantic inference tasks. Semantic inference can be used to create extensible knowledge-bases, which are expected to have many applications in Question Answering systems. In this paper, we propose a novel method for textual entailment rule extraction from a given text corpus. The proposed system learns lexical patterns from Wikipedia and uses hit counts from a Web search engine to judge whether any entailment rule existed between each pattern pairs. We also explain in detail the algorithms for assessing the confident score of an entailment rule. Finally, we present some experimental results of these algorithms and discuss about future direction of this research.

1. はじめに

近年，応用的な知識の意味推測に利用可能な，Textual Entailment という研究分野に注目が集まっている．Textual Entailment とは,「与えられたテキストから特定の知識を推測できるか」を認識することを目的とする意味推論 (Semantic Inference) の理論的枠組みである．例えば，“alcohol reduces blood pressure” というテキストが与えられるとする．我々人間はこのテキストから，用いる動詞は違うものの，“alcohol affects blood pressure” という知識を推論することができる．このような推論を計算機を用いて自動的に行う研究分野が，Textual Entailment である．Textual Entailment が注目を集めている理由の一つに，質問応答システム (Questiong and Answering system) へ応用が可能である，という点が挙げられる．例えば先の例の知識を持っていれば，“What affects blood pressure?” という質問に対し，“alcohol reduces blood presure” という文から答えを推測し，出力することができる．

Textual Entailment を扱うシステムは，内部に推論の元となる抽象化された知識を持つ必要がある．例えば先の例で持っておくべき知識は，“X reduce Y → X affects Y” である (X, Y は名詞が入る変数)．このように語彙パターンで抽象化された知識を持っていれば，テキストコーパスの中からパターンに該当する文を探し出すことができ，推測を行うことができる．語彙パターン間に推論が成り立つかどうかの知識は Entailment ルールと呼ばれる．本稿では，この語彙パターン間の Entailment ルールを抽出するシステムを考える．

Entailment ルールは，T_1，T_2 をそれぞれ語彙パターンと変数の 3 つ組とすると，“$T_1 \rightarrow T_2$” のように表現される知識であるが，一般的に Entailment ルールは，T_1 が真のとき T_2 が真になる，ということを表すものである．以後，この定義に合致するものを Entailment ルールとして扱う．

2. 関連研究

Berant ら [Berant 10] は，特に上位下位関係，及び言い換え関係にある動詞間での Entailment ルールの抽出を行うために，Entailment グラフ上での線形計画法による最適化問題を解くことで，どの動詞間に Entailment ルールが存在するかを求めるアルゴリズムを提案した．具体的には，まず動詞の上位下位関係が分かる WordNet を利用して正例と負例の訓練データを作成することにより，Entailment ルールの分類器を学習する．次に，Entailment ルールを抽出したい動詞の集合が与えられた時，分類器を元にしたルールの存在の可否と，動詞の 3 項間には Entailment のループが存在してはいけない，などのいくつかの束縛条件の下，与えられた動詞間で大域的な最適化問題を定義して解くことにより，Entailment ルールの抽出を実現している．

また，Web テキストを対象として Entailment ルールの抽出を行った研究として，Schoenmackers らの研究 [Schoenmackers 10] がある．この研究では 1 次ホーン節を Web から教師無しで学習することを目的としている．この研究ではまず，関係抽出の研究で広く用いられている手法により “City” などのクラス名と，“New York” などのクラス内のインスタンスを抽出する．次に，抽出したインスタンス間に存在する関係を抽出し，インスタンスのペアとその間の関係，という 3 つ組を得る．得られた 3 つ組間に存在するルールを，条件付確率と帰納論理プログラミングの手法を用いて抽出している．

これらの研究に対し本研究では，Entailment ルールが成立する背景にある意味領域に注目する．すなわち，どの意味領域においてどの Entailment ルールが成り立つのか，意味領域を限定した場合としない場合についての獲得できる Entailment ルールの違いを検証する．また，“$T_1 \rightarrow T_2$” で表わされる Entailment ルール成り立つ 3 つの場合 (因果・推移関係，上位下位関係，言い換え関係) それぞれについて，各関係を区別できる指標を提案する．

