

Document Ranking Applied to Second Language Learning

Rodrigo Wilkens(✉), Leonardo Zilio, and Cédrick Fairon

Université catholique de Louvain, Louvain-la-Neuve, Belgium
{rodrigo.wilkens,leonardo.zilio,cedrick.fairon}@uclouvain.be

Abstract. This paper addresses the needs of language learners and teachers by combining keyword-based search and language level information on an algorithm that can rank documents by pertinence to the required topic (keywords) and adequacy to the user's language level. We conducted several experiments using the EF-CAMDAT corpus (annotated for topic and level) and we observed that the best ranking results were an average of BM25 and linguistic information. We also saw that the grammar of level C1 is the best indicator for level. Finally, we proposed a customization for prioritizing beginner or intermediate levels at the top of the rank.

1 Introduction

The retrieval of interesting documents based on a given set of keywords is a trivial event in the everyday life of almost everyone and it also applies to the activity of learning a new language, especially when this kind of behavior is encouraged or requested by the language teacher [1]. As indicated by [2], the pedagogical approach of flooding learners with language information is extensively used by language teachers. However, the search for relevant reading material takes time and effort, and, as a result, teachers often resort to schoolbooks, which are readily available, but have a limited text choice and are possibly not up-to-date and hardly in line with students interests.

When the task concerns only the search for keywords, most search engines can provide the users with a good set of texts that will suffice for their needs, but some of the texts will not be accessible to those with a lower language proficiency [3]. As such, when dealing with the need for documents that, at the same time, are relevant according to the given keywords and match certain constraints of text complexity, then further re-ranking of the documents may be needed to address these new criteria. This could be the case, for instance, when the user has a low-level reading capacity, has a disease that affects reading, or is a language learner. The latter case is the main focus of this paper.

By retrieving documents that are relevant in terms of both topic and language complexity it would be possible to address the needs of a broader range of users. From the teachers' or learners' perspective, it would be possible to retrieve texts that are better tailored to their needs, whether being more engaging for the

© Springer International Publishing AG, part of Springer Nature 2018
G. Pasi et al. (Eds.): ECIR 2018, LNCS 10772, pp. 618–624, 2018.
https://doi.org/10.1007/978-3-319-76941-7_53

learners or presenting relevant information for an extra class task. As such, the use of level and topic-oriented document retrieval could better inform the acquisition of second language in both fronts.

Taking all of this into consideration, this paper aims at presenting a system that retrieves documents based on a keyword-oriented query and ranks them according to topic and language level. We further hypothesize that both text content and level are complementary for the ranking process.

This paper is divided as follows: Sect. 2 presents search engines that address some level of language complexity together with topic; Sect. 3 describes the corpus that was used and the methodology employed in the experiments; Sect. 4 shows the results of our experiments; and, finally, Sect. 5 contains our final remarks and perspectives for future work.

2 Related Work

There are several systems that attempt to retrieve texts based on topic of interest while also supporting language learning activities. Here we present a selected few, which address topic and, in varying degrees, language level: SourceFinder [4], REAP (Reader-Specific Lexical Practice for Improved Reading Comprehension) [5], FLAIR (Form-Focused Linguistically Aware Information Retrieval) [2], and READ-X [6]. These systems are similar in purpose, but vary in their methods, which can be better compared by observing the following dimensions: text source, text representation, language model, and target user.

Regarding the text source, they can be divided in systems based on the Web (REAP, FLAIR, and READ-X) or on closed corpus (SourceFinder). The Web-based systems delegate the document retrieval to a Web-based search engine, and then re-rank the results. In contrast, closed corpus systems run their own search criteria over the whole corpus. The closed corpus approach can have a lot of information annotated off-line, but disposes of fewer documents in comparison to the Web-based approach.

