語彙知識を用いた日本語テキスト含意認識評価セット構築と認識実験

村松 祐希 , 山本 和英
長岡技術科学大学 電気系
E-mail:{muramatsu,yamamoto}@jnlp.org

1　はじめに

近年、テキスト含意認識の研究が注目を集めている。テキスト含意認識とは本文（text,t）と仮説 (hypothesis,h) を与え、仮説から本文が推論可能であるかを判断するタスクである。以下に例を示す。（本稿では推論可能であるものを含意していると呼ぶ。）

例 1) テキスト含意認識の例
t:ロシア外務次官は１３日夜、６カ国協議で合意した北朝鮮への重油５万トン支援への参加を見合わせることを明らかにした。
h:重油５万トン支援、ロシアは参加せず・6 カ国協議合意
含意判定:真

上記のように t から h の内容が可能であるなら、含意判定は真（ならなければ偽）となる。

テキスト含意認識の研究を活発化させる動きとして、大規模な評価型ワークショップが行われている[1]。このワークショップは過去に RTE-1〜4 が行われている（現在、RTE-5 が開催中である）。また、ワークショップの回数を重ねるごとに要求されるタスクがより高度なものになっている。例えば、「Entailment」,「Unkown」,「Constradiction」の 3 値判定の評価セットが追加されている。Giampiccolo ら [1] の報告によると 2008 年に行われた RTE-4 の成績トップのシステムの精度は 2 値判定が約 75%、3 値判定が約 70%であった。

これらのワークショップで使用されるテキスト含意認識の評価において 3 つの問題点が考えられる。1 つ目は評価セットの構築方法である。既存の評価セットは、質問応答、情報抽出、要約、情報検索の手法を元に作られている。つまり、テキスト含意認識の問題を先の 4 つの手法と置き換えている。しかし、テキスト含意認識の問題の中で、どの程度網羅しているか分かりにくいことが挙げられる。2 つ目は評価方法である。現在、評価方法は精度と Confidence Weighted Score(CWS) がある。CWS は情報検索などで使用する平均適合率を参考にしている。平均適合率を用いる際、評価セットの重要度を考慮しなければならない。しかし、いくつかの手法を用いて評価セットを構築した場合、適切な重要度を考えることは難しい。3 つ目は言語の壁である。公開されている評価セットは一般的に英語が多い。よって、他の外部情報が日本語である場合、言語の壁が発生し、扱いにくいことが挙げられる。日本語の評価セットを構築する手法は僅かに存在する。しかし、評価セットの分類観点や再現性の高さという面に課題がある。評価セットの重要度を考慮していない点や分類観点のごく一部にしか対応出来ていない点も問題である。

これらの問題を解決するための一環として、本稿では語彙知識を用いた日本語評価セットの構築を提案する。語彙知識を使うことで従来よりもテキスト含意認識の問題を明確化させ、再現性の高い評価セットを構築することが可能である。さらに、語彙知識の使用した回数と形態素の連接確率から評価セットの重要度が計算可能となり、評価方法の改善になる。

また、本稿で構築した評価セットの網羅性について調査する必要がある。そこで、過去の手法を参考に 2 つの手法を考案し、構築したデータの含意認識実験を行った。参考にした手法は本文と仮説の形態素の合致率を考慮した手法と共起を使用した手法を用いた。

2　関連研究

テキスト含意認識の評価セットを構築した報告として Dagan et al.[2] の研究がある。彼らは主に新聞コーパスから抽出した文を本文とし、7 つの手法 (質問応答、文章読解、情報抽出、機会翻訳、換言、情報検索、類似文章) により、本文から仮説を作成している。以下に例を示す。

例 2)Dagan et al. の手法 (情報検索)
t:Google files for its long awaited IPO.
h:Google goes public.
含意判定:真

