MESH classification of clinical guidelines using conceptual embeddings of references

Johan Eklund, David Gunnarsson Lorentzen, and Gustaf Nelhans

{johan.eklund, david.gunnarsson_lorentzen, gustaf.nelhans} @hb.se
University of Borås, Swedish School of Library and Information Science, S-501 90 Borås (Sweden)

Abstract
In this study, we investigate different strategies for assigning MeSH (Medical Subject Headings) terms to clinical guidelines using machine learning. Features based on words in titles and abstracts are investigated and compared to features based on topics assigned to references cited by the guidelines. Two of the feature engineering strategies utilize word embeddings produced by recent models based on the distributional hypothesis, called word2vec and fastText. The evaluation results show that reference-based strategies tend to yield a higher recall and F1 scores for MeSH terms with a sufficient amount of training instances, whereas title and abstract based features yield a higher precision.

Introduction
This paper builds on previous attempts to use a combination of cited references as text for machine learning purposes to develop new means for combining text "mining" with scientometric citation-based methods. (Eklund & Nelhans, 2017) Such an approach, if rendered stable, would provide means for developing a joint methodology, between different specialities in information science and to broaden the scope of analysis of citation-based data.

In previous studies, we used the Latent Dirichlet Allocation (LDA) algorithm (see e.g. Blei, 2003) for topic modelling of references, using an approach in which references are treated as "words" and reference lists as "sentences" (or documents) of such "words". We demonstrated that the topical structure of document collections could be studied using a combination of citation network properties and text-based methods.

In this study, we take a somewhat different stance by generating semantic representation vectors of the cited documents treated as semantic compositions of the MeSH terms assigned to the cited documents. The ultimate objective of this approach is to be able to classify a heterogeneous set of non-standardized documents, hence the choice of clinical guideline documents that are extracted from different sources, covering many different languages, having different sets of metadata content, but all containing a matched list of citations. The techniques introduced are generic and are applicable to other kinds of professional and policy documents.

Our approach answers a simple question that could be framed as

- RQ 1. How can we in a meaningful way classify a set of documents for which we do not have access to their texts, but instead their sets of cited references?

Furthermore, we want to evaluate our method by comparing it to a more traditional approach using representation vectors containing term weights of words appearing in titles and/or abstracts.

- RQ 2. How well do reference-based feature vectors perform as compared with a text-based method in a collection of bibliographic records containing titles and abstracts?

Word embedding and semantic composition
The distributional hypothesis of semantics (see e.g. Sahlgren, 2008) entails the notion that words that tend to occur in the same contexts are also semantically related. The idea of

generating representations of words that capture their co-occurrence patterns and yield semantic representation vectors has been investigated for several decades in the fields of computational linguistics, information retrieval, and document classification. Among the early methods generating semantic representation vectors based on co-occurrence data can be mentioned Hyperspace Analogue to Language (Lund & Burgess, 1996), latent semantic analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), and random indexing (Kanerva, Kristoferson, & Holst, 2000).

In recent years a new generation of word embedding methods based on neural network learning has emerged. One of those models is commonly called *word2vec* and was published by Mikolov, Chen, Corrado, & Dean (2013). This algorithm utilizes a shallow feed-forward neural network to implement two different approaches for encoding word context, called continuous bag-of-words (CBOW) and continuous skip-gram respectively. With regard to the CBOW model, the objective is to predict the target word, given an input consisting of a set of context words. Conversely, the objective of the continuous skip-gram model is to identify a set of context words, given an input word. The state vectors of the hidden layer obtained through back-propagation are subsequently used as representation vectors for the vocabulary pertaining to the training collection of documents.

Bojanowski, Grave, Joulin, & Mikolov (2017) present an extension to the continuous skip-gram model based on incorporating morphological information in the learning of word representation vectors. This is accomplished by associating each word with a set of character *n*-grams. For example, the 3-grams contained in the word (up to the word boundaries) *smoke* are *sm*, *smo*, *mok*, *oke*, and *ke*. A representation vector for each character *n*-gram is learned separately and finally, a representation vector for each word in the vocabulary is established by computing the sum of the representation vectors for the *n*-grams associated with the word. A potential advantage of this method is that it facilitates learning of representation vectors in collections with many infrequent words, and consequently only a few instances available for learning. Since this algorithm is implemented in the fastText library (FastText, n.d.) for text classification and representation learning, it is named the *fastText* algorithm in this paper.

