情報科学論文のための意味関係検索システム

建石 由佳　　宮尾 祐介　　相澤 彰子
国立情報学研究所
{yucca,yusuke,aizawa}@nii.ac.jp

1　はじめに

学術論文のオンライン出版化が広まるに伴って出版量が増え，読む側では自動検索が必須のツールとなり，知的な検索が望まれるようになっている．例えば「CRF（conditional random field）」を検索するときも，CRF とは何か，CRF は何に使われるか，あるいは CRF を改善するにはどうしたらよいか，など，さまざまな検索要求があり，それに沿った結果に容易に辿りつけることが望ましい．しかし，現在の検索システムの多くはキーワード（フレーズを含む）の集合を検索キーとして，実際の意図とは関係なくそのキーワードあるいはその異形を含む文献をすべて結果とするため，結果が要求に沿うものであるかどうかは検索者が判断しなければならない．

ここで，検索結果を表示する際に何らかの意味的なグループ分けができれば判断の助けになるであろう．そのためには，キーワードと文中の他の語の意味関係を知ることが重要になると考えられる．例えば，「CRF を用いた形態素解析」では CRF を利用したこと，「CRF における素性削減」では CRF そのものを改良することに関する記述であることが読み取れる．ただし，「用いる」「おける」などの語との共起のみを手掛かりにするのでは，先の例と「Confusion Network を用いた CRF」 のように CRF そのものの仕組み関する記述を区別できないので不十分である．

そこで，我々は文中の語と語を関係づけることによって検索対象のキーワードが文中でどのような意味的役割を果たしているかを検出し，それにより検索結果を文単位で分類することを試みる．この目的のために我々は情報科学論文に意味に基づく関係をアノテートしたコーパスを作成し，そのコーパスに基づく教師あり学習により論文中の文を分類するシステムを試作したので報告する．

2　関連研究

従来，関係をとらえるために，構文解析が用いられてきた．例えば，ACL Searchbench[4] では HPSG パーザーが論文のトピックの検出と，Predicate-Argument 関係に基づく結果の絞り込みに利用される．Gupta ら [2] は，論文の Focus（トピック），Domain（Focus が適用される対象），Technique（Focus を達成するための方法）を抽出するために Dependency Parsing にもとづく特定の動詞との依存関係を利用している．しかし，意味的な依存関係が直接構文上の依存関係ですべてとらえられるとは限らないため，我々は直接に意味関係を抽出する方法をとる．日本では，[1] など，論文や特許文書の中の Technology(技術), Effect（技術を使用した結果を表す Attribute と Value のペア）に対応するエンティティを同定する研究があるが，エンティティの同定にとどまり，相互の関係付けは行われていない．

意味的関係を直接抽出する研究は，生命科学分野で盛んに行われており，タンパク質相互作用等のイベント，細胞内局在関係などの関係など様々な情報をアノテートしたコーパスが作られ，情報検索システムが作成されている（例えば [3] など）．生命科学分野の大きな特徴は，オントロジー（Gene Ontology[1]など），標準語彙集（UMLS[2]など）などの言語資源の整備が進んでいることである．そのため，多くの研究では特定の現象について，それにかかわる物質や作用などをオントロジー等へのマッピングを経由して抽出することが行われている．

我々の対象である情報科学分野はそのようなオントロジー等が存在せず，また，日々新しい技術が生み出されていくという情報科学（あるいは工学一般）の特質上，標準オントロジーの作成は困難と考えられる．したがって，文中に出てくるものごとともものごとの関係を構造化するのに必要なアノテーションスキーマを実際のアノテーションを行いながら定義していくという方針を取った．

3　関係アノテーション付きコーパス

我々は情報処理学会論文誌のアブストラクトをベースとして，論文中の語と語の関係をほぼもれなく構造化したコーパスを作成した [6, 8]．検索システムの作成を最終目的とするため，抽出対象としての特定のトピックを仮定せず，論文に書かれた情報を網羅的に構