例 2 は情報検索の手法を用いている。情報検索の手法では新聞コーパスから抽出した文を仮説とし、Web 検索エンジンを使用することで本文を取得している。しかし、独立した 7 つの手法から構築される評価セットに対して重要度を決めることは難しい。国内の報告では小谷ら [3] の研究がある。彼らは、含意判定のための推論要因を「包含」、「語彙 (体言)」、「語彙 (用言)」、「構文」、「推論」の 5 つに分類した。それぞれの分類に下位分類を設け、日本語テキスト含意評価セットを構築した。しかし、各分類の説明に曖昧性が存在する。例えば、「語彙 (体言)」の説明は「t にある名詞の意味や性質が h の真偽の情報が与えられるようなデータである。」となっている。そのため、手法を再現することが困難と考える。含意判定には国語辞典と Web テキストを使用し、同義・上位下位関係の語や句によって解いている。含意文の生成では大西ら [4] の研究がある。動詞意味分類に対して格交代の情報を付与し、含意する規則を考えている。生成した文が妥当であるかを求めるために、「名詞」、「格助詞」、「動詞」の自己相互情報量から計算している。しかし、作成された含意文を我々が観察した所、含意文が 6 形態素以内の短い文しか存在しなかった。この観察結果から含意する規則が長い文に対して使用することが難しい。

認識実験の関連研究として Perez and Alfonseca[5] の研究がある。本文、仮説から含意判定を行うため、本文、仮説の合致率を計算している。具体的には BLEU を用いて、予め決めた閾値より BLEU 値が高ければ真、低ければ偽という含意判定を行っている。閾値は経験的な値を使用している。共起語を使用した手法として Glickman et al.[6] の研究がある。全ての仮説の形態素から最も共起確率の高い本文の形態素を計算する方法である。共起確率と仮説の確率から計算した値を積算し、閾値より高ければ真としている。共起確率には Web 検索エンジンを使用して計算している。

日本語を対象とした含意認識手法では小谷ら [7] の報告がある。換言表現を述語項構造解析の観点から正規化することで、含意判定を行っている。正規化には形態素解析器 JUMAN の辞書に含まれる代表表記と構文解析器の格解析の結果を使用している。また、LFG 解析を用いた手法として外池 [8] らの手法がある。LFG(Lexical Functional Grammar) とは生成文法の 1 つである。文をあらかじめ用意した文法に汎化させ、含意判定を行っている。従来は機能表現の意味を考慮していなかったが、現在は機能表現データベースを LFG 文法に取り入れることで解決している。しかし、LFG 文法は再現率が低いため、彼らの手法に関しても課題が残る。

[1] http://www.nist.gov/tac

3 語彙知識

本稿で対象としている語彙知識は含意関係として考える語の集合である。過去の手法で、含意関係を階層型辞書における上位下位関係と考える場合が多く、本稿も同様に考える。そのため、階層型辞書を使用することで、語彙知識を獲得することが出来ると考えた。そこで、Bond et al.[9] が構築した日本語 WordNet[1)] を使用する。日本語 WordNet は独立行政法人情報通信研究機構 (NICT) が開発した階層型辞書であり、一般公開されている。以下に日本語 WordNet の例を示す。

表 1: 日本語 WordNet の構成例

synset		entailment(05780885-n)
definitions	eng	something that is ～ implications
word	jpn	暗示、含み、内容、含意、含蓄
	eng	entailment, implication, deduction
relations	hype	inference

表 1 は日本語 WordNet の構成例である。synset は概念を表す。この場合、概念は entailment である。definitions は概念の語釈文である。word は synset に含まれる単語である。word には jpn(日本語) と eng(英語) が存在する。また、複数の単語が登録されている場合、それらは同義語を意味する。(「暗示」、「含み」、「内容」、「含意」、「含蓄」は同義語として考える。) relation では他の概念との関係を示す。例えば、hype は上位概念を意味する。他には下位概念を意味する hypo なども存在する。この場合、entailment の上位概念は inference となる。本稿は日本語 WordNet の上位概念に含まれる単語と下位概念に含まれる単語の関係を上位下位関係とする。よって、表 1 の日本語の上位語は上位概念の inference に含まれる単語と考える。

4 評価セットの構築

我々の提案する評価セットの構築は大きく分けて 4 段階からなる。評価セットの構築の手順について以下の図に示す。

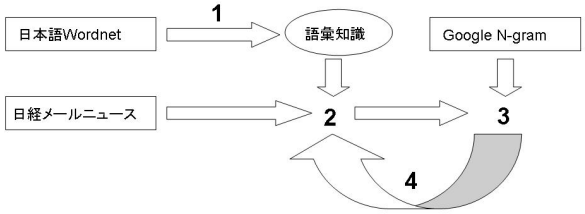

図 1: 評価セット構築方法

1. 語彙知識の獲得
2. 本文中の語を選定
3. 仮説の生成
4. 仮説のフィードバック

本文に対して僅かに語彙を変化させた仮説が良い仮説と考える。これは含意判定として難しい問題が出来るからと考える。本手法は日本語 WordNet、日経ニュースメール [2)]、Web 日本語 N グラム第 1 版 [3)] を資源として利用した。日本語 WordNet から語彙知識を獲得し、日経ニュースメールの記事を本文として入力する。語彙知識によって置き換える語を本文中から選択する。語彙知識の中から最も自然な文となる語を n-gram 確率を用いて計算し、仮説を生成する。仮説を新しい本文として、フィードバックさせることにより、新しい仮説を生成する。

4.1 語彙知識の獲得

本節はテキスト含意認識の評価セットを構築するために日本語 WordNet から語彙知識を獲得する。語彙知識は日本語 WordNet に含まれる上位下位語、同義語とする。上位下位語は親子関係まで考慮する。

4.2 本文中の語を選定

本節は本文の中の語を語彙知識と置き換えることで仮説を生成する。我々は本文として日経ニュースメールを採用した。日経ニュースメールは事実となる情報を 1～3 文で記述されており、評価セットとして適していると考えた。日経ニュースメールの例を示す。

例 3) 日経ニュースメール
◆大企業景況感、２期ぶりマイナス (見出し)
10～12 月内閣府などが発表した１０～１２月の大企業の景気判断指数はマイナス１．９。デフレ・円高を警戒、マイナスは２期ぶり。(本文)

日経ニュースメールは 1 文と「見出し」と 1～3 文の「本文」で構成される。本稿はこの「本文」を本文として用いる。ここから本文の中の語の選び方について述べる。日経ニュースメールの「本文」を形態素解析器 MeCab[4)] で形態素解析を行う。「本文」の全ての形態素に対して、語彙知識の中に同じ形態素が存在するかの判定を行う。語彙知識に含まれると判断された形態素の中で無作為に 1 つ選択する。選択された形態素が仮説の置き換え元となる。無作為に選択する理由は 4.3 節で計算する置き換え先の語の計算量を減らすためである。

4.3 仮説の生成

本節では 4.1 節で獲得した語彙知識と 4.2 節で選定した本文中の形態素を置き換えることで仮説を生成する。しかし、選定した本文中の形態素に対して語彙知識の形態素が複数存在する場合、なんらかの観点で差別化する必要があると考える。そこで、本手法は形態素の連接確率を用いた。語彙知識によって置き換える元と先の連接確率の向上率が高いほど、人間が読んで自然な文であると考えた。向上率にした理由は語彙知識によって入れ替える前の文に対して平等に扱うのが良いと考えたからである。連接確率には統計量として信頼できる Web 日本語 N グラム第 1 版を使用する。Web 日本語 N グラム第 1 版は誰でも入手可能で、1～7gram のデータがある。本手法では計算コストを考え 3-gram のみを用いる。よって、向上率は 4.2 節で選定された形態素と前後 2 形態素を用いて向上率を計算する。Web 日本語 N グラム第 1 版に存在しないデータは出現回数 1 回として計算する。選択された形態素を w_s と考える。また w_s に対応している語（上位下位語、同義語）の集合を D とする。($w_d \in$ D) 最も向上率の高い形態素 w_d は次の式により計算する。

$$argmax(w_d) = \frac{p(w_d|_{w_{d-1}}^{w_{d-2}})p(w_{d+1}|_{w_{d-1}}^{w_d})p(w_{d+2}|_{w_d}^{w_d+1})}{p(w_s|_{w_{s-1}}^{w_{s-2}})p(w_{s+1}|_{w_{s-1}}^{w_s})p(w_{s+2}|_{w_s}^{w_s+1})} \quad (1)$$