連絡先: 田中翔平，tanaka@mi.ci.i.u-tokyo.ac.jp

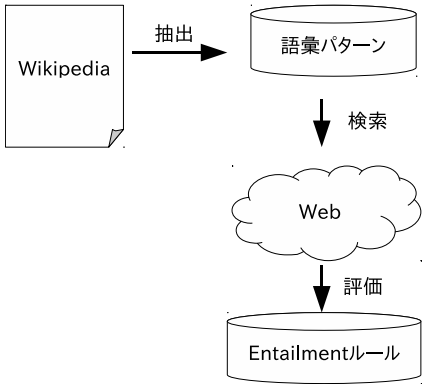

図 1: Entailment ルール抽出システム

3. 提案システム

図 1 に，システム全体図を示す．Entailment ルールの抽出は，以下の 3 つのステップで行う．

1. 語彙パターン抽出
   ルール抽出の対象となる語彙パターンを抽出する．
2. Web 検索
   ルールの評価に必要な情報を Web から取得する．
3. Entailment ルール抽出
   語彙パターン間に Entailment 関係が存在するかを評価し，Entailment ルールを抽出する．

3.1 語彙パターンの抽出

本研究では，Wikipedia[*1]をコーパスとして用いる．本システムでは特定のカテゴリに注目した Entailment ルールを抽出するため，入力としてカテゴリ名のペアを与える．語彙パターンの抽出は入力カテゴリに属する記事のみを対象とし，抽出するパターンも入力カテゴリペア間の関係を記述した語彙パターンのみである．例えば，入力として (company, product) を与えると，company のカテゴリに属する記事と，product のカテゴリに属する記事のみを語彙パターンの抽出対象とし，さらに抽出するパターンも company-product の関係を記述したものに限定する．

抽出する語彙パターンは，以下のルールに該当する単語の並びである．

- タイトルと共参照関係にある単語と，Wikilink との間に挟まれる．
- アルファベットと「-」,「'」のみで構成される単語の系列で，6 単語以内である．

タイトルと共参照関係にある単語 (タイトル自身，及び代名詞など，文中でタイトルのことを言及するフレーズ) とは，次のルールに該当する単語である．

- タイトルそのもの．
- タイトルをスペースで区切ったそれぞれのトークン．
- he, she, they, the *noun* の中で最も多く出現したフレーズ．

2 つ目のルールは,「Apple Inc.」がしばしば本文中で「Apple」のみで言及されるような，省略による共参照表現を解決するルールである．

次に，パターンのフィルタリングと抽象化の処理を行う．まず，前置詞や be 動詞，定冠詞などのストップワードのみからなる語彙パターンは削除する．例えば，“is a” や “and the” などの語彙パターンがこれに該当する．次に，be 動詞，定冠詞を “*”(ワイルドカード) で置き換える．これは，例えば “was built by” と “were built by” を “* built by” に抽象化するように，動詞の活用によるパターンの揺れを吸収するためである．

最終的に Wikipedia から抽出される company-product 語彙パターンの例を表 1 に示す．なお，語彙パターン抽出の際に用いた，パターンの前後に出現するエンティティのペア (すなわち，タイトルと Wikilink のペア) も同時に抽出する．

表 1: Wikipedia から抽出した company-product 関係の語彙パターンとエンティティペアの例．(X: company, Y: product)

Pattern	Count
X introduced Y	29
X's Y	23
X released Y	19
Y * built by X	19
X launched Y	17
...	...

X	Y
Toyota	Prius
Honda	Insight
Microsoft	Xbox
Sony	Playstation3
Nvidia	GeForce 256
...	...

3.2 Web 検索

次に，Entailment ルールの抽出と評価に必要な情報を Web 検索エンジンを用いて獲得する．本研究では，Yahoo!デベロッパーネットワーク[*2]の Web 検索 API を用いた．

Web 検索で獲得する情報は以下の 2 つである．

- 検索クエリのヒット件数
- 検索結果のスニペット

スニペットとは，検索エンジンの検索結果ページに表示される，ヒットした各ページの内容の切り抜きのことである．

検索に用いるクエリは，以下の 2 種類である．

- “X P Y” ただし，(X, Y)$\in \mathcal{S}$; P$\in \mathcal{P}$
- “X P Y” “X P' Y” ただし，(X, Y)$\in \mathcal{S}$; P, P'$\in \mathcal{P}$

