複数の客観的手法を用いたテキスト含意認識評価セットの構築

宇高 邦弘　　山本 和英
長岡技術科学大学 電気系
{udaka,yamamoto}@jnlp.org

1　はじめに

本文（text,t) 及び仮説（hypothesis,h）を持つペアが存在するとき、本文の持つ意味を仮説が含み得るか否かを機械的に判定するタスクをテキスト含意認識とよぶ。以下にテキスト含意認識の例を示す。

例 1) テキスト含意認識
t:坂口安吾は『信長』や『白痴』などの小説を書いた。
h:坂口安吾は小説を書いた。
含意判定:含意

テキスト含意認識は質問応答、要約、機械翻訳など、自然言語処理における幅広いタスクにおいて様々な役割を果たす。例えば、機械翻訳においては翻訳精度の指標として、質問応答では質問の答えを得る手法として、応用することが可能である。海外では大規模評価型ワークショップ [1] がこれまでに 6 回開催される (RTE-1〜RTE-6) など、活発に研究が行われており、処理対象とする言語表現や処理内容などもより高度になりつつある。また、ワークショップが開催されるにつれて新しい評価セットが公開され、内容も RTE-1 の頃に比べ、高度な処理を必要とするものになっている。

このように海外ではテキスト含意認識が注目を集めているため、現在公開されている評価セットは英語で記述されたものが多い。日本語での評価セットを構築する手法は僅かしか存在しないため、公開されている評価セットは少ない。また、既存の評価セットは作成手法や分類基準が明確でないため再現性が低い。そして様々な含意認識の問題を含むため難易性が統一されておらず、含意認識を行うシステムに入力として用いた場合に問題点を議論し難い。

これらの問題を解決するために、本稿では明確な手法を用いて複数の評価セットを構築する。具体的には、含意認識の問題として含まれる換言、要約分野で使用される手法を構築時に 1 種類のみ用いて個々の評価セットを構築する。これにより個々のテストセットには再現性があり、作成される個々のペアの難易性が変化することがない。そのため、含意認識システムの問題点を容易に把握、検討することが可能である。

2　関連研究

テキスト含意認識の評価セット構築を目的とした研究は、海外では Dagan et al.[2] の研究がある。彼らは新聞コーパスに存在する文を本文として使用し、質問応答、情報抽出、情報検索、複数文書要約、換言獲得、機械翻訳、文書読解の手法を用いた。各手法を用いて本文を加工したものを仮説とすることで評価セットを構築している。

日本語による評価セット構築の研究として、小谷ら [3] の研究がある。小谷らは含意判定のための推論要因を 5 つに分類（包含、語彙（体言)、語彙（用言)、構文、推論）し、それぞれに下位分類を設けることで、網羅性の高い評価セットを構築した。この評価セットは一般公開されており、誰もが使用可能である。しかし、小谷らの評価セットには次のような問題点が挙げられる。

まず、本文と仮説ペア数が分類によって偏りがある。小谷らの評価セットは計 2674 ペアで構成されている。その中で最もペア数の多い分類は 868 ペア、最もペア数の少ない分類は 219 ペアで構成されている。従って、この評価セットを含意認識システムの入力として用いた場合、多くのペアが存在する分類について正しく含意認識するだけで高精度となってしまう。

また、分類基準が曖昧である。例えば、語彙（用言）は「t にある用言の意味や性質から h の真偽の情報が与えられるようなデータである。」とされている。よって分類ごとのペア数の偏りを減らすために新しく評価セットを構築する場合、再現性が低く困難である。

そして、細かく分類されているため網羅性は高いが、様々な含意認識における問題を混ぜて作成されているため、個々の本文と仮説のペアにおいて難易性が統一されていない。このため、この評価セットを入力とした含意認識結果の問題点が議論しにくい。

我々は要約、換言などに用いられる手法を使用して評価セットを構築した。これにより各ペアの難易性が統一されるため含意認識結果の問題点が議論しやすい。また、明確な手法によって評価セットを構築するため再現性がある。

