様式 C-7-2

自己評価報告書

平成２３年４月１４日現在

機関番号：１２６０１
研究種目:基盤研究(A)
研究期間：2008～2012
課題番号：２０２４０００８
研究課題名（和文）　ＷＥＢ文書の動的な読解支援システムに関する研究

研究課題名（英文）　The Studies on Language Interfaces for Browsing Documents

研究代表者
石井　久美子（田中　久美子）（ISHII KUMIKO/TANAKA KUMIKO）
東京大学・大学院情報理工学系研究科・准教授
研究者番号：１０３２３５２８

研究分野：自然言語処理, 言語インターフェース
科研費の分科・細目：情報学、メディア情報学・データベース
キーワード：ユーザインターフェース, 自然言語処理, Web 情報処理, 情報抽出, 情報検索

１．研究計画の概要

文書を閲覧する際、語学能力や知識の不足により、文書の内容を完全には理解することができなかったり、解釈判断に困ることがある。本研究の目的は、動的な文書処理を用いて、ユーザの文書閲覧上の支援を行う言語技術を構築することである。

本研究目的には、クライアントとしてのユーザインターフェースからの研究と、サーバとしての自然言語処理に基づくソフトウエアの研究の二つの側面がある。

(1)クライアントに関しては、ブラウザ閲覧時にユーザアクションに基づき、閲覧を補助する情報を提示する web mash up のための汎用インターフェースの研究を行う。

(2)サーバに関しては、汎用インターフェースと共に用いるもので、翻訳、文書の有害性判定、マーケッティングの三つの応用を提案した。

２．研究の進捗状況

(1)クライアントに関しては、初年度に共に研究した修士学生が２年目に起業し、web mash up のための汎用インターフェースは現在その会社より無料で利用できる状況にある。この意味で、予定よりも早くの研究目的のうちの一つは達成された。この点での成果は、起業をサポートする上で、論文発表は不利になると判断したため、特許のみを申請するに留め、国際特許取得の可能性を現在有している。

(2)サーバについては、研究を開始した 2008 年当初が web mash up 萌芽の時期に重なったため、当初掲げた三つのサービスのうち、マーケッティングについては、現在では各社すでにさまざまな試みがあるため中止し、当初の目的を遂行するための要素技術研究に代えた。このため、残りの二つの翻訳と文書の有害性判定の二つのサービスに関して研究を行っている。

- 翻訳に関しては、単語や熟語の動的辞書引きなど単純なサービスは商業的なものがすでにあちこちで見られる。本研究では、これとはまったく視点を変え、ユーザインターフェースと翻訳サーバを組み合わせ、スマートフォンや iPad でコミュニケーション支援を行うソフトウエアを、NICT との共同で開発している（資金的には独立）。本アイデアは必ずしもブラウザ上での文書閲覧に特化したソフトウエアではないが、そのようなものとしての応用も可能である。平成２３年の２月には、英日翻訳のプロトタイプが国際会議ＩＵＩで Best Paper Award を得た。
- 有害性判定については、当初は猥褻や暴力といった文書部分の除去を考えていたが、この点は昨今は商用のウィルスソフトウエアやファイヤーウォール、専用ハードウエアとして実現されている。一方、大学などアカデミアでは、別の有害性として、剽窃や捏造が問題となっている。そこで、剽窃に関し、ユーザが論文など閲覧する際に、剽窃が疑われる部分を動的に指摘するサービスを研究している。このように「有害性」の内実は変化したが、必要となる技術や研究方向は当初と目指す技術は、本質的には変わらない。現在、国内大会レベルの成果が上がっており、今後はこれを発展させる。

