様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２６　日現在

機関番号：　１３９０４
研究種目：　挑戦的萌芽研究
研究期間：　2015～2017
課題番号：　１５Ｋ１２０９７

研究課題名（和文）異なる言語で記述されたテキスト間の含意関係認識

研究課題名（英文）Textual Entailment Recognition of Cross Lingual Passages

研究代表者
土屋　雅稔（TSUCHIYA, Masatoshi）

豊橋技術科学大学・情報メディア基盤センター・准教授

研究者番号：７０３７８２５６

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究課題では，複数文からなるテキストPと，テキストPとは異なる言語で記述された単文Hの含意関係認識について検討した．最初に，Wikipedia の日英対訳コーパスを利用して，日英の言語横断単文含意関係コーパスの作成を行った．この方法では，含意関係にある文対は大量に収集することができたが，含意関係にない文対を収集する場合には，特に論理的な矛盾を含む文対の収集に難点があることを示した．加えて，論理的に矛盾を含む文対を収集する方法について，英語単言語大規模含意関係コーパスを対象とする分析を行い，クラウドソーシングを用いて仮説文を収集するという単純な方法には問題があることを明らかにした．

研究成果の概要（英文）：This research project aimed to resolve the cross lingual textual entailment recognition. First of all, this research proposed the method to build the large cross lingual textual entailment corpus using a parallel corpus which was designed for machine translation. The experimental construction of the corpus of English-Japanese sentence pairs using wikipedia English-Japanese parallel corpora showed that the proposed method to build the textual entailment corpus was effective to collect a large number of positive examples, but it was weak to collect negative examples. Especially, it was difficult to collect negative examples which contains logical contradictions. And more, the experimental result showed that the simple crows saucing method to collect negative examples cannot avoid the hidden bias.

研究分野：　自然言語処理

キーワード：　含意関係認識

１．研究開始当初の背景

ウェブは現代の情報交換において非常に重要であるが，ウェブ上には玉石混交の情報があふれており，これらの情報を参照・利用する場合には，人手によって当該情報の信頼性を確認する必要がある．利用者が当該情報の分野に詳しくない場合には特に，どの程度の根拠情報を調査すれば良いか明らかではないため，信頼性の確認作業は大きな労力を要する．そのため，情報の信頼性を自動的に確認する，または少なくとも確認を支援する方法が強く求められている．

情報の信頼性を確認するには，情報間の隠れた関係性を明らかにすることが有効である．その基本技術として，あるテキスト P が与えられた時に，別のテキスト H の真偽を判定する含意関係認識が広く研究されている．しかし，ウェブにおける日本語の情報には，英語などの外国語の情報を根拠としている場合も数多いため，日本語テキストのみを対象とする含意関係認識は不十分であり，異なる言語で記述されたテキストを対象とする言語横断な含意関係認識が必要である．

２．研究の目的

本申請課題では，ある言語で記述されたテキストが与えられた時に，別の言語で記述されたテキストの内容の真偽を判定するという，2 つの異なる言語で記述されたテキスト間の含意関係認識に取り組む．

３．研究の方法

既存の日英対訳コーパスを対象として，人手によるアノテーションを加えて，日本語-英語の含意関係コーパスを自動的に作成する方法を健闘する．得られたコーパスを利用して，異なる言語で記述されたテキスト間の含意関係認識を，機械翻訳することなく行う方法を研究する．

４．研究成果

本申請課題では，上記の研究の方法にしたがい，最初に，日英単文含意関係コーパスの自動構築方法を研究した．日英単文含意関係コーパスの作成にあたっては，機械翻訳用の言語資源である既存の日英対訳コーパスに注目した．

日英対訳コーパスは，日本語文 J_i と，(ほぼ)同じ意味の英語文 E_i のペア (J_i, E_i) が N 個からなる集合である．この定義より，日英対訳コーパスは，含意関係が成り立っている日本語-英語の文ペアの集合と見なすことができ，含意関係の正例を収集することができる．

収集対象となる Wikipedia 対訳コーパスに含まれる文の文長（単語数）と文数の関係を，図 1 および図 2 に示す．Wikipedia 対訳コーパスは，京都に関する日本語のページを対象として，人手翻訳により作成された対訳コーパスであり，見出しなどの単なる単語列が多数含まれていることが，図中の文長 4 以下のピークより分かる．これらの単語列は，主語や述語を含まず，含意関係認識の対象とする文としては短過ぎるため，今回のコーパス作成の対象から除外する．つまり，単語数が 5 以上であり，かつ同時に末尾にピリオドが出現している英語文と，当該英語文の訳文（日本語文）からなる文対を，本研究のコーパス構築対象として用いる．