上記の式より w_s に対して最も高くなる w_d を置き換えることにより、仮説を生成する。本手法は 1 つの本文に対して仮説の数も 1 つに限定する。

4.4 仮説のフィードバック

本手法は 4.3 節で生成した仮説を新しい本文としてフィードバックさせる。(4.2 節から再び繰り返す。) フィードバックさせた本文を元に新しい仮説を生成することを繰り返すことで、大量に評価セットを構築出来ると考えた。フィードバックする回数は任意の回数とする。

5 認識実験

本稿では過去のテキスト含意認識の手法を元に 2 つの手法を考案して認識実験を行った。

5.1 ROUGE

1 つ目は本文と仮説の形態素の合致率を使用した手法である。Perez and Alfonseca は BLEU から含意判定を行った。しかし、BLEU の計算式には合致率を計算する文の表層が異なる場合、ペナルティ加算される。また、4-gram まで考慮して計算を行う場合、1〜4-gram まで考慮される。本手法で構築した評価セットは 2 文の表層が似ているため、1〜3-gram を考慮しても差が出にくいと考える。そこで、本手法は ROUGE を使用して計算を行った。ROUGE は指定された n-gram のみ考慮するため、BLEU よりも良いと考える。任意に設定された閾値より高ければ真、低ければ偽と判定する。共起確率は従来 Web 検索エンジンの検索ヒット件数が用いられた。しかし、長距離の共起が発生し、ノイズになる可能性が高くなると考えた。そこで、本実験では Web 日本語 N グラム第 1 版 7-gram のデータを使用した。よって共起のウィンドウサイズは 7 とした。

5.2 相互情報量

2 つ目は本文と仮説の形態素の共起頻度を使用した方法である。Glickman et al. は本文の形態素の出現頻度と本文と仮説の形態素の共起頻度から含意判定を行った。しかし、仮説の形態素のみの出現頻度を考慮していない。そこで本手法は相互情報量を用いた手法を提案する。以下の式から計算する。ただし、本文の形態素を t=$\{v_1, \cdots, v_n\}$ とし、仮説の形態素を h=$\{u_1, \cdots, u_m\}$ と定義する。p() は () に含まれる形態素の出現確率を示す。() に複数含まれる場合は、共起確率を示す。

$$S_{MI} = \frac{1}{m} \sum_{u \in h} max_{v \in t} \frac{p(u,v)}{p(u)p(v)} \quad (2)$$

全ての仮説の形態素に対し、相互情報量が最大値となる本文の形態素を相加平均する。S_{MI} が任意の閾値より高ければ真と判定し、低ければ偽と判定する。

6 評価実験

6.1 評価セット構築

評価セット構築に日経ニュースメール 32062 件を入力として使用した。語彙知識のために使用した日本語 WordNet は最新版 (0.92) を使用した。収録概念数、単語数、語義数（概念と単語の対）は 49,655 概念、87,133 単語、146,811 語義であった。語彙知識は上位語、下位語、同義語に分けて実験を行った。全ての日経ニュースメールに対して本文を生成し、本文に対する仮説を生成した。仮説のフィードバック回数は、本実験で 5 回とした。生成した仮説のスコアの高い順から上位 100 件の評価セットに対して人手評価を行った。評価は本文と仮説の対が真、偽の 2 値判定を行った。

6.2 認識実験

認識実験では 6.1 で人手で 2 値判定した結果をタグ付けし、評価セットを入力として使用した。実験結果で示す CWS には評価セットの重要度を考慮しなければならない。そこで、評価セットの重要度を次のように決めた。まず、フィードバックの回数が少ない評価セットほど重要と判断した。これは日経ニュースメールの内容に近く、事実を述べている可能性が高いからと考える。フィードバックの回数が同じ評価セットでは仮説の向上率を表す w_d が高いほど重要とした。これは可読性が高い文であるほど重要な文と考えた。

