# 情報科学論文における<br>用語の意味クラスおよび役割のアノテーション

建石　由佳*　　宮尾　祐介*†　　相澤　彰子*‡
*国立情報学研究所 †総合研究大学院大学 ‡東京大学
{yucca, yusuke, aizawa}@nii.ac.jp

## 1　はじめに

情報科学を含む工学系の技術文書のコーパスでは，用語を役割に基づく意味クラスに分けるのが一般的である（[1],[2],[8] など）．これは，「どんな技術があり，どのような目的に使われるか」など，技術の使われ方に主眼をおいた検索要求に基づくものであり，ものの性質やふるまいを主な興味の対象とする生命科学分野のコーパスとは異なった方向性を持つものである．

役割ベースのクラス分けでは，全く同一の物に対する言明であってもクラスが文脈によって変わりうる．従って，役割ベースのクラスは，語そのものに付随するというよりも，文脈に相当する，文中あるいは文章中の他の語に対する関係であるととらえることができる．この考えのもと，用語にはクラスを設けないか非常に浅い分類にとどめ，同一文中の他の語との関係をアノテートする方法がとられること（[4], [6], [11], [12]）もある．

我々は，新しい枠組みとして，日本語情報科学論文の用語間の関係アノテーションスキーマ [11] に関係付けられるものの性質に基づいた用語の意味クラスを導入することを試みる．Information Artifact Ontology (IAO) [9] に基づく意味クラスを定義したアノテーションスキーマを作成し，それに基づいた英語アブストラクトのアノテーションを行った．また，アノテートされた文献からの情報抽出を試みた．

## 2　アノテーション

我々のアノテーションでは，ものの性質をそのものを指す用語の意味クラス，役割を用語間の関係であらわすことを試みている．用語のアノテーションには IAO に基づくクラスを使用し（ただし，クラス名は適宜簡略化した），また，いくつかのトライアルの結果，必要なことが分かったクラスを追加した．クラスと対応する語句の例を表 1 に示す．「あいまいなクラス」は，文脈上判定が難しいケースで，応用上あえて厳密な判定をしなくてもかまわないと判断したものである．

関係は [11] に基づき，表 2 の関係を用いた．

このスキーマに基づき，情報学分野の英文アブストラクトに用語と関係のアノテーションを行った．用語には冠詞を含まぬこととし，関係のトリガーとなる語もマークすることとした．図 1 にアノテーション例（brat システム [10] による表示）を示す．現在，ACL

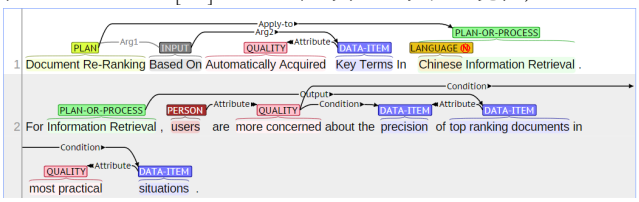


図 1: アノテーション例

Anthology より 250 件（第 4 節の実験に利用した 100 件を含む），SEMEVAL 2010 Task 5 Set より 150 件に対しアノテーションが完了している．現在のデータは `https://github.com/mynlp/ranis` にて公開しており，随時更新する．

## 3　自動抽出

この用語と関係について三輪らの認識器 [5] を用いた自動認識実験を行った．この認識器は用語と関係を同時に学習するもので，CONLL-2004 のデータセット [7] に対し関係抽出に対して (P, R, F)=(0.837,0.599,0.698) の精度を持つ．第 4 節で使用したものを除く 300 件のうち 250 件上の 10-fold cross validation の結果，用語に対しては (P,R,F)=(0.629, 0.628, 0.629)，関係に対しては (P,R,F)=(0.543, 0.452, 0.493) の精度が得られた．

