Assessing the Semantic Difficulty of Queries*

Discussion Paper

Guglielmo Faggioli[1], Stefano Marchesin[1]

[1]University of Padova, Padova, Italy

Abstract

Traditional Information Retrieval (IR) models, also known as lexical models, are hindered by the semantic gap, which refers to the mismatch between different representations of the same underlying concept. To address this gap, semantic models have been developed. Semantic and lexical models exploit complementary signals that are best suited for different types of queries. For this reason, these model categories should not be used interchangeably, but should rather be properly alternated depending on the query. Therefore, it is important to identify queries where the semantic gap is prominent and thus semantic models prove effective. In this work, we quantify the impact of using semantic or lexical models on different queries, and we show that the interaction between queries and model categories is large. Then, we propose a labeling strategy to classify queries into semantically hard or easy, and we deploy a prototype classifier to discriminate between them.

1. Introduction

The semantic gap is a long-standing problem in *Information Retrieval (IR)* that refers to the difference between the machine-level description of document and query contents and the human-level interpretation of their meanings [2]. In other words, it represents the mismatch between users' queries and the way retrieval models understand such queries [3].

The semantic gap affects any domain, but it is prominent in medical search [4, 5, 2]. For instance, a query containing the word "tumor" might not be effectively answered if the retrieval model does not identify the synonymy relationship between "tumor" and, for example, "neoplasm". Conversely, given a query containing the term "cold", a retrieval model might retrieve erroneous documents if it does not distinguish between the different meanings the term "cold" assumes depending on the context. These queries are known as *semantically hard* queries [6].

Traditional IR models, which are known as lexical models, fail to effectively address semantically hard queries. Semantic models were thus introduced to bridge the semantic gap [7] and to overcome the limitations of lexical models. However, semantic models have been shown to provide complementary signals to lexical models that prove effective for semantically hard queries, but less for other queries [8]. Thus, it becomes necessary to identify what category of models – between lexical and semantic – best suits a user query given the document collection at hand. In other words, we need to understand what are the inherent features of query and

* The full paper has been originally presented at DESIRES 2021 [1]

IIR 2021 – 12th Italian Information Retrieval Workshop, June 29th – July 1st, 2022, Milano, Italy

Table 1
Mean Average Precision (MAP) of the models on OHSUMED and TREC-COVID collections.

	Lexical					Semantic				
	TF-IDF	**BM25**	**QLM**	**DFR**	**DFI**	**W2V**	**NVSM**	**SAFIR$_s$**	**SAFIR$_p$**	**SAFIR$_{sp}$**
OHSUMED	0.524	0.62	0.577	0.641	0.592	0.568	0.595	0.604	0.61	0.612
TREC-COVID	0.362	0.488	0.434	0.496	0.467	0.482	0.455	0.463	0.461	0.466

documents that make lexical or semantic models more effective. To this end, we address the following research questions: **RQ1** How and to what extent does the semantic gap impact query performance? **RQ2** What features determine the prominence of the semantic gap within queries? For **RQ1**, we investigate and compare the impact of lexical and semantic models on different topics. How large is the interaction between topics and model categories? To what extent does this interaction reflect in the different topic formulations (i.e., queries)?
For **RQ2**, we explore a set of well-known features that relate to lexical and semantic models. In particular, we seek to understand whether pre-retrieval features can be used to categorize queries as semantically easy or hard.

2. Experimental Analysis

We consider two collections in the following analyses: OHSUMED [9] and TREC-COVID (Round 1) [10]. Regarding lexical models, we consider TF-IDF [11], BM25 [12], QLM [13], DFR [14], and DFI [15]. As for semantic ones, we consider W2V [16], NVSM [17], and the three variants of SAFIR [6]. We evaluate models using AP. Table 1 reports the performance on both collections.

2.1. RQ1: Topic and Category Interaction

Several works have shown that queries strongly interact with retrieval models in determining their performance [18, 19, 20]. This means that two models might have similar average performance on a set of queries but, when looked at the query-level, their performance might vary greatly. Such consideration also applies to lexical and semantic models. Some queries are best suited to semantic models, while others to lexical ones [8, 6]. We are thus interested in quantifying the interaction between queries and model categories. To determine whether the models category – that is, lexical or semantic – has a significant effect on performance, we conduct an ANOVA on the runs obtained with the considered retrieval models. ANOVA is a well-known statistical technique that allows identifying statistically significant differences among experimental conditions. Several works in IR applied ANOVA to determine the effect of different factors on the overall performance of an IR system [18, 21, 19, 22]. ANOVA models the explained variable, which in our case is *Average Precision (AP)*, as a linear combination of the effect of each factor in the experimental setup, plus an error component. The error term accounts for the variance in the data unexplained by the model. From the ANOVA on our data, we observe that the effect of the sole models category is not significant (p-value$>$0.05) – which means that lexical and semantic categories are not statistically significantly different. We cannot say that either lexical or semantic models perform best in absolute terms. The topic-category