Moving on to the dimension of text representation, some systems (REAP, FLAIR, and SourceFinder) index and rank the documents according to a series of linguistic information, while systems like READ-X use only one type of measure. Linguistic information can involve, among others, part-of-speech tags, length-based counts, and readability measures. Topic annotation is usually automatic in these systems, but they differ in the source and number of tags.

An important dimension for all of the systems is how they model the learners language skills. Most of them model the learners capacity in terms of a readability measure, but others further explore this dimension by using a language learning framework (e.g. FLAIR is strongly based on the CEFR [7]). The readability measures are used in different ways: for instance, READ-X annotates the texts with classic measures and presents them to the user as parameters for queries, while REAP applies readability measures that use dictionary information.

Finally, there is the intended target user. READ-X does not specify its target users, but they need to understand about readability measures for interpreting

the output and selecting the adequate thresholds. SourceFinder and FLAIR focus on a teachers audience, providing support in their activities, while REAP focuses on both sides of the spectrum, and is designed for both teachers and learners. All of these systems require that the users have a deep understanding of their language model. While this assumption is acceptable for language teachers, it is not for the learners.

3 Methodology

In this paper, we want to use document ranking to address the needs of learners and teachers for texts that correspond to a topic, but that are also adequate for a specific language level. In this section, we present the procedures that we carried out and the resources we used for achieving our main goal.

The European Union developed the Common European Framework of Reference for Language Learning (CEFR) [7], which provides guidance for the teaching of foreign languages. The CEFR presents a description of communication goals that a learner should achieve in each of six main levels: A1, A2, B1, B2, C1, and C2, which are further subdivided into three levels.

For evaluating the system's ranking algorithm, we selected the EF-CAMDAT [8] corpus, which presents a collection of texts produced by learners of English from different levels of proficiency, according to the CEFR. The corpus is divided according to the CEFR and contains a total of 532 thousand documents (33 million tokens) by 83,385 learners from 137 countries. Each document is linked to a specific topic (e.g., writing an invitation, or describing an experience), and similar topics are generally addressed on a single CEFR level.

For testing our hypothesis that text content and level are complementary for ranking documents, we selected only documents that belong to the same topic, but spreading across different, non-adjacent CEFR levels. This selection was to ensure that the documents were not too close to each other in terms of level, so as to avoid errors derived from bias in human judgment. By applying these strong constraints for the selection of documents, we obtained three valid topics. From each of them, we extracted the first 10 noun keywords based on the higher log-likelihood [9].

For ranking the documents by keywords, we opted for BM25 [10], which is a bag-of-words retrieval function that ranks documents using a query of terms. We used the Lucene implementation of BM25 [11], but, for ranking the documents by language level, we substituted the score function of BM25 for linguistic information. The full corpus was annotated with linguistic information, such as grammar level, vocabulary level [1], and three readability measures (Dale-Chall Score [12], Flesch-Kincaid measure [13], and Gunning Fog Index [14]). After that, we proceeded with three experiments for ranking the documents:

[1] For the grammar and vocabulary of level C1, we used the inversed score, because we expect that an A-level document would present them in a lesser degree.

Ranking score: We evaluated the results of different ranking algorithms (BM25 alone, linguistic information alone, and our model, which is the average of the previous scores) to observe which one renders the best results.

Linguistic information: We evaluated which linguistic information could best predict the language level of a document, a task that is hard to approximate automatically.

Search customization by language level: We proposed a method for ranking the documents in a way that a selected language level is given preference.

All of these experiments were evaluated by Spearman's rank correlation according to: (1) the rank of BM25 for topic, and (2) the rank of documents by level assessment scores. We also evaluate the results using precision at ten (p@10): for topic, we identified as correct when a document had the expected topic, and, for level, we used the lowest language level retrieved. As terms of comparison, we considered both extremes: the standard BM25 score, which prioritizes the topic, and the scores from linguistic information, which prioritize the level.

4 Evaluation

The results for the three experiments can be seen in Table 1. The table presents correlation and precision at ten (p@10) of a given feature in relation to the topic of the retrieved documents and their language levels.

