様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　１６　日現在

機関番号：　３３９２４
研究種目：　若手研究(B)
研究期間：　2013～2015
課題番号：　２５７３０１２９

研究課題名（和文）生命医学文献からの高被覆な事象の抽出

研究課題名（英文）Wide-coverage event extraction from biomedical texts

研究代表者
三輪　誠（Miwa, Makoto）

豊田工業大学・工学（系）研究科（研究院）・准教授

研究者番号：００５２９６４６

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：生命医学文献からの事象抽出には一つの注釈付けされたデータ（コーパス）を利用する教師あり学習を用いたシステムが主流である．しかし，このようなシステムはコーパスに書かれた限定的な事象しか抽出できず，また，多くの事象を被覆するように注釈付けをすることも莫大なコストがかかる．これらの問題に対処するために，複数の注釈付けされたデータから一つのモデルを学習することで高被覆なモデルを作成する手法，また，注釈付けされていないテキストから注釈付けする候補を見つける手法について提案・評価を行い，複数コーパスからの学習による精度向上，注釈付けされていないテキストの利用可能性を明らかにした．

研究成果の概要（英文）：Biomedical event extraction systems often employ supervised machine learning approaches to learn from an annotated corpus. Such systems, however, can only extract limited types of events due to the limited annotated information, and it is costly to annotate a large amount of texts to cover a wide variety of event types. To deal with these problems, we propose and evaluate a method to build a wide-coverage model from several corpora and a method to find annotation candidates from unannotated texts. We show that the learning from several corpora can improve the event extraction performance, and we also present the possibility to use unannotated texts in event extraction.

研究分野：　自然言語処理

キーワード：　事象抽出　複数コーパス　半教師あり学習　教師なし学習

１．研究開始当初の背景

生命医学文献からの事象 (イベント) 構造抽出は，自然言語処理，バイオインフォマティクスの分野で注目されている．従来広く研究されてきた高精度な事象構造抽出システムは，統一的に注釈付けされた 1 つのデータ (注釈付きコーパスと呼ぶ) を教師データとした機械学習を用いたものであり，注釈付きコーパスの対象とする事象については 60% 程度の精度 (F 値) を得ることができるようになっていた．しかし，コーパスに注釈されている事象は限定的であるため，生命医学アプリケーションに用いるには十分な被覆率が得られなかった．また，単一のコーパスで多くの事象を被覆するように注釈付けすることはそのような文書の発見の難しさ，人手でタグ付する文書の量，両方の面から困難であったため，複数のコーパスに別々の事象が付与されており，統一的に利用することが難しかった．

２．研究の目的

生命医学分野における文献からの事象構造抽出において，複数の注釈付きコーパスを利用して高被覆な事象構造の抽出を行い，注釈なしテキストを利用して，生命医学アプリケーションに利用可能な事象を注釈付けするための注釈付けの候補を提示することで，高被覆な事象抽出を行うことである．

３．研究の方法

研究目的の達成に向けて，複数の注釈付きコーパスを同時に利用できる学習手法の開発と，注釈なしテキストを援用した事象候補の発見を行った．また，それに付随して起こる問題の検証，提案システムの実問題への利用・適用を行った．

(1) 複数の注釈付きコーパスを同時に利用できる学習手法については，一つのコーパスに注釈付けされている事象に関連しているように見えながらそのコーパスに注釈付けされていない表現を，そのコーパスにおいて信頼できる負例として取り出すことで，複数のコーパスから信頼できる教師データを抽出して，複数のコーパスから同時に１つのモデルを学習する手法を提案した．

(2) 注釈なしテキストを援用した事象候補の発見については，事象の手掛かりとなる単語を発見することを中心に行った．このために，その周辺に現れる単語や生命医学用語，係り受けにある用語などを用いて，文中の単語を特徴づけし，単語ベクトルやトピックモデルを用いて単語をクラスタリングする手法を用いて，単語を分類する手法を評価した．

