科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月　 ５ 日現在

機関番号：１２６０１
研究種目：基盤研究(C)
研究期間：2010～2013
課題番号：２２５３０５３１

研究課題名（和文）スペイン語圏の社会的事件の通時データベース作成と政策決定への応用

研究課題名（英文）Developing chronological databases of social events in the Hispanic regions for policy applications

研究代表者
和田　毅（Wada, Takeshi）

東京大学・総合文化研究科・准教授

研究者番号：２０５３４３８２

交付決定額（研究期間全体）：（直接経費）　　3,200,000円、（間接経費）　　960,000円

研究成果の概要（和文）：スペイン語圏の通信社が配信するニュース記事を、配信と同時にリアルタイムで収集するシステムと、収集した記事を自然言語処理技術を用いて自動的にコード化するプログラムを開発した。実用的な政治・経済・社会的事件の通時データベースを作成するという長期的目標の達成には至っていないが、学会発表と論文発表を積極的に行った。特筆すべきは、そのインパクトが、言語情報学、言語学、中南米地域研究、政治学、社会学など多くの学術分野に及んでいることである。アメリカ、チリ、スペイン、イタリアなどへ赴き国際学会で頻繁に発表を行い、また、研究成果発表の大部分を英語とスペイン語で行っているため、国際的な認知度は非常に高い。

研究成果の概要（英文）：We have developed a software program which collects, on a real-time basis, news articles published by major news agencies in Latin America. By employing natural language processing techniques, we have also created a software program which assigns syntactical and semantic tags to the texts automatically. Our ultimate goal is to generate a chronological database of political, social, and economic events worldwide. While our software programs have not produced such a database of practical use yet, we have already made a major impact through our presentations and publications in a variety of academic fields such as linguistics, language and information sciences, Latin American studies, political science, and sociology. Our project's international visibility is quite high because our journal articles and presentations are mostly in English or Spanish and we have participated rather aggressively in international conferences in countries such as the United States, Chile, Spain, and Italy.

研究分野：社会科学

科研費の分科・細目：社会学

キーワード：データベース　社会運動　中南米地域研究

１．研究開始当初の背景

（１）国際関係論や政治学では、多数の犠牲者を生みかねない暴動や抗争を未然に防ぐために、その発生を予測するシステムの開発努力が続けられてきた。しかし、既存のシステムの多くは、民族・宗教的多様性、社会経済格差、経済発展レベル、政治体制の特徴など、「静的で構造的な変数」を用いているため、暴動が勃発するタイミングを予測できないという欠点があった。本研究は、政治社会学や社会運動論の分野で近年有力になってきた「動的な政治過程」に着目するアプローチを予測システムに取り入れることで、この理論的限界を克服していく。具体的には、暴動や内乱が勃発する過程に頻繁に生じる「イベント」（政治・経済・社会的事件）を特定することによって、暴動や紛争の勃発を予期する精度を高めようというのである。

（２）動的アプローチを用いて将来を予測するためには、現在起きているイベントの情報をリアルタイムで入手し、即座に分析する必要がある。既存のシステムは、手作業で情報収集を行うため分析までに時間がかかりすぎていた。この問題を解決するため、ロイター通信が配信する記事を自動コード化するシステムが考案されてきた。しかし、このシステムには、①英語の記事しか扱えない、②ロイター通信の記事しか扱えない、③イベントの起きた場所やイベントの主な争点をうまく抽出できないなどの限界があり、実用的なデータだとはいえなかった。有用なイベント・データベースを構築することは急務であった。

２．研究の目的

（１）本研究の目的は、スペイン語圏の通信社が配信するニュース記事を、配信と同時にリアルタイムで自動コード化するプログラムを開発し、政治・経済・社会的事件の通時データベースを作成することである。将来、社会学、言語学、政治学、国際関係論、マス・メディア論など様々な学術分野の研究に利用できるようなデータベースの構築を目指した。