(3)動的な文書閲覧に必要な要素技術研究を二つ遂行している。

- 言語汎用の構文解析を動的に行う要素技術の研究を翻訳や語学学習の動的な閲覧支援を行うために行っている。提案方法では、文のうち、解析しやすいところから、徐々に解析するため、途中結果をユーザインターフェースから動的に利用することができ、また日本語に限らずどのような言語の構文でも解析可能である。従来手法と網羅的に比較を行っておりその高性能化を探究している。
- 文書分割処理は、一文書内に混じっている異なる種の文書部分を分割する処理である。現在の web 上の文書には、一つの文書に、異なる言語部分や、異なる著者が書いた文書が、入り乱れたものとなっており、これを原因として閲覧が困難となったり、剽窃などの問題につながる。このため、本研究の目的を遂行するには、文書の「切り分け」が重要である。昨年始めに言語での分割研究に着手したが、有害性も同様の要素技術を要することに着目し、現在はより汎用な視点をとり、与えられた文書を異種の文書部分に分割することを目指している。

３．現在までの達成度

②総じて、おおむね順調に進展している。
計画以上に進展している部分、計画どおりの部分、計画どおりにはいかなかった部分の三つに分かれる。また、成果として本来得られるはずであった論文も、起業などの関係から特許に代えざるをえなかった部分などある。

４．今後の研究の推進方策

(1)については、学生が起業した会社は黒字化して順調であり、目的は十分に達成された。
(2)のサーバに関しては、翻訳については、現在提案中のものは、国際的には Best Paper Award と評価が高いこともあり、クライアントとサーバ間の通信による遅延を改善したり、多言語化を行う計画である。さらに、国際会議論文２編、雑誌論文１編を目指している。文書の有害性判定については、残りの期間では研究が終了しない可能性もあるが、最終的には学内利用の剽窃検知システムを目指す。本研究の範囲内では、要素技術について目処を立て、サーバとしてのプロトタイプ構築を目指す。
(3)の要素技術研究については、言語汎用の動的な構文解析に関しては、昨年度の成果を受け、現在、提案手法の性能の比較を行っている。本研究の範囲内で２編程度の学術論文を執筆する。また、文書分割手法については、言語による分割では圧縮、剽窃に関しては極地統計量と手がかりがあり、残りの期間で目処をつけたい。学術論文２編程度が成果として考えられる（うち１編は、(2)の有害性と重なる)。

５． 代表的な研究成果

〔雑誌論文〕（計 3 件）

① Tanaka-Ishii, Kumiko and Jin, Zhihui. From Phoneme to Morpheme ---Another verification using Corpus in English and Chinese---, Studia Linguistica, volume 62, number 2, pages 224-248, 2008.

〔学会発表〕（計 11 件）

① Song, Wei and Finch, Andrew and Tanaka-Ishii, Kumiko and Eiichiro, Sumita. picoTrans: An Icon-driven User Interface for Machine Translation on Mobile Devices. Proceedings of the Intelligent User Interface, pages 23--32, 2011 年 2 月 14 日, 査読あり, Palo Alto, U.S. **Best Paper Award**.
② 溝江将, 田中久美子. 単独剽窃検知--所与の一文書だけから剽窃箇所を推定する--. 言語処理学会第 17 回年次大会, 2011 年 3 月 19 日.豊橋.
③ Kitagawa, Kotaro and Tanaka-Ishii, Kumiko. Tree-Based Deterministic Dependency Parsing ---An Application to Nivre's Method. Proceedings of the Association for Computational Linguistics, pages 189--193, 2010 年 7 月 13 日. 査読あり.Uppsala, Sweden.
④ Yogatama, Dani and Tanaka-Ishii, Kumiko. Semi-supervised Multilingual Spectral Clustering Using Document Similarity Propagation, International Conference on Empirical Methods for Natural Language Processing, pages 871--879. 2009 年 8 月 7 日.査読あり. Singapore.

〔産業財産権〕

○出願状況（計 3 件）
名称：確認システム、情報提供システム、ならびに、プログラムを記録したコンピュータ読取可能な情報記録媒体 / Check System, Information Providing System, and Computer-readable Information Recording Medium Containing a Program
発明者：石井久美子 / ISHII Kumiko
権利者：東京大学
種類：特許
番号:PCT/JP2008/061729, WO 2009/001926 A1
出願年月日：2008.06.27
国内外の別：PCT 出願
○同内容の特許を日本と米国でも別に出願