Automatic Summarization for Terminology Recommendation: the case of the NCBO Ontology Recommender

Pablo López-García[1,2], Stefan Schulz[1], and Roman Kern[2]

[1] Medizinische Universität Graz - Institut für Medizinische Informatik, Statistik und Dokumentation. Auenbruggerplatz 2, 8036 Graz (Austria)
[2] Know-Center GmbH. Inffeldgasse 13/6, 8010 Graz (Austria)

Abstract. The National Center for Biomedical Ontology (NCBO) ontology recommender helps users choose a biomedical terminology by analyzing a submitted document. Submitting a single document might not be representative and result in poor recommendations, while submitting a large sample might be expensive, sometimes unfeasible. In this paper, we investigate the effectiveness of two well-researched automatic summarization techniques as an alternative: topic modeling using Latent Dirichlet Allocation and keyword extraction using TextRank. In our case study, both techniques proved to be extremely valuable, dramatically boosting performance without significantly affecting terminology recommendations ($r = 0.83$–0.98).

Keywords: biomedical terminology, automatic summarization, ontologies, TextRank, topic modeling

1 Introduction

Selecting one or more domain terminologies that are best suited for a given application has proved to be a hard task [14]. Especially in the biomedical field, terminology systems (vocabularies, classification, nomenclatures, ontologies) highly vary in scope, size, architecture, granularity, and purpose [4]. For instance, SNOMED CT provides controlled terms for virtually every aspect of health care [2], while others are highly specialized, such as the Foundational Model of Anatomy (FMA), the National Cancer Institute (NCI) Thesaurus, the Gene Ontology, or the Medical Subject Headings (MeSH).

To help users choose a suitable terminology in text annotation applications, the National Center for Biomedical Ontology (NCBO) [10] released the NCBO ontology recommender web service [7]. After analyzing the structure and terms of a document submitted by a user and the candidate terminologies in Bioportal, the recommender suggests a list of terminologies. It is expected that the terminology ranked first is the most appropriate for annotating that particular document and others with a similar context. Bioportal is an open repository of biomedical terminologies from the NCBO that currently reports nearly 6 million

biomedical terms distributed in 370 terminologies, most of which are based on an ontological foundation [11].

Unfortunately, techniques that are widespread in the field of recommender systems, such as collaborative filtering [8], can rarely be applied when recommending biomedical terminologies. On the one hand, the content of biomedical terminologies is much harder to model than the content of books, movies, or songs—the prototypical target items of recommender systems. On the other hand, user feedback in biomedical terminology is scarce. In Bioportal, for example, users can rate the usability, coverage, quality, formality, correctness, and documentation of terminologies, but in most cases the number of ratings is negligible (only one for SNOMED CT[1]).

There are several limitations, however, when using a document submitted by a user as context for making recommendations. Firstly, the submitted document might not accurately represent the context of the user's document collection, misleading the recommender. Secondly, getting recommendations is expensive: our experience shows that a single recommendation can take over 30 seconds when using a full clinical document as input. Thirdly, even with a substantially improved performance of the system, an intensive use with numerous submissions of full texts to the recommender web service might result in degraded performance. Therefore, it would be desirable to minimize both the number and size of submitted documents, while maintaining their informational value.

Summarizing the context of a collection before submitting it to a recommender has proved to be a useful technique to improve efficiency without substantially influencing recommendations [1]. On the one hand, Hariri et al. showed that topic modeling a collection using Latent Dirichlet Allocation (LDA) [3] was useful for building a query-driven recommender for song recommendations [5]. Topic modeling finds clusters of related keywords in documents that usually make sense to humans, e.g., "paracetamol", "aspirin", and "ibuprofen" identified as a cluster in a collection of medical records would be generally associated with the topic of analgesics. Once the a collection has been topic modeled, each document is represented as a weighted mixture of topics, from more to less prevalent. On the other hand, keyword extraction using TextRank [9], a graph-based ranking model technique, provides an efficient and concise way of summarizing a document that might be used for the same purpose.

1.1 Objectives

The main objective of this paper is to study the effectiveness of (a) topic modeling using Latent Dirichlet Allocation and (b) keyword extraction using TextRank as summarization strategies in a context-based biomedical terminology recommender, the NCBO ontology recommender.

