ISS-SP-337

CRF による学術論文からの実験情報抽出の一手法

平井　久貴† 新妻　弘崇†† 太田　学††
† 岡山大学工学部情報工学科　†† 岡山大学大学院自然科学研究科

1. はじめに

共通の研究目的をもつ関連研究の性能を比較するために，実験環境や実験対象データ，評価指標などの実験情報に注目することは多い．これらの実験情報は，その成果を報告する学術論文に記載されているが，記載方法の自由度が高く，論文中から効率良く探すのは手間がかかる．そこで本研究では論文から実験情報を自動抽出する手法を提案する．

2. 実験情報

本研究では，図表，図表キャプション，本文中の段落，の論文構成要素を論文からの抽出対象とする．また，データセット，評価指標，実験結果を実験情報の属性とする．そしてこれらの属性をもつ論文構成要素を実験情報と定義し，論文中から抽出する．

3. 提案手法

本研究では，論文 PDF から変換した XML より取り出した素性を実験情報抽出に利用する．論文 XML には複数のレイアウトタグがあるが，本研究では，通常複数行からなる文のかたまりを表す BLOCK，図を表す IMAGE に CRF++[1]を利用してラベルを付与する．利用する素性は BLOCK の幅といったレイアウト素性や，特徴的な単語の有無などの言語的素性である．特徴的な単語とは，例えば“Table”，“Figure”，“dataset”，“metric”，“result”といった実験情報を示唆するものである．これらの素性を用いて，それぞれ属性に分けた図，表，図キャプション，表キャプション，段落のいずれかの実験情報ラベルを，BLOCK または IMAGE に付与する．

4. 評価実験

実験データにはNTCIR-9[2]の論文合計 91 件を利用する．実験では 3 分割交差検定を行い，再現率，適合率，F 値を算出した(表 1)．ここで，再現率は，抽出すべき実験情報のうち正しく抽出できたものの割合を示し，適合率は抽出した実験情報のうち正しく抽出できた割合を表している．F 値は再現率と適合率の調和平均である．いずれも 1 に近いほど良い．また，表中に数値の記載されていない箇所は，データ中に存在しなかった実験情報である．

表 1 より，平均では再現率 0.409，適合率 0.330，F 値 0.334 となることが分かる．また，表キャプション(データセット)は殆ど抽出できていない．これは，表キャプション(データセット)が，実験データ中にあまり見られない実験情報であったことが一因と考えられる．全体的に適合率が低いので，言語的素性についてさらに検討したい．

表 1 実験情報抽出結果

	再現率	適合率	F 値
図(データセット)	–	–	–
図(評価指標)	–	–	–
図(実験結果)	0.670	0.372	0.473
表(データセット)	0.055	0.039	0.044
表(評価指標)	–	–	–
表(実験結果)	0.844	0.410	0.541
図キャプション(データセット)	–	–	–
図キャプション(評価指標)	–	–	–
図キャプション(実験結果)	0.457	0.581	0.473
表キャプション(データセット)	0.223	0.089	0.110
表キャプション(評価指標)	–	–	–
表キャプション(実験結果)	0.816	0.578	0.694
段落(データセット)	0.092	0.205	0.122
段落(評価指標)	0.332	0.350	0.319
段落(実験結果)	0.199	0.303	0.232
平均	0.409	0.330	0.334

5. まとめ

本研究では，CRFを用いて実験情報を自動抽出する手法を提案し，91 件の論文データからの抽出実験を行い，提案手法の性能を評価した．実験情報抽出実験の結果の F 値は 0.334 だった．

参考文献

[1]CRF++, http://crfpp.sourceforge.net/
[2]NTCIR-9, http://research.nii.ac.jp/ntcir/index-ja.html