(3) (2)に付随して，注釈なしテキストを利用して作成した情報の利用可能性を探るため，従来の新聞記事からの事象抽出において精度の向上が認められている，教師あり学習手法に教師なし学習で得られた単語の情報を素性として追加して利用する手法について調査を行った．また，この結果から新聞記事における事象抽出と生命医学文献からの事象抽出の違いについて調査を行った．

(4) 提案システムの外部評価と普及を目的として，国際共通タスク BioNLP Shared Task 2013 への参加，デモシステム・Web サービスへの登録・公開を行った．また，実問題への利用の例として，生物医学文献データベースである PubMed に登録されている論文の抄録にシステムを適用し，事象を抽出した．

４．研究成果

複数の注釈付きコーパスを同時に利用できる手法と，注釈なしテキストの援用である．このため，前者については手法そのものの提案，応用タスクでの評価，デモ，実データへの利用を行い，提案・評価については雑誌論文や学会等において発表した．後者については，注釈なしテキストの利用可能性の調査，他の新聞コーパスとの違いの解析を行い，解析結果を学会において発表した．

(1) 複数の注釈付きコーパスを同時に利用できる学習手法については，7 つの既存のコーパスを利用して評価を行った．信頼できる情報のみをそれぞれのコーパスから取り出す提案手法は，コーパスを単独で用いる手法（単一コーパス)，単純にコーパスを組み合わせる手法（単純な組み合わせ)，他のコーパスのモデルを元に１つのコーパスのモデルを作る精度向上を行うモデルの積み重ね（積み重ね)，他のコーパスのモデルを他のコーパスに適応させる領域適用手法（領域適応）と比較して，表 1に示す通り 7 つのコーパス全体で最も高い精度（F 値）を達成できた．また，提案手法は１つのコーパスに対して１つのモデルを学習する積み重ねや領域適応と異なり，7 つのコーパスに対して１つのモデルを学習する手法であり，新しいデータにも柔軟に利用可能である．さらに，モデルが１つであるにもかかわらず，他の手法に比べてコーパスの違いに影響を受けにくいこともわかった．それぞれのコーパスに共通した情報を同時に利用できることから，利用した全てのコーパスにおいて精度の向上を確認することができ，そ

のうち比較対象システムのある 3 つのコーパスについては．世界最高精度を達成した．

表 1　7 つのコーパスにおける F 値 (%)

設定	F 値 (%)
単一コーパス	50.5
単純な組合せ	50.5
積み重ね	51.3
領域適応	52.5
提案システム	53.0

(2) 新たに定義された事象について事象抽出システムの精度を競う国際共通タスク BioNLP Shared Task 2013 に参加した．複数の注釈付きコーパスを利用する学習手法を利用することで，一つのコーパスで学習したモデルよりも高精度なモデルが学習できることを新たなコーパスで改めて確認した．提案システムの結果を投稿したところ， Cancer Genetics (CG) タスクと Pathway Curation (PC) タスクにおいて，それぞれ 2 位 (6 チーム)，1 位 (2 チーム) を達成した．さらに，共通タスク後にモデルの改良を行った結果を表 2 に示す．参考までに CG において 1 位，PC において 2 位であった TEES-2.1 システムの精度 (F 値)も載せた．この結果より，国際的な基準での提案手法・提案システムの精度の高さ・位置づけを明らかにした．

表 2 BIONLP2013 SHARED TASK データにおける F 値 (%)

	CG	PC
提案システム	53.33	52.47
TEES-2.1	55.41	51.10

(3) 注釈なしテキストを援用した事象候補の発見については，トピックモデルや単語ベクトルなどの最新の単語表現学習を利用した半教師あり学習・教師なし学習を用い，単語の表現を獲得した．この結果を元に，まず，得られた表現を素性として教師あり学習に利用した．しかし，この方法では，どの表現を用いても，元の教師あり学習の精度を超えることができないことが分かった．この結果から，事象構造抽出の教師あり学習に単語の半教師あり学習・教師なし学習の情報を追加するのではなく，同時に学習する統一的なモデルを開発する必要があることが分かった．一方で，注釈なしテキストを用いた単語表現そのものについては，頻度の高い単語については，事象の候補を提示するのに利用可能な情報となりうることが分かり，未知の事象を抽出する際の手掛かりとして利用可能であることが分かった．

