先端科学技術研究科　修士論文要旨

所属研究室 (主指導教員)	自然言語処理学 (渡辺　太郎 (教授))		
学籍番号	1911294	提出日	令和 3年 1月 25日
学生氏名	山口　泰弘		
論文題目	End-to-end Information Extraction with Imbalanced Data Learning Technique ラベルの不均衡を考慮したEnd-to-End情報抽出モデルの学習		

要旨

自然言語で書かれた文書から構造化された情報を構築する技術を情報抽出と呼ぶ.
情報抽出は, 固有表現認識, 共参照解決, 関係抽出といった複数のタスクから構成され, 各タスクにおいてそれぞれ研究が進められてきた.
情報抽出システムを構築する際は, これらのタスクをつなぎ合わせたパイプライン処理として抽出を行う.
しかし, パイプライン処理においては上流のタスクでの誤りが下流のタスクまで伝播し, システム全体の性能を低下させる問題が生じることが知られている.

そこで, 最近では情報抽出における複数のタスクを同時にひとつのモデルで処理するEnd-to-End情報抽出の手法が研究されている.
固有表現認識, 共参照解決, 関係抽出のタスクは, スパンの選択とスパン・スパンペアの分類の問題に帰着することができ, これまで提案されているEnd-to-End情報抽出モデルでは, 文中の全ての可能なスパンを列挙した後, それらのスパンについて分類を行うという手法が採用されている.

しかし, 全ての可能なスパンを列挙した場合, 抽出の対象となるスパンの数より多くのスパンを評価する必要があり, 分類問題を解く上ではラベルの不均衡が問題となる.
機械学習モデルにおいては, 学習データのラベルの不均衡はモデルの性能を低下させることが知られていて, End-to-End情報抽出モデルにおいてもこの不均衡が予測性能の低下を引き起こしていると考えられる.

本研究では, End-to-End情報抽出モデルにおける学習データの不均衡を解決するために, Under Sampling, Over Sampling, Hard Example Samplingの3つのサンプリング手法による学習を提案し, 予測性能の比較を行った.
実験の結果, Hard Example Samplnigにおいてベースラインモデルと比較して最も大きな改善が見られた.