[1] `http://bioportal.bioontology.org/ontologies/SNOMEDCT`, as of September 2014

2 Materials and Methods

The NCBO ontology recommender web service suggests the most appropriate biomedical terminology for annotating biomedical documents. The recommender analyzes a document submitted by the user and applies three criteria for making recommendations: coverage, connectivity, and size of the candidate terminology, taking into account all 370 terminologies from Bioportal [7]. Recommendations are offered both via a web interface and via a REST API.

As a representative document collection we selected discharge summaries from an Intensive Care Unit (ICU), reporting events of a hospitalization (e.g., admitting and discharge diagnoses, physical examinations, and past and follow-up medications). These text address a number of topics of interest in biomedical informatics (e.g., anatomy, drugs, and diseases). The documents were obtained from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) research database[2], a collection of de-identified data from an ICU [13]. 26,657 discharge summary texts were extracted from the *text* field in the *noteevents* table of the MIMIC-II database. Table 1 shows an excerpt of a discharge summary.

ADMISSION DIAGNOSIS: End stage renal disease, admitted for transplant surgery.
HISTORY OF PRESENT ILLNESS: The patient is a 65 year-old woman with end stage renal disease, secondary to malignant hypertension. She was started on dialysis in (...)
PAST MEDICAL HISTORY: End stage renal disease, secondary to malignant hypertension on dialysis. History of anemia following gastric angiectasia (...)
ALLERGIES: No known drug allergies.
MEDICATIONS: Unknown.
SOCIAL HISTORY: Married, lives with her husband. She has a history of a half pack of cigarettes per day for 20 years. Occasional alcohol.
PHYSICAL EXAMINATION: The patient was afebrile. Vital signs were stable. Blood pressure was 124/58; heart rate 76; weight 160 pounds. Abdomen soft and nontender (...)
HOSPITAL COURSE: On [**3389-7-7**], the patient went to the operating room for living donor kidney transplant, performed by Dr. [**Last Name (STitle) 593**] and assisting by (...)
DIAGNOSES: End stage renal disease, status post renal transplant. Arterial thrombosis. Deep venous thrombosis. Resolving hypertension.

Table 1. Excerpt of a discharge summary from the MIMIC-II database.

For topic modeling our collection of discharge summaries, we used a topic modeling tool[3] based on LDA. We used all 26,657 documents from MIMIC-II, 200 iterations of Gibbs sampling, and 10 topics, which we termed *Topic A*, *Topic B*... *Topic J*. For each document, we were only interested in the two most prevalent topics and their associated keywords (termed *primary* topic and *secondary* topic, respectively). As a per-document summarization strategy, we used an in-house improvement of TextRank. Table 2 shows the obtained keywords using both topic modeling and TextRank summarization when applied to the free text from Table 1.

Our main goal was to evaluate how effective topic modeling using LDA and keyword extraction using TextRank were, in comparison to submitting full texts

[2] `http://mimic.physionet.org/database.html`

[3] `http://code.google.com/p/topic-modeling-tool/`

Topic1		**Topic2**		**TextRank**		
blood	**patient**	continued	neumonia	arterial	femoral	q
day	post	failure	pulmonary	**blood**	good	**renal**
discharge	postoperative	fluid	**renal**	**day**	**history**	right
history	rate	negative	started	disease	lower	transplant
mg	status	**patient**	tube	extremity	normal	ultrasound

Table 2. Keywords for document in Table 1 using topic modeling and TextRank. Topic1 and Topic2 represent the primary and secondary topics, respectively. Matching keywords using different methods are marked in bold.

to the recommender. For that purpose, we considered the recommender as a black box and took a sample of 20 documents from the MIMIC-II database[4]. Figure 1 shows our approach for getting the recommendations in each case.

