Local Document Relevance Clustering in IR Using Collocation Information

Leo Wanner*, Margarita Alonso Ramos*

* ICREA and Pompeu Fabra University
Passeig de Circumval·lació, 8; 08003 Barcelona, Spain
leo.wanner@upf.edu
* Faculty of Philology, University of La Coruña
Campus de Zapateira 15071 La Coruña, Spain
lxalonso@udc.es

Abstract

A series of different automatic query expansion techniques has been suggested in Information Retrieval. To estimate how suitable a document term is as an expansion term, the most popular of them use a measure of the frequency of the co-occurrence of this term with one or several query terms. The benefit of the use of the linguistic relations that hold between query terms is often questioned. If a linguistic phenomenon is taken into account, it is the phrase structure or lexical compound. We propose a technique that is based on the *restricted lexical cooccurrence* (*collocation*) of query terms. We use the knowledge on collocations formed by query terms for two tasks: (i) document relevance clustering done in the first stage of local query expansion and (ii) choice of suitable expansion terms from the relevant document cluster. In this paper, we describe the first task, providing evidence from first preliminary experiments on Spanish material that local relