様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　４　月　１４　日現在

機関番号：　６３８０１
研究種目：　基盤研究(C)（一般）
研究期間：　2012～2015
課題番号：　２４５００１９３

研究課題名（和文）知識の再利用性向上に向けた文書の箇条書き化

研究課題名（英文）Simplifying Complicated Sentences for Information Extraction from Text Documents

研究代表者
原　一夫（HARA, KAZUO）

国立遺伝学研究所・生命情報研究センター・特任研究員

研究者番号：３０４６７６９１

交付決定額（研究期間全体）：（直接経費）　　4,100,000円

研究成果の概要（和文）：教科書などの解説記述（ストーリー）から，知識を個別に切り出し，文脈に依存しない単純文（箇条書き）に変換することによって，これまでストーリーの中に閉じ込められていた個々の知識を要素化し，再利用性を高めることを試みた．かつて冊子体の形式のみで存在した文書が電子化されたことで，近年情報の流動性が向上したが，本研究はこれをさらに進め，文書内に閉じ込められた個々の知識の流動性を高めることを目標とした．解剖学の教科書テキストを用いた実験で，単純文化する前と比較して，単純文化した後では約7倍の数の知識を切り出すことに成功した．

研究成果の概要（英文）：In discourse contexts, the frequent use of cohesive ties such as reference expressions and coordinated phrases not only troubles the function of automated systems (i.e., natural language parsers) to extract knowledge from the resulting complicated sentences, but also affects the identification of mentions of Named Entities (NEs). We propose to revamp the prose style of anatomical textbooks by transforming cohesive discourse into itemized text, which can be accomplished by annotating reference expressions and coordinating conjunctions. We demonstrate that, compared to the original text, the transformed one is easy for machines to process and hence convenient as a way of identifying mentions of NEs and their relations. Since the transformed text is human readable as well, we believe our approach provides a promising new model for language resources accessible by both human and machine, improving the computational reusability of textbooks.

研究分野：　自然言語処理

キーワード：　単純文化　脱文脈化　構文解析　意味解析

１．研究開始当初の背景

医学知識の形式表現の一つとして，臓器や身体部位などの解剖学用語の概念に，標準人体中の位置と形状を手作業で与え，ポリゴンデータ化する研究を，われわれは行ってきた（BodyParts3D: 3D structure database for anatomical concepts, N. Mitsuhashi and K. Okubo et al., Nucleic Acids Research, Vol. 37, Database-Issue, pp. 782-785, 2009）．この研究を進める中で気付いたことは，標準人体のモデル（カノニカルモデル）が持つべき特徴を作業者が漠然とイメージするのではなく，モデルの特徴要素を根拠と共にリストすることがなければ，作成したポリゴンデータは科学の道具とならないことである．

そこで，根拠の出典となる教科書記述を，半自動で固有名詞同定等を行いながら手作業で箇条書きに変換し，知識ベース化する作業を始めたところ（http://togodb.dbcls.jp/contextfreegray1），教科書記述の箇条書き化は自然言語処理分野での並列構造解析および（文内文間）照応解析が中心となるタスクであることが明らかになった．さらに，医学文書は，専門用語が数多く出現する一方で述語の種類は少ないことから，人間には難解でも計算機では文構造を識別しやすいと考えられる．このため，教科書記述の箇条書き変換は，研究代表者らが数年前から提案し改良を続けている自然言語処理の技術を用いるなどすれば，機械化も可能であると考えるに至った．

一方，生命医学分野全体を顧みると，生命医学は，膨大な数の文献（PubMed 収録数は 2000 万件を超える）の中に記述された知識と（実験から得られる）爆発的に増加する分子データを照合する，発見研究が主体である．このため，文献知識の総体を機械援助で利用するための基盤(知識ベース)作りが盛んである．しかし，その多くは知識総体の一部だけを理解している専門家が宣言的に作り上げる知識の形式（ルール，オントロジーなど）に依存する．このため，知識の根拠出典の提示や異なる形式間の比較・マージ等ができず，常に増大し修正が要求される知識総体の変化に対応しにくい．

よって，われわれは，文献内に元々一律に自然言語で記された知識を，自然言語の形のまま，機械で比較・マージ等（まとめや更新の知的作業に相当）ができるような単純文（箇条書き，すなわち，われわれが Simplified Sentence と呼ぶ形式）に変換することが，むしろ最初に挑むべき課題であると考えるに至った．