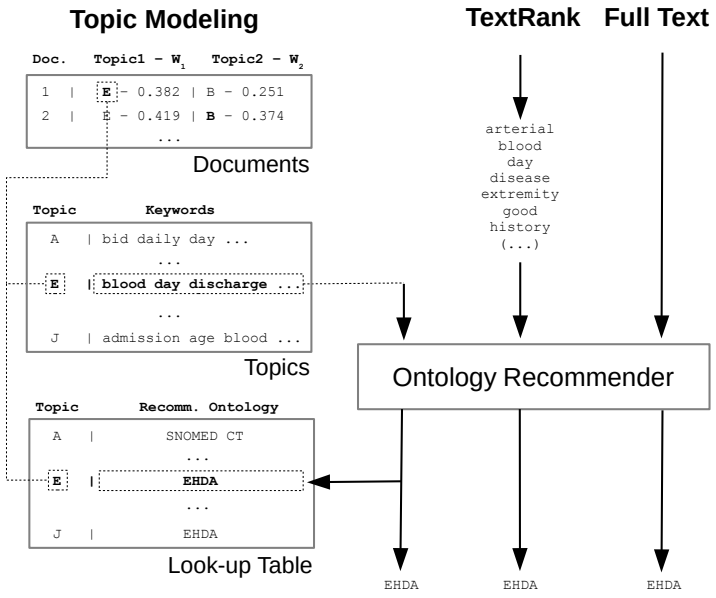

Fig. 1. Recommendations using topic modeling, TextRank, and full texts.

For topic modeling, we submitted each topic's keywords to the recommender and stored the top recommended terminology for each topic, storing an association between topics, keywords, and recommended terminologies in a look-up table for future use. For keyword extraction (TextRank), we submitted the keywords representing the summary of each document from the sample. As gold standard for comparison, we used the full text of each document. We applied a limit of 7,000 characters to every document submitted, as our preliminary experiments showed that the recommender was not able to process long documents.

[4] The first 20 documents retrieved by our PostgreSQL installation.

We recorded recommendation times, including pre-processing when applicable (e.g., time spent summarizing a document using TextRank).

3 Results

The discovered keywords using topic modeling (Table 3) suggest several contexts in the documents, such as: medication administration (A), cardiology (C, I), and diagnostic tests (G). However, only two terminologies were recommended: SNOMED CT and EHDA. Surprisingly, EHDA, focused on developmental stage-specific anatomical structures of the human, was recommended for 7 of the 10 topics, including diagnostic tests (G).

	Keywords	**Terminology**
A	bid daily day disp mg po refills sig tablet times	SNOMED CT
B	continued failure fluid negative patient pneumonia pulmonary renal started tube	SNOMED CT
C	aortic cm left mildly mitral normal regurgitation systolic valve ventricular	EHDA
D	bilaterally ct discharge head hemorrhage history intact left normal patient	EHDA
E	blood day discharge history mg patient post postoperative rate status	EHDA
F	admission discharge history home hospital medications mg normal pain patient	SNOMED CT
G	blood ct glucose hct neg plt pm pt rbc wbc	EHDA
H	chest contrast ct evidence fracture impression left pain small tube	EHDA
I	artery cardiac chest coronary disease heart left mg pain patient	EHDA
J	admission age blood day discharge infant life normal respiratory weeks	EHDA

Table 3. Topics, keywords, and recommended terminologies using topic modeling.

Table 4 shows the primary and secondary topics and their weights in each document from the sample. Topic E was the most frequent overall, appearing in half of the documents.

#	**Topic1**	W_1	**Topic2**	W_2	$W_1 + W_2$	#	**Topic1**	W_1	**Topic2**	W_2	$W_1 + W_2$
1	E	0.382	B	0.251	0.663	11	F	0.541	G	0.220	0.761
2	E	0.419	B	0.374	0.793	12	E	0.557	C	0.109	0.666
3	**J**	**0.907**	**E**	**0.055**	**0.962**	13	C	0.274	A	0.274	0.548
4	E	0.297	B	0.271	0.568	**14**	**I**	**0.243**	**J**	**0.162**	**0.405**
5	A	0.667	F	0.211	0.878	15	E	0.318	D	0.171	0.489
6	D	0.500	E	0.179	0.679	16	E	0.581	B	0.203	0.784
7	F	0.619	D	0.157	0.776	17	I	0.327	G	0.261	0.588
8	D	0.409	H	0.273	0.682	18	H	0.750	G	0.083	0.833
9	F	0.229	G	0.217	0.446	19	E	0.384	F	0.282	0.666
10	H	0.486	D	0.159	0.645	20	I	0.467	E	0.298	0.765

Table 4. Topics and associated weights. Maximum and minimum scores are in bold.

