COLING2012 参加報告（その６）
– 関係獲得技術に関する論文紹介 –

山田一郎[†]

1　はじめに

COLING2012 で報告された情報抽出・知識獲得に関する研究から，属性関係，イベント—時間表現の関係，エンティティ間の関係といった関係獲得技術に関する以下の 3 つの論文を紹介する．詳細は原著論文をご参照頂きたい．

- Attribute Extraction from Conjectural Queries (Pasca 2012)
- Improved Temporal Relation Classification using Dependency Parses and Selective Crowdsourced Annotations (Ng and Kan 2012)
- Unsupervised Discovery of Relations and Discriminative Extraction Patterns (Akbik, Visengeriyeva, Herger, Hemsen, and Loser 2012)

2　論文紹介

Attribute Extraction from Conjectural Queries

概要：ある事柄が正しいか否かを問いかける conjectural search query (例えば，“is millennium stadium heated ?” など) を利用した属性抽出手法に関する論文．従来から属性抽出に関する研究は数多く取り組まれており，Wikipedia や Freebase などのリソースに明示されている属性を抽出する手法や，テキストやクエリから獲得する手法などが提案されている．(Pasca and Durme 2007) では，属性 (A) とインスタンス (I) が “of” で結ばれた “A of I” (例えば “(seating capacity)$_A$ of (millennium stadium)$_I$”)，または “What <be> A of I” というパターンをクエリから獲得して，インスタンスに対する属性を抽出している．このクエリは fact-seeking query と呼ばれ，インスタンスの属性に対する属性値を質問している．このような従来手法では，属性として獲得できる語句は名詞句に限定される．提案手法の conjectural search query は，インスタンスに対して属性が当てはまるかを確認する質問で，形容詞 (heated?) や複数の単語から成る表現 (open to the public?)，名詞句 (a retractable roof?)，主観的 (funny?)，客観的 (a true story?) など，従来手法に比べて幅広く属性を抽出できる．

[†]NHK 放送技術研究所, NHK Science & Technology Research Laboratory

手法：まず属性を抽出したクラスを決定し，このクラスに属するインスタンスを特定する．インスタンスの特定では Wikipedia のカテゴリを利用し，例えば，Chemical elements のクラスに対して，Radon, Scandium, Europium, Xenon, Oxygen などのインスタンスが特定される．次に，クエリに対してパターン”<be> I A”と”why <be> I A”とのマッチングをとり，このパターンにおいて各インスタンス (I) と共起する属性 (A) を抽出する．この際，インスタンスの曖昧性解消も行う．最後にクラスごとに抽出した属性に対し，a) 該当クラス内で多くのインスタンスとマッチし他のクラスでは少しのインスタンスとしかマッチしない，b) クエリに頻出する，という指標を利用してランキングを行う．

実験：5 億のクエリを対象として 40 のカテゴリを設定し，属性と共起するインスタンスが 5 個より少ないものは信頼できないと判断して属性候補から削除．各クラスに対して抽出された 25 個の属性をランダムサンプルして，著者自身により属性として正しさを評価．この際，vital(正解)，okay(ある程度正解)，wrong(不正解) の 3 つのラベルを用い，okay と判定した場合は適合率計算時に 0.5 の値を加算する (正解の場合は 1.0 加算)．評価の結果，全体の適合率は 0.84 と良好な値となり手法の有効性を示した．2 つのパターンを利用したが，“why <be> I A” の方がより信頼できる属性が抽出できることも確認した．抽出した属性を，手作業により名詞化して従来手法で抽出した属性と比較したところ，全体の 31%の属性は名詞化することができず，さらに 28%の属性は名詞化できたが従来手法では抽出できなかったことを確認し，提案手法は従来手法では抽出できないような属性を多く含んでいることを実証．

感想：属性とは何かを考えさせられる論文．従来の概念では「属性値」に該当する部分まで「属性」としている．例えば，Films のクラスで獲得された属性 funny などは，従来手法などでは属性 genre に対する属性値と考えられる．この funny のような属性を conjectural attribute と呼び，この属性の重要性を議論している．QA などにおける質問文では “Is the genre of Toy Story funny?” といった冗長な表現は使われず，“Is Toy Story funny?” が使われる．そのため，funny を Films のクラスの属性として扱うほうが直接的に回答を導きだせる．提案手法は単純であるが，獲得できる属性はとても有益と感じる．

Improved Temporal Relation Classification using Dependency Parses and Selective Crowdsourced Annotations

概要：文中に含まれているイベントと時間表現間の関係を分類する temporal relation classification に関する論文．temporal relation classification は TempEval-2 Task C で取り上げられているタスクで，文中の動詞 (イベント) と時間表現の関係を，Overlap, Before, After, Before-or-overlap,

Overlap-or-after, Vague のいずれかに分類することを目的とする．例えば以下の文を考える．

> Two top aides to Netanyahu, political adviser Uzi Arad and Cabinet Secretary Danny Naveh, left for Europe *on Sunday*, apparently to investigate the Syrian issue, the newspaper said.

