Which Word Embeddings for Modeling Web Search Queries? Application to the Study of Search Strategies

Claire Ibarboure, Ludovic Tanguy and Franck Amadieu
CLLE: CNRS & University of Toulouse, France
{*firstname.lastname*}*@univ-tlse2.fr*

Keywords: Word Embeddings, Information Retrieval, Queries, Search Strategies.

Abstract: In order to represent the global strategies deployed by a user during an information retrieval session on the Web, we compare different pretrained vector models capable of representing the queries submitted to a search engine. More precisely, we use static (type-level) and contextual (token-level, such as provided by transformers) word embeddings on an experimental French dataset in an exploratory approach. We measure to what extent the vectors are aligned with the main topics on the one hand, and with the semantic similarity between two consecutive queries (reformulations) on the other. Even though contextual models manage to differ from the static model, it is with a small margin and a strong dependence on the parameters of the vector extraction. We propose a detailed analysis of the impact of these parameters (e.g. combination and choice of layers). In this way, we observe the importance of these parameters on the representation of queries. We illustrate the use of models with a representation of a search session as a trajectory in a semantic space.

1 INTRODUCTION

This study is part of a project which aims to investigate information retrieval (IR) strategies based on experimental data. Ultimately, the aim is to distinguish behaviours that may vary according to criteria such as the users' level of knowledge, the type of the task or socio-demographic criteria. Therefore, we aim to automate the analysis of the language data involved in an IR session (queries, result pages, documents and verbalisations of the intentions formulated by the users) with the queries in the foreground. In this experiment, we use the CoST dataset (Dosso et al., 2021): an experimental dataset in French in which several participants were asked to perform the same IR tasks with a web search engine.

Our more precise question here is to know which NLP (Natural Language Processing) models can best be used to represent these data and especially queries given their characteristics. Indeed, queries are mostly expressed by keywords, even though more and more natural language formulations are submitted to web search engines (White et al., 2015). As a result, we seek here to identify how pretrained word embedding vector models, and in particular those dealing with word sequences (transformer models like BERT (Devlin et al., 2019)), can be used to effectively represent queries, knowing that they do not match the training data of language models and that we do not have sufficient data for specific retraining or finetuning. We also want to apply an agnostic method, without focusing on a precise characteristic for queries. More specifically, we seek here to test the ability of these models to capture the semantic relationships (in broad sense) between queries within a search session. The question can be asked at two levels: locally, to identify the different types of reformulation (e.g. to detect whether the user is trying to specify or generalise his search (Rieh and Xie, 2006) and if so in which direction, or whether he is staying in the same search space or initiating a new one), and more globally, to identify the global strategy and the dynamics of the search. Our long-term objective is to identify user profiles based on the study of behavioural variations in experimental data. In particular, we are looking at how users explore the semantic space of a search session. As a result, we are seeking to construct a neutral representation without pre-training models to bring out broad behavioural variations and explore these variations.

Table 1[1] shows three sessions corresponding to the same information need (concerning the comparison of different NLP methods for plagiarism detection[2]). We can note the usual characteristics of

[1] These sessions have been partially modified and simplified for the purposes of illustration.

[2] Task: As part of your Masters internship, you aim to

Ibarboure, C., Tanguy, L. and Amadieu, F.
Which Word Embeddings for Modeling Web Search Queries? Application to the Study of Search Strategies.
DOI: 10.5220/0012177600003598
In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023) - Volume 1: KDIR*, pages 273-280
ISBN: 978-989-758-671-2; ISSN: 2184-3228

queries (orthographic approximation, free syntax), rewriting operations (correction, paraphrases) but also reformulations with different levels of semantic variation. Semantic variation is one of the elements that we can analyse in the study of user behaviour, in order to identify user strategies. Above all, in these examples, we observe varied global behaviours: in session 1 the user concentrates on the central problem, which he completes by modifying part of the query. Conversely, in session 2, the user scans the different notions mentioned in the task statement, not delving into any of these notions except from query 4 to 5. The last session shows the user's interest in the "plagiarism detector", interrupted by a query concerning "thesaurus", to return to the previous notion.