生命医学分野における知識総体の形式表現としては，医科学では，ジーンオントロジーをはじめとするオントロジーが盛んにつくられてきた（http://www.obofoundry.org/）．解剖学では，ワシントン大メジノ博士らの FMA(Foundation Model of Anatomy) や米国病理学会の SNOMED が有名である．また，酵素間の関係として，生化学知識をグラフ表現した京都大学の金久らによる KEGG も，形式的知識表現の例である．しかし，これらはいずれも専門家による宣言に基づいており，比較・マージ・更新などに課題を残している．本研究課題は，知識のソースである文書テキストを文脈非依存の箇条書きとしてバラバラに素子化することで，知識総体の比較・マージ・更新・再配列等を容易にするためのアプローチであり，専門家の理解と宣言を要しない．このようなアプローチは，われわれの知る限り，他に見られない．

２．研究の目的

生命医学分野の文書内にストーリー（あるいはディスコース）として記述された自由文（英文）を，文脈に依存しない単純文の集まりに分解すれば，ストーリーに埋め込まれているため流動性の欠けた知識に，再利用性を与えることが可能になるのではないか，というアイデアの実証を行う．かつて冊子体の形式のみで存在した文書が電子化されたことで，近年情報の流動性が向上したが，本研究課題はこれをさらに進め，文書内に閉じ込められた個々の知識の流動性を高めることを最終的な目標とする（図１参照）．

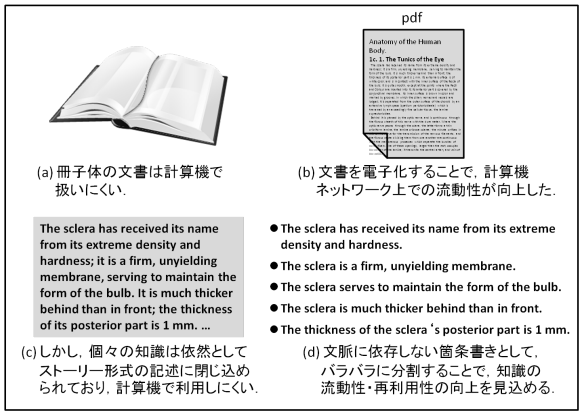

図1：本研究課題の背景(a)(b)と目的(c)(d)

３．研究の方法

生命医学分野の教科書，特に人体の構造を記した解剖学の教科書から知識を個別に切り出し，単純文（箇条書き）に変換する．単純文への変換は，まずは人手作業で行うが，自然言語処理の技術を用いた自動化も行う．ここでは，客観的事実として成り立つ知識のみを記述する解剖学の教科書を題材とすることで，モダリティ解析にひとまず立ち入ることなく，箇条書きコーパスの作成と箇条書き変換自動化手法の開発を行う．

教科書は凝縮された知の結晶であるが，読む以外の利用法が発明されていない．多くの教科書が紙から電子化され(たとえば Project Gutenberg archived in Internet Archive)，または利用自由の community books（http://en.wikipedia.org/wiki/Category:Wikipedia_books_(community_books)）として公開される中で，本研究は教科書に材料としての利用法を与え，無数の教科書を多角的に利用可能なひとつの知識源としてまとめる第一歩である．

４．研究成果

（1）テキストを単純文（Simplified Sentence 形式）に変換する手順の確立：

この手順は，ストーリーの中に圧縮表現されたテキストの解凍・展開，すなわち，著者による意味的圧縮を展開するステップ（脱文脈化）と，構文的圧縮を展開するステップ（単文化）に大別できる．前者については指示代名詞（“it”や“that”など）を対応する先行詞に置換する操作を，後者については等位接続詞（“and”やカンマなど）により並列される句を同定し文を分割する操作を，ウェブ上でのアノテーションシステムとして作成した．さらに，関係代名詞，分詞構文の単文化・脱文脈化も進めた．

特に，等位接続により並列される句の範囲をコンピュータで自動同定するのは困難なことが知られているが，その精度を向上させるために，われわれがかつて開発した並列構造解析の手法（Hara et al., ACL-IJCNLP 2009）を依存構造解析と融合することを試みた．その成果の一部を国際会議 IWPT で報告した．