interaction is significant and the ω^2 value for the strength of association of 34.7% indicates a large effect. This means that the category significantly impacts on how good the results on a specific topic will be. Such a finding suggests that the semantic gap is an inherent property of the topics, less related to the specific retrieval models and more on their category. To further support this intuition, the interaction between the topic and the category is larger than the effect of the sole model. Thus, if we understand when a topic is lexical or semantic, we can achieve large performance improvements. As for TREC-COVID, each topic is represented by four different formulations: *query*, *description*, *narrative* and *concatenation* of query and description. Each formulation of a topic can only be used in relation to that topic thus formulations have to be treated as a nested factor inside the topic. From the results on TREC-COVID we observe that both the topic and its formulations have a large effect. The importance of the formulation factor indicates that, with an appropriate topic formulation, the performance on the topic can change greatly. ANOVA shows that the interaction between the topic and the models category is large (ω^2 =39%) – larger than the effect of both the sole category (2.1%) and the model (30.4%). Also the interaction between the topic formulation and the models category is large (ω^2 =19.7%), although not as large as the one between topic and category. This suggests that the semantic gap relates more to the underlying information need than the different topic formulations.

We hypothesize that the relation between topics and model categories, highlighted by ANOVA, links to the semantic gap and to the association of a topic with its relevant documents. For instance, if a topic has many relevant documents containing synonyms of the query terms, then a semantic model might be best suited. In fact, in this case, most of the topic formulations do not contain all the possible query synonyms and will thus be affected by the semantic gap. Conversely, topics that can be easily represented by few keywords – likely present in relevant documents – have less ambiguous formulations, which are best suited to lexical models.

2.2. RQ2: Features Importance for the Semantic Gap

Section 2.1 showed the impact of choosing the proper category depending on the target query. If we could classify queries as semantically hard or easy, we might also adopt an IR model from the right category. To train a classifier for doing that, we need *i)* to label queries as "semantic" or "lexical", and *ii)* to find a set of features that correlate with such aspects of the queries.

The first aspect we address is the labeling of queries as "semantic" or "lexical". The absence of a rigorous definition of *semantically hard* or *easy* for a query prevents us from manually labeling queries as "semantic" or "lexical". Therefore, we propose to label queries according to how the two models categories perform on them. To this end, we first compute the average performance of each model. Then, for each query, we perform the following three steps. First, we compute for each model the relative improvement over its average performance. Secondly, we determine whether the relative improvement is, on average, greater for lexical or semantic models. Finally, we label the considered query as "semantic" if the improvement over the average model performance is greater for semantic models than for lexical ones; vice versa, we label the query as "lexical". Note that we do not consider absolute performances to label queries, since even a poorly performing lexical method like TF-IDF (cfr. Table 1) might prove effective when the query is semantically easy. Thus, we focus on relative improvements, which provide more robust signals to performance outliers. To address the second aspect of **RQ2**,

Table 2
Classifiers performance. We report mean and standard deviation over 3- and 5-folds for OHSUMED and TREC-COVID, respectively. † indicates statistical significance over the random classifier.

	OHSUMED		TREC-COVID	
	Accuracy	**F1**	**Accuracy**	**F1**
DTr	0.626 (0.089)	0.586 (0.057)	**0.668 (0.093)**†	**0.659 (0.141)**†
SVM	0.687 (0.074)	0.611 (0.079)	0.623 (0.053)	0.610 (0.136)
MLP	**0.740 (0.081)**	**0.675 (0.146)**	0.628 (0.217)	0.590 (0.269)

we explore two different sets of pre-retrieval features: Lexical- and Semantic-oriented features. Lexical-oriented features are based on query and corpus statistics and depend on the distribution of terms within the collection. Regarding semantic-oriented features, we first perform semantic indexing on OHSUMED and TREC-COVID collections as in [6]. Then, we adopt features similar to those proposed by Mothe and Tanguy [23], but, instead of considering only query-based features, we take into account both query- and corpus-based features. The considered features are reported and described in the original paper [1]. Consequently, we employ three well-known classification models to understand the effectiveness of the considered pre-retrieval features when used to classify queries into lexical and semantic categories. The adopted models are: Decision Tree (DTr), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). To perform experiments, we label queries using the process described above. For each classifier, we perform grid search with cross-validation to obtain the best hyper-parameters. We adopt 5-fold cross-validation for TREC-COVID, whereas we use 3-fold cross-validation for OHSUMED to avoid obtaining single-class folds due to the low number of samples. The results of the different classifiers are reported in Table 2, where we report mean and standard deviation over the different folds. To determine results significance (marked as †), we apply a randomization test with Bonferroni correction for multiple comparisons [24].

The preliminary – yet promising – results highlight that the considered lexical- and semantic-oriented features relate with models categories. Therefore, they can be used as a starting point to investigate the presence of the semantic gap within test collections and to build better approaches for category selection.