| クラス | 説明/例 |
|---|---|
| IAO-Thing | |
| Ocurrent (`Occurrent`) | 時間に依存して存在するもの |
| Process (`Processual entity`) | 動作一般 |
| Time (`temporal region`) | 時間・時刻 |
| Continuant (`Continuant`) | 時間に依存せず存在するもの |
| Location (`Spatial region`) | 場所 |
| Artifact (`Object`) | 人工物で物理的実体を持つもの |
| Person (`Object`) | 人名 |
| Plan (`Directive information entity`) | 動作指令:プログラム，手順書，レシピ |
| Data-item (`Data Item, Textual entity`) | データ,（命令・指令ではない）文書 |
| IAO で定義されないクラス | |
| Quantity | 数量（単位の有無を問わない） |
| Quality | 性質 |
| Modality | モダリティ |
| Reference | 照応表現 |
| Citation | 文献参照 |
| Language | 自然言語 |
| Organization | 組織名 |
| Domain | 研究対象の分野：「NLP」 |
| Formula | 数式 |
| あいまいなクラス | |
| Plan-or-Process | コンピュータプログラム（`Plan`）ともその実行（`Process`）とも解釈できるもの： 「Web search」 |
| Intelligent-agent | 人間とも人間の行動を模倣したコンピュータプログラムとも解釈できるもの：「player (of video games)」 |
| Judging-Process | システムなどの動作であるが，筆者の価値判断を伴う記述：「outperform」 |

表 1: 用語のクラス:カッコ内は対応する IAO のクラス

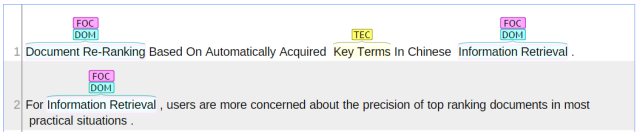


図 2: FOCUS-DOMAIN-TECHNIQUE アノテーションを brat 形式に変換したもの

## 4 情報抽出への応用

我々のアノテーションを用いて，[2] の FOCUS（論文の主題），DOMAIN（対象），TECHNIQUE（利用される技術）の抽出を試みた．彼らのコーパス (図 2) からランダムに選んだ 100 件を用い，人手でタグ付けしたセット (Gold) と第 3 のモデルを用いて自動的にタグ付けしたセット (Auto) を作成し，使用した．なお，Auto と Gold を比較すると用語について (P,R,F)=(0.680, 0.706, 0.693)，関係に関して (P,R,F)=(0.416,0.523,0.463) の精度であった．

アブストラクトを Stanford parser (version 3.4.1) [3] で Tokenize したのち各 Token について各カテゴリ (FOCUS, DOMAIN, and TECHNIQUE) に属する/属さないの 2 値分類問題として python scikit-learn 0.17 の svm.SVC (線型カーネル) を用いて解いた．素性としては以下のものの組み合わせを検討した：1) 品詞 2) (構文上の依存関係，Head か Argument か，相手の品詞) の 3 つ組 3) その Token が属する語の意味クラス 4) (意味関係の種類，Head か Argument か，相手のクラス) の 3 つ組 5) Token の位置：タイトルか，アブストラクトの最初の文か，最後の文か 1)，2) は Standford Parser の出力, 3), 4) は我々のアノテーションのエンティティの型と関係をそれぞれ使用した．例として表 3 に図 2 の 2 行目（アブストラクトの最初の文）の最初の 5 Token に対する素性を揚げる．各クラスに属する Token の数が属さない Token の数に比べ少ないので試行の結果ウェイトを 1 ： 4 とした．

結果 (10-fold cross validation での F 値) を表 4 に示す．表中 FOC, DOM, TEC はそれぞれ FOCUS, DOMAIN, TECHNIQUE であり，P, D, C, R, L は表 3 と同じである．太字はベースライン (P, D, L 素性のみ使用した場合) より向上した値を示す．直接の比較はできないがこの結果は [2] でのルールベースの結果とほぼ同等である．