(4) (3)の結果をより深く解析し，注釈なしテキストを利用して作成した単語の情報の利用可能性を探るため，教師あり学習手法に素性として追加して利用した結果について調査した．この調査のため，新聞記事からの事象抽出と生命医学文献からの事象抽出の事象抽出における差異を評価し，事象の構造やタスク，評価設定の共通点・違いについて調査を行った．これは，これまで議論されてこなかった分野間の事象の違いについての初めての調査である．調査の結果として，従来の新聞記事からの事象抽出問題においては(すでに文献で報告されている通り)精度が向上する一方で，生命医学文献からの事象抽出においては精度が低下することが分かった．また，元としている複数コーパスを利用しない事象抽出システムは新聞記事コーパスにおいても表 3 に示したとおり，世界最高精度に近い精度を達成できることも示した．この表には事象抽出システムの生命医学文献コーパスにおける精度と新聞記事における精度 (F 値) を示し，また，それぞれのコーパスにおける本研究のシステムを除く世界最高精度のシステムの精度を示している．この検証における比較結果より，生命医学・新聞両分野に同じ事象抽出システムを適用することができること，また提案システムが両分野において高精度な事象抽出を行うことができることを世界で初めて示した．

表 3 生命医学文献 (GENIA) と新聞記事 (ACE2005) における F 値 (%)

	GENIA	ACE2005
本研究における 事象抽出システム	52.71	52.1
EVEX	50.97	
Li et al. (2013)		52.7

(5) Web 上にテキストを入力して投稿することで，作成した事象抽出システムのモデルを利用して，入力したテキストから事象を抽出する事象抽出システムのデモに提案したモデルを追加した．また，実際にこのシステムを利用したアプリケーションやサービスに組み込むための Web サービスとしての利用も可能とした．デモシステムについてはログインを必要としないオープンなシステムであるため，ユーザ数は不明であるが，Web サービスについては，世界中から 60 名以上のユーザ登録があり，国際的にも広く関心を引くことができていることがわかった．

(6)提案した事象抽出モデルを，分散並列環境を利用して，PubMed に含まれる約 2,000 万件の文献抄録に適用することで，21,306,332 件の事象を抽出した．従来の 1 つのコーパスを用いたシステムを用いて得られた事象は 800 万件程度であり，作成したシステムを用いることで 2 倍以上の事象が抽出できることを確認した．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　2　件）

① Makoto Miwa and Sophia Ananiadou. Adaptable, High Recall, Event Extraction System with Minimal Configuration. BMC Bioinformatics, 査読有, Vol. 16, Suppl 10, S7, 2015

DOI: 10.1186/1471-2105-16-S10-S7

② Makoto Miwa, Sampo Pyysalo, Tomoko Ohta and Sophia Ananiadou. Wide coverage biomedical event extraction using multiple partially overlapping corpora. BMC Bioinformatics, 査読有, Vol. 14, No. 1, P175, 2013.

DOI: 10.1186/1471-2105-14-175

〔学会発表〕（計　3　件）

① Makoto Miwa and Yutaka Sasaki. Modeling Joint Entity and Relation Extraction with Table Representation. The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). 2014 年 10 月 28 日. Doha (Qatar).

② Makoto Miwa, Paul Thompson, Ioannis Korkontzelos and Sophia Ananiadou. Comparable Study of Event Extraction in Newswire and Biomedical Domains. The 25th International Conference on Computational Linguistics (COLING 2014). 2014 年 8 月 29 日. Dublin (Ireland).

③ Makoto Miwa and Sophia Ananiadou. NaCTeM EventMine for BioNLP 2013 CG and PC tasks. BioNLP Shared Task 2013 Workshop. 2013 年 8 月 9 日. Sofia (Bulgaria).

〔その他〕

ホームページ等

① EventMine demonstrator
http://www.nactem.ac.uk/EventMine/demo.php

６．研究組織

(1)研究代表者

三輪　誠（MIWA MAKOTO）

豊田工業大学・工学（系）研究科（研究院）・准教授

研究者番号：00529646