In this article, we propose an exploratory approach to represent the semantic variations between queries in order to identify the overall strategies used at the session level with the aim of proposing a typology. To bring out this type of global strategies, the first step we detail in this article consists in identifying the way in which queries will be represented by vectors (one per query). Our objective is to find a representation of the queries that allows us to compute the different facets of a search session.

To do this, we propose two ways of comparing models on their ability to:

1. distinguish the broad domains associated with queries,
2. evaluate the correlation between a manual annotation of reformulations that could correspond to a measure of semantic distance, and a measure of similarity calculated automatically by vector models.

In this article we present a brief review of the state of the art dealing with the variation of behaviours through the study of queries, vector representations of queries in IR, and knowledge of the inner mechanics of contextual vector models. We then present the methodology of our studies and the data used. We then look at the results before concluding with a discussion.

Table 1: Example of search session for computer science decision making tasks - (Translation of sessions in italics).

	Session1
1	analyse texte methodologie
2	analyse texte thesauraus
3	analyse texte anti plagiat methode
4	analyse texte anti plagiat méthodes
5	analyse texte anti plagiat n-grammes
1	*text analysis methodology*
2	*text analysis thesauraus*
3	*text analysis anti plagiarism method*
4	*text analysis anti plagiarism methods*
5	*text analysis anti plagiarism n-grams*
	Session 2
1	détection de plagiat méthode
2	n-grammes
3	thesaurus
4	word embedding
5	word embedding n-grammes
1	*plagiarism detection method*
2	*n-grams*
3	*thesaurus*
4	*word embedding*
5	*word embedding n-grams*
	Session 3
1	détecteur de plagiat
2	programme détecteur de plagiat
3	conception programme détecteur de plagiat
4	thesaurus plagiat
5	détection de plagiat
1	*plagiarism detector*
2	*plagiarism detector program*
3	*design plagiarism detector program*
4	*thesaurus plagiarism*
5	*plagiarism detection*

2 RELATED WORK

Several authors are interested in search sessions, in particular through the study of queries. The interest of being able to represent search sessions, and by extension to understand the interactions between the user and the search engine, is to allow the improvement of browsers interfaces, particularly in relation to autocompletion or query prediction.

2.1 Study of Search Behaviour Through Queries

Search sessions provide a space for analysing behaviour and, in particular, variation.

Queries have often been analysed to study variation in user behaviour. Some studies propose ty-

develop a plagiarism detection program. You would like to set up a text analysis methodology but you are hesitating between the simple use of text morphology (words, n-grams, sentences etc.) or the use of external resources (dictionaries, thesaurus, Word embeddings). After outlining the advantages and disadvantages of each type of analysis, select the method that seems best to you and justify your choices.

pologies of queries enabling analysis of semantic or surface variation (Rieh and Xie, 2006; Huang and Efthimiadis, 2009; Adam et al., 2013). These typologies provide a linguistic criterion (e.g. use of semantic relationships (hyperonym, synonym, etc.)) on which to study behavioural variations.

Recent research is attempting to study user behaviour during reformulation phases in order to propose the best search techniques and subsequently personalised search engines for users (Chen et al., 2021). Some works try to understand the intentions and behaviours of users during an IR activity in order to provide new applied prospects (Liu et al., 2019).

To study these variations, we seek to qualify search sessions with vector models. Some studies have already dealt with the vector representation of IR elements (e.g. sessions, URLs, etc.)

2.2 Representation of Queries in IR

Studies have generalised the use of NLP tools such as neural models to study search sessions. Vector models are used for various studies on sessions and on reformulations made by users on search engines.

Many studies have focused on how to represent queries. Mitra (2015) is interested in the vector representation of queries and reformulations as part of work on query prediction or auto-completion. Mehrotra and Yilmaz (2017) uses the context of the search task to try to provide a better representation of queries. Other works will also build enriched semantic spaces including different elements associated with the session, such as URLs for web searches (Bing et al., 2018).

To represent our search sessions and queries, we are testing contextual vector models (BERT-type). We know that these models have complex structures that have been of interest to a number of researchers.

2.3 Transformers Architecture

Transformer neural models have now become the main tool used in NLP for processing any kind of data for any task. Most of the time pretrained models are integrated in a classifier and trained on a specific configuration. But transformer models can also be used for their ability to provide a vector representation of the input text data.

In fact, transformer-based models provide access to the weights of the different hidden layers of the neural network. A lot of work continues to be published on the study of BERT like models in order to understand how capture the different aspects of language and in particular at which layer(s) of the neural network certain linguistic information is acquired during training. To this end, many studies have focused on the analysis of the different layers of BERT (i.e. ”BERTology”).

The 0^{th} layer (input) is described by Ethayarajh (2019) as a non-contextualised layer which is used as a reference for the comparison of other layers in contextualisation work. We know that surface information is found in the early layers of BERT (Jawahar et al., 2019). From the work of Lin et al. (2019), cited by Rogers et al. (2020), up to the fourth BERT layer of the base model, the latter relies in particular on linear word order. The middle layers refer in particular to syntactic features (Rogers et al., 2020; Jawahar et al., 2019). From the upper layers, semantic features emerge and the models will have representations particularly related to the context (Jawahar et al., 2019; Mickus et al., 2020; Ethayarajh, 2019).

In our study in which we apply pretrained transformer models to search queries, it is therefore very important to carefully choose the layers used (and how they are combined).

3 METHODOLOGY

We will first present the data, then the different vector models used to represent them. Our scheme compares the models by considering the queries from two different levels: at a higher level, they are associated with a general domain (search topic set by the task) and more locally according to the semantic relations between two consecutive queries (reformulation).

3.1 Experimental Data

In this study, we work on experimental data. This allows us to have clearly delimited search sessions with a clear beginning and an end, since the user has to perform a complex search around a predefined topic. These data are an interesting alternative to ecological data from search logs that require a significant effort to reconstruct sessions (Gomes et al., 2019) and we are able to better control the divergences and discontinuities of a multi-task IR activity (Mehrotra and Yilmaz, 2017).

The CoST dataset was collected by Dosso et al. (2021) in the context of a work in cognitive psychology and ergonomics. For this purpose, the authors set up a protocol requiring the completion of fifteen information search tasks of different complexities in three research topics: computer science, cognitive psychology/ergonomics and medicine. Participants had to perform a web search using the Google engine, and all their actions were tracked and timed (queries, number of result pages observed, URLs visited).

From the available data, we selected nine tasks (three per domain) among the most complex ones in order to maximise the size of the sessions studied. These were the multi-criteria (Bell and Ruthven, 2004), problem solving and decision making tasks (Campbell, 1988). 18 participants were randomly selected to reduce the computation load. The dataset used here represents a total of 162 search sessions, or 1262 queries. These sessions varied in size depending on the complexity of the task, but also according to the search domain. On average, the sessions are made up of 8 queries, the longest has 48 and 27 of the sessions contain only one query.

The sessions (and therefore the queries they contain) are divided into three very distinct topics, which involve very different notions, lexicons and themes. It is this first level of distinction between queries that we use to compare embeddings.

The second level concerns the pairs of consecutive queries of the sessions: it is a qualification of the reformulation operation. The annotation available in the dataset is based on the distinction proposed by Sanchiz et al. (2020) between *exploration* and *exploitation*. Exploration is qualified as the initiation of a new search space, represented by a significant semantic ”jump” between two queries. Conversely, exploitation is seen as the pursuit of a search path. The annotation applies a further distinction between exploitation and narrow exploitation. In the end, each pair of consecutive queries in the same session is qualified on a 4-level ordinal scale: exploration (large ”jump”), exploitation (intermediate ”jump”), narrow exploitation and surface reformulation when no semantic change is made (e.g. spelling correction), as illustrated in Table 2.

Table 2: Example of annotations.