3　各評価セットの構築方法

本文と仮説の対を作成するために、換言、要約分野で見られる手法を用いた。これらの手法は、仮説作成時にヒューリスティックを使用しないなど、再現性があるため、作成される個々のペアは同等の難易性を持つ。これにより、作成された評価セットに含意認識を行うことで得られる結果から、含意認識システムの問題点や認識可能なペアの特徴などを把握、議論しやすい。加えて、これらの手法を組み合わせて評価セットを作成することで、より難易性の高い評価セットを作成可能となるため、難易性の操作も可能である。また、要約、換言は含意認識を行う上で必要な技術である。要約は含意認識する場合に重要な情報を得ることが可能であり、換言は構文、単語が変化する場合にも含意認識が可能となる技術である。以上から、今回は換言、要約分野で使用される 4 種類の手法を用いた。

各手法において、入力する本文は 日経ニュースメール(1) を使用した。日経ニュースメールは 1 記事が 1〜3 文で構成されており、1 記事中で話題が変化することはない。また新聞記事に見られる特殊な文体で記述されているものの、Web テキストに比べ誤用や一般的でない表現が少ない。以上の点から、評価セット構築に適していると考えた。

個々の手法において、形態素解析には $ChaSen$[2] を使用し、構文解析には $CaboCha$[3] を使用した。

最後に、作成した個々のペアについて人手で含意判断を行った。含意判断基準として「真」と「偽」を用意した。偽と判定されるものには、主語の欠如など含意判断を行うために十分な情報を仮説が持たないペアも含まれている。語順が不適切など、日本語として正しくない文が仮説として生成される場合や、固有名詞を変化させている場合があるが、そのようなペアは含意判定時に人手で省いた。個々の評価セットは真と判定されるペアが 500、偽と判定されるペアが 500 の計 1000 ペアで構成されている。

3.1 複文の単文化による評価セット構築

動詞や形容詞など、用言が名詞を修飾する文節を連体修飾節と言う。連体修飾節は意味的関係から「内の関係」と「外の関係」に分類できる。[4]

(a):階段を登る男性
(b):階段を登る足音

(a) は被修飾名詞「男性」と連体修飾節中の用言「登る」の間に格助詞「が」を補うことで「男性が登る」という単文が作成出来る。(b) は「足音」と「登る」の間にそのような格助詞も補えない。以上を基に、被修飾名詞と連体修飾節中の用言との間に格助詞を補うことで単文を生成し、仮説とした。具体的には以下の方法で本文と仮説のペアを作成した。

1. 本文を CaboCha で構文解析
2. 構文解析結果から述部でない動詞、形容詞、形容動詞が文末ではない文節にかかるとき、その文節の先頭が名詞か否かを確認
3. 文節の先頭が名詞ならば、動詞、形容詞、形容動詞と名詞との間に以下の格助詞を補うことで 9 種類の 3-gram を作成

   が、を、に、で、へ、と、から、より、まで
4. 作成した 9 種類の 3-gram について、Web 日本語 N グラム第 1 版[4] から出現頻度を獲得し、最も出現頻度の高い 3-gram を選択
5. 2 の動詞、形容詞、形容動詞が存在する文節にか係る文節を、4 で選んだ 3-gram の先頭に追加することで仮説を作成

以下に、この手法で作成可能である本文と仮説のペアの例を示す

例 2) 複文の単文化によって作成されるペア
t:ＡＴ＆Ｔは高速ネット接続を可能にするＣＡＴＶ網を他の通信会社に開放する。
h:高速ネット接続をＣＡＴＶ網が可能にする

3.2 述部に係らない文節の削除による評価セット構築

益岡ら [5] によると、「述語」は文の中心的な要素であり、特定の事態を表現する。しかし、主語と述語だけで構成された文は仮説として用いた場合に、判定に必要な情報を十分に持たないことがある。以下に例を示す。

例 3) 含意判定時に必要な情報が欠けている仮説と本文
t:インテルは１ギガヘルツのＭＰＵ「ペンティアム３」の出荷を始めたと発表
h:インテルは発表

そこで、述部に係る文節以外を削除することで仮説を生成した。これにより、含意判定するのに必要な情報をある程度残しつつ、仮説を生成できる。具体的には以下の方法で本文と仮説のペアを作成した。