（２）さらに、このデータベースを活用して、世界各地の暴動、民族浄化、集団虐殺、内戦などの「政治的暴力」の発生を予測する。政治社会学者と応用言語学者の共同研究により、紛争予防のための政策決定に貢献する画期的なシステムを開発する。

３．研究の方法

研究方法は、①自動記事収集システム開発、②自動コード化ソフトウエア開発、③イベント・データ解析の３つの作業から成る。

（１）自動記事収集システムは、インターネット上に公開されたスペイン語圏各国の主要な通信社が配信する記事を自動的に収集して「記事データベース」に保存するシステムである。記事の形式は通信社によって異なるため、<headline> <leadParagraph>等の共通のタグを付けて保存し、以後の作業にて同一手順で扱えるようにする。

（２）自動コード化ソフトウエアとは、（１）で作成した記事データベースに収集された配信記事を自動解析して、「イベント・データベース」を作るためのソフトウエアである。様々な自然言語処理(Natural Language Processing: NLP)技術を適用し、記事テキストから「①いつ(日時)、②どこで(場所)、③だれが(アクター)、④だれに対して(ターゲット)、⑤なぜ(問題)、⑥なにをした(行動)」という６つのイベント要素を抽出し、データ化する。これにより、どのアクターとターゲットが対立もしくは協力していたか、どのアクターがどのような行動戦略を用いていたか、どのような問題がどのターゲットに向けられていたか等、政治・社会集団間の関係を示す情報を数値化することが可能になる。人と人、集団と集団の関係を研究する社会科学にとって、このデータの有用性が飛躍的に高まる。

（３）このイベント・データベースを解析して、政治的暴力予測モデルを作成する。過去に多大な人的被害を引き起こした紛争・暴動に関する先行研究をもとに、どのイベントが、どういう順序で起きると、暴動を引き起こす確率が上がるかを調べる。予測モデルの精度は、自動コード化ソフトウエアの精度にも依存するため、今回実用的なモデルを構築するのは困難であるが、将来の実用的モデルの基礎となるものを今回の研究期間内に作成する。

４．研究成果

（１）自動記事収集システムの開発。スペイン語圏の国々の主要な通信社の記事を配信と同時に自動的に収集して記事データベースに保存するシステム（ラッパー、wrapper）を作成した。収集作業を常時自動的に実施する必要があるため、サーバーを契約した。FeedGator ソフトウエアの機能を利用して、RSS 形式で配信される記事を収集することに成功した。しかし、通信社によっては頻繁にエラーを生じるものもあり、その修正や調整にかなりの労力を費やさなければならないことも判明した。さらに、RSS 形式では記事を配信していない通信社の記事を収集することはできないなどの問題もあった。これらの限界を鑑みて、今後はオンライン記事の購入を検討していくこととした。

（２）記事データベースから必要な情報を抽出し、自動コード化ソフトウエアへ入力するシステムの構築。記事データベースに保存さ

れた配信記事には、各種の HTML タグや不要な情報が含まれている。不要な情報を取り除き、文ごとに区切りを入れ、単語やその他の文字列（トークン、token）を自動的に抽出するシステムを作成した。この結果をテキストファイル形式で保存し、次の自動コード化ソフトウエアの入力データとする。