	Session	Annotation
1	plagiarism detector	–
2	plagiarism detector program	Exploitation (2)
3	plagiarism detector design	Narrow exploitation (3)
4	anti-plagiarism text analysis method	Exploration (1)
5	anti-plagiarism text analysis methods	Spelling correction (4)

3.2 Pretrained Embeddings

We tested two types of vector models to represent queries. To begin with, we used FastText (Grave et al., 2018), a so-called ”static” (or type-level) model where a single vector is associated with a word in the vocabulary without taking into account its context. The interest of Fastext among other static models (such as Word2Vec) lies in its ability to propose a representation for the frequent out of vocabulary (OOV) query terms in our data (proper nouns, typos, etc.). The model used was trained on Common Crawl (around 12M words) and Wikipedia with the CBOW method and the following hyperparameters: 300 dimensions, character n-grams of length 5 and a window of size 5. To represent the queries, we computed the average vector of the tokens that compose it based on a simple tokenisation on spaces. It should be noted that the corpus used is a generic corpus. As a reminder, we did not pre-train the models because we are taking an exploratory approach to see how the models represent queries.

We also used contextual models. Unlike static models, these models assign a vector to a word according to its context, i.e. the other words in the sentence and therefore for us in the query, by exploiting their relative position. We decided to use the basic models of CamemBERT and FlauBERT, two variants of BERT (Devlin et al., 2019) pretrained for French. These are the most commonly used lighweight generic French models. The CamemBERT base model was trained on the OSCAR corpus extracted from Common Crawl (Martin et al., 2020). The FlauBERT base model was trained on different sources: Wikipedia articles, novels or texts from Common Crawl (Le et al., 2020). In both cases we used the base model (12 layers of 768 dimensions), and we did not perform any fine-tuning. Indeed, as we said earlier, there is no specific data large enough to allow us to fine-tune the models. Our goal is not to identify the best model to use, but to determine their ability to build a query representation from their original training data.

To represent each query by a single vector, we considered two different strategies commonly used. The first is to use the vector corresponding to the [CLS] token which can be used as a support for the global representation of the word sequence it delimits; the second is, as for static models, to compute the average vector of the query tokens (Rogers et al., 2020).

Given the specificity of our data, which are not necessarily sentences (and rarely complete sentences), the question of which layers are most relevant remains and we therefore wanted to test a large number of configurations. In the end, we tested all possible subsets of the 13 layers (including the input layer) of the CamemBERT and FlauBERT models. For combinations of 2 or more layers, we tested the mean and the concatenation of the vectors. In total we compared 65477 different ways of representing the queries.

3.3 Comparison Methods

As mentioned above, we confronted the query representations with two external features (search domains, and similarity between reformulations). At this stage our aim is not to build a predictive model by training a model, but rather to understand how the features presented above are correlated to the representation of queries. In both cases, we need to define a similarity measure between the vectors, and we considered two standard metrics: Euclidean distance and cosine distance.

We measure the abilities of the models to capture the coarse-grained topic of a given query because we assume that query terms belong to different and clearly delimited semantic lexical classes. By clustering queries on these topics, we aim to see whether these semantic classes are captured by the models. For the clustering by topic, we performed a hierarchical ascending classification (HAC) of the queries into three clusters that we compared to the three domains of the dataset (psychology, computer science, medicine) using the Rand index. We only considered distinct queries within the same domain and removed duplicates. This left us with three groups of queries of almost similar size: 332 queries for psychology, 348 for computer science and finally 313 for medicine. The Rand index score is a simple measure of the aligment between these 3 groups and the 3 clusters obtained based on query similarity.

For the second task, we used the manual annotations described above and compared, for each pair of consecutive queries in the same session (1101 in total). We believe that the annotation scale can be seen as a semantic distance between two queries. We consider that exploration corresponds to a greater semantic distance than exploitation, which corresponds to a smaller distance. We rate the annotations on a scale from 1 (exploration) to 4 (surface correction). The comparison between these annotations and the distance between the vectors of the two queries is measured by the Spearman correlation coefficient.

4 RESULTS

4.1 Clustering by Topics

The highest score, normalised Rand index of 0.61, is achieved by the FastText model (with Euclidean distance). The highest score achieved by the contextual embeddings is 0.53. It is impossible in this study to have a perfect rand index as there are common queries (e.g. so-called navigational queries, e.g. *google scholar*) across the different search domains.