ここで，$\mathcal{S}$ は抽出されたエンティティペアの集合， $\mathcal{P}$ はパターンの集合である．例えば，エンティティペアが (Toyota, Prius) であれば,「“Toyota introduced Prius”」というクエリや,「“Toyota introduced Prius” “Toyota's Prius”」というクエリでの Web 検索を行う．なお，ワイルドカード “*” で抽象化された語彙パターンについても,「“Prius * built by Toyota”」のように，そのままクエリを生成し検索を行う．この検索クエリの場合，“*” の部分に何らかの単語が入るフレーズを含む Web ページがヒットする．すなわち，例えば「“Prius is built by Toyota”」というフレーズを含む Web ページがヒットする．

3.3 ルール抽出

次に，Web 検索で獲得した情報を元に，Wikipedia から抽出した語彙パターン間に Entailment 関係が存在するかを検証し，ルールを抽出する．本研究では特に，以降述べる 3 つのタイプへと Entailment ルールを分類し，それぞれで異なるアプローチを行う．

3.3.1 因果・推移関係

「語彙パターン P_i と P_j が因果・推移関係にある」とは，(X, P_i, Y) が原因で (X, P_j, Y) が生じる場合や，(X, P_i, Y) と (X, P_j, Y) の間に時間的な遷移が認められる場合である．ここで，これらのエンティティとパターンの 3 つ組をそれぞれ T_i, T_j で表現する．T_i と T_j が因果・推移関係にある場合，T_i が生じた時，T_j が生じる確率 $P(T_j|T_i)$ が高いことが推測できる．条件付確率 $P(T_j|T_i)$ は本来求めたい確率 $P(T_i \rightarrow T_j)$ とは異

*1 http://www.wikipedia.org

*2 http://developer.yahoo.co.jp/

なるが，Schoenmackers らの研究 [Schoenmackers 10] を参考にし，T_i と T_j の間の統計的な依存度を計算することで $P(T_i \rightarrow T_j)$ を近似的に求めることを試みる．$P(T_j|T_i)$ は Web 検索のヒット件数から近似的に求める．すなわち，

$$P(T_j|T_i) = \frac{P(T_i\&T_j)}{P(T_i)} \approx \frac{hit(T_i\&T_j)}{hit(T_i)} = \frac{hit(\text{“X}P_i\text{Y}''\text{“X}P_j\text{Y}'')}{hit(\text{“X}P_i\text{Y}'')}$$

により計算する．ここで，$hit(Q)$ は検索クエリ Q の Web ヒット件数である．このスコアのエンティティペアあたりの平均値 $(\sum_{X,Y\in\mathcal{S}} \frac{hit(\text{“X}P_i\text{Y}''\text{“X}P_j\text{Y}'')}{hit(\text{“X}P_i\text{Y}'')})/|\mathcal{S}|$ の高い語彙パターン間には因果・推移関係が存在すると考え，因果・推移関係を示す Entailment ルールとして抽出する．

3.3.2 上位下位関係

「語彙パターン P_i と P_j が上位下位関係にある」とは，T_i:(X, P_i, Y) が T_j:(X, P_j, Y) を意味的に包含する場合を言う．この場合，T_i が上位，T_j が下位であり，T_i の方が T_j よりも広い (様々な) 文脈で使える表現であることになる．従って上位下位関係の抽出には，語彙パターンが用いられる文脈を解析して Entailment ルールを抽出する必要がある．本研究では，Web 検索で獲得した語彙パターンの文脈ベクトルの要素から，上位下位関係の抽出を試みる．

まず，文脈ベクトル v_i を定義する．v_i の各次元は P_i を使った Web 検索で得られたスニペットに出現する各単語に対応し，値はその単語のスニペット内での出現数である．次に，文脈ベクトルの重複度を，以下で定義する．

$$overlap(v_i, v_j) = \frac{|v_i \cap v_j|}{|v_j|}$$

ここで，$|v_i \cap v_j|$ は v_i と v_j に共通して出現する 0 でない要素 (単語) の数であり，$|v_j|$ は v_j の 0 でない要素の数である．つまり，P_j による検索で得られたスニペットの集合に出現する単語の中で，P_i による検索で得られたスニペットの集合の中にも出現するものの割合を計算することができる．語彙パターン P_i，P_j が上位下位関係にあるとき，P_i が用いられる文脈では P_j を用いることができる ($O(v_j \in v_i)$ が高い) が，P_j が用いられる文脈では必ずしも P_i を用いることができない ($O(v_i \in v_j)$ が低い)．したがって，P_j が上位，P_i が下位の上位下位関係を検証する score 関数は以下で定義する．

