複数文にまたがる関係抽出における構文情報の効果

三浦 康秀 †,‡　外池 昌嗣 †　大熊 智子 †　増市 博 †　篠原 (山田) 恵美子 ‡
荒牧 英治 ††,‡‡　大江 和彦 ‡

† 富士ゼロックス株式会社 研究技術開発本部　‡ 東京大学 医学部附属病院
†† 東京大学 知の構造化センター　‡‡ 科学技術振興機構 さきがけ
{yasuhide.miura, masatsugu.tonoike, ohkuma.tomoko, hiroshi.masuichi}@fujixerox.co.jp,
emiko-tky@umin.ac.jp, eiji.aramaki@gmail.com, kohe@hcc.h.u-tokyo.ac.jp

1　はじめに

複数の要素 (entity) 間の関係を抽出する手法は，テキストから有益な情報を得るための技術として広く研究されている．関係抽出を取り上げている評価型ワークショップの ACE[2]，BioCreative[4] では，「人名と組織名の関係の抽出」，「蛋白質間の相互作用の抽出」等のタスクが実施されている．関係抽出の中でも，1 文中に出現する 2 要素間の関係の抽出は最も基本的なタスクであり，関係の抽出に効果的な様々な素性が明らかになっている．特に，係り受け構造中の 2 要素を結ぶ最短パス [1] や 2 要素を含む部分木 [5] 等の部分構造は有効な素性となることが知られている．しかし，2 要素が別々の文に現れる関係抽出問題では，これら情報は同一文中のときのようには適用できない．

我々は電子カルテの自由記述テキストから医薬品の副作用関係を自動的に抽出する問題 [6] に取り組んでいる．医薬品の副作用は医療行為において必ず発生する現象ではなく，電子カルテのテキスト上において副作用の記述は頻出しない．数少ない副作用関係を出来る限り抽出するためには，抽出を 1 文で完結している副作用関係に限定するのは望ましくない．

本稿では，複数文にまたがる 2 要素の関係抽出問題における構文情報の効果を，特に以下の 2 点に着目して確認する．

(1) 2 要素を結ぶ最短パスは複数文にまたがる関係抽出問題でも有効かどうか

(2) 同一文中で完結している関係と複数文にまたがる関係を抽出する際に有効な構文情報に違いはあるのかどうか

これらを確認するため，隣接する 2 文にまたがる副作用関係を抽出する評価実験を行った．最短パスを 2 文の ROOT を結んだ係り受け構造より構築したところ，関係抽出の性能向上が得られ，複数文にまたがる関係においても最短パスの有効性が伺えた．また，本稿で提案する距離毎の係り受けパスや関連要素の係り受けパス用いたところ，提案素性により F 値で約 0.06 の精度向上が得られた．

図 1: ACE に則った関係アノテーションの例

2　取り扱う “関係”

要素および関係をどのように表現するかについては，図 1 に示される ACE の表現方法に則る[1]．例では，要素「人」“Card” が要素「教育機関」“the University of South Carolina” と所属関係にあることを示している．本研究では複数文にまたがる関係の抽出に着目するため，4 章で述べる実験においては副作用関係コーパス [6] を更新[2]したものを用いる．副作用関係コーパスでは，1 文書内に記述される副作用関係を同一文内に限らずアノテートされている．

副作用関係コーパスの仕様:　副作用関係コーパスは，医療に関連する要素およびそれら要素間の関係がアノテートされている．副作用関係は，図 2 の例で示され

[1] ACE のアノテーションガイドライン (http://projects.ldc.upenn.edu/ace/docs/English-Relations-Guidelines_v6. 2.pdf) より引用．