At the top of the table we have information about the ranking score experiment. Comparing to BM25, the average model was a bit worse in terms of topic, but improved the level ranking in correlation and had 100% of p@10 for both level and topic. On the theoretical level, this confirms our hypothesis that averaging BM25 and linguistic information improves the document ranking for the purposes of addressing language level. In a more practical approach, the next experiment addresses the extent to which linguistic measures can approximate language level.

Regarding the linguistic information experiment, in which we used linguistic features to model the level, we can see that, in terms of p@10, many features present 100% for both level and topic. As such, for a learner that is only interested in the top documents, the average model can achieve excellent results using several features individually. However, when we consider the correlation, which takes into account all of the retrieved documents, we see that the best feature, specially for level, was the grammar of C1. Comparing to the pure linguistic model, the average model also consistently improves the scores both in correlation and in p@10. These results further support the hypothesis for the average model, even in a practical setting, when level assessment is not provided and linguistic features are the only clues for approximating it.

For the last experiment, the ranking using a specific level, we focused on beginner and intermediate levels[2]. We modeled the beginner level in the same

[2] In this work we did not address advanced-level learners, because they are supposed to use traditional search engine systems.

way as in the previous experiments, and, for the intermediate level, we averaged the score from the beginner level with its inverse, which is assumed to be the score of advanced level, for targeting the middle of the CEFR spectrum. The results showed that we had mainly intermediate-level documents as best representatives of the topic queries, and, as such, the BM25 score achieves better results both in correlation and in p@10. Even so, if we want to ensure that intermediate-level documents are prioritized, the average model presented an improvement over isolated linguistic features. In a more practical scenario, the best feature to approximate intermediate level would be Dale-Chall score [12]. The results (on the columns Beginner Level and Intermediate Level of Table 1) show that the ranking equation did not have a negative impact in terms of topic, and achieved a high correlation in terms of level.

Table 1. Evaluation of document retrieval by topic and level

Model	Beginner level								Intermediate level							
	Avg model				Linguistic model				Avg model				Linguistic model			
	Correlation		p@10 (%)		Correlation		p@10 (%)		Correlation		p@10 (%)		Correlation		p@10 (%)	
	Topic	Level	Topic	Level	Topic	Level	Topic	Level	Topic	Level	Topic	Level	Topic	Level	Topic	Level
$BM25$	**100**	56	**100**	56	**100**	33	**100**	33	**100**	**100**	**100**	**67**	**100**	**100**	**100**	**33**
Level	96	**69**	**100**	**100**	86	**75**	87	**100**	96	69	**100**	50	86	74	87	7
$A1_{gram}$	95	61	**100**	**100**	82	59	**100**	**100**	81	26	**100**	**67**	45	−2	63	23
$A1_{vocab}$	94	57	**100**	**100**	77	49	**100**	**100**	81	26	**100**	**67**	42	−4	30	3
$A2_{gram}$	75	26	90	**100**	14	−10	43	**100**	79	50	**100**	60	27	31	20	7
$A2_{vocab}$	93	56	**100**	**100**	76	48	93	**100**	83	31	97	**67**	46	1	3	3
$B1_{gram}$	74	25	70	**100**	11	−11	37	**100**	80	44	**100**	43	30	21	17	3
$B1_{vocab}$	87	42	**100**	**100**	54	22	67	**100**	79	23	97	33	33	−10	33	3
$B2_{gram}$	68	30	**100**	**100**	−5	−4	90	**100**	80	60	83	57	28	48	10	20
$B2_{vocab}$	80	35	**100**	**100**	29	6	70	**100**	83	51	83	50	38	34	7	3
$C1_{gram}$	95	**69**	87	**100**	79	76	13	**100**	79	21	97	**67**	36	−11	17	3
$C1_{vocab}$	90	**69**	90	**100**	64	72	10	**100**	81	39	90	53	36	13	10	3
[12]	87	41	**100**	**100**	55	19	93	**100**	92	60	**100**	**67**	70	55	87	13
[13]	94	56	**100**	**100**	79	49	**100**	**100**	93	48	**100**	60	73	35	30	20
[14]	92	50	**100**	**100**	71	37	**100**	**100**	94	54	**100**	**67**	79	46	**100**	**33**

