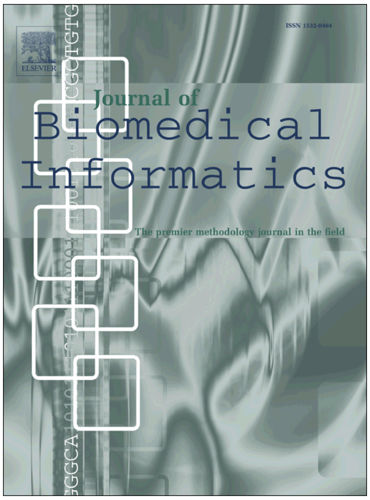

# A Part-Of-Speech Term Weighting Scheme for Biomedical Information Retrieval

Yanshan Wang, Stephen Wu, Dingcheng Li, Saeed Mehrabi, Hongfang Liu

PII: S1532-0464(16)30112-5
DOI: http://dx.doi.org/10.1016/j.jbi.2016.08.026
Reference: YJBIN 2634

To appear in: *Journal of Biomedical Informatics*

Received Date: 27 April 2016
Revised Date: 30 August 2016
Accepted Date: 31 August 2016

Please cite this article as: Wang, Y., Wu, S., Li, D., Mehrabi, S., Liu, H., A Part-Of-Speech Term Weighting Scheme for Biomedical Information Retrieval, *Journal of Biomedical Informatics* (2016), doi: http://dx.doi.org/10.1016/j.jbi.2016.08.026

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

00 (2016) 1–16

# A Part-Of-Speech Term Weighting Scheme for Biomedical Information Retrieval

Yanshan Wang[a,*], Stephen Wu[b], Dingcheng Li[a], Saeed Mehrabi[a], Hongfang Liu[a,*]

[a]*Department of Health Sciences Research*
*Mayo Clinic*
*Rochester, Minnesota, USA*
[b]*Department of Medical Informatics & Clinical Epidemiology*
*Oregon Health and Science University*
*Portland, Oregon, USA*

## Abstract

In the era of digitalization, information retrieval (IR), which retrieves and ranks documents from large collections according to users' search queries, has been popularly applied in the biomedical domain. Building patient cohorts using electronic health records (EHRs) or searching literature for topics of interest are some IR use cases. Meanwhile, natural language processing (NLP), such as tokenization or Part-of-Speech (POS) tagging, has been developed for processing clinical documents or biomedical literature. We hypothesize that NLP can be incorporated into IR to strengthen the conventional IR models. In this study, we propose two NLP-empowered IR models, POS-BoW and POS-MRF, which incorporate automatic POS-based term weighting schemes into bag-of-word (BoW) and Markov Random Field (MRF) IR models, respectively. In the proposed models, the POS-based term weights are iteratively calculated by utilizing a cyclic coordinate method where golden section line search algorithm is applied along each coordinate to optimize the objective function defined by mean average precision (MAP). In the empirical experiments, we used the data sets from the Medical Records track in Text REtrieval Conference (TREC) 2011 and 2012 and the Genomics track in TREC 2004. The evaluation on TREC 2011 and 2012 Medical Records tracks shows that, for the POS-BoW models, the mean improvement rates for IR evaluation metrics, MAP, bpref, and P@10, are 10.88%, 4.54%, and 3.82%, compared to the BoW models; and for the POS-MRF models, these rates are 13.59%, 8.20%, and 8.78%, compared to the MRF models. Additionally, we experimentally verify that the proposed weighting approach is superior to the simple heuristic and frequency based weighting approaches, and validate our POS category selection. Using the optimal weights calculated in this experiment, we tested the proposed models on the TREC 2004 Genomics track and obtained average of 8.63% and 10.04% improvement rates for POS-BoW and POS-MRF, respectively. These significant improvements verify the effectiveness of leveraging POS tagging for biomedical IR tasks.

*Keywords:* biomedical information retrieval, natural language processing, part-of-speech, bag-of-word, markov random field

## 1. Introduction

The widespread adoption of Electronic Health Records (EHRs) has enabled the secondary use of EHRs for healthcare analytics or clinical decision making [1][2]. Information retrieval (IR), which retrieves and ranks documents

*Corresponding author
*Email addresses:* wang.yanshan@mayo.edu (Yanshan Wang ), wst@ohsu.edu (Stephen Wu), li.dingcheng@mayo.edu (Dingcheng Li), mehrabi.saeed@mayo.edu (Saeed Mehrabi), liu.hongfang@mayo.edu (Hongfang Liu )