3. Conclusion

We investigated the impact of the semantic gap on query performance, which features can be used to determine this gap, and whether we can exploit them to classify query as semantically easy ("lexical") or hard ("semantic"). Using ANOVA we studied the interaction between IR models and information need, observing that the semantic gap relates more to the underlying information need than the different topic formulations. Then, we proposed a labeling strategy, based on relative improvements, to annotate queries as "semantic" or "lexical". Finally, we explored two different sets of pre-retrieval features and we deployed a prototype classifier to understand the effectiveness of such features when used to classify queries. We obtained promising results, which suggest a link between the used features and the models categories.

References

[1] G. Faggioli, S. Marchesin, What makes a query semantically hard?, in: Proc. of the Second International Conference on Design of Experimental Search & Information REtrieval Systems, Padova, Italy, September 15-18, 2021, volume 2950 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 61–69.

[2] B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, M. Lawley, Information retrieval as semantic inference: a Graph Inference model applied to medical search, Inf. Retr. Journal 19 (2016) 6–37.

[3] R. Zhao, W. I. Grosky, Narrowing the semantic gap - improved text-based web document retrieval using visual features, IEEE Trans. Multimedia 4 (2002) 189–200.

[4] T. Edinger, A. M. Cohen, S. Bedrick, K. H. Ambert, W. R. Hersh, Barriers to Retrieving Patient Information from Electronic Health Record Data: Failure Analysis from the TREC Medical Records Track, in: AMIA 2012, American Medical Informatics Association Annual Symposium, AMIA, 2012.

[5] B. Koopman, G. Zuccon, Why Assessing Relevance in Medical IR is Demanding, in: Proc. of the Medical Information Retrieval Workshop at SIGIR co-located with the 37th annual international ACM SIGIR conference (ACM SIGIR 2014), volume 1276 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2014, pp. 16–19.

[6] M. Agosti, S. Marchesin, G. Silvello, Learning Unsupervised Knowledge-Enhanced Representations to Reduce the Semantic Gap in Information Retrieval, ACM Trans. Inf. Syst. 38 (2020) 38:1–38:48.

[7] H. Li, J. Xu, Semantic Matching in Search, Found. Trends Inf. Retr. 7 (2014) 343–469.

[8] S. Marchesin, A. Purpura, G. Silvello, Focal elements of neural information retrieval models. An outlook through a reproducibility study, Inf. Process. Manag. 57 (2020) 102109.

[9] W. Hersh, C. Buckley, T. J. Leone, D. Hickam, Ohsumed: An interactive retrieval evaluation and new large test collection for research, in: Proc. of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994, Springer London, London, 1994, pp. 192–201.

[10] E. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, L. L. Wang, TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection, SIGIR Forum 54 (2021).

[11] W. B. Croft, D. Metzler, T. Strohman, Search Engines: Information Retrieval in Practice, Addison-Wesley, Reading (MA), USA, 2009.

[12] S. E. Robertson, U. Zaragoza, The Probabilistic Relevance Framework: BM25 and Beyond, Found. Trnd. Inf. Retr. 3 (2009) 333–389.

[13] C. Zhai, Statistical Language Models for Information Retrieval. A Critical Review, Found. Trnd. Inf. Retr. 2 (2008) 137–213.

[14] G. Amati, C. J. van Rijsbergen, Probabilistic Models of Information Retrieval based on measuring the Divergence From Randomness, ACM Trans. Inf. Syst 20 (2002) 357–389.

[15] İ. . Kocaba ş, B. T. Din ç er, B. Karao ğ lan, A nonparametric term weighting method for information retrieval based on measuring the divergence from independence, Information Retrieval 17 (2014) 153–176.

[16] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in

Vector Space, in: Proc. of the 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, 2013.
[17] C. Van Gysel, M. De Rijke, E. Kanoulas, Neural vector spaces for unsupervised information retrieval, ACM Trans. Inf. Syst. 36 (2018) 1–25.
[18] D. Banks, P. Over, N.-F. Zhang, Blind Men and Elephants: Six Approaches to TREC data, Information Retrieval 1 (1999) 7–34.
[19] N. Ferro, G. Silvello, Toward an Anatomy of IR System Component Performances, J. Assoc. Inf. Sci. Technol. 69 (2018) 187–200.
[20] J. S. Culpepper, G. Faggioli, N. Ferro, O. Kurland, Topic difficulty: Collection and query formulation effects, ACM Transactions on Information Systems 40 (2021).
[21] E. Voorhees, D. Samarov, I. Soboroff, Using Replicates in Information Retrieval Evaluation, ACM Trans. Inf. Syst 36 (2017) 12:1–12:21.
[22] G. Faggioli, N. Ferro, System effect estimation by sharding: A comparison between anova approaches to detect significant differences, in: Proc. of the 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Springer International Publishing, Cham, 2021, pp. 33–46.
[23] J. Mothe, L. Tanguy, Linguistic Features to Predict Query Difficulty, in: Proc. of the Predicting query difficulty-methods and applications workshop, co-located with the ACM Conference on research and Development in Information Retrieval, SIGIR 2005, 2005, pp. 7–10.
[24] P. Sedgwick, Multiple significance tests: the bonferroni correction, Bmj 344 (2012).