In collections yielding sparse feature vectors, term representations that captures co-occurrence patterns may also yield document representations that are related by means of conceptual dimensions, rather than by term dimensions. The traditional vector space model as presented by Salton, Wong, & Yang (1975) is based on the mechanism of assigning document vectors as linear combinations of a sequence of term weights and the orthonormal basis of the Euclidean space. In other words, each document is represented as a weighted sum of a set of mutually orthogonal (unit) term vectors. In this study, we treat each reference as the semantic composition of the MeSH terms assigned to that reference. By the *semantic composition* of terms, we denote structures that are essentially lists of words, such as phrases and sentences. According to Blacoe & Lapata (2012), representational models for structures have received less attention than the semantic modelling of single words. Two basic approaches for generating representation vectors of semantic compositions are by means of vector addition, and elementwise multiplication (the so-called Hadamard product). In line with one of the strategies investigated by Blacoe & Lapata (2012), we then generate representation vectors for every reference document r_j as

$$\mathbf{r}_j = \sum_{\forall k_i \in d_j} \mathbf{k}_i$$

where $\mathbf{k}_i$ is the representation vector of term k_i. In other words, we treat the list of MeSH terms assigned to a document as the semantic composition of those terms. Likewise, we treat each

clinical guideline g_k as the semantic composition of the references contained in g_k and produce a corresponding representation vector as

$$\mathbf{g}_k = \sum_{\forall r_j \in g_k} \mathbf{r}_j$$

Using such additive semantic representation models facilitates accumulative and distributed computations of the available training data, since addition is a commutative as well as associative operation.

Methodology

In total 6 different strategies for the representation of clinical guidelines were investigated and compared:

A1. Title
A2. Abstract
A3. Title combined with abstract
R1. MeSH terms represented by binary vectors containing 0's in all positions except one unique position containing the number 1; so-called one-hot encoding.
R2. MeSH terms represented by word embeddings (n = 300) generated by the word2vec algorithm.
R3. MeSH terms represented by word embeddings (n = 300) generated by the fastText algorithm.

A dataset consisting of 285 bibliographic records of clinical guidelines has been used. These records have been enriched by MeSH terms and abstracts downloaded from PubMed through the Entrez Programming Utilities API. The guidelines were selected as to ensure that MeSH terms and abstracts are available for each guideline. The complete reference list identifiable by PMIDs were mined from the full texts and meta data, including MeSH terms, were collected for each cited reference. The titles and the abstracts have been preprocessed using the *tm* package for R (Feinerer, Hornik, & Meyer, 2008) by converting the texts to lower case, removing punctuation marks and numeric characters and filtering out common English stopwords. All the details of the classification experiments such as feature selection, training, and cross-validation have been implemented using the *mlr* package for R (Bischl et al., 2016).For each experiment involving title, abstract, or a combination of the two, the texts have been assigned document vectors by means of tf-idf weighting. Among many possible term weighing strategies available, we have opted for the *tfx* strategy as described by Salton & Buckley (1988), which amounts to tf-idf weighting by means of an unnormalized tf component, an ordinary idf component, and no vector length normalization.

For the reference-based strategies as well as the strategy using only words in guideline titles, a chi-square metric (see e.g. Zheng, Wu, & Srihari, 2004) was used to rank and select k terms for each target class. After some experimentation, it was decided that the value k = 100 yields comparatively good performance and was therefore used for the experiments.

The classification algorithm used for the experiments is *logistic regression*, which is a binary classification algorithm modelling the log-odds of class probabilities as a linear function of the document features. Previous comparative studies using logistic regression for text categorization (Zhang & Oles, 2001; Zhang, Jin, Yang, & Hauptmann, 2003) have shown a performance comparable to state-of-the-art algorithms like the support vector machine (SVM). Since the training of logistic regression classifiers does not require the tuning of hyper-

parameters (unlike, for instance, the SVM algorithms) it was selected for this study. An L2 penalty term has been used to regularize the regression coefficients and prevent overfitting.