クラス (C) と関係に基づく素性 (R) は精度向上に寄与しており特に TECHNIQUE においては P,D,L を利用しなくてもベースラインよりも高い精度を得ている．これは，TECHNIQUE が「ある技術を実現す

| 関係名 | 説明 | 例 |
|---|---|---|
| APPLY-TO($A$, $B$) | 手法 (A) と目的 (B) の関係 | $CRF_A$-based $tagger_B$ |
| RESULT($A$, $B$) | 前提 (A) と結論または (意図しない) 結果 (B) の関係 | $multimodal\ interface_A$ led to *3.5 fold speed* $improvement_B$ |
| AGENT($A$, $B$) | 行為 (A) と行為者 (B) の関係 | a frustrated $player_B$ of a $game_A$ |
| INPUT($A$, $B$) | プロセス (A) とその入力の関係．行為 (A) とそれに使用されるもの (B) の関係 | $corpus_B$ for $training_A$ |
| OUTPUT($A$, $B$) | プロセス (A) とその出力 (B) の関係．行為 (A) とそれにより生成されるもの (B) の関係 | an $image_A$ is $displayed_B$ |
| IN_OUT($A$, $B$) | B が同時に A の INPUT でも OUTPUT でもあるとき． | a $modified_A$ $annotation\ schema_B$ |
| TARGET($A$, $B$) | $B$is the target of an action $A$, which does not suffer alteration | to $drive_A$ a $bus_B$ |
| ORIGIN($A$, $B$) | 行為 (A) とその起点 (B) の関係 | the $projcect_B$ started in $2011_A$ |
| DESTINATION($A$, $B$) | 行為 (A) とその着点 (B) の関係 | an image $displayed_A$ on a $palm_B$ |
| ORI_DEST($A$, $B$) | B が同時に A の ORIGIN でも DESTINATION でもあるとき | $oscillate_A$ between $two\ numbers_B$ |
| CONDITION($A$, $B$) | B が A の付帯条件であるという関係 | a $survey_A$ conducted in $India_B$ |
| ATTRIBUTE($A$, $B$) | B が A の属性あるいは特性であるという関係 | $accuracy_B$ of the $tagger_A$ |
| POSESSION($A$, $B$)<br>COMPARE($A$, $B$) | あるもの (A) とその所有者 (B)<br>A は B の評価における比較対象 | $LDC_B$'s $corpora_A$<br>$F\text{-}score_A$ compared to the $baseline_C$ |
| MEMBER-COLLECTION($A$, $B$) | 集合 (A) とその要素 (B) の関係 | a $sentences_B$ in $PTB_A$ |
| COMPONENT-OBJECT($A$, $B$) | 全体 (A) と部分 (B) の関係 | a $back\ button_B$ in the $toolbar_A$ |
| EQUIVALENCE($A$, $B$) | 文中で定義された同義語（略称，訳，読みなど） | $DoS_B$ ($denial\text{-}of\text{-}service_A$) attack |
| COREFERENCE($A$, $B$) | 照応表現 (A) とその参照先 (B) の関係 | retrieve the $documents_B$ and store $them_A$ |
| IS-A($A$, $B$) | 上位下位関係 | $services_A$ such as $Google_B$ |

表 2: 関係の種類

| Token | 品詞 (P) | 構文に基づく素性 (D) | 用語の意味クラス (C) | 関係に基づく素性 (R) | 位置 (L) |
|---|---|---|---|---|---|
| For | IN | | | | first |
| Information | NNP | nn/arg/NNP | PLAN-OR-PROCESS | Output/head/DATA-ITEM | first |
| Retrieval | NNP | nn/head/NNP,<br>prep_for/arg/VBN | PLAN-OR-PROCESS | Output/head/DATA-ITEM | first |
| users | NNS | nsbjpass/arg/VBN | PERSON | Attribute/head/QUALITY | first |
| are | VBP | auxpass/arg/VBN | | | first |