This result is not necessarily very surprising, as domain membership is directly related to the lexicon present in the queries, with little need for disambiguation or contextual representation. If we look in more detail at the results of the contextual models, we see that the highest score is obtained by the FlauBERT model with the average vector of tokens in a query and a combination by average of layers between 0 (input layer) and 3. For CamemBERT, we see the best results with low layers, although we also have some good results with high layers.

To conclude on this task, it is therefore the static models whose representation of the queries comes closest to a categorisation by topic since they are essentially based on the similarity between the isolated words. As for contextual models, those based on the lower layers have a similar behaviour. The representations based on the high layers, which we know are capable of capturing a certain abstraction of the content of the queries, lead to a grouping of queries on different bases than the simple domain.

However, this task remains trivial compared with the second task which is much more qualitative in its approach.

4.2 Distance Between Consecutive Queries

As a reminder, we calculate the correlation between the vector based similarity measure and the manual annotation that may correspond to a semantic similarity scale, to compare the representation of queries by models.

The highest correlation coefficient obtained between the similarity of the vectors and the semantic annotations of the query pairs is 0.77. This is achieved with the FlauBERT model, which slightly exceeds the static model, which obtained a coefficient of 0.75 (score obtained with the Euclidean distance and cosine distance). FlauBERT gives several results around this maximum value of 0.77 with mostly low layers averaged between 0 and 6 for the mean vector of tokens and Euclidean distances or cosine distance.

In order to provide a more detailed analysis, we applied a multiple linear regression on all the data. The dependent variable was the correlation score between the manual annotation and the similarity measure, and the explanatory variables were all the parameters described above and the models detailed in Table 3.

We observed a *t-value* of -735.86 for FlauBERT compared to CamemBERT. This shows that FlauBERT has an overall negative impact on

Table 3: Linear regression analysis: t-values of all studied parameters.

Parameters	***t-value***
FlauBERT (vs CamemBERT)	**-735.861**
Euclidean distance (vs cosine distance)	4.452
[CLS] token vector (vs average query token vector)	**-365.965**
Concatenation of layers (vs mean)	-141.862
Layer 0 included	**43.197**
Layer 1 included	**139.634**
Layer 2 included	**110.573**
Layer 3 included	**90.677**
Layer 4 included	**74.765**
Layer 5 included	**67.597**
Layer 6 included	**30.632**
Layer 7 included	-29.334
Layer 8 included	-61.279
Layer 9 included	-75.202
Layer 10 included	-69.398
Layer 11 included	-72.712
Layer 12 included	-267.334

correlation, unlike CamemBERT, which reports better results overall. In terms of vector type, we had a *t-value* of -365.97 for the vector [CLS] token. It is therefore preferable to use the average vector of the query tokens to represent the queries. For similarity measures, it seems more appropriate to use a Euclidean distance. It is preferable to average the layers rather than concatenate them (*t-value* = -141.86). For best results, it is advisable to use mainly layers 0 to 5, and a bit of the 6th layer, with priority given to layers 1 and 2, which have *t-values* of 139.63 and 110.57 respectively.

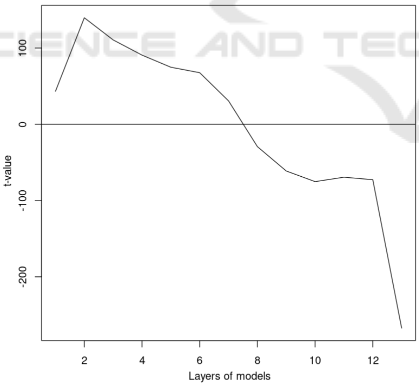

Figure 1: Overall impact of each layer on the correlation score.

Figure 1 represents the *t-value* for each layer. We can see that from layer 5 onwards the results decrease, crashing completely with the higher layers.

In general, the vector representation which are most distant to manual annotation are the [CLS] vector and the FlauBERT model for this task. To get closer to the manual annotations it is generally preferable to build the representations by taking the average of the low layers (between 0 and 5, and a bit 6) for the mean vector and the Euclidean distance.

