坂口智洋 （黒橋禎夫教授）
「Anchoring Events to the Time Axis toward Storyline Construction」
（ストーリーライン生成のための時間と事象情報の対応付け）
平成 31 年 3 月 25 日授与

Web 上に情報が蓄積されて 20 年あまりが経った。人々の価値観や社会の状況は刻々と変わっており、現在では当たり前となっていることでも、その当時大きなインパクトがあったことは数多くある。逆に、当時は大して注目を集めなかったが現在になって有名になったものもある。モノや出来事を多角的に理解する上で、それらが現在までにどのような変遷を辿ってきたのかを知ることは重要である。今日、Web は最新の情報を知ることができるだけでなく、過去の出来事や考えをその当時の記事を通して知ることができる情報空間となっており、情報を統合・比較・要約して読者に提示する情報編纂技術は今後ますます重要になると考えられる。

あるトピックに関連する一連の出来事を物語として捉え、複数のテキストから登場人物、場所、事象、そしてそれらの間の関係性などの情報を抽出して時間軸上で表現したものはストーリーラインと呼ばれる。これまでに様々なストーリーライン表現方法が提案されてきたが、その最も核となる構造は、事象を時間軸上で表現するタイムライン構造である。これまでテキストの時間情報解析では、主に事象表現に関連する相対的な時間関係に着目してきた。具体的には、事象表現間、事象表現と時間表現間の時間的前後関係や包含関係などである。このような相対的な時間表現方法は、小説など時間設定が不明瞭なテキストにおいても用いることができ汎用性が高い。一方で、新聞記事など事実に基づくテキストでは、事象表現を直接時間軸に対応付ける方がより正確で効率的な表現が可能である。本論文では、事象表現を時間軸に対応付けるための基盤技術とデータ構築方法を提案している。

本論文では、まず、時間表現をテキストから検出し正規化する手法を提案している。時間表現の正規化とは、時間表現のもつ時間情報を定められた形式に変換することである。例えば図1では、「昨日」を2017-12-05に正規化している。時間表現の語順はしばしば入れ替わり様々な表現が存在しているが、既往研究では時間表現のパターンを予め用意して正規化を行っていたため、多様な表現に十分対処できないという問題があった。本論文では時間表現の構成性に着目し、時間表現に含まれる基本的な語彙ルールを用意し、これを組み合わせることで多様な語順や並列構造に対処している。

次に、事象表現を時間軸に対応付ける時間情報コーパスの構築について述べている。タイムラインの学習や評価を行うためには、事象表現を時間軸に結びつけたタグ付けデータが必要である。本論文で提案しているタグ付け基準は従来研究と比較して2つの特徴をもつ。1つは幅広い表現をタグ付け対象としていることである。従来研究では一時性の強い表現を対象としてタグ付けを行っていたが、本論文では一時性の弱い表現の時間情報も重視し、これらをタグ付け対象に含めている。もう1つは、従来では扱わなかった、頻度や期間などの多様な時間情報を扱うためのタグを新たに導入したことである。このタグ付け基準を用いて日本語新聞にタグ付けを行い、113 文書 4,534 表現からなる時間情報コーパスを構築した。

最後に、事象表現を時間情報に対応付けることで、複数のテキストからあるトピックに関連するタイムラインを生成している（図 1）。提案手法は従来研究と比べて幅広い文脈を考慮するもので、2段階の機械学習からなる。第1段階では局所的な情報に基づいて事象表現を時間軸に対応付け、第2段階では大域的な情報を用いてこれを修正する。実験の結果、既往研究よりも高精度なタイムラインを生成できること、第2段階の処理が有効であることが示されている。

今後は、事象間の関係性や事象に対する意見・評価など、事象表現に関連する多様な情報を時間軸と対応付けられるように研究を発展させたい。

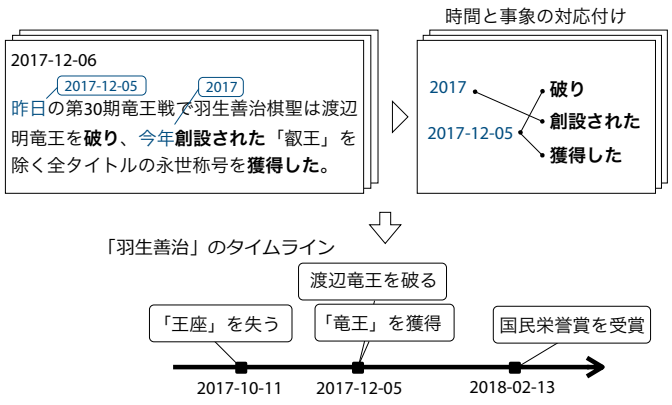

図 1 タイムライン生成システム