特別講義Ⅱ 非構造化データの活用に向けて レポート
提出日：11/14
学籍番号：223337
名前：田川幸汰

＜非構造化データ＞
非構造化データとは明確なデータ構造や特定のスキーマ（データの定義やフォーマット）に従わないデータを指す。非構造データにはテキストや表（Table）がある。非構造化データの一例として議会会議録と有価証券報告書の活用に向けた取り組みを説明する。
＜地方議会会議録のデータ活用＞
地方議会会議録は各自治体で異なるフォーマットで管理され、発言日、会議名、発言者、発言内容がまとめられているが、発言者を一意に識別するには課題がある。そこで、「ぎ〜みる」という都道府県議会会議録検索システムでは、議員IDを追加することで名前の省略、芸名の使用、かな表記や旧字体による表記揺れに類する名揺れ問題を解決した。「Otaru Open City」では、非構造化データの連携として、字幕とWordCloudを活用した動画検索、法律名の抽出と外部知識源との接続、特徴語の抽出と関連発言の検索が行われている。自然言語処理のもっともよい手法を明らかにするために、SQuADというWikipediaの文章を基にした膨大な質問と、その答えからなるデータセットを用いて、NTCIR-16というShared taskを開催している。これは質問と答弁の対応づけ、政党ごとの賛否推定、予算項目と議論の連結といったタスクを含んでいる。
＜有価証券報告書のデータ活用＞
有価証券報告書は、金融商品取引所に上場している企業が、投資家保護や公正な取引を目的として事業年ごとに公開する報告書である。この報告書には、XBRLというタクソノミとインスタンスを用いて関連する要素をひも付ける財務報告用の標準化されたXMLベースの言語が使用され、インラインXBRL（ix）タグとして含まれている。特に、TOPIC100企業の報告書には平均220の表が含まれており、これらのタグによって情報を効率的に利用できる。有報に含まれる表やテキストを対象として、構造化情報を抽出する技術を開発するために、UFO TaskというShared taskを開催している。これは、有価証券報告書の表から項目と値の抽出(TDE)、有価証券報告書に含まれる表の項目名や数値と関連するテキストの連結(TTRE)といったタスクを含んでいる。表形式データに基づいて質問に答えるタスク(TableQA)では、異なる表現の質問に対しても同じ内容であれば一貫した回答が得られるロバスト性が求められる。
<興味深い内容>
地方議会会議録のデータ活用は、選挙の際に重要な役割を果たす可能性があると感じた。選挙で投票する際、多くの人が候補者の公約を基に判断するが、実際に公約を実行しない候補者も多く、信頼性に欠ける場合がある。会議録データを活用することで、候補者が公約内容と一致する発言を行っているかを確認でき、その内容がリンクしていれば、候補者の信頼性を高めることができると考えられる。自身の研究では、自己位置推定の精度向上を目指しており、ユーザーのフィードバックをテキストデータとして取り入れることが有効であると考えている。例えば、「左手に●●が見える」「正面に●●が見える」といった情報を活用することで、位置推定の範囲をその周辺に絞り込み、より正確な位置を推定できる可能性がある。