[1] http://www.geneontology.org/
[2] http://www.nlm.nih.gov/research/umls/

造化することにした．また，前述のように，実際のテキストにアノテーションを行いながらスキーマを定めていった．現行スキーマは情報処理学会論文誌アブストラクト 41 件を 2 名でタグ付けしながら定めたもので，16 種類の関係と, 3 種類のエンティティ（関係の「項」）を持つ．

3.1 エンティティ

エンティティを示す語句に対しては, TERM, OBJECT, MEASURE のいずれかのタグを付与する．前述のように情報科学分野では標準的な意味体系が存在しないので，現段階で細かい意味分類を行うことはせず,「固有名であるかどうか」「評価にかかわる表現であるかどうか」の区別のみをすることにした．また，エンティティは名詞 (句) であるとは限らない．これは，情報科学分野における概念，とくに，システム等の機能やシステムを用いて解決すべき問題などが単純な名詞句では表現されないことが多いためである．例えば,「ドキュメントを安全に取扱えるプロトタイプシステムを構築した」という文では，システムの機能は「ドキュメントを安全に取扱える（こと)」であり,「ドキュメント」の他に動詞「取扱える」，形容動詞「安全な」をとらえた上,「ドキュメント」「取り扱える」「安全な」の関係がとらえられなければならない．このような必要性から，エンティティには名詞 (句) 以外も含めることにした．

TERM は下記の OBJECT, MEASURE 以外である．前述のように，動詞等も含む．OBJECT は, システム名など実体を持つものの名前である．専門分野に限らず, 国名, 人名なども OBJECT とする．ただし, アルゴリズム, スキーマ, 言語の名前は OBJECT とせず, 中に人名などを含んでいても全体で TERM とする．MEASURE は評価, 価値判断を示す語（「良い」,「高性能」など）, 数値・数量表現のほか, 可能性（「できる」,「不可能」など）, 必要性（「したい」,「必要」）を表す語に対して付与される．

エンティティは入れ子にせず, 複合語の場合は最も長い範囲を取ることとする．ただし, 可能性・必要性を表す語は前後から分離する．

3.2 エンティティ間の関係

エンティティの種類をあらかじめ細かく分類することができないのと同様，情報科学論文に記述される関係をあらかじめ定めておくのは困難である．しかし，論文の多くは，問題を提起する，解決手法を提案する，解決手法を実現するためのシステムやデータを作成する，それら作成物を通じて解決手法を問題に応用する，そしてその結果を評価する，という記述要素を持っている．そこで我々は，まず，手法やシステムを問題に「適用する」場合に手法などと問題の間に付与する APPLY_TO 関係と，その結果を評価した際に評価対象と評価結果の間に付与する EVALUATE 関係を中心に，コンピュータシステムに付随する「入力（INPUT)」と「出力（OUTPUT)」の関係などを加えた．

また，例えば，何かを作成するというとき，作成物は作成という行動の出力と同等のものと考えられるので，作成物と「作成」の間に OUTPUT 関係をつける，などの汎化を行った．これによって，プログラムの機能などを表す動詞句内部の格要素の関係を構造化することにした．

現行スキーマにおける関係タグは

システムの振る舞いに関するもの PERFORM (動作主体), APPLY_TO (アルゴリズム, 手段, 意図した結果, 目的), RESULT (意図しない結果, 副作用, 因果関係), INPUT (入力, 材料), OUTPUT (出力, 成果物), TARGET (動作対象で, それ自身が変化しないもの), ORIGIN (起点), DESTINATION (着点)

性質に関するもの CONDITION（実験条件，状況), ATTRIBUTE (属性, 特徴), STATE (（筆者以外の）他のものに関する評価・感情)

評価に関するもの EVALUATE (（筆者による）評価結果), COMPARE (評価の際の比較対象)

その他 SUBCONCEPT (上位-下位, 全体-部分, クラス-インスタンス), EQUIVALENCE (定義, 略称, 照応), SPLIT (カッコなどで分断された語句の前後をつなぐ)

の 16 種類である．関係はエンティティ間の有向関係で, すべて二項関係あるいはその組み合わせとしてあらわす.