Evaluation

For the evaluation of the classifier performance we have used stratified 3-fold cross validation with 100 iterations to obtain stable performance figures. The use of stratified cross-validation was employed to ensure that all folds contain positive examples, which is a critical factor for classes with only a few examples. The evaluation measures used are precision, recall, and the F1 score. In order to compute the average F1 score (which itself is the harmonic mean of precision and recall) for the repeated cross-validation evaluation there are several possible procedures that could be considered. Forman & Scholz (2010) discuss three different definitions:

1. $F_{avg} \coloneqq \frac{1}{k}\sum_{i=1}^{k} F_i$ where F_i is the F1 score for each fold.
2. $F_{\mathrm{pr,re}} \coloneqq 2 \cdot \frac{\mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}}$ where Pr and Re are the average precision and recall respectively over all folds.
3. $F_{\mathrm{tp,fp}} \coloneqq (2 \cdot \mathrm{TP}) / (2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN})$ where TP, FP, and FN are the sum of the true positives, false positives, and false negatives respectively over all folds.

Based on experimental evidence, Forman & Scholz (2010) find that the $F_{\mathrm{tp,fp}}$ strategy is the least biased estimator of the F1 score for cross-validation (especially for datasets with a high degree of class imbalance) and this strategy has also been selected for this study.

Twelve MeSH terms have been used as target terms for the evaluation of each strategy. These terms were chosen on the basis a selection of topics (pregnancy and birth, dietary supplements, smoking, cardiovascular diseases, and mental health), as well as having a sufficient amount of instances in the training set.

Findings

In tables 1-3 we present the performance score (average precision, average recall, and average F1 respectively) for each MeSH term and feature strategy.

Table 1. Average precision and recall for the selected terms and the six feature strategies.

term	*title*		*abstract*		*title+abstract*		*binary*		*word2vec*		*fastText*	
	p	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*
Pregnancy	0.83	0.52	**0.93**	0.69	**0.93**	0.70	0.92	**0.89**	0.79	0.85	0.83	0.86
Infant, Newborn	**0.56**	0.34	0.51	0.23	0.53	0.24	0.55	**0.48**	0.38	0.37	0.42	0.40
Infant	0.56	0.22	**0.95**	0.25	0.92	0.24	0.59	0.37	0.46	0.40	0.50	**0.46**
Pregnancy Outcome	**0.82**	0.47	0.75	0.43	0.73	0.41	0.49	0.46	0.43	0.45	0.46	**0.51**
Dietary Supplements	0.76	**0.79**	**0.87**	0.52	0.88	0.55	0.73	0.61	0.59	0.65	0.59	0.60
Vitamins	**0.68**	**0.54**	0.66	0.39	0.66	0.39	0.60	0.44	0.40	0.45	0.38	0.42
Smoking Cessation	0.79	**0.85**	0.84	0.62	0.80	0.67	**0.90**	0.76	0.77	0.64	0.74	0.62
Smoking	0.35	0.12	**0.97**	0.38	0.91	0.36	0.46	**0.44**	0.29	0.30	0.28	0.31
Stroke	0.71	0.36	0.78	0.38	**0.80**	0.42	0.50	0.41	0.52	**0.54**	0.50	0.46
Cardiovascular Diseases	0.99	0.67	**1.00**	0.71	**1.00**	**0.71**	0.42	0.34	0.43	0.48	0.40	0.42
Depression	0.02	0.09	0.00	0.00	0.00	0.00	**0.28**	**0.27**	0.16	0.25	0.14	0.22
Anxiety	**0.74**	**0.35**	0.00	0.00	0.00	0.00	0.34	0.34	0.16	0.30	0.17	0.30

What is clearly noticeable in table 1 is that the precision score tends to be higher for the feature strategies based on title and abstract respectively. This in turn indicates that those types of metadata fields tend to contain information that is more focused on the content of the guidelines than the corresponding references.

The average recall for the different strategies, also presented in table 1, follows a different pattern than what can be observed with regard to the average precision. The reference-based strategies tend to yield a higher recall than the title and abstract based strategies, in particular for the more frequent MeSH terms in the dataset. This is an indication that the use of references and their corresponding MeSH terms tends to increase the coverage of the induced classifier.