また、各手法において任意の閾値を決める必要がある。本実験では全ての評価セットで計算した手法別の各スコアの中で最も高い評価となったスコアを各手法の閾値として設定した。

7 評価結果

6 章で述べた評価実験の結果を以下に示す。

7.1 評価セット構築

使用した語彙知識とフィードバックの実験結果を以下の表 2 に示す。

表 2: 評価セット構築

	フィードバック回数				
	1 回	2 回	3 回	4 回	5 回
上位語	68%	41%	41%	39%	21%
下位語	46%	26%	43%	32%	32%
同義語	81%	60%	43%	37%	30%

上記の結果は 6.1 節で述べた方法で評価を行った結果、真と判定された割合である。全体的な傾向として、フィードバックの回数を増やすと真が少なくなる傾向であった。しかし、下位語を用いた場合、フィードバック回数が 2 回から 3 回にすると 17%向上した。真の評価が最も多かった条件は同義語を用いてフィードバック回数が 1 回のときであった。最も少なかったのは上位語を用いてフィードバック回数が 5 回のときであった。フィードバック回数 5 回までの累計で真のデータが多く生成された知識は同義語＞上位語＞下位語になった。6.2 節で使用する評価セットは上記となり、評価セット数は上位語の真が 210 件（500 件中）、下位語の真が 179 件（500 件中）、同義語の真が 251 件（500 件中）となった。

7.2 認識実験

6.1 節で構築した評価セットを使用し、認識実験を行った。以下表 3 に結果を示す。

表 3: 認識実験結果

	上位語		下位語		同義語	
手法	精度	CWS	精度	CWS	精度	CWS
Perez and Alfonseca	56%	46%	64%	60%	51%	71%
ROUGE	56%	48%	65%	61%	51%	71%
Glickman et al.	56%	48%	64%	63%	51%	59%
相互情報量	56%	56%	65%	61%	53%	71%

評価に用いた精度と CWS は以下の方法で計算を行った。以下に式を示す。

$$\text{精度} = \frac{D}{All} \quad (3)$$

$$CWS = \frac{1}{D} \sum_{1 \leqq k \leqq N} r_k \cdot precision(k) \quad (4)$$

$$precision(k) = \frac{1}{k} \sum_{1 \leqq i \leqq k} r_i \quad (5)$$

All=全評価セット数,D=全正解データ数,N=最後の正解データ番号（ただし、もし k が正解なら r_k=1、不正解なら r_k=0）

4 つの手法を比較した結果、精度と CWS 共に相互情報量の手法が下位語の CWS を除いて最も高い結果となった。また、ROUGE は Perez and Alfonseca とほぼ等しい結果となった。全ての手法において下位語の精度が最も高い結果となり、上位語の CWS が最も低い結果となった。精度と CWS を比較した場合、上位語と下位語は精度が CWS より高い結果となったが、同義語は低い結果となった。精度と CWS で最も差があったのは同義語となり、無かったのは下位語となった。

8 考察

7.1 節評価セット構築と 7.2 節認識実験についてそれぞれ考察を行った。

8.1 評価セット構築

構築した評価セットについて考察を行った。以下に例を示す。

例 4) 上位語の例
t:米ボーイングは２９日、次期超大型旅客機の開発延期を発表。
h:米ボーイングは２９日、次期超大型旅客機の開発中止を発表。
含意判定:偽

例 4 は上位語の語彙知識を用いて評価セットを構築した例である。日本語 WordNet に存在する「延期」の上位語に「中止」が含まれていたため、含意していない評価セットとなった。改善点として、連接確率の要素数を大きくするが挙げられる。同義語を用いた含意していない例として以下のような例がある。

例 5) 同義語の例
t:米アカデミー賞授賞式が始まり、２度目の受賞が期待された宮崎駿監督の「ハウルの動く城」は惜しくも受賞を逃した。
h:米アカデミー賞授賞式が始まり、２度目の優勝が期待された宮崎駿監督の「ハウルの動く城」は惜しくも優勝を逃した。
含意判定:偽