情報処理学会論文誌アブストラクト 30 件を用いた予備実験 [8] の結果，スキーマ策定にかかわった 2 人のアノテータの間で F-Score でエンティティでは 93%, 関係では 59.8%の一致が見られた．この実験で不一致のあった個所を整理し，アノテーションガイドラインを整備するとともに，不一致箇所を訂正したコーパスを作成し，次に述べる実験でのゴールドスタンダードとした.

3.3 アノテーション実験

このようなスキーマが一般のアノテータにとっても一般のアノテータにとっても容易に理解し作業できるものであることの検証のため，スキーマの策定経緯を知らないアノテータ 1 名と前節で述べたゴールドスタンダードとの間で一致率を測った．アノテータは，生命科学分野でのイベントアノテーションの経験を持つが，今回のスキーマ自体に関する予備知識はない．ア

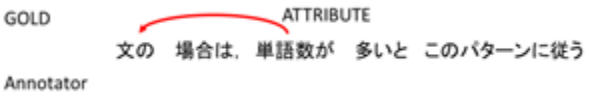

図 1: 揺れの例

ノテータにはあらかじめ，同じスキーマでアノテートした別の 100 アブストラクトとマニュアルを渡したが，他の教示は行わなかった．また，自動前処理もおこなわなかった．F-Score による一致率はエンティティで 96.4%，関係で 79.1%であった．小規模な実験結果ではあるが，この結果は，現在のスキーマが，第三者にも安定したアノテーションを可能にすることを示唆する．また，アノテータからは，判定に迷うケースで実際のアノテート例を見ることが非常に役に立ったとのフィードバックを得ており，高い一致率と合わせて，アノテーションガイドラインに実例が重要であるということを示している．

30 件中最初の 10 件についてアノテータの結果とゴールドスタンダードを比較し，エラーの傾向を調べた．エンティティの不一致は 20 か所あったがそのうちの 16 か所は「問題がある」,「課題である」などにおける「問題」「課題」などを MEASURE 扱いするかどうかに関するものであった．これらの表現はスキーマの設計時にも揺れが多くみられ，ガイドラインにいくつかの例を追加したが，まだ基準が十分クリアでないようである．関係の不一致 (64 か所）についてはさまざまな物があったが，目立ったパターンとして,「文の場合は，単語数が多いとこのパターンに従う」の「文」と「単語数」の間の関係（この場合は ATTRIBUTE 関係）のように，関係づけられるもの同士の関係が読点を挟んだり遠い場合の不一致が見られた (図 1)．この例は，英訳すると “A sentence follows this pattern when the number of words in it is large” のように代名詞が現れるため，その共参照関係（我々のスキーマでは EQUIVALENCE 関係）を介して “sentence” と “number of words” を関係づけることができるが，和文では代名詞が陽に表れないために関係があることが意識されないのだという，日本語に特有の問題であると考えられる．

実験に使った 30 件と，例示に使った 100 件の計 130 件はリクエストベースで公開している．

4 論文検索システム

作成したコーパスを使って，SVM による関係抽出器[3]を制作し，その結果に基づいて論文中の文を分類するシステムを試作した．システムはウェブブラウザ上で動作[4]し，情報処理学会論文誌アブストラクト約

[3]詳細に関する報告はは準備中
[4]現在 URL は公開していない

図 2: 検索システム：キーワードの役割による分類

3500 件からの検索が行える．図 2 と図 3 は現在のシステムのインターフェースを示している．現在，システムを使用しながら，機能，インターフェースの両面での問題点を洗い出している．