1. 本文を CaboCha で構文解析
2. 構文解析結果から、述部に直接係る文節以外を削除
3. 削除した文が本文と同一でない場合、これを仮説とする

以下に、この手法で作成可能である本文と仮説のペアの例を示す

例 3) 述部に係らない文節の削除によって作成されるペア
t:ＮＴＴは電話線を使う高速ネット「ＡＤＳＬ」を月８００円で開放する
h:ＮＴＴは「ＡＤＳＬ」を８００円で開放する

3.3 副詞の削除による評価セット構築

益岡らによると、「副詞」には「様態の副詞」、「程度の副詞」、「量の副詞」などがあり、さまざまな働きをする。これらの働きは、副詞は語や文の意味を詳しくするが、語や文が示す事態には大きな影響を与えないと考える。 以上から、文中に出現する副詞を削除することで、文の示す事態を変化させずに仮説を生成した。また、名詞の中には副詞として用いることが出来るものがあるため、それらも削除の対象とした。

具体的には以下の方法で本文と仮説のペアを作成した。

1. 本文を ChaSen で形態素解析
2. 文節内に副詞が存在するなら、文中に出現するすべての副詞、名詞-副詞可能を削除。
   このとき、副詞の後に助詞、助動詞が存在する場合はそれらも削除
3. 削除した文が本文と同一でない場合、これを仮説とする

以下に、この手法で作成可能である本文と仮説のペアの例を示す

例 3) 副詞の削除によって作成されるペア
t:東証のベンチャー向け新市場「マザーズ」に２２日、ネット関連２社が初めて上場
h:東証のベンチャー向け新市場「マザーズ」に２２日ネット関連２社が上場

3.4 接頭辞の削除による評価セット構築

益岡らによると、「接頭辞」は「語幹（派生語幹）の前に付加して独立の語を派生する」働きをもつ。接頭辞も副詞と同様に語や文の意味を詳しくするが、語や文が示す事態には大きな影響を与えないと考える。以上から、文中に出現する接頭辞を削除することで、文の示す事態を変化させずに仮説を生成した。

具体的には以下の方法で本文と仮説のペアを作成した。

表 1:今回作成した評価セットに対する認識実験結果

評価セット分割番号	複文の単文化による評価セット		述部に係らない文節の削除による評価セット		副詞の削除による評価セット		接頭辞の削除による評価セット	
	Glickman et al. の認識精度	Muramatsu et al. の認識精度	Glickman et al. の認識精度	Muramatsu et al. の認識精度	Glickman et al. の認識精度	Muramatsu et al. の認識精度	Glickman et al. の認識精度	Muramatsu et al. の認識精度
1	50%	55%	50%	59%	49%	55%	50%	49%
2	50%	54%	48%	66%	50%	57%	50%	51%
3	50%	54%	50%	54%	50%	57%	50%	50%
4	49%	59%	50%	54%	49%	57%	49%	49%
5	49%	54%	50%	59%	51%	58%	49%	50%
6	50%	60%	50%	54%	50%	56%	50%	50%
7	49%	69%	49%	61%	49%	58%	49%	50%
8	50%	54%	50%	60%	50%	56%	50%	49%
9	50%	56%	49%	56%	50%	60%	50%	49%
10	50%	57%	50%	54%	48%	57%	50%	49%
標準偏差	0.21	19.76	0.44	13.77	0.64	1.69	0.21	0.44

1. 本文を ChaSen で形態素解析
2. 文節内に接頭辞が存在するかを確認 し、存在するなら文中の接頭辞を全て消去
   このとき、否定の意味を持つ以下の接頭辞は消去対象としない

   反、未、非、無、不
3. 出来た文が本文と異なるなら、ペアとして出力

以下に、この手法で作成可能である本文と仮説のペアの例を示す

例 4) 接頭辞の削除によって作成されるペア
t:ジー・オー巨額詐欺事件で大神源太被告ら５人の初公判が２０日、東京地裁で開かれた
h:ジー・オー巨額詐欺事件で大神源太被告ら５人の公判が２０日、東京地裁で開かれた

4 認識実験の方法

3 章で構築した各評価セットの難易性変化量を調べるために、2 種類の含意認識システムに各評価セットを入力として用いた。含意認識手法としては Glickman et al.[6] の手法及び、Muramatsu et al.[7] が用いた Subpath Set に基づく手法を使用した。