Table 5 shows the recommended terminologies and scores when submitting the full texts and their TextRank versions. When using the full texts, 4 different terminologies were recommended, with SNOMED CT and EHDA recommended for 85% of the documents. Using TextRank, a terminology not identified using the full texts was suggested[5].

[5] Bone Dysplasia Ontology – `http://bioportal.bioontology.org/ontologies/BDO`

#	Full Text	Score	TextRank	Score	#	Full Text	Score	TextRank	Score
1	EHDA	4378.80	EHDA	1303.18	11	SNOMED	1851.21	NCIT	124.93
2	SNOMED	2539.19	SNOMED	114.75	12	EHDA	1453.65	RH-MESH	97.72
3	SNOMED	1086.31	BDO	108.68	13	SNOMED	1306.00	EHDA	153.02
4	EHDA	5436.65	EHDA	1846.39	14	EHDA	423.34	NCIT	35.81
5	NCIT	183.62	SNOMED	37.12	15	RH-MESH	2068.63	EHDA	81.61
6	RH-MESH	222.07	SNOMED	39.80	16	SNOMED	1783.57	SNOMED	97.44
7	SNOMED	1983.63	SNOMED	146.16	17	EHDA	1734.18	EHDA	285.63
8	EHDA	5926.30	EHDA	688.57	18	EHDA	5949.76	EHDA	884.94
9	EHDA	2695.63	EHDA	612.06	19	SNOMED	1838.72	EHDA	137.71
10	EHDA	4054.92	EHDA	1096.61	20	SNOMED	1979.53	EHDA	1150.17

Table 5. Recommended terminologies and associated scores for the sample.

Figure 2 shows the distribution of recommended terminologies. In all cases, EHDA best represented the sample, followed by SNOMED CT. The correlation between terminology distributions respect the gold standard (full texts) was very high (r = 0.83–0.98). Topic modeling the MIMIC-II database took 5 minutes 17 seconds (11 ms per document) and 7 seconds per topic were spent getting a recommendation. When submitting documents, a recommendation took 27 seconds per full text, 11 seconds using TextRank (including summarization).

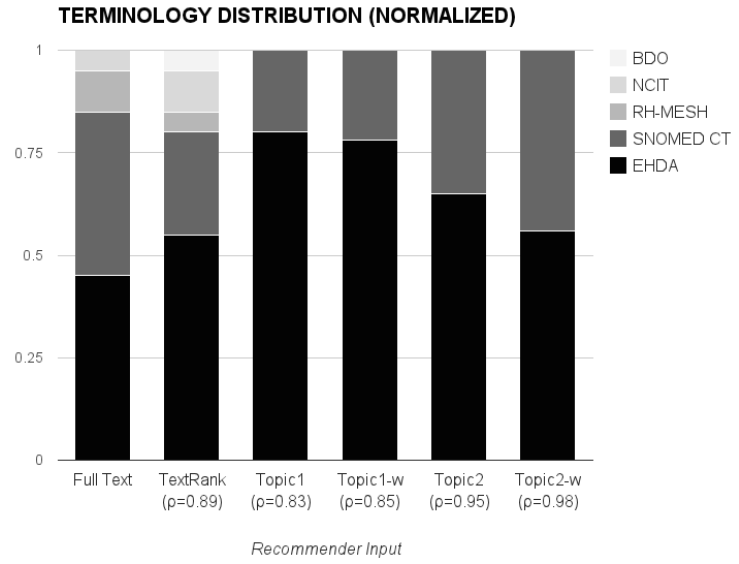

Fig. 2. Distribution of recommended terminologies for the sample using full text, TextRank, and Topic modeling (-w = weighted).

4 Discussion

EHDA and SNOMED CT were recommended for the majority (85%) of documents in the sample, EHDA being preferred. Why EHDA was the most recommended terminology when submitting discharge summaries as context needs to be carefully studied, as discharge summaries contain a broad range of topics (discharge diagnoses, physical examinations, past and follow-up medications, etc.) that are not covered by EHDA. Even in the case of anatomy, FMA [12] seems more appropriate, as EHDA is focused mainly on tissue development [6]. Although assessing the validity of the recommender was not the goal of our study, the inexplicable prevalence of EHDA in the recommendations suggests possible shortcomings in the recommender that would inevitably limit the significance of our results.