表 3: 素性の例

るための手法」であり，手法を適用する目的を示す語との関わりにより決まるものであるから，我々のような意味関係が有効にとらえることができるものだといえる．C のみ利用した場合よりも R のみを利用したほうが精度がよいが，これは，情報科学分野では手法（TECHNIQUE）も手法の適用対象（DOMAIN）もプログラム，システム等意味的に近いカテゴリにあることが多く，文脈すなわち他の語との関係がないと区別が難しいことと関連しているのであろう．しかし，R 単独では C と R を組み合わせたときの精度は得られないことから，関係のみでなく，語そのものの意味クラスも必要な情報であることが示されている．

また，FOCUS では位置 (L) の寄与が大きい．これは FOCUS が「研究の主題」をあらわし，タイトルに書かれることが多いことと対応している [1]．

| | Gold | | | Auto | | |
|---|---|---|---|---|---|---|
| 素性 | FOC | DOM | TEC | FOC | DOM | TEC |
| C | 0.336 | 0.330 | **0.358** | 0.303 | 0.335 | **0.321** |
| R | 0.353 | 0.353 | **0.375** | 0.301 | 0.264 | **0.328** |
| C,R | 0.403 | **0.392** | **0.383** | 0.329 | 0.358 | **0.327** |
| L,C,R | 0.456 | **0.418** | **0.387** | 0.426 | 0.366 | **0.328** |
| P,D,C,R | 0.439 | **0.416** | **0.403** | 0.415 | 0.411 | **0.370** |
| P,D,L,C,R | **0.475** | **0.432** | **0.403** | 0.460 | **0.413** | **0.374** |
| ベースライン | | | | | | |
| P,D,L | 0.462 | 0.381 | 0.319 | | | |

表 4: FOCUS, DOMAIN, TECHNIQUE の抽出

[1] [2] でも「他のルールで FOCUS が見つからなければタイトルを FOCUS とする」というルールを採用している

## 5 おわりに

オントロジーに基づく用語の意味クラスと用語間の関係をタグ付けしたコーパスを作成し，論文からトピックを抽出する実験に利用した．実験から，用語間の関係はある技術が別の技術を実現する手法なのか，別の技術の適用対象なのかの役割を区別するのに有効であることがわかった．今後の課題としては自動抽出の精度向上，他の情報抽出への利用等があげられる．

## 謝辞

本研究の一部は情報・システム研究機構データ中心科学リサーチコモンズ事業の助成を受けている．スキーマの作成・アノテーションに協力いただいた大田朋子，Sampo Pyysalo の両氏に感謝する．

## 参考文献

[1] Peter Anick, Marc Verhagen, and James Pustejovsky. Identification of technology terms in patents. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014.

[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] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423–430, 2003.

[4] Adam Meyers, Giancarlo Lee, Angus Grieve-Smith, Yifan He, and Harriet Taber. Annotating relations in scientific articles. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014.

[5] Makoto Miwa and Yutaka Sasaki. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1858–1869, 2014.

[6] Jumana Nassour-Kassis, Michael Elhadad, and Arnon Sturm. Building conceptual maps from scientific articles. In *Israeli Seminar on Computational Linguistics*, 2015.

[7] Dan Roth and Wen-Tau Yih. A linear programming formulation for global inference in natural language tasks. In *HLT-NAACL2004 Workshop: Eight Conference on Computational Natural Language Learning*, pp. 1–8, 2007.

[8] Michael Roth and Ewan Klein. Parsing software requirements with an ontology-based semantic role labeler. In *Proceedings of the 1st Workshop on Language and Ontologies*, pp. 15–21, London, United Kingdom, April 2015.

[9] Alan Ruttenberg. Information artifact ontology. `https://bioportal.bioontology.org/ontologies/IAO`, 2014. Accessed: 2014-04-07.

[10] 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.

[11] Yuka Tateisi, Yo Shidahara, Yusuke Miyao, and Akiko Aizawa. Annotation of computer science papers for semantic relation extraction. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 1423–1429, Reykjavik, Iceland, May 2014.

[12] Behrang Zadeh and Siegfried Handschuh. Evaluation of technology term recognition with random indexing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014.