$$score(v_i \rightarrow v_j) = O(v_j \in v_i) - O(v_i \in v_j)$$

本アプローチでは，このスコアの高い語彙パターンのペア間に，上位下位関係が存在するとして Entailment ルールを抽出する．

3.3.3 言い換え

言い換え関係は，上位下位関係の特殊な場合と考えることができる．すなわち，ルール抽出対象となる語彙パターンの両方から，ともに意味的な包含が成り立つ場合である．文脈ベクトルでそのような場合を考えると，文脈ベクトルが類似していて，かつ語彙パターン同士があまり同一文書内で記述されない場合，言い換え関係が成り立つと言える．つまり，cosine 類似度の高いパターン同士で，両者の Web 上での共起が少ないペア間に，言い換え関係の Entailment ルールを抽出する．

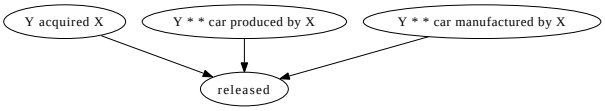

図 3: 上位下位関係 (car manufacturer-vehicle)

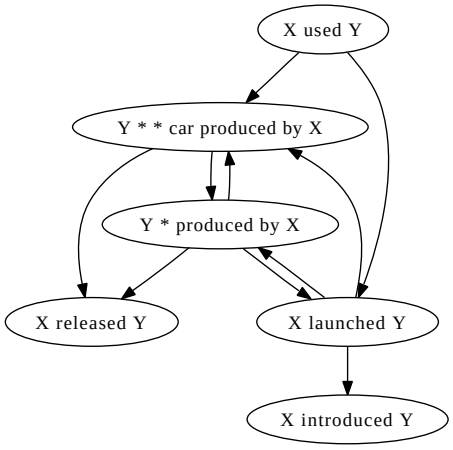

図 4: 因果・推移関係 (campany-product)

4. 実験

4.1 概要

ここでは，本稿で言及した基礎的なアプローチを利用した Entailment ルールの抽出実験の概要と，その結果を示す．まず，提案システムは語彙パターンの抽出に Wikipedia を用いるが，本実験では英語版の Wikipedia 記事 (2011 年 10 月 11 日にダンプされたデータ) を利用する．次に，実験の対象とするカテゴリは会社-製品カテゴリ (company-product) と車メーカー-車カテゴリ (car manufacturer-vehicle) である．後者は前者の意味領域をさらに限定したカテゴリとなっており，両カテゴリでの Entailment ルール抽出の比較を行う．会社カテゴリは記事の属するカテゴリ名の中に “companies” が含まれる記事とし，製品カテゴリは記事の属するカテゴリ名に “device”, “introduction”, “product” が含まれる記事とする．車メーカーカテゴリは記事の属するカテゴリ名の中に “car manufacturer” が含まれる記事とし，車カテゴリは記事の属するカテゴリ名に “vehicle” が含まれる記事とする．実験では対象となる記事から本稿で紹介した手法を用いて語彙パターンを抽出し，提案システムを用いて各パターン間に存在する Entailment ルールを抽出した．

表 2: 言い換え表現 (company-product)

Pattern1	Pattern2
X producces Y	X produced Y
X developed Y	Y * developed by X
X produces Y	Y * produced by X
X * * manufacturer of Y	Y * manufactured by X
Y * * car produced by X	Y * produces by X
...	...

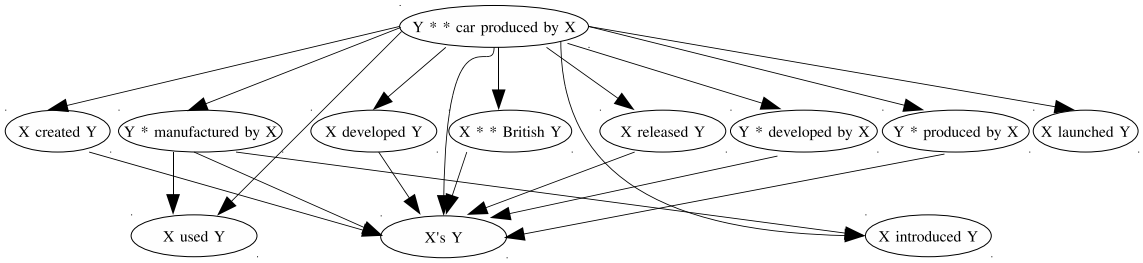

図 2: 上位下位関係 (company-product)