[2] 以前と比べて，アノテーションの追加およびエラーの修正が行われている．

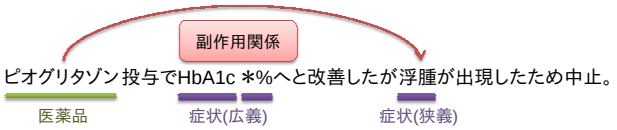

図 2: 副作用関係の例

表 1: 文距離と副作用関係の数

文距離	副作用関係数 (%)
0	893 (20.0%)
1	892 (20.0%)
2 以上	2,673 (60.0%)
計	4,458

るように，要素「医薬品」と要素「症状」間の関係として表現している[3]．

副作用関係コーパス中の要素および関係の数: 現在までに副作用関係コーパスには，4,894 の要素「医薬品」，27,215 の要素「症状」，そして 4,458 の副作用関係がアノテートされている．また，本稿では複数文にまたがる関係の抽出に着目するため，2 要素がそれぞれ属する文の間の距離と副作用関係数の関係を表 1 に示す．文距離が 0，つまり 1 文中で完結している副作用関係の割合は約 20%にとどまり，複数文にまたがる副作用関係を抽出しないことは，再現率の上限を大きく制限することになる．

3 関係の学習手法

本研究では，関係を持ちうる要素ペアをコーパスより抽出し，関係にある場合を正例，ない場合を負例として機械学習手法を用いて学習する．以下，要素 A と要素 B 間の関係を抽出すると仮定した場合の手続きを示す．

[ステップ 1: 要素ペアの抽出] 要素 A と要素 B 間の関係情報が付与されたコーパスから，同一文中および隣接する 2 文にまたがって出現する全ての要素 A と要素 B のペア $\langle a, b\rangle$ を関係候補ペアとして抽出する．ペアのラベルは，副作用関係にあるペアを正，そうでないペアを負とする．

[3]本稿の例文では，個人情報保護の観点から全ての数値表現を“＊”に置き換えている．

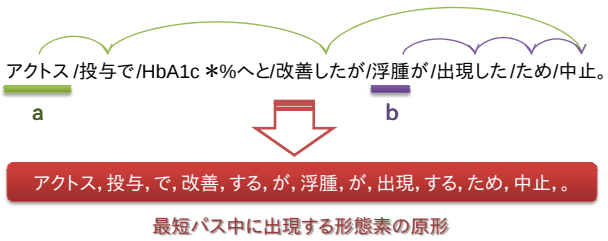

図 3: ペア間係り受け最短パス素性の例

[ステップ 2: 要素ペアからの素性の抽出] 関係ペアの素性としては，表 2 の素性を用いる．構文情報を用いた素性の例を示すと，ペア間係り受け最短パス素性では，図 3 に示すような a を含むチャンク (句，文節) と b を含むチャンクの係り先が合流するまでのパス中に含まれる形態素を用いる．副作用関係が隣接する 2 文にまたがる場合は，各文を個別に解析し ROOT を結んだ構造 (第 1 文の最後のチャンクが第 2 文の最後のチャンクに係る) の最短パスを抽出する．素性の組み合わせは BASIC, +S.PATH，+AB.PATH，+AB.PATH2 の 4 種類の設定を用いる．BASIC は構文情報を用いないベースラインとしての設定であり，+S.PATH は BASIC に最短パスを加えている．+AB.PATH は BASIC に a, b それぞれの係り受けパスを加えており，+S.PATH と類似しているが，2 要素 a, b を結び付けるパスは考慮していない．+AB.PATH2 は+AB.PATH に a, b の距離を限定した係り受けパスおよび a, b の関連要素である a', b' の係り受けパスを加えている．

[ステップ 3: 関係の学習] 正負のラベルと要素ペアの素性を用いて，要素 A と要素 B 間の関係を Support Vector Machine (SVM) で学習する．

4 実験

提案手法の性能を確認するため，副作用関係コーパスを用いて評価実験を行った．

実験に用いた医薬品・症状ペア: 副作用関係コーパスから，同一文中もしくは隣接した 2 文にまたがる 40,342 医薬品・症状ペアを抽出した．40,342 ペアのうち，1,190 ペアを含む文で係り受け解析がエラーになったので，これらのペアは評価実験から除外した．実験に用いた 39,158 ペア中の正例と負例の数は，それぞれ 1,761 と 37,397 であった．