Glickman et al. は本文に含まれる形態素の出現確率と、本文及び仮説に含まれる形態素の共起確率から含意判定を行った。また、Muramatsu et al. は市川ら [8] が文の構文類似度を求めるために使用した Subpath Set を基に、日本語 $WordNet^{(5)}$ を使用して同義語まで考慮した構文類似度を用いて含意判定を行った。Muramatsu et al. の手法は含意認識において表層情報を扱う手法を用いた場合の含意認識結果を得るために、Glickman et al. の手法は表層以外の情報を扱う手法を用いた含意認識結果を得るために使用した。

認識実験では、構築した個々の評価セットついて 10 分割交差検定により精度を求めた。10 分割した評価セットは、含意判定時に真と判定されるものが 50 ペア、偽と判定されるものが 50 ペアで構成されている。10 分割した評価セットについて精度の標準偏差を求めた。標準偏差が 0 に近いほど精度のばらつきが少ないため、難易性が変化しないと考える。評価に用いた精度は以下の式で算出した。

$$\text{精度} = \frac{RP}{AllP} \quad (1)$$

RP:正解ペア数,AllP:使用した評価ペア数

構築した個々の評価セットを 10 分割した場合、真と判定されるデータが 50、偽と判定されるデータが 50 の計 100 ペアずつに分割される。

5 認識結果

表 1 に、今回作成した 4 種類の評価セットについて Glickman et al. 及び Muramatsu et al. の手法を用いて含意認識した場合の精度を示す。 表 1 を見ると、Glickman et al. の手法において、どの評価セットでも標準偏差が小さい傾向にある。Muramatsu et al. の手法は、複文を単文化することで構築された評価セット及び述部に係らない文節の削除によって構築された評価セットにおいて大きな標準偏差を示している反面、接頭辞及び副詞の削除によって構築された評価セットにおいては小さい標準偏差を示している。また、分割した評価セットの個々の精度を見ると、Glickman et al. の手法においてはどの評価セットでも精度に大きな変化は見られない。しかし、Muramatsu et al. の手法においては、複文を単文化することで構築された評価セット及び述部に係らない文節の削除によって構築された評価セットにおいて、大きなばらつきを示している。

6 考察

各評価セットに対する Glickman et al. の手法での含意認識結果から、今回作成した 4 種類の評価セットは難易性が変化していないと考える。これについて、各評価セットの本文と仮説のペアは出現する形態素の変化が少ない。複文を単文化することで構築された評価セットは格助詞を１文字もしくは 2 文字補う程度の変化であり、述部に係らない文節の削除によって構築された評価セットは本文中に現れる形態素をそのまま使用する。接頭辞及び副詞の削除は本文から 1〜2 形態素を削除した程度の変化となる。Glickman et al. の手法を用いて含意認識した場合、本文中に出現する形態素同士の共起確率を使用するため、どのペアでも同様の

共起確率となり、含意認識結果に大きな差が現れないと考える。Muramatsu et al. の手法での含意認識結果から、接頭辞及び副詞の削除においては難易性が変化していないと考える。副詞の削除及び接頭辞の削除による評価セットは本文と仮説で 1 文字程度の変化となる。Muramatsu et al. の手法は、構文情報から得た部分木や形態素の一致度から含意認識を行うため、構文や形態素に大きな変化がない場合は一致率も大きく変化しない。そのため、含意認識結果に大きな差が生じなかったと考える。

しかし、複文の単文化による評価セットと述部に係らない文節の削除による評価セットにおいて、Muramatsu et al. の手法による含意認識結果で標準偏差が大きくなることが示された。これについて、複文を単文化することで構築された評価セットは仮説を生成する時に構文を大きく変化させる。従って、形態素数や部分木の数によって大きく含意判定が左右される Muramatsu et al. の手法はこの評価セットを入力とした場合、正しく認識が可能なペア数にゆれが生じると考える。また、述部に係らない文節の削除によって構築された評価セットにおいて、仮説の形態素及び部分木数は本文に比べ大きく縮小される。よって部分木が 1 つでも一致しないだけで一致度が大きく変化するため、含意認識結果にも影響を与える。そのため、Muramatsu et al. の手法の入力としてこの評価セットを用いた場合、含意認識精度に大きなばらつきが見られると考える。