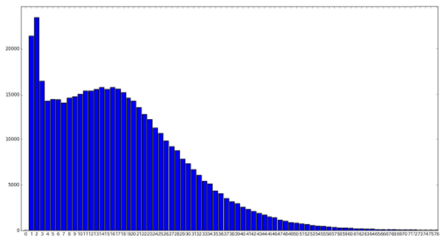

図 1　対訳コーパスに含まれる英語文の文長と文数

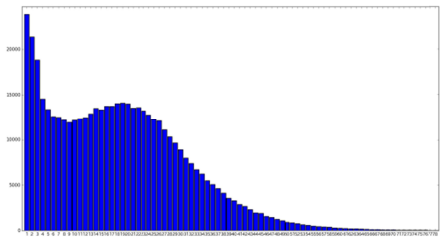

図 2　対訳コーパスに含まれる日本語文の文長と文数

日英含意関係認識器の評価を適切に行うためには，含意関係の正例だけでは不十分であり，含意関係の負例も同程度に収集する必要がある．そこで，ある日本語文 J_i に対して，その訳文以外の N-1 個の英語文 $E_k\ (k \neq i)$ からなる文ペア (J_i, E_k) に注目する．これらの文ペアは，含意関係の負例として利用できる可能性があるが，全てを負例として採用すると，(i) 含意関係の正例が紛れ込む可能性がある，(ii) 負例が正例の N-1 倍の数となりバランスが悪化して適切な訓練や評価が行えなくなる，という 2 つの問題がある．

これらの問題を回避するため，本研究では，以下のような手順により負例の収集を行った．最初に，ある日本語文 J_i と，よく類似しているが異なる日本語文 J_k を，同一の Wikipedia 記事中からランダムに 1 文選択し，日本語文 J_k に対する訳文 E_k と日本語文 J_i のペア (J_i, E_k) を負例候補とする．なお，文の類似度としては，tfidf によって重み付けされた cosine 類似度を用いる．次に，英語文 E_k の元となった日本語文 J_k と日本語文 J_i の文対 (J_i, J_k) を対象として，

NTCIR-RITE2 タスクにおいてベースラインモデルとして提供された日本語単言語含意関係認識器を用いて含意関係認識を行い，混入した正例を除去する．

Wikipedia 対訳コーパスから，先に述べた条件に合致しない単語列を取り除いた残り 669711 文を対象として，含意関係コーパスを自動構築した結果を表 1 に示す．表 1 より，同一ページ中の類似文を用いたため，片方向含意(F)と判定される正例が半数近く混入していることが分かる．無関係(I)と判定される負例だけでなく，矛盾(C)と判定される負例がかなり含まれている．

	F	B	C	I
訓練データ	310066	0	46634	313011

表 1 自動構築したコーパスのラベル分布（自動判定）

この自動構築したコーパスが，通常の含意関係コーパスとどの程度異なるかを検討するため，ランダムに取り出した一部の文対（1000 個）を対象として人手により含意関係ラベルを付与した結果を表 3 に，比較相手となる NTCIR 含意関係コーパスのラベル分布を表 2 に示す．2 つの表から，自動構築した含意関係コーパスは，NTCIR 含意関係コーパスに比べて，無関係(I)と判定される文対がやはり過大であり，矛盾(C) と判定される文対が不足していることが分かる．次に，自動構築したコーパスの例を図 3 に示す．片方向含意(F)と判定された文対および無関係(I)と判定された文対については，かなり自然な文対が得られているが，矛盾(C) と判定される文対については，無理がある文対が多数を占めていた．以上より，対訳コーパスから，高品質の含意関係コーパスを自動構築するには，矛盾(C) と判定される文対を獲得する方法に課題があることが分かった．

	F	B	C	I
訓練データ	207	83	65	193
テストデータ	205	70	61	212

表 2 NTCIR コーパスのラベル分布

	F	B	C	I
訓練データ	292	0	79	629

表 3 自動構築したコーパスのラベル分布（人手判定）

F	干し芋はサツマイモを蒸して乾燥させた食品である．
	適度な水分を含む為，粘度のある噛み応えとサツマイモらしい甘みが特徴的である．
C	永禄 9 年 12 月 29 日，勅許を得て徳川氏に改姓．
	二代将軍徳川秀忠の次男（嫡男）．
I	織田軍は劣勢の中，重臣・森可成と信長の実弟・織田信治を喪った．
	その事業は，大方向を示したところで重臣の一人・明智光秀の裏切りに遭い，自刃に追い込まれたことによって頓挫した．

図 3 自動構築したコーパスの例

この検討に基づいて，矛盾(C)と判定される文対の関係性をより細かく分析した．具体的には，英語単言語の大規模含意関係認識コーパスを対象とする分析を，Naïve-Bayes モデルを用いて行った．

含意関係コーパスにおけるラベルの偏りは，ラベルそのものの頻度分布に陽に現れる偏りと，その文対において使われている単語などの用語法に陰に含まれる偏りの 2 つに大きく分けることができる．陽に現れる偏りについては，オーバーサンプリングやダウンサンプリングなど各種の対処法が，機械学習手法として提案されており，部分的にではあっても対処することが可能である．それに対して，陰に含まれる偏りについては，まずその偏りの存在を見出すことが問題となる．