We also used linear regression, focusing on one model at a time. The results are quite similar to the results for the whole data set. We can only note that the layers to be favoured for CamemBERT are included between 0 and 3, and for FlauBERT we can use layers from 0 to 6. Figure 2 shows the distribution of results between the models. We can see that the best results obtained by FlauBERT are atypical cases. Conversely, the worst results obtained by CamemBERT are atypical cases. Overall, we can therefore conclude that CamemBERT correlates better with manual annotation.

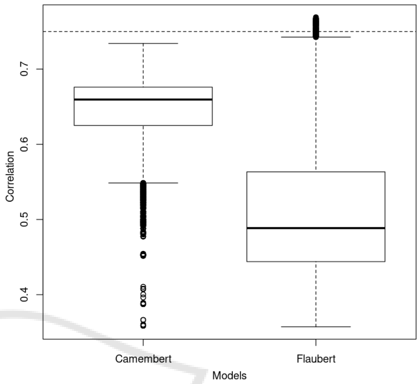

Figure 2: Comparison of models by correlation - dotted line for FastText result.

To conclude on this task of correlating a manual annotation with a similarity measure calculated automatically, CamemBERT seems to give better results overall. It is advisable to use mainly low layers and average vector of query tokens to represent queries. However, these models are highly dependent on the parameters to be selected. In addition, FastText is both very similar in terms of semantic similarity and is less expensive to use.

5 CONCLUSION

To conclude, pre-trained models do capture the similarity between search queries and therefore can be used for more refined explorations, even if they have been trained on generic language data that can differ widely from our target.

Mainstream transformer-based models can be used, but they are highly dependent on a number of choices that need to be made. We have observed that the average vector of query tokens and the use of the lower layers are more correlated with the semantic similarity between consecutive queries. For the type of data and benchmark we used, the abstraction computed by the upper layers of the transformers are not relevant.

These results can perhaps be explained by representations of these layers that are too abstract for our data. This can also be explained by the non-canonical character of the word sequences that form the queries, but more precisely by the fact that some queries are very precisely composed of juxtaposed words without any explicit syntagmatic link (e.g. ”thesaurus plagiarism”, as occurrences in a standard corpus would require at least a preposition).

We showed that using the default average of upper layers (or even the sole output layer) of transformer models is not the most efficient way to obtain semantic-aware embeddings of search queries.

At this stage we will therefore favour the representations proposed by Fastext, averaging the type-level vectors of each query terms. This method also has the decisive advantage of being much less expensive to compute.

6 FUTURE WORK

Returning to our exploratory objective, a search session (minimally defined as a sequence of queries) can be very simply visualised as a trajectory in a semantic vector space, along the lines of what was proposed by Mitra (2015). Figure 3 shows the sessions presented in Table 1 in this form, using a principal component analysis to project the different vectors in two dimensions (with a represented variance of 19.89%), with arrows connecting successive queries and colours distinguishing the different sessions.

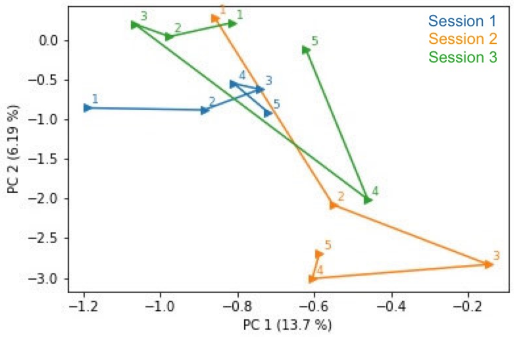

Figure 3: Examples of sessions projected into vector space.

To understand this figure, we can try to retrace the path of the sessions. If we look at session 1 (in blue in the figure), we can see that it is very close to queries 3 and 4. This represents the very close similarity between the two requests, which differ by one morphological variation and one spelling correction on one of the words (”*methode*” – ”*méthodes*”). Overall, this session does not extend over a large part of the semantic space. Session 2 (in orange in the figure), on the other hand, occupies a larger area. We can see quite significant variations between the first three queries, compared with queries 4 and 5 which are very similar (differing by only one term: ”word embedding” – ”word embedding n-grams”). Finally, session 3 (in green in the figure) is represented by a loop. In fact, after the first three fairly similar localised queries, we have a significant variation with the 4th query corresponding to a major change with the change from ”detector program design” to ”thesaurus”. However, the user seems to return to the notion of detection in the first query, represented by a trajectory from the query to the origin of the session.

