氏名	NGUYEN THI HONG NHUNG			
学位の種類	博士(情報科学)			
学位記番号	博情第 312 号			
学位授与年月日	平成 26 年 12 月 24 日			
論文題目	Unsupervised Relation Extraction from Biomedical Literature Using Deep Syntax（深い統語構造を利用した生命科学文献からの関係抽出）			
論文審査委員	主査	東条　敏	北陸先端科学技術大学院大学	教授
		飯田　弘之	同	教授
		Nguyen Minh Le	同	准教授
		鶴岡　雅慶	東京大学	准教授
		三輪　誠	豊田工業大学	准教授

論文の内容の要旨

The explosive growth of published biomedical research provides scientists with the chance to find correlations or associations between biomedical concepts from the literature. Therefore, there is a growing demand to convert information in free text into more structural forms. This demand motivates many researchers and scientists to work on *relation extraction*, an information extraction task that aims to extract semantic relations between important biomedical concepts. However, most of the previous studies on this topic have focused on specific or predefined types of relations, which inherently limits the types of the extracted relations. To overcome this limitation, we propose a relation extraction system that attempts to locate all possible relations present in the input documents.

In building such a general relation extraction system, we face two challenges: (1) there is no available tool that is trained on a gold standard corpus to recognize all named entities while dictionary-based tools may generate many false positives; and (2) there is no available annotated corpus for such a general schema of relations that can be used for training the extraction model. Our proposed system relies on a dictionary-based named entity recognizer and performs some post-processing to discard spurious entities. It then deals with the second challenge by employing predicate-argument structure (PAS) patterns, which are well-normalized forms that represent deep syntactic relations. In this dissertation, we introduce six PAS patterns for binary relations. After matching the patterns to extract candidates of relations, the system checks the semantic types according to a semantic network to find true relations. Our manual evaluation on a set of 500 sentences randomly selected from MEDLINE has shown a reasonable level of performance of the system (a pseudo F-score of 55.89% on average) compared with other state-of-the-art systems, including ReVerb, OLLIE and SemRep. Our system can detect broader types of relations but less precisely than SemRep, a rule-based semantic interpreter for biomedical text. The evaluation in another setting on pre-defined relations has also shown its wider coverage.

We then have applied our system to the entire MEDLINE corpus and produced more than 137 million semantic relations. The extraction results are useful in their own right, but they also provide us with a quantitative understanding of what kinds of semantic relations are actually described in MEDLINE and can be ultimately extracted by (possibly type-specific) relation extraction systems. The entire collection of the extracted relations is publicly available in machine-readable form, so that it can serve as a potential knowledge base for high-level text-mining applications in the biomedical domain.

When using the extracted relations as an underlying database, the text-mining applications would meet the problem of spurious mismatches caused by the diversity of natural language expressions. Therefore, the second task in our dissertation is to detect synonymy between relational phrases that represent the relations to alleviate the problem of mismatches. Most of the previous work that has addressed this task uses similarity metrics between relational phrases based on textual strings or dependency paths, which, for the most part, ignore the context around the relations. To overcome this shortcoming, we employ a word embedding technique to encode relational phrases. We then apply the *k*-means algorithm on top of the distributional representations to cluster the phrases. Our experimental results show that this approach outperforms state-of-the-art statistical models including Latent Dirichlet Allocation and Markov Logic Networks.

Keywords: relation extraction, open information extraction, predicate-argument structure, relational phrase clustering, word embeddings.

論文審査の結果の要旨

近年の様々な文献の電子化に伴い，文献データベースのサイズは巨大化し，したがって検索における精度を向上させることは重要な研究課題となっている．このようなテキスト・マイニングの技術は自然言語処理，機械学習，データベースなど従来的な情報技術分野を総合的に研究対象とするものであり，その精度向上のためには個々の分野に内在する問題を適切に解く必要がある．特に平テキストで貯えられたコーパスから構造化された用語間の関係を発見することはデータ管理上も有益であり，本研究の目的とするところはこのような用語間関係の高い精度による抽出である．本研究ではこの対象となる文献データベースを生物医学分野とした．同分野は昨今の研究の進展がめざましく，そのデータベースも需要が高いものである．したがって検索システムも数多く開発されており，成果の比較が行いやすい．本研究ではこの対象データベースとして MEDLINE を用いた．

本研究では，まず自然言語処理技術によって対象テキストを構文解析する．このような解析は従来的なマイニング・システムが関心を払わなかったところであるが，本研究では構文を見ることによって，その意味として述語構造 (predicate argument structure; PAS)

を取り出すことができる点に着目した．この PAS はその引数として主語・目的語に対応する二つの語を抱えるため，それらが述語によって関係づけられていると見ることができる．本システムは生物医学コーパス中に現れる専門用語関係を述語構造に捉えられた引数の関係として抽出した．まず手作業によりランダムに選んだ MEDLINE の 500 文に対しては，比較対象とする先行システムと同等（F 値において 55.89%）の精度が得られた．次にこの手法を MEDLINE のフル・コーパスに適用し，従来システムを大きく上回る 137,000,000 余の意味的関係が得られた．これらは MEDLINE 内の構造関係を見る上でも有益であると同時に，同コーパスの潜在的意味内容をも抽出したと解釈できる．

次に本システムでは得られた意味的関係の中から重複する類義語を排除するために，用語埋め込みによる検証を行った．用語のクラスター分類においては k-平均法を用いた．この結果，類義語認識は従来的な方法である，潜在的ディリクレ割り振りやマルコフ論理ネットワークを上回る精度を実現することができた．

以上，本論文は生物医学データベースの検索について，述語構造を援用したことによって精度を高めたものであり，学術的に貢献するところが大きい．よって博士(情報科学)の学位論文として十分価値あるものと認めた．