When analyzing performance, our results suggest that it might not be feasible for users to submit a large number of documents as a representative context, as getting recommendations for a sample of 20 documents with full texts (limited to 7,000 characters) took nearly 10 minutes. The keywords obtained using topic modeling were less in number than the keywords using TextRank. This should, in principle, make recommendations using topic modeling less correlated to the ones obtained using the full texts, but the opposite was true. This fact might be explained because all 26,657 documents from MIMIC-II were used when modeling the topics, providing a much more accurate context.

5 Conclusions and Future work

In this study, we have proposed and evaluated two well-researched automatic summarization techniques for summarizing a large collection of clinical documents used as input to the NCBO ontology recommender: topic modeling the collection using LDA, and per-document TextRank keyword extraction. When comparing both approaches to our gold standard (full texts) in the evaluation, we found out that recommendation times improved considerably. In all cases, the distributions of recommended terminologies were highly correlated with the gold standard distribution ($r = 0.83$–0.98). The high correlation shows that both TextRank and topic modeling are valuable techniques to summarize the context provided by the full texts and boost recommendation performance without seriously affecting the overall recommendation results.

As future work, we plan to: (i) use a larger sample of documents to investigate if our results are consistent, (ii) select a collection of documents from other domain to generalize our results, and (iii) investigate potential quality issues in the recommender, given the prevalent but inexplicable recommendations of the EHDA terminology when submitting discharge summaries as input.

Acknowledgments

The authors thank H. Ziak, A. Rexha, G. Hammer, C. Martínez-Costa, M. Kreuzthaler, and G. A. Uribe Gómez for their contributions; the NCBO for providing the ontology recommender and Bioportal; and the MIT and the Beth Israel Deaconess Medical Center for providing the MIMIC-II database. This work was developed within the EEXCESS project funded by the European Union FP7/2007-2013 under grant agreement number 600601.

Bibliography

[1] Adomavicius, G., Tuzhilin, A.: Context-Aware Recommender Systems. In: Recommender Systems Handbook, pp. 217–253. Springer (2011)
[2] Benson, T.: Principles of Health Interoperability. HL7 and SNOMED. Springer (2010)
[3] Blei, D.M., et al.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
[4] Freitas, F., et al.: Survey of Current Terminologies and Ontologies in Biology and Medicine. RECIIS—Electronic Journal in Communication, Information and Innovation in Health 3(1), 7–18 (2009)
[5] Hariri, N., et al.: Query-Driven Context Aware Recommendation. In: 7th ACM Conference on Recommender Systems. pp. 9–16. ACM (2013)
[6] Hunter, A., et al.: An Ontology of Human Developmental Anatomy. Journal of Anatomy 203(4), 347–355 (2003)
[7] Jonquet, C., et al.: Building a Biomedical Ontology Recommender Web Service. Journal of Biomedical Semantics 1(Suppl 1), S1 (2010)
[8] Linden, G., et al.: Amazon.com Recommendations: Item-to-Item Collaborative Filtering. Internet Computing, IEEE 7(1), 76–80 (2003)
[9] Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Texts. In: Conference on Empirical Methods in NLP. pp. 404–411. ACL (2004)
[10] Musen, M., et al.: The National Center for Biomedical Ontology. Journal of the American Medical Informatics Association 19(2), 190–195 (2012)
[11] Noy, N.F., et al.: BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse. Nucleic Acids Research 37(suppl 2), W170–W173 (2009)
[12] Rosse, C., Mejino Jr, J.L.: A Reference Ontology for Biomedical Informatics: the Foundational Model of Anatomy. Journal of Biomedical Informatics 36(6), 478–500 (2003)
[13] Saeed, M., et al.: MIMIC-II: a Public-Access Intensive Care Unit Database. Critical Care Medicine 39(5), 952 (2011)
[14] Tan, H., Lambrix, P.: Selecting an Ontology for Biomedical Text Mining. In: Workshop on Current Trends in Biomedical NLP. pp. 55–62. ACL (2009)