日本語に比べて非常に大規模な含意関係認識コーパスが利用できるため，英語を対象とする分析を行った．分析対象としては，Stanford Natural Language Inference (SNLI) コーパスと，Sentence Involving Compositional Knowledge (SICK) コーパスの 2 つを用いた．この 2 つのコーパスの諸元を，表 4 に示す．この 2 つのコーパスは，写真に対して人手でタイトルを付与した Flicker コーパスに基づいて開発された含意関係コーパスであるため，文長や語彙の分布などはかなり類似したコーパスである．ただし，SNLI コーパスは仮説文を得るためにクラウドソーシングを利用しているのに対して，SICK コーパスはラベル判定のみにクラウドソーシングを利用している点で異なる．更に，表 5 に 2 つのコーパスの含意関係ラベル分布を示す．表 5 より，SNLI コーパスは含意関係ラベルがバランス良く含まれるように設計されているコーパスであるのに対して，SICK コーパスはそのような設計は行われていないことが分かる．言い換えれば，SICK コーパスには陽に偏りが含まれているのに対して，SNLI コーパスには陽な偏りは含まれていない．

	SNLI	SICK
学習セット文対数	55K	4,500
開発セット文対数	10K	500
テストセット文対数	10K	4,927
前提文平均単語長	14.1	9.8
仮説文平均単語長	8.3	9.5
学習セット語彙数	36,427	2,178
テストセット語彙数	6,548	2,178

表 4 英語単言語含意関係コーパスの諸元

	SNLI	SICK
含意	33.4%	28.9%
中立	33.3%	56.4%
矛盾	33.4%	14.5%

表 5 英語単言語含意関係コーパスのラベル分布

陰に含まれている偏りを検出するため，次式によって定義される Naïve-Bayes モデルを用いる．

$$\hat{y} = \operatorname{argmax} P(y) \prod P(x_i|y)$$

ここで，y は含意関係ラベル，x_i は素性である．素性としては，仮説文に出現する全ての単語 unigram を用いる．この Naïve-Bayes モデルは，前提文によって与えられる文脈情報を一切用いることなく，仮設文のみから含意関係ラベルを推定している．言い換えれば，陰に含まれる偏りのみを用いて，含意関係ラベルを推定するモデルとなっている．

この Naïve-Bayes モデルを用いて，仮説文のみから含意関係ラベルの推定を行ったところ，SNLI コーパスについては精度 63.3%，SICK コーパスについては精度 56.7% を得た．また，2 つのコーパスに対して，Naïve-Bayes モデルによって推定されたラベルと正解ラベルの混同行列を表 6 および表 7 に示す．SICK コーパスを対象として推定を行っている時には，Naïve-Bayes モデルは単純に最頻出ラベル（この場合は中立）を出力しているだけである．それに対して，SNLI コーパスを対象として推定を行っている時には，Naïve-Bayes モデルは，少なくとも最頻出ラベルを出力するような単純な動作はしていない．すなわち，この 2 つの混同行列より明らかに，SNLI コーパスについては陰に偏りが含まれている．

このように，含意関係コーパスには，含意関係ラベルの分布から存在が明らかな偏りだけでなく，別種の偏りが陰に含まれている場合がある．そのため，矛盾(C)を含む文対を収集する方法も，クラウドソーシングを用いて人で作成した文を収集するなどの単純な方法では不十分であることが明らかになった．

推定ラベル	正解ラベル		
	含意	中立	矛盾
含意	2275	644	706
中立	508	1976	563
矛盾	585	599	1968

表 6 SNLI コーパスの混同行列

推定ラベル	正解ラベル		
	含意	中立	矛盾
含意	3	3	2
中立	1411	2790	718
矛盾	0	0	0

表 7 SICK コーパスの混同行列

５．主な発表論文等

〔雑誌論文〕（計２件）

Masatoshi Tsuchiya. Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment. Proceedings of the 11th International Conference on Language Resources and Evaluation, pp.1506-1511, 2018.

Masatoshi Tsuchiya, Ryo Minamiguchi. Automatic Alignment between Classroom Lecture Utterances and Slide Components. Proceedings of Interspeech2017, pp.2904-2908, 2017.

〔学会発表〕（計３件）

土屋雅稔. 含意関係認識コーパスの偏りによる性能評価への影響. 情報処理学会研究報告, Vol.2017-NL-233, No.5, 2017.

Ryo Minamiguchi, Masatoshi Tsuchiya. Developing Corpus of Lecture Utterances Aligned to Slide Components. Proceedings of The 12th Workshop on Asian Language Resources, pp.30-37, 2016.

Daiki Hayakawa, Masatoshi Tsuchiya, Hitoshi Isahara. Developing Corpus of Japanese-English Singular Sentence Textual Entailment. Proceedings of The 2016 International Conference on Advanced Informatics: Concepts, Theory and Application, 2016.

６．研究組織

(1)研究代表者

土屋雅稔（Masatoshi Tsuchiya）
豊橋技術科学大学・情報メディア基盤センター・准教授
研究者番号：70378256