本手法では 2 種類の含意認識システムでの結果のみでのみ評価セットの難易性変化を測ったが、今後はこれら以外の含意認識システムを用いての検討を行う必要があると考える。

7 おわりに

本稿では、本文及び仮説で構成される各ペアに難易性の変化がなく、かつ再現性の高い評価セットを構築した。構築手法として、複文の単文化、述部に係らない文節の削除、副詞の削除、接頭辞の削除の 4 種類の手法を個々に用いて評価セットを構築した。個々の評価セットについて、Glickman et al. 及び Muramatsu et al. の含意認識手法を用いて 10 分割交差検定により含意認識を行った。その結果、Glickman et al. の手法では各評価セットにおいて、精度のばらつきが少なかった。Muramatsu et al. の手法では、副詞の削除による評価セット及び接頭辞の削除による評価セットにおいて、精度のばらつきが少なかった。反面、複文の単文化による評価セット及び述部に係らない文節の削除による評価セットにおいて、評価セットの作成手法と含意認識手法との関係から精度に大きな差が見られた。

今後の課題として、テストセットの種類数の少なさが挙げられる。含意認識に含まれる問題は推論や換言など、多くの知識と自然言語処理の応用が必要である。今回作成した評価セットはそれら全ての問題を網羅していない。以上から含意認識に含まれる問題を個別に認識出来ることを目標とし、さらに評価セットの種類を増やす予定である。また、個々の評価セットの難易性が変化しないか否かをさらに詳しく調べるために、今回使用した手法とは異なった含意認識システムを用いて含意判断を行う予定である。

使用した言語資源及びツール

(1) 日経ニュースメール,
https://letter.goo.ne.jp/nkgmail/member.cgi

(2) 形態素解析器「ChaSen」, Ver.2.3.3, 奈良先端科学技術大学院大学 松本研究室,
http://chasen.naist.jp/hiki/ChaSen/

(3) 構文解析器「CaboCha」 ,Ver.0.52, 奈良先端科学技術大学院大学 松本研究室,
http://chasen.org/~taku/software/cabocha/

(4) Web 日本語 N グラム第 1 版,
http://www.gsk.or.jp/catalog/GSK2007-C/

(5) 日本語 WordNet, 独立行政法人情報通信研究機構
http://nlp222.nict.go.jp/wn-ja

参考文献

[1] TAC 2010 workshop,http://www.nist.gov/tac/2010/workshop/

[2] Ido Dagan,Oren Glickman and Bernardo Magnini. The PASCAL Recognizing Textual Entailment Challenge. *In Proceeding of the PASCA Challenge Workshop on Recognizing Textual Entailment*, 2005

[3] 小谷 通隆, 柴田 和秀, 中田 貴之, 黒橋 禎夫. 日本語 Textual Entailment のデータ構築と自動獲得した類義表現に基づく推論関係の認識. 言語処理学会 第 14 回年次大会 発表論文集,pp.1140-1143,2008.

[4] 阿部川 武, 奥村 学. 日本語連体修飾節と被修飾名詞間の関係の解析. 自然言語処理,Vol.12,No.1,pp.107-123,2005.

[5] 益岡 隆志, 田窪 行則. 基礎日本語文法ー改訂版ー, くろしお出版

[6] Oren Glickman,Ido Dagan and Moshe Koppel. Web Based Probablistic Textual Entailment. *In Proceeding of the PASCAL Recognizing Textual Entailment Challenge*,pp.33-36, 2005

[7] Yuki Muramatsu, Kunihiro Udaka and Kazuhide Yamamoto. Textual Entailment Recognition using Word Overlap, Mutual Information and Subpath Set. *Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon*,pp.18-27, 2010

[8] 市川 宙, 橋本 泰一, 徳永 健伸, 田中 穂積. テキスト構文構造類似度を用いた類似文検索手法. 情報処理学会研究報告. 情報学基礎研究会報告 2005(42), pp.39-46, 2005