Monolingual and Cross-lingual information retrieval in cultural Microblog at CLEF 2018

Chedi Bechikh Ali[1] and Hatem Haddad[2]

[1] Institut supérieur de gestion, Université de Tunis, Tunisia
chedi.bechikh@gmail.com
[2] Université libre de Bruxelles, Belgium
hatem.haddad@ulb.ac.be

Abstract. For CLEF 2018, we focus on cultural microblog search. The aim of this work is to find relevant microcritics in a monolingual and cross lingual context about films. This task is challenging due to the short length of the query and of the documents. For the monolingual context we propose to expand the query using a probalistic weighting scheme. For the french-english cross language task, we used a state of the art approach based on query transation.

Keywords: Microblog search · Query expansion · Query translation.

1 Introduction

The Cross Language cultural microblog search is the first task[3] from the lab multilingual cultural mining and retrieval (MC2).

The goal of this task is to find relevant microblogs in different languages from MC2 corpus using 75 topics related to films from a dataset of 70 000 000 microblogs. This corpus is collected between May and September 2015 and is about the keyword Festival. The topics used are selected by the task organizers from VodKaster website and represent a selection of microcrtitcs in french about film and cinema festivals. The topics are composed by a film title, a narrative field containing a microcritic about the film and a third field containing a list of expressions extracted manually from the microcritics.

Figure 1 shows an example of the query structure.

In this work we choose to intervene at the topics level, because it is simpler than to change the indexing shceme. For the monolingual search we based our work on a query expansion approach, for the cross lingual search we used query translation.

This paper is organized as follow, in section 2 we describe related work on query expansion and cross language information retrieval. Then in section 3 we describe the proposed approach used for query expansion and the query translation techniques used. We conclude this work in Section 4.

[3] https://mc2.talne.eu/

< *topic* >
< *id* > 201800 < /*id* >
< *title* > Phantom of the Paradise < /*title* >
< *narrative* > Palma d'or pour le festival de Swan < /*narrative* >
< *nugguets* > Palma d'or;festival de Swan < /*nugguets* >
< /*topic* >

Fig. 1. topic number 201800

2 Related work

In this paper we propose to use query expansion for monolingual microblog search, so we present a brief state of the art about query expansion. We present also a state of the art about the different approaches used for cross language information retrieval.

2.1 Query expansion

Given that user queries are usually short and that some words can be ambiguous, the use of the simple retrieval model based on the matching between query and document is prone to errors and omissions. Also the users of an information retrieval system can use other words than those present in relevant documents, this lead to the issue of term mismatch. Researchers proposed to use query expansion to resolve the problems caused by short queries and term mismatch.

Different approach were proposed [2]:

- Interactive Query Refinement
- Relevance feedback
- Word Sense Disambiguation in IR
- Search Results Clustering

In our work we rely on relevance feedback. For Relevance feedback, an initial retrieval run is performed using the initial formulation of the query. In addition, some number of top-ranked documents are mined for additional query terms to be added to the initial query. The content of the assessed documents is used to adjust the weights of terms in the original query and/or to add words to the query.

2.2 Cross language information retrieval

For cross language information retrieval (CLIR), given a query in language A, the goal is to retrieve documents in language B. CLIR can be useful in different contexts. For example, relevant documents may not exist in the query's language, so the system must be able to retrieve relevant documents in other languages. Many approaches were used for CLIR: query translation, document translation,

or the translation of both documents and query. Most researchers rely on query translation because it is easier than the translation of all the documents. For this purpose different approaches were used [1]:

- Machine translation
- Dictionary
- Parallel or comparable corpora
- Inter-language representation

These different approaches can be used for cross language microblog search and adapted to our task context. The translated queries are then executed against the target collection in a monolingual way.

3 Methodology

The documents search in microblogs presents many difficulties: the queries are short, so the retrieval system doesn't have enough contexts to understand user's needs. This can lead to different problems: word ambiguity and word mismatch between queries and documents. Also, we can notice that the documents (microcritics) are also short, so they lack of context.

We submitted 3 runs:

- A French monolingual run: "Be-Ha-submission-fr-fr"
- Tow french-english Cross lingual runs:
  - "Be-Ha-Submission3Fr-English-dictionary"
  - "Be-HaSubmussion2-FR-English"

For the monolingual run we used query expansion based on the probalistic BM25 model [4]. This process is based on many steps:

- Step 1: Original queries are used to retrieve the top 50 microcritics for each query. These retrieved microcrtics are used as a set of pseudo-relevant documents.
- Step 2: For each query, we merge the retrieved documents set in a single document.
- Step 3: Since expansion terms are selected from this set of supposed relevant documents, the new set of documents is used to compute the weight of each expansion term. The top 30 terms with the highest score are selected as expansion terms and added to the original query.
- Step 4: The new query is used to match the collection of microblogs.

We choose the BM25 model to attribute the score for each terms, because it is one of most accurate model for information retrieval. This model is based on a weighting scheme defined by this formula [4]:

$$Score(R|D) = \sum_{t \in Q} w^{(1)} \frac{(k_1 + 1)\, tf}{k1 \left((1 - b) + b \frac{|d|}{|d_{avg}|} \right) + tf} \frac{(k3 + 1) qtf}{k3 + qtf} \qquad (1)$$

where qtf is the number of times that the term t is present in the query, tf is the frequency of the term t in the document, $|d|$ is the number of terms in the document, $|d_{avg}|$ is the average length of a document, $k1$ is a parameter that controls the saturation of tf, $k3$ is a parameter that controls the saturation of qtf and b is the frequency normalization parameter.

$w^{(1)}$ is similar to the inverse document frequency (idf) and is defined by [**?**]:

$$w^{(1)} = log\frac{N - df + 0.5}{df + 0.5} \tag{2}$$

Where N is the number of documents in the collection and df is the number of documents where the term t occurs in the collection.

For cross language microblog search, we translate the query using tow techniques:

- A french-english dictionary
- Bilingual Wordnet.

For the run "Be-Ha-Submission3Fr-English-dictionary" we used a dictionary translation approach to translate the queries from the french to the english language.

A bilingual dictionary was used for query translation and it is constructed from an online dictionary. It consists of 33k distinct English words and 28k distinct French words, which constitutes 76k translation pairs. It contains lemmatized forms of content words (nouns, verbs, adjectives, adverbs).

For the run "Be-HaSubmussion2-FR-English" we used an inter-lingual representation based on english and french Wordnet. WordNet is a large lexical database of English that was extended to many languages. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept [3]. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

In wordnet each word is mapped to an indentifier, so words with the same meaning have the same number. For exemple the french word 'acteur' (actor) has as identifier "09765278-n", wich is the samed identifier for the word "actor", "actress", "performer", etc.

4 Conclusion

This paper describes our first participation on the first task Cross language Microblog search of the MC2 lab at CLEF 2018. The aim of this work is to find relevant microblogs given title and microcritic about films. In this work we submitted 3 runs, one french monolingual run and two cross lingual runs based on query translation. This work is still in progress and needs more investigations for future work, so other expansion approaches must be proposed, also the indexing process must be studied.

References

1. Bechikh-Ali, C., Haddad, H., Slimani, Y.: Cross-language information retrieval based on bilingual formal concept mining. In: 14 IACS/IEEE International Conference on Computer Systems and Application (AICSSA 2017), Hammamet, Tunisia, October 30-November 3, 2017. pp. 1–7 (2017)
2. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Comput. Surv. **44**(1), 1:1–1:50 (2012)
3. Fellbaum, C. (ed.): WordNet: an electronic lexical database. MIT Press (1998)
4. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994. pp. 109–126 (1994)