5ZB-6

機械学習による論文アブストラクトのセグメンテーション

広畑堅治　　岡崎直観　　石塚満

1　はじめに

実証的研究の論文には,「背景・目的」→「関連研究」→「提案手法」→「結果」→「結論」に代表されるような，章立ての定型パターンが存在する．同様に，実証的研究を報告するアブストラクトにも，定型的で予測可能な構造が存在する [1]．論文アブストラクトの文に対し,「提案手法」「結論」などのラベルを割り当て，アブストラクトの談話構造を認識できれば，論文の主張点と思われる箇所を推定したり，論文間の関係を推定することに役立つ．

これまで，Support Vector Machine (SVM) による機械学習アプローチで，論文アブストラクトの談話構造を推定する研究が行われてきた [3, 2]．論文アブストラクトには定型の流れ（例えば「結果」の後に「結論」が来るなど）があるので，本研究では，系列ラベリング手法である条件付確率場（Conditional Random Fields, CRF）を適用し，論文アブストラクトの談話構造を行う．

2　提案手法

医学系の論文データベースである MEDLINE には，約 1,600 万件の論文情報が登録されており，そのうちの約 800 万件のレコードには，論文の概要（アブストラクト）が登録されている．収録されている論文アブストラクトの約 9 %（140 万件）には，論文著者の手によって文にラベル付けが行われている．図 1 に，ラベル付けされたアブストラクトの例を示す．本研究では，あらかじめラベル付けされている 1 万件の論文を学習データとした．

分類クラス　アブストラクトには，多様なラベルが論文の著者により用いられている．その中でも，FINDINGS と RESULTS のように，似たような働きを持つラベルが存在する．そこで，ラベルの出現頻度上位 100 位以内のものを抽出し，その中で似たようなラベルをグループ化し，以下の４クラスに集約させた．

[**OBJECTIVE**] : To determine the reliability of clinical assessments of nutritional status in surgical patients . [**METHOD**] : Prospective observer assessment . [**RESULTS**] : In this study , clinical estimates of nutritional well-being have been compared with objective markers of nutritional status. Clinicians tended to under-diagnose the extent of nutritional depletion

図 1: MEDLINE のアブストラクトの例

- INTRODUCTION (導入, 目的及び論文の背景)
- METHOD (方法)
- RESULTS (結果)
- CONCLUSION (結論)

ただし，以上の４クラスいずれにも属さないと判断されるラベルを含むアブストラクトは，学習・評価コーパスから除外した．

本研究は，機械学習アプローチでアブストラクト中の文を分類するので，学習コーパスから素性を取り出す必要がある．今回，学習に用いた素性は，以下の通りである．

- 文に含まれる n-gram (unigram のみ, bigram のみ, または unigram と bigram の混合)
- 文に含まれる数値表現の数
- アブストラクト内における文の相対位置（先頭から末尾までを５分割）
- 前後 w 文の素性 ($w = \{1, 2, 3\}$)

分類器	sentence(%)	abstract(%)
SVM: mixed 200000	80.1	16.5
CRF: mixed 200000	81.0	24.9

表 1: n-gram のみを組成にした時の分類結果

n-gram　アブストラクト中の文に対して，GENIA タガー（英語の品詞タガー）[1] を適用し，単語と品詞の列を取り出す．単語列の中から *a, the* などのストップワードを取り除き，n-gram を生成する．全ての n-gram を学習に用いると，素性の数が膨大となるので，それぞれの分類クラスにおいて偏って出現する n-gram を，χ^2 検定で選出する．素性の数 m を変化させたときに，分類性能がどのように遷移するか調べる．

評価方法　ラベル付けされている MEDLINE アブストラクトから，ランダムに 10,000 件のアブストラクト（計 110,465 文）を抽出し，学習コーパスとした．また，残りの MEDLINE アブストラクトから，1,000 アブストラクト（計 10,861 文）を評価コーパスとした．タスクの正解の基準として，アブストラクト中の文単位で正解をカウントする場合と，アブストラクト単位で正解をカウント（アブストラクトに含まれる文全てが正しく分類された場合に正解をカウント）の 2 通りを用意した．

3　実験

学習器として，SVM$^{\text{light}}$[2] と FlexCRFs[3] を用いた．表 1 は，unigram と bigram の混合 n-gram を用いて素性を生成し，SVM と CRF の分類性能を比較するものである．紙面の都合で省略するが，n-gram は unigram と bigram を混合させ，χ^2 検定上位 200000 件を素性に用いたときが，最もよい性能を示した．この結果から分かるように，文単位での正解率は大差がないが，アブストラクト単位での正解率は，CRF の方が高く，論文アブストラクトの談話構造を CRF が捉えられていることを示唆している．

これまでの素性に加えて，前後 w 文の内容を素性に入れることにより，性能が向上するかどうか調べた．

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/
[2] http://svmlight.joachims.org/
[3] http://flexcrfs.sourceforge.net

w	sentence(%)	abstract(%)
$w = 0$	80.1	24.9
$w = 1$	88.9	46.9

表 2: 前後の文の利用

素性	sentence(%)	abstract(%)
n-gram のみ	88.9	46.9
n-gram+(1)	88.9	46.4
n-gram+(2)	90.3	49.8
n-gram+(1)+(2)	**90.7**	**50.5**

表 3: 素性の比較

表 2 は，前後の文に含まれる素性を使うことの効果を示している．

前後 $w = 1$ 文の素性を加え，さらに (1) 文内に含まれる数値表現の数，(2) アブストラクト内における文の相対位置という素性を新たに加え，性能の変化を調べた（表 3）．位置情報が分類性能の向上によく貢献していることが分かり，数値表現の数を素性に加えることにより，さらなる性能向上が見られた．

参考文献

[1] Elizabeth DuRoss Liddy. The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing and Management*, Vol. 27, No. 1, pp. 55–81, 1991.

[2] Masashi Shimbo, Takahiro Yamasaki, and Yuji Matsumoto. Using sectioning information for text retrieval: a case study with the medline abstracts. *In Proc. Second International Workshop on Active Mining (AM'03)*, pp. 32–41, 2003.

[3] Yasunori Yamamoto and Toshihisa Takagi. A sentence classification system for multi biomedical literature summarization. *International Conference on Data Engineering Workshops*, Vol. 0, p. 1163, 2005.