（2）人手作業によるテキストの単純文変換の実施：

解剖学の知識を持つ医師が上記アノテーションシステムを使用し，Henry Gray 著“Anatomy of the Human Body”（脳解剖の章）を単純文に変換した．792 文を 1876 の単純文に変換した．

（3）単純文化した複数のドキュメントの機械による自動比較・マージの実施：

単純文の比較・マージを機械で行うためには，単純文の主語述語目的語トリプルを機械で自動同定できることが必要になると考えられる．これを上記の 1876 文を用いて確認した．具体的には，構文解析器 Enju を用い主語述語目的語トリプル（解剖学用語を含むものに限定）の自動同定を試みたところ，単純文化する前では 45 トリプルしか同定できなかったのに対し，単純文化後では 310 トリプルを同定することに成功した．この成果を，国際会議 KDIR で発表した．

（4）単語の類似度の計算方法を開発：

単純文の比較・マージを機械で行うためには，単純文の類似度の計算方法の開発が必要になる．このため，文の構成要素である単語の類似度の計算方法を開発した．開発した手法は，ベンチマークデータで既存手法よりも高精度を得ることに成功した（自然言語処理分野の国際会議 COLING での発表論文，および人工知能学会論文誌への投稿論文を参照）．

（5）ハブ現象に関する調査：

単語の類似度を計算する際に使用したデータセットを含む，大規模データセットには，一般に，他の多くのデータと高い類似度を持つデータ（ハブデータ）が生じることを，われわれは発見した．ハブデータは，データの近傍検索・分類の妨げとなることがあり，データセットの価値を低下させる．そこで，ハブ現象を人工知能（機械学習）と情報検索の観点から考察し，成果を国際会議 AAAI, SIGIR, EMNLP で発表した．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① 原 一夫，鈴木 郁美，新保 仁，松本 裕治:文法的・意味的共起を利用した単語類似度の計算，人工知能学会論文誌，Vol. 28，No. 4，pp. 379-390，2013，DOI: 10.1527/tjsai.28.379，査読有

〔学会発表〕（計６件）

① Kazuo Hara, Ikumi Suzuki, Kei Kobayashi and Kenji Fukumizu: Reducing Hubness: A Cause of Vulnerability in Recommender Systems. In Proc. the 38th Annual ACM SIGIR Conference (SIGIR), pp. 815-818, Santiago, Chile, 2015年8月11日，査読有（採択率 31%）

② Akifumi Yoshimoto, Kazuo Hara, Masashi Shimbo and Yuji Matsumoto: Coordination-aware Dependency Parsing (Preliminary Report). In Proc. the 14th International Conference on Parsing Technologies (IWPT), pp. 66-70, Bilbao, Spain, 2015年7月22日，査読有

③ Kazuo Hara, Ikumi Suzuki, Masashi Shimbo, Kei Kobayashi, Kenji Fukumizu and Miloš Radovanović: Localized Centering: Reducing Hubness in Large-Sample Data. In Proc. the 29th AAAI Conference on

Artificial Intelligence (AAAI), pp. 2645-2651, Austin, USA, 2015年1月29日, 査読有 (採択率 27%)
④ Kazuo Hara, Ikumi Suzuki, Kousaku Okubo and Isamu Muto: Annotating Cohesive Statements of Anatomical Knowledge Toward Semi-Automated Information Extraction. In Proc. the 6th International Conference on Knowledge Discovery and Information Retrieval (KDIR), pp. 342-347, Rome, Italy, 2014年10月23日, 査読有
⑤ Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, Marco Saerens, and Kenji Fukumizu: Centering Similarity Measures to Reduce Hubs. In Proc. the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 613-623, Seattle, USA, 2013年10月19日, 査読有 (採択率 26%)
⑥ Kazuo Hara, Ikumi Suzuki, Masashi Shimbo and Yuji Matsumoto: Walk-based Computation of Contextual Word Similarity. In Proc. the 24th International Conference on Computational Linguistics (COLING), pp. 1081-1096, Munbai, India, 2012年12月14日, 査読有 (採択率 30%)

６．研究組織

(1)研究代表者
原　一夫（HARA, KAZUO）
国立遺伝学研究所・生命情報研究センター・特任研究員
研究者番号：3046769

(2)連携研究者
大久保　公策（OKUBO, KOUSAKU）
国立遺伝学研究所・生命情報研究センター・教授
研究者番号：40233069