5 Conclusion

In this paper, we combined language learning with traditional document retrieval. We used a language learners corpus, which contains documents annotated with language level and topic. The experiments with ranking documents by topic and level showed us that averaging a traditional information retrieval algorithm and linguistic information renders better results from a language learning perspective. Among all linguistic features, we observed that the grammar of level C1 had better results for both topic and level in terms of correlation. Finally, when trying to customize the document ranking for prioritizing an intermediate language level, we concluded that the use of Dale-Chall score can approximate the best result, while also largely ensuring a correspondence with the language level order.

In general, the results of our experiments show that by averaging topic and language level, we have a complementary document ranking process. A more language level-oriented ranking of the documents tends to not affect much the ranking by topic, while allowing for a result that is useful for a learner.

In future studies, we intend to investigate more informative linguistic measures by applying a level assessment classification of documents (e.g., [15]). We also want to research the filtering of documents by language level assessment.

Acknowledgements. We would like to thank the Walloon Region (Projects BEWARE n. 1510637 and 1610378) for support, and Altissia International for research collaboration.

References

1. Purcell, K., Rainie, L., Heaps, A., Buchanan, J., Friedrich, L., Jacklin, A., Chen, C., Zickuhr, K.: How teens do research in the digital world. Pew Internet & American Life Project (2012)
2. Chinkina, M., Kannan, M., Meurers, D.: Online information retrieval for language learning. In: 2016 ACL, p. 7 (2016)
3. Vajjala, S., Meurers, D.: On the applicability of readability models to web texts. In: 2013 ACL, p. 59 (2013)
4. Passonneau, R., Hemat, L., Plante, J., Sheehan, K.M.: Electronic sources as input to gre® reading comprehension item development: sourcefinder prototype evaluation. ETS Res. Rep. Ser. **2002**(1) (2002). 66 pages
5. Collins-Thompson, K., Callan, J.: Information retrieval for language tutoring: an overview of the reap project. In: 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 544–545. ACM (2004)
6. Miltsakaki, E., Troutt, A.: Read-x: Automatic evaluation of reading difficulty of web text. In: E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, Association for the Advancement of Computing in Education (AACE), pp. 7280–7286 (2007)
7. Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N., North, B.: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press, Cambridge (2009)
8. Geertzen, J., Alexopoulou, T., Korhonen, A.: Automatic linguistic annotation of large scale l2 databases: the ef-cambridge open language database (EFCAMDAT). In: Cascadilla Proceedings Project on 31st Second Language Research Forum. Somerville, MA (2013)
9. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: Workshop on Comparing Corpora, pp. 1–6. ACL (2000)
10. Robertson, S.E., Walker, S.: Okapi/keenbow at TREC-8. In: TREC, pp. 151–162 (1999)
11. Pérez-Iglesias, J., Pérez-Agüera, J.R., Fresno, V., Feinstein, Y.Z.: Integrating the probabilistic models bm25/bm25f into lucene. preprint arXiv:0911.5046 (2009)
12. Dale, E., Chall, J.S.: A formula for predicting readability: instructions. Educ. Res. Bull. **27**, 37–54 (1948)
13. Kincaid, J.P., Fishburne Jr., R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch (1975)

14. Gunning, R.: The Technique of Clear Writing. Mcgraw-Hill, NY (1968)
15. Heilman, M., Collins-Thompson, K., Callan, J., Eskenazi, M.: Combining lexical and grammatical features to improve readability measures for first and second language texts. In: HLT 2007: The Conference of the NAACL, pp. 460–467 (2007)