（3）自動コード化ソフトウエア（自然言語処理パイプライン）の開発。自動コード化ソフトウエア開発には、information extraction など様々な自然言語処理技術を応用した。スペイン語言語情報処理学の専門家の上田とルイズ・ティノコ、そして、自然言語処理学を専門とする博士課程院生が中心となって、Linux 上で開発を進めた。具体的には、主に以下の４つの作業を実施し、これらを「自然言語処理パイプライン」として連結することで、自動コード化を図った。
①入力テキストデータに、スペイン語の品詞タグ（名詞、動詞、形容詞など）を自動的に付与する品詞タグ付ソフト（Spanish Part-Of-Speech Tagger）を作った。
②品詞タグが付与されたテキストに、スペイン語の活用形や男性形・女性形などの形態素解析を行い、タグを付与する形態素タグ付きソフト（Spanish Morphological Tagger）を作成した。
③上記のタグ付きテキストをインプットとして、スペイン語構文解析（Spanish Syntactic Parsing）を行うソフトを作成した。これにより、主語、述語、目的語など、文章の中で各単語が果たす文法的役割を特定し、文法タグを付与できることになった。
④上記の文法タグが付与されたテキストをインプットとして、さらに述語との意味関係を分類してタグを付与する意味役割ラベリング（Semantic Role Labeling）ソフトを作成した。イベントの６つの要素の中で中心的な意味を持つ「行動（なにをした）」に関する情報が述語によって表現されている文章であれば、この意味役割タグを活用することで、イベントの他の要素に関する情報も特定できる可能性を開いた。

（4）自動コード化ソフトウエア開発の今後の展望。現段階の自動コード化ソフトウエアは、「行動（なにをした）」に関する情報が述語によって表現されている文章であれば、それ以外の要素を特定できる可能性が高い。しかし、必ずしも述語によって「行動」が表現されるとは限らない。たとえば、「労働者のストライキが勃発した」という文の場合、述語は「勃発した」であるが、行動を示しているのは「ストライキ」という名詞である。名詞で行動が表現されている場合、様々なパターンでアクターやターゲットが示されるため、このパターンを解析しなければならない。現在機械学習（Machine Learning）法を導入して、この開発を進めているが、コード化の精度を高めるためには、行動・アクター・ターゲットを正しく特定した「正解例」を用意して、コンピュータ・プログラムに学習させる必要がある。正解例を大量に作成するために、人海戦術を用いる必要があり、その作業を行うためのインターフェイスは作成済である。今後は、大学院生を雇用して機械学習作業を進めたいと考えている。

（5）イベント・データの解析手法の開発。自動コード化ソフトウエアによるイベント・データベースが実用段階には至っていないため、既存のイベント・データを利用して、解析方法を模索し、いくつかのモデルを作成した。例えば、アメリカ社会学会で発表した論文（学会発表の⑥）は、18-19 世紀イギリスのイベント・データを用いて、暴力的な行動を予測しようとしたものである。具体的には、アソシエーション・ルール分析（Association Rule Analysis）を用いて、暴力的な行動の前に頻繁に行われている行動パターンを発見することで、予測に役立てようと試みた。

（6）国内外における位置づけとインパクト。「日々のニュースを正確にコード化したイベント・データベースを提供する」「暴動のリスクが高い場所を的確に予測し紛争の予防に貢献する」といった究極の目標を達成する形で国内外に貢献する段階にはまだ至っていないが、この研究プロジェクトの過程で多大なインパクトを内外に及ぼすことができた。とくに、自動コード化ソフトウエア開発過程で習得した自然言語処理をはじめとする技術を応用して、学会発表と論文発表を積極的に行った。特筆すべきは、そのインパクトが、ひとつの学術分野にとどまらず、言語処理学、言語学、ラテン・アメリカ地域研究、政治学、社会学など多くの分野に及んでいることである。そして、研究成果発表の大部分を英語とスペイン語で行い、積極的にアメリカ、チリ、スペイン、イタリアなどへ赴き国際学会で発表を行っているため、国際的な認知度（Visibility）は非常に高いといえる。

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　９件）

① Wada, Takeshi. 2014. “Who are the active and central actors in the ‘rising civil society’ in Mexico?” *Social Movement Studies*. 13:127-157. 査読有.
② Wada, Takeshi. 2013. “Legacies and extensions of Charles Tilly: A relational and dynamic event analysis of contentious repertoires in Great

Britain.” *The International Journal of Conflict & Reconciliation*. 1(2):1-40. 査読有.

③ 和田　毅. 2013. 「ここ 20 年間の社会学による中南米地域研究への貢献度の潜在意味解析」『ラテン・アメリカ論集』 ラテン・アメリカ政経学会 47:1-23. 査読無.