この文には，イベントとして left，investigate，said，時間表現として on Sunday が含まれ，left は on Sunday に対して Overlap，investigate と said は on Sunday に対して After の関係を持つ．このようなイベントと時間表現の関係を統計的に推定する．

手法：TempEval-2 のタスクでは学習セットが少量 (959 インスタンス) であることが原因で，十分な精度が得られていない．提案手法では，SVM によりイベントの時間表現に対する関係を分類する．この際，a) 素性の次元を減らす，b) アノテーションデータ量を増加させるという 2 点を考慮することにより精度の向上を図る．従来手法では素性として lexical cues(単語，品詞など)，context(イベントや時間表現の属性など)，係り受け構造など，多くの特徴を組み合わせて素性を生成する研究が多かったが，提案手法では素性の次元を減らすために，「イベントと時間表現間の係り受け構造のパス」と「時間表現の依存構造」の 2 種類の特徴のみを利用する．

また，Croudflower というクラウドソーシングサービスにより学習データを増加させる．さらにデータに対する効果的なアノテーション対象を検討．各イベントに対して判定の難易度を定義し，容易に判定できるイベントに対してはアノテーション対象から除外する処理を行う．イベントに対する難易度は，イベントと時間表現間との係り受け構造のパスの長さで決定し，時間表現と直接係り受け関係にあるイベントは除外する．

実験：TempEval-2 の実験データを利用．提案した素性を用いた手法の適合率は 67.4%と，TempEval-2 のタスクにおける最高精度 (65.0%) を上回ることを示した．また，クラウドソーシングサービスにより学習データを 8,851 インスタンス生成して実験した結果，適合率が 73.2%に向上した．さらに，容易なアノテーション対象を学習データから除外した実験では，5,576 インスタンスの学習データで，適合率 73.2%の結果が得られ，37%の学習データ削減を実現．

感想：難しい事例に対して正解データを与えて学習を行うプロセスは，SVM を逐次学習などで用いる際に超平面の近くにあるデータ (判定困難なデータ) に対してアノテーションを行うと良い精度が得られやすいという従来から使われているヒューリスティックスと類似している．クラウドソーシングなどの専門家が行わないアノテーションにおいても，難しい事例を増やすことが効果的という知見は興味深い．難しい事例を増やすと誤りが増えて問題が生じそうだが，学習データ数増加の効果のほうが上回ることを示している．

Unsupervised Discovery of Relations and Discriminative Extraction Patterns

概要：教師無しによるエンティティ間の関係獲得に関する論文．従来から行われている関係獲得は，あらかじめ決められた関係に対して，そのインスタンスを学習データとして人手で与え，新たな構文パターンやエンティティペアが決められた関係であるか否かを判定する識別問題として扱われている．一方，教師無しによる関係獲得は，エンティティペアのクラスタリングを行い，各クラスタを一つの関係と解釈する．本論文のポイントは以下の 2 点．

- 文の依存構造における素性選択
- 各関係におけるパターンに対する重み付け

手法：文の依存構造における素性選択では，Stanford dependency parser により構文解析を行い，エンティティ間の係り受け関係の最短パスにある単語群を core tokens とし，さらに，core tokens と以下の関係を持つ単語群を optional tokens として抽出する．

> *nn* (noun compound modifier)，*neg* (negation modifier)，*prt* (phrasal verb particle)，*poss* (possession modifier)，*possessive* (possessive modifier)，*nsubj* (nominal subject)，*nsubjpass* (passive nominal subject)

この core tokens と，optional tokens のパワーセット (全ての可能な組み合わせ集合) を，該当するエンティティペアの素性とする．この素性によりエンティティペア間の類似度を定義し，k-means アルゴリズムによりクラスタリングを行う．

各クラスタにおけるパターンに対する重み付け処理では，エンティティ間の依存構造パスに対して以下の 2 点を考慮した distinctiveness の計算式を定義．

- クラスタに含まれる多くのエンティティペアと共起するパターンは重要
- 多くのクラスタに出現するパターンは曖昧

実験：YAGO のデータから関係を持つエンティティペアを取り出し、このペアが同一文中で出現するテキストを Web から獲得．このテキストに対して人手でアノテーションを行い，関係が explicit に記述されているもの，implicit に記述されているもの，記述されていないものに分類．関係が explicit に記述されているものだけからなる gold-standard data と，関係が implicit に記述されているものも含めた silver-standard data を作成．このデータを基準としてクラスタリング結果を評価し，依存構造の選択を行わない従来手法と比較して良好な結果であることを確認．また，各関係におけるパターンの重み付け結果を利用した関係分類実験においても良好な結果を確認．

感想：クラスタリング処理で，エンティティ間の係り受け構造を素性として使う手法に新規性は無いが，係り受け構造から抽出される素性をうまく選択することにより，従来手法と比べて素性数を減らし，精度を大幅に向上出来ている点は興味深い．

参考文献

Akbik, A., Visengeriyeva, L., Herger, P., Hemsen, H., and Loser, A. (2012). "Unsupervised Discovery of Relations and Discriminative Extraction Patterns." In *Proceedings of the 24th International Conference on Computational Linguistics*, pp. 17–32.

Ng, J.-P. and Kan, M.-Y. (2012). "Improved Temporal Relation Classification using Dependency Parses and Selective Crowdsourced Annotations." In *Proceedings of the 24th International Conference on Computational Linguistics*, pp. 2109–2124.

Pasca, M. (2012). "Attribute Extraction from Conjectural Queries." In *Proceedings of the 24th International Conference on Computational Linguistics*, pp. 2177–2190.

Pasca, M. and Durme, B. V. (2007). "What you seek is what you get: Extraction of class attributes from query logs." In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 2832– 2837.

略歴

山田一郎（正会員）：1993 年名古屋大学大学院修士課程修了．同年 NHK 入局．2008 年から 2011 年（独）情報通信研究機構出向．現在 NHK 放送技術研究所主任研究員．博士（情報科学）

（2012 年 11 月 30 日依頼）
（2013 年 1 月 21 日受付）