例 5 の場合、「受賞」と「優勝」が日本語 WordNet で同義語に登録されているため含意していない結果となった。解決策として、連接確率で考慮されていない形態素に対し、共起確率を用いて、連接確率と線形補間することが挙げられる。例えば例 5 の場合、「賞」や「受賞」共起確率と「賞」と「優勝」を比較することで解決できる可能性があると考える。

8.2 認識実験

7.2 節の結果から BLEU (Perez and Alfonseca) と ROUGE(本手法) はほぼ等しい結果となった。本手法で構築した評価セットは本文と仮説の表層が非常に似ているため、どちらの手法も含意判定することに不向きであったと考える。改善するためには表記ゆれや編集距離などを考慮した方法が考えられる。また、本手法の評価セットは本文と仮説の文の数が同じであるが、文の数が異なる場合、ROUGE のスコアは下がる傾向があるため、注意する必要があると考える。

相互情報量を用いた方法が一部を除いて従来手法 (Glickman et al .) よりも上回った理由について考察する。従来は本文中の確率と本文と仮説の共起確率しか考慮していない。このことより、仮説中の確率が高いほど真に判定する傾向があったと考える。本手法は相互情報量を使用したことにより、仮説中の確率を考慮されたため、従来よりも上回ることが多かったと考える。

本手法では構文解析の結果や機械学習を用いた手法を考慮していないため、今後はこれらの手法を検討する必要があると考える。

9 おわりに

入力された新聞コーパスに対して含意認識のための評価セットを構築し、認識実験を行った。手法は語彙知識の獲得、本文中の語を選定、仮説の生成、仮説のフィードバックの四段階で構築した。本手法では日本語 WordNet を語彙知識として扱った。評価セット構築の結果、上位語、下位語、同義語の評価セットを構築することができた。また、認識実験では相互情報量を用いた手法が一部を除いて最も高い結果となった。今後は構文情報や機械学習を用いた手法を検討する予定である。

使用したツール及び言語資源

1) 日本語 WordNet, 独立行政法人情報通信研究機構, http://nlpwww.nict.go.jp/wn-ja
2) 日経ニュースメール, https://letter.goo.ne.jp/nkgmail/member.cgi
3) Web 日本語 N グラム第 1 版, http://www.gsk.or.jp/catalog/GSK2007-C/
4) 形態素解析器 MeCab,Ver.0.98, http://mecab.sourceforge.net/

参考文献

1] Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dangan, and Bill Dolan. The Fourth PASCAL Recongnizing Textual Entailment Challenge. *In Proceedings of the Fourth PASCAL Recongnizing Textual Entailment Challenge*, 2008

2] Ido Dagan , Oren Glickman and Bernardo Magnini. The PASCAL Recongnizing Textual Entailment Challenge. *In Proceedings of the PASCAL Recongnizing Textual Entailment Challenge*, 2005

3] 小谷通隆, 柴田知秀, 中田貴之, 黒橋貞夫 日本語 Textual Entailment のデータ構築と自動獲得した類似表現に基づく推論関係の認識. 言語処理学会第 14 回, pp. 1140–1143, 2008.

4] 大西良明, 乾健太郎, 松本裕治. 事態間関係知識の整備と含意文生成への応用. 言語処理学会第 14 回, pp. 1152–1155, 2008.

5] Diana Perez and Enrique Alfonseca Application of the Bleu algorithm for recognising textual entailment. *In Proceedings of the PASCAL Recongnizing Textual Entailment Challenge*, 2005

6] Oren Glickman, Ido Dagan, and Moshe Koppel. Web Based Probabilistic Textual Entailment. *In Proceedings of the PASCAL Recongnizing Textual Entailment Challenge*, 2005

7] 小谷通隆, 柴田知秀, 黒橋貞夫. 言い換え表現の述語項構造への正規化とテキスト含意関係認識での利用. 言語処理学会第 15 回, pp. 260–263, 2009.

8] 外池昌嗣, 梅基宏, 大熊智子, 増市博. LFG 解析を利用した日本語含意関係判定における機能表現の取り扱い. 言語処理学会第 15 回, pp. 376–379, 2009.

9] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki. Enhancing the Japanese WordNet. *in The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP*, 2009