④ Fisher, Andrew R. M., Guðmundur Oddsson, and Takeshi Wada. 2013. “Policing Class and Race in Urban America.” *International Journal of Sociology and Social Policy*. 33(5/6):309-327.査読有.

⑤ Ruiz Tinoco, Antonio. 2013. “Twitter como corpus para estudios de geolingüística del español.” *Sophia Lingüística* LX:147-163.査読無.

⑥ Ueda, Hiroto. 2013. “Una nota sobre el método de taxonomía cuantitativa de grandes datos: Coeficientes de asociación aplicados a los variantes del Diccionario de americanismos.” *Dialectologia*. IV:221-235.査読無.

⑦ 上田　博人. 2013. 「広域スペイン語語彙バリエーション研究における新しい数量化の試み－日本語計量言語地理学の方法に学ぶ－」『日本語・日本学研究』3:59-90. 査読無.

⑧ Wada, Takeshi. 2012. “Modularity and transferability of repertoires of contention.” *Social Problems*. 59(4):544-571. 査読有.

⑨ Ruiz Tinoco, Antonio. 2011. “Variación léxica y gramática del español peninsular e hispanoamericano.” *The Korean Journal of Hispanic Studies*. 3:29-53.査読無.

〔学会発表〕（計　17 件）

① Wada, Takeshi. “Rigidity and flexibility of repertoires of contention.” The 2013 La Società Italiana di Scienza Politica (SISP) Conference. University of Florence, Firenze, Italy. 2013 年 9 月 14 日.

② Ueda, Hiroto. “Analizador lingüístico común con parámetros de gramática, diccionario y cadenas de aplicación.” V Congreso Internacional de Lingüística de Corpus. Univerisaid de Alicante, Spain. 2013 年 3 月 15 日.

③ Ruiz Tinoco, Antonio. “Variación de la anteposición de más con adverbios de negación.” V International Conference on Corpus Linguistics. Universidad de Alicante, Spain. 2013 年 3 月 14－16 日.

④ Ruiz Tinoco, Antonio. “Variación sintáctica en Twitter en el español fronterizo de Estados Unidos.” 24th Conference on Spanish in the United States and 9th Conference on Spanish in Contact with Other Languages. The University of Texas Pan American, USA. 2013 年 3 月 6－9 日.

⑤ 和田毅.『ここ 20 年間の社会学による中南米地域への研究と展望』 第 49 回ラテン・アメリカ政経学会全国大会. 東洋大学白山第 2 キャンパス. 2012 年 11 月 10 日.

⑥ Wada, Takeshi. “A Study of Charles Tilly’s Data on Contentious Gatherings in Great Britain (BRIT).” The American Sociological Association annual meetings. Las Vegas, USA. 2011 年 8 月 21 日.

⑦ Ueda, Hiroto. “Nuevo método para la recogida de datos de variación léxica. Encuestas en web en el proyecto de Varilex.” XVI Congreso de la Asociación de Lingüística y Filología de América Latina. Alcala de Henares, Spain. 2011 年 6 月 6－9 日.

⑧ Ueda, Hiroto. “Base de datos de conflictos sociales recogidos en la prensa hispana.” Congreso Internacional sobre Lengua e Inmigración. Universidad de Alcalá, Spain. 2010 年 3 月 24 日.

⑧ Ueda, Hiroto. “La lengua española en la proyección mundial de Internet.” Congreso Internacional de Lengua Española. Valparaíso, Chile. 2010 年 2 月 28 日.

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
和田　毅（WADA, Takeshi）
東京大学・大学院総合文化研究科・准教授

研究者番号：20534382

(2)研究分担者

(3)連携研究者
上田　博人（UEDA, Hiroto）
東京大学・大学院総合文化研究科・教授
研究者番号： 20114796

ルイズ・ティノコ　アントニオ
（RUIZ TINOCO, Antonio）
上智大学・外国語学部・教授
研究者番号：80296889