図 2 では，左上に示されているウィンドウに検索キーワードを入力すると，右側のウィンドウに検索キーワードと関連付けられた語が表示され，同時に左側のウィンドウにキーワードを含む文が表示される．キーワード（SVM）は図上では黄色であらわされている．結果はまとめて表示することも可能であるが，キーワード入力ウィンドウの下のボタンにより，関係により分類することが可能である．図 2 では左側のウィンドウに分類された結果が表示されており，関係づけられた相手の語が緑で表示されている．また，図 2 の画面で右側のウィンドウの特定の語をクリックすることにより，その語と関係づけられている場合のみを表示するような絞り込みができる（図 3)．検索キーワードには，正規表現を含めることができ，あいまい検索も可能にしている．また，検索結果の文ごとに，原文（情報処理学会論文誌アブストラクト）と自動アノテーション結果 (図 4) へのリンクを持つ．

図 3 では,「SVM」(左ウィンドウ黄色）と「構築」(左ウィンドウ緑）が関係づけられている文を，関係ごとに表示している．最初の文（「本研究の目的は SVM を用いて高い累積正解率を持った多重仮説を構築することである」）では SVM を利用して他のものを構築することが「SVM」と「構築」の間の APPLY_TO 関係により示されている．2 番目の文（「... パラメトリックな確率モデルおよび SVM を構築する」）では SVM 自体を構築することが「SVM」が「構築」（という行為）の OUTPUT であるという関係で示されている．現在の実装は，コーパス上の関係ラベルをそのまま使って関係とその方向性をもとに分類するものであるが，これが検索者にとって本当に使いやすい分類であるのか，また，使いにくい場合，どのような提示をするのがよいのか，ということが現在検討中の課題で

図 3: 検索システム：他の語との関係による分類

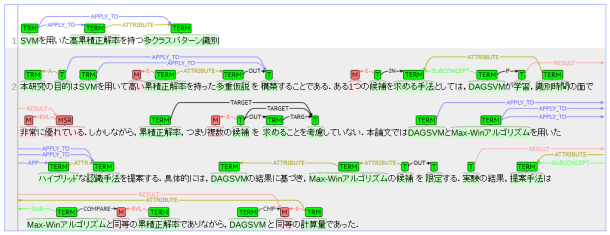

図 4: 図 3 の最初の文を含むアブストラクトのアノテーション（brat[5] による表示）

ある．

5 おわりに

情報科学論文中に意味関係をアノテートしたコーパスを作成し，それに基づく機械学習により，キーワードで示される語句を含む文を，キーワードと他の語との関係に基づいて分類し表示するシステムを試作した．

我々の枠組みは文中に示されるものと他のものや行為との関係を示すことにより，そのものが論文に示される研究の中で果たしている役割にもとづいた文の分類・検索を可能にしているといえる．ただし，我々の枠組みでは，文に記述されたことが新たに分かったことなのか，問題として指摘されていることなのか，すでに分かっていたことなのか，などの区別をしていない．これらは Zoning と呼ばれる枠組みで研究がおこなわれており（[7] など），今後これらと組み合わせるなどして，より柔軟な検索を可能にすることを目指す．

参考文献

[1] Satoshi Fukuda, Hidetsugu Nanba, and Toshiyuki Takezawa. Extraction and visualization of technical trend information from research papers and patents. In *Proceedings of the 1st International Workshop on Mining Scientific Publications*, 2012.

[2] Sonal Gupta and Christopher D Manning. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th IJCNLP*, 2011.

[3] Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. Overview of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 1–7, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[4] Ulrich Schäfer, Bernd Kiefer, Christian Spurk, Jörg Steffen, and Rui Wang. The ACL anthology searchbench. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pp. 7–13, 2011.

[5] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL*, 2012.

[6] Yuka Tateisi, Yo Shidahara, Yusuke Miyao, and Akiko Aizawa. Relation annotation for understanding research papers. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 140–148, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[7] Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1493–1502, 2009.

[8] 建石由佳, 仕田原容, 宮尾祐介, 相澤彰子. 情報科学論文からの意味関係抽出に向けたタグ付けスキーマ. 言語処理学会第 19 回年次大会講演論文集, pp. 702–705, 2013.