We see that the shape of the trajectory is a telling indicator of the overall strategy adopted by the user. However, it remains difficult to associate interpretable semantic operations with these trajectories.

At present, we are planning to deepen our approach to study behavioural variations. We want to study variations in user processing of thematic spaces. We define a thematic space as a search axis referring to a precise theme with a precise semantic content. We are building a new dataset in French with complex search tasks. In the statements of these tasks, two thematic spaces are distinguished (e.g. a task requiring detailed information on both Greek mythology and Italian Renaissance painting). We observe how these spaces are processed by users at the session level. We study, for example, the presence or absence of these spaces, their chronology of appearance or the separation of the spaces across queries.

This approach may enable us to identify user profiles for the exploration and organisation of the thematic search space. With a neutral representation as presented in this paper, we can continue our exploratory approach with the study of thematic spaces and thus see how the models capture these different spaces.

REFERENCES

Adam, C., Fabre, C., and Tanguy, L. (2013). Etude des relations sémantiques dans les reformulations de requêtes sous la loupe de l’analyse distributionnelle. In *SemDis (enjeux actuels de la sémantique distributionnelle) dans le cadre de TALN 2013*, page (publication en ligne), Sables d’Olonne, France.

Bell, D. J. and Ruthven, I. (2004). Searcher’s assessments of task complexity for web searching. In *European conference on information retrieval*, pages 57–71. Springer.

Bing, L., Niu, Z.-Y., Li, P., Lam, W., and Wang, H. (2018).

Learning a unified embedding space of web search from large-scale query log. *Knowledge-Based Systems*, 150:38–48.

Campbell, D. J. (1988). Task complexity: A review and analysis. *The Academy of Management Review*, 13(1):40–52.

Chen, J., Mao, J., Liu, Y., Zhang, F., Zhang, M., and Ma, S. (2021). *Towards a Better Understanding of Query Reformulation Behavior in Web Search*, page 743–755. Association for Computing Machinery, New York, NY, USA.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dosso, C., Moreno, J. G., Chevalier, A., and Tamine, L. (2021). *CoST: An Annotated Data Collection for Complex Search*, page 4455–4464. Association for Computing Machinery, New York, NY, USA.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Gomes, P., Martins, B., and Cruz, L. (2019). Segmenting user sessions in search engine query logs leveraging word embeddings. In *Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings*, page 185–199, Berlin, Heidelberg. Springer-Verlag.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Huang, J. and Efthimiadis, E. (2009). Analyzing and evaluating query reformulation strategies in web search logs. pages 77–86.

Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Lin, Y., Tan, Y. C., and Frank, R. (2019). Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

Liu, J., Mitsui, M., Belkin, N. J., and Shah, C. (2019). Task, information seeking intentions, and user behavior: Toward a multi-level understanding of web search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, page 123–132, New York, NY, USA. Association for Computing Machinery.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Mehrotra, R. and Yilmaz, E. (2017). Task embeddings: Learning query embeddings using task context. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 2199–2202, New York, NY, USA. Association for Computing Machinery.

Mickus, T., Paperno, D., Constant, M., and van Deemter, K. (2020). What do you mean, BERT? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290, New York, New York. Association for Computational Linguistics.

Mitra, B. (2015). Exploring session context using distributed representations of queries and reformulations. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 3–12, New York, NY, USA. Association for Computing Machinery.

Rieh, S. Y. and Xie, H. I. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management*, 42(3):751–768.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sanchiz, M., Amadieu, F., and Chevalier, A. (2020). An evolving perspective to capture individual differences related to fluid and crystallized abilities in information searching with a search engine. In Fu, W. T. and van Oostendorp, H., editors, *Understanding and Improving Information Search: A Cognitive Approach*, pages 71–96. Springer International Publishing, Cham.

White, R. W., Richardson, M., and Yih, W.-T. (2015). Questions vs. Queries in Informational Search Tasks. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 135–136, New York, NY, USA. Association for Computing Machinery.