

論文 QA のための画像処理～グラフを読む～

磯崎 秀樹 中野 仁登 浅川 護 荒木 良元

岡山県立大学

1 はじめに

我々は、論文を対象とした質問応答システムの研究を行っており、これを「論文 QA」[1]と呼ぶ。現在、ACL Anthology(<http://aclweb.org/anthology/>) の論文 3 万弱を利用し、自然言語処理に関する質問に答えられることを目指している。

例えば以下のような質問に答えられることを目指して、各モジュールの整備を行なっている。日本語の質問を英語のクエリに変換するのは、日本語の論文などから自動生成した専門用語翻訳辞書 [2] を用いている。

1. 線形計画法を初めて使った論文は？

Lucene をベースにした検索エンジンで、論文の出版年が一番古いものを出力している。

2. C-value の定義は？

深層学習を用いた数式画像抽出ツール [3] によって得られた数式画像を回答候補として、距離によるスコアを与えている。

3. 英語品詞タグづけの最高精度は？

数値を表す文字列を候補として抽出し、近傍の単語など素性とした機械学習によって、精度を表す数値でないものを除去している [4]。

論文の中には、グラフが多数あり、これらが重要な意味を持っていることが多い。

そこで本稿では、論文中のグラフを全自動で数値化する方法について説明する。グラフを数値化する手法としては、人間が手伝う半自動的な手法と全自動で読み取る手法が考えられる。前者は簡単に実装できるが、論文 QA では何万もの論文を扱うので、全自動でやりたい。グラフを読み取る研究は視覚障害者への情報提示手法として研究が行われている [5]。

いずれも、OpenCV(<http://opencv.org>) で画像処理を行い、Tesseract-OCR(<https://github.com/tesseract-ocr>) で文字を読むアプローチを採用している。なお、グラフを読む研究は、chart recognition[6] あるいは chart image recognition[7] と呼ばれることがある

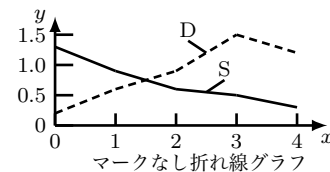
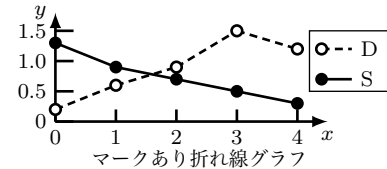


図1 折れ線グラフのマークあり/なし

2 手法

2.1 折れ線グラフ

折れ線グラフの中には、●▲■★などのマークがグラフの各頂点に描かれているもの（マークあり）と、描かれていないもの（マークなし）がある（図1）。

マークありのグラフについては、伊藤 [8] が、グラフの凡例部分からマークを切り出して、それをテンプレートとしてテンプレートマッチングを行う手法を実装した。しかし、マークなしの場合は、この手法を利用できない。

また、マークがある場合にも、周りにある線や文字などの影響により、テンプレートがマッチングしないことがある。その場合にも、線をたどることで、読み取れる可能性がある。

そこで、マークのない折れ線グラフを読む手法を実装した。当初、OpenCV の HoughLines で線分を検出する手法を試したが、グラフの一部しか取れないので、scipy.misc.imrotate で画像を実際に回転させて線分の向きごとのヒストグラムを作ることで線分を検出する方法に変更した。

これにより、実線は読み取れるようになったが、破線が読み取れないので、布田ら [9] の手法を参考に破線をつなぐと、読み取れるようになった。

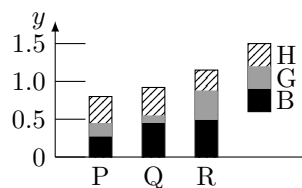


図2 積み上げ棒グラフ

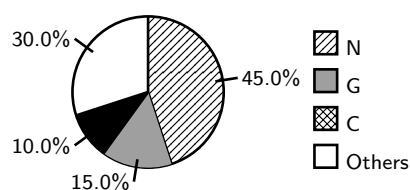


図3 円グラフの読み取り

2.2 棒グラフ

色で塗り分けられた棒グラフを認識するのは難しいが、中には、図2の上のような「積み上げ棒グラフ」もあり、処理が複雑になる。詳細は人工知能学会全国大会で発表予定であり、割愛する。

2.3 円グラフ

円グラフは大小比較が難しいとされ、科学技術論文では使うべきでないとされているが、円グラフを含む論文もある。例えば、ACL Anthology の W11-4106, W14-3105, P08-3012 などが円グラフを含んでいる。

円グラフの読み取りは [7] などで議論されている。

我々は、OpenCV の Hough 変換によって円グラフの中心の座標 (x, y) と半径 r を求めた。しかし、円グラフの周りの文字などの影響により、本来の値とは若干ずれた値になることがある。そこで、 (x, y, r) の3次元空間で、得られた値の周りの近くを探索して、実際に円グラフの円周とぴったり重なる値を求めた。

これで得られた中心点から延びる実線を求めて、扇形と扇形の境界線を求めた。ただし、この手法では、黒く塗りつぶされた領域が全て境界線として得られるので、これらをまとめる処理が必要である。

次に、凡例から塗りつぶしパターンを求めてテンプレート・マッチングを行うことで、対応する領域を求めた。

3 おわりに

論文に関する質問応答システム「論文 QA」を作成している。本稿では、論文に含まれるグラフを自動で認識し、数値化するプログラムについて説明した。今後はデータを増やして実験を行いたい。本研究は JSPS 科研費 26330366 の助成を受けたものです。

参考文献

- [1] 磯崎秀樹, 伊藤圭汰, 荒木良元: 論文 QA のための画像処理 表を読む, 言語処理学会年次大会, pp. 139-142 (2015).
- [2] 中嶋達弥: 論文 QA のための専門用語辞書の自動生成, 岡山県立大学修士論文 (2016).
- [3] 伊達信太郎, 磯崎秀樹: 深層学習と OCR を併用した論文画像からの数式領域の検出, 電子情報通信学会 NLC (2015).
- [4] 菊川貴志: 論文 QA のための最良スコアの回答候補抽出, 岡山県立大学修士論文 (2016).
- [5] Greenbacker, C. F., Wu, P., Carberry, S., McCoy, K. F., Elzer, S. and Demir, and Daniel Chester Sand eniz D. D. M.: Improving the Accessibility of Line Graphs in Multimodal Documents, in *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pp. 52-62 (2011).
- [6] Liu, Y., Lu, X., Qin, Y., Tang, Z. and Xu, J.: Review of Chart Recognition in Document Images, in *Proc. of SPIE-IS&T Electronic Imaging* (2013).
- [7] Weihua Huang, C. L. T. a. K. L.: Model-based Chart Image Recognition, in *Graphics Recognition. Recent Advances and Perspectives, LNCS-3088* (2003).
- [8] 伊藤圭汰: 質問応答のためのグラフ読み取り, 岡山県立大学卒業論文 (2015).
- [9] 布田寿康, 大町真一郎, 阿曾弘具: 連結成分追跡による文書画像中の折れ線グラフの認識, 電子情報通信学会論文誌 D-II, Vol. J86-D-II, No. 6, pp. 825-835 (2003).