表 2: 関係抽出に用いる要素ペア $\langle a, b \rangle$ の素性

素性	説明	BASIC	+S.PATH	+AB.PATH	+AB.PATH2
文字距離	a と b 間の文字数.	√	√	√	√
形態素距離	a と b 間の形態素数.	√	√	√	√
文距離	a および b が属する文の間の距離．同一文中であれば 0，隣接した 2 文にまたがっていれば 1.	√	√	√	√
出現順序	a の後に b が現れる場合は 1，逆であれば 0.	√	√	√	√
要素の細分類	a および b の細分類．副作用関係コーパスの例を挙げると，症状の細分類として “症状”，“変化表現”，“検査名”，“検査値” がある.	√	√	√	√
ペア間形態素	a と b の間に現れる形態素の原形.	√	√	√	√
ペア間係り受け最短パス	a と b を結び付ける係り受け関係の最短パス中に含まれる形態素の原形.		√		
係り受けパス a	a を含むチャンクから，文末までの係り受けパス中に含まれる形態素の原形.			√	√
係り受けパス b	b についての係り受けパス素性.			√	√
距離毎係り受けパス a	距離 d までの係り受けパス a に含まれる形態素の原形．$d \in \{1, 2, 4\}$ それぞれについて独立した素性として扱う.				√
距離毎係り受けパス b	b についての距離毎係り受けパス素性.				√
係り受けパス a'	a および b が属する文中の a 以外の要素 A を a' とした場合の，a' を含むチャンクから，文末までの係り受けパス中に含まれる形態素の原形.				√
係り受けパス b'	b' を a' と同様に求めた場合の係り受けパス素性.				√

表 3: 各設定での精度

設定	適合率	再現率	F 値
BASIC	0.2792	0.2699	0.2703
+S.PATH	0.3219	0.3138	0.3129
+AB.PATH	0.3455	0.3338	0.3346
+AB.PATH2	0.3763	0.3716	0.3682

実験設定　形態素に基づく素性の抽出には，形態素解析器 JUMAN version 6.0 [4] の結果を用いた．係り受け構造に基づく素性の抽出には，係り受け解析器 KNP version 3.01 [5] の結果を用いた．SVM の実装には LIBSVM version 2.89[6]を用いた．カーネル関数には polynomial を用い，各種パラメータには $degree = 3, gamma = 1, coef = 1, C = 1$ を用いた．また，SVM 出力の確度を得るため，LIBSVM の probability estimates オプションを有効にした．評価時には，5 分割交差検定で確率の閾値を各分割における正例の割合 (4–5%の間) に対応する値に設定して精度を測定した．結果として，各設定で表 3 の精度 (適合率，再現率，F 値) が得られた．

[4]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
[5]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html
[6]http://www.csie.ntu.edu.tw/~cjlin/libsvm

5　考察

複数文にまたがる関係抽出における最短パスおよび提案素性の効果を確認するために，2 つの分析を行った．

同一文と隣接 2 文個別の精度の比較　4 章の実験において，2 要素が同一文に属する場合および 2 要素が隣接 2 文にまたがる場合それぞれの F 値を表 4 に示す．どの設定においても，同一文の精度が隣接 2 文の精度を大幅に上回っている．また，隣接 2 文の精度は，BASIC より+S.PATH が F 値で約 0.02 高くなっており，僅かではあるが隣接 2 文における最短パスの効果が伺える．しかし，同一文では+S.PATH より F 値で約 0.04 高い+AB.PATH だが，隣接 2 文の精度ではほとんど BASIC と違いが見られず，同一文の精度を向上させる素性が必ずしも隣接 2 文に対して有効ではないことも伺える．+AB.PATH2 では，隣接 2 文の精度が+AB.PATH と比較して F 値で約 0.02 上昇しており，追加された距離毎の係り受けパスや関連要素の係り受けパスの素性が，同一文と隣接 2 文双方の抽出に有効であること示唆している．