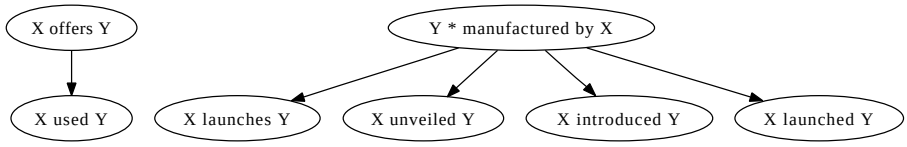

図 5: 因果・推移関係 (car manufacturer-vehicle)

表 3: 言い換え表現 (car manufacturer-vehicle)

Pattern1	Pattern2
X launched Y	X launches Y
X produced Y	X * produces Y
Y * * car manufactured by X	Y * manufactured by X
Y * * car produced by X	Y * * car manufactured by X
X produced X	Y * * car produced by X
...	...

4.2 結果

因果・推移関係の Entailment ルールの抽出結果を図 4，図 5 に示す．このグラフは Entailment グラフと呼ばれるものであるが，矢印の根元の語彙パターンが矢印の先のパターンを entail していることを示している．図 4 はスコアの高い順に 10 個のルールを描いたものである．図 5 は条件付確率の閾値に図 4 の中での最低値を採用して描いたグラフである．図 4 では produced,released,introduced という時間推移関係が抽出でき，図 5 では特に車という製品で使われやすい manufacture という動詞の周辺のの時間推移関係が抽出できた．

図 2，図 3 に，提案システムが抽出した上位下位関係の Entailment ルールを示す．図 2 は，スコアの高い順に 20 個のルールを描いたものであり，図 3 は閾値として図 2 の中での最低値を採用したグラフである．company-product カテゴリでは “’s” を頂点とした上位下位関係が抽出できたが，car manufacturer-vehicle カテゴリでは produce-release という時間推移関係に相当するルールが抽出された．これは文脈ベクトルの重複度の計算により正しく上位下位関係を抽出できなかったのが原因だと考えられる．本研究では文脈ベクトルは検索結果のスニペットから生成するが，スニペット自体が Web ページのほんの一部であり，膨大にある Web テキストをごく一部のテキストで近似していると言える．従って実際の Web テキストまで踏み込んで解析を行うことで，より情報の多い文脈ベクトルが生成できることが期待できる．また，Wikipedia から語彙パターンを抽出する際に不必要な語彙パターンを抽出してしまった例がある．それは図 2 中の “X * * British Y” や，“X used Y” などである．前者・後者ともに X, Y に相当するエンティティが抽出したい company，product のペアで無かったのが原因である．これはカテゴリ指定のために用いたキーワードによる誤りであり，今後はカテゴリ構造を考慮し，抽出対象とするエンティティをより洗練する必要があると考えている．

表 2，表 3 に言い換え表現として抽出された語彙パターンのペアを示す．掲載しているのは，類似度の高い上位 5 件である．それぞれのカテゴリで同じ動詞の変化形間のルールが抽出される一方で，car produce-car manufacture のようなカテゴリ内で限定的に成り立つ言い換え関係も抽出できた．今後は manufacture-produce のような動詞が異なるが，意味的に類似する言い換え関係を抽出するために，周辺文脈のより詳細な解析を行うつもりである．

5. おわりに

本稿では，Wikipedia から獲得した語彙パターン間に存在する Entailment ルールを，Web 上の情報を利用して抽出するシステムを提案した．また，2 つのカテゴリについて提案システムを用いて Entailment ルールの抽出を行い，その結果を示した．今後は，今回の実験結果で判明した問題点を踏まえ，まず各手順での精度の向上を行うつもりである．また，カテゴリの限定と Entailment ルールの成立について，更なる調査を行うつもりである．そして，RTE などの応用的なタスクについても，本研究の適用可能性を探る予定である．

参考文献

[Berant 10] Berant, J., Dagan, I., and Goldberger, J.: Global learning of focused entailment graphs, in *Proc. of the 48th Annual Meeting of the ACL*, pp. 1220–1229 (2010)

[Schoenmackers 10] Schoenmackers, S., Etzioni, O., Weld, D. S., and Davis, J.: Learning first-order Horn clauses from web text, in *Proc. of the 2010 Conference on EMNLP*, pp. 1088–1098 (2010)