Table 2. Average F1 score for the selected terms and the six feature strategies.

term	*title*	*abstract*	*title+abstract*	*binary*	*word2vec*	*fastText*
Pregnancy	0.64	0.79	0.80	**0.90**	0.82	0.84
Infant, Newborn	0.42	0.32	0.33	**0.51**	0.37	0.41
Infant	0.32	0.40	0.38	0.46	0.43	**0.48**
Pregnancy Outcome	**0.60**	0.55	0.53	0.48	0.44	0.48
Dietary Supplements	**0.78**	0.66	0.68	0.66	0.62	0.59
Vitamins	**0.60**	0.49	0.49	0.51	0.43	0.40
Smoking Cessation	0.82	0.72	0.73	**0.82**	0.70	0.68
Smoking	0.18	**0.54**	0.51	0.45	0.30	0.29
Stroke	0.48	0.51	**0.55**	0.45	0.53	0.48
Cardiovascular Diseases	0.80	0.83	**0.83**	0.38	0.45	0.41
Depression	0.03	0.00	0.00	**0.27**	0.19	0.17
Anxiety	**0.47**	0.00	0.00	0.34	0.21	0.22

The average F1 score for the different strategies is displayed in table 2. A slight advantage can be noticed for the title and abstract based features when taken as a group of strategies. If we instead compare the individual strategies, we find that the title based and binary reference-based strategies display a comparable performance with 4 top scores each. There is, however, no clearly discernible trend with regard to the top score and the number of guidelines indexed by each respective MeSH term. However, a natural question arises in connection with the observed results, namely what the correlation is between the performance of each strategy and the number of instances available in the dataset for each MeSH class.

Table 3. Rank correlation between the number of instances of MeSH terms and performance scores.

	title	*abstract*	*title + abstract*	*binary*	*word2vec*	*fastText*
precision	0.26	0.22	0.35	0.76	0.69	0.81
recall	0.25	0.35	0.34	0.73	0.57	0.69
F1	0.26	0.31	0.32	0.81	0.61	0.81

In table 3 we presented the rank correlation, as measured by Spearman's rho, between the number of instances of each MeSH term in the dataset and the average performance scores for each strategy. There is a clearly discernible difference between the title and abstract based strategies versus the reference-based strategies, in the sense that the reference-based strategies tend to perform better when more instances are available.

Discussion

McJunkin (1995) argues that title words should be considered when performing keyword-based searches. While controlled vocabularies, such as the Library of Congress Subject Headings, are

useful when specific entry terms or word order is not known, the author states that subject-rich terms in titles could be used favorably for keyword-based searching since these are often more current than established controlled vocabulary. (ibid.) The results of the present study indicate that the use of title words alone provides a classifier performance with regard to precision that is comparable to that of words appearing in abstracts. What can also be observed is that reference-based features tend to yield a higher recall, but there is no clear evidence in this study that the use of semantic word embedding yields a more performant representation of the content of the references. This may be explained in terms of the limited amount of data used for producing word embeddings together with the observation that models utilizing neural network training, such as word2vec, generally need large training sets to perform well (Mikolov et al. 2013).

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770531.

References

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., ... & Jones, Z. M. (2016). mlr: Machine Learning in R. *The Journal of Machine Learning Research, 17*(1), 5938-5942.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Blacoe, W., & Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 546-556). Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135-146.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*(6), 391-407

Eklund, J., & Nelhans, G. (2017). Topic modelling approaches to aggregated citation data. Presented at the *22nd International Conference on Science and Technology Indicators*, Paris, September 6-8.

FastText. (n.d.). Retrieved February 08, 2019, from https://fasttext.cc/.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software, 25*(5), 1-54.

Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *SIGKDD Explorations, 12*, 49-57.

Kanerva, P., Kristoferson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 22, No. 22).

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2), 203-208.

McJunkin, M. C. (1995). Precision and recall in title keyword searches. *Information Technology and Libraries, 14*(3), 161.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv*:1301.3781.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics, 20*, 33-53.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513-523.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Zhang, T., & Oles, F. J. (2001). Text categorization based on regularized linear classification methods. *Information retrieval, 4*(1), 5-31.

Zheng, Z., Wu, X., & Srihari, R. (2004). Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter, 6*(1), 80-89.