同一文のみで関係を学習した場合との違い　同一文のみで学習した場合と，同一文+隣接 2 文で学習した場合の，同一文の抽出精度の比較結果を表 5 に示

表 4: 同一文と隣接 2 文の個別の精度

設定	F 値 (同一文)	F 値 (隣接 2 文)
BASIC	0.3447	0.1776
+S.PATH	0.3978	0.2037
+AB.PATH	0.4360	0.2021
+AB.PATH2	0.4686	0.2252

表 5: 同一文の抽出に対する隣接 2 文の効果

設定	同一文のみで学習	同一文＋隣接 2 文で学習
BASIC	0.3173	0.3447
+S.PATH	0.3802	0.3978
+AB.PATH	0.3753	0.4360
+AB.PATH2	0.4127	0.4686

す．+AB.PATH，+AB.PATH2 では F 値で約 0.06 の精度向上が得られており，複数文にまたがる関係の学習が同一文の関係の識別に有効であることを伺わせている．+AB.PATH と比較して，+S.PATH は同一文+隣接 2 文の精度上昇の程度が小さい．類似した設定の+S.PATH と+AB.PATH で違いが生じている理由は分からないが，+S.PATH で行っている ROOT を結ぶ処理が必ずしも正しくない可能性が考えられる．

6 まとめ

本研究では，複数文にまたがる関係抽出における構文情報の効果を確認した．評価実験により，副作用関係コーパス中の同一文中および隣接 2 文にまたがる副作用関係の抽出精度を測定したところ，距離毎の係り受けパスおよび関連要素の係り受けパスを加えた際に，F 値で約 0.3682 の精度が得られた．また，結果を分析したところ，隣接 2 文の関係においても最短パス，距離毎の係り受けパス，関連要素の係り受けパスが有効であることが伺えた．さらに，隣接 2 文の関係を学習データに加えることにより同一文の関係抽出の精度が向上していることが確認され，複数文にまたがる関係の学習が同一文の関係の抽出に有効であることが伺えた．

7 今後の課題

深い構文情報の活用　評価実験では構文情報を用いた素性を追加することにより，関係抽出の性能が向上した．しかし，現在の素性は係り受けパス中に出現する形態素を用いているのみであり，深い構文解析結果に基づく係り受けパスも考慮している Bunescu らの手法 [1] との違いは大きい．今後は，動詞項構造等のより深い構文情報の利用を検討している．

より離れた関係への対応　評価実験では同一文中および隣接 2 文にまたがる副作用関係を対象とした．これは，副作用関係コーパス中の副作用関係の約 40%に相当する．離れた文間の関係抽出を対象とした従来手法としては Hirano らの Centering Theory を用いた手法 [3] があり，さらなる再現率の向上のためにも文間の関係等を用いた手法に取り組む予定である．

参考文献

[1] R. Bunescu and R. Mooney. A shortest path dependency kernel for relation extraction. In *Proc. of HLT'05*, pp. 724–731, 2005.

[2] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. The automatic content extraction (ACE) program - tasks, data, and evaluation. In *Proc. of LREC 2004*, pp. 837–840, 2004.

[3] T. Hirano, Y. Matsuo, and G. Kikui. Detecting semantic relations between named entities in text using contextual features. In *Proc. of ACL 2007*, pp. 157–160, 2007.

[4] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, Vol. 9(Suppl 2):S4, , 2008.

[5] M. Zhang, J. Zhang, J. Su, and G. Zhou. A composite kernel to extract relations between entities with both flat and structured features. In *Proc. of ACL-44*, pp. 825–832, 2006.

[6] 三浦康秀, 荒牧英治, 大熊智子, 外池昌嗣, 杉原大悟, 増市博, 大江和彦. 電子カルテからの副作用関係の自動抽出. 言語処理学会第 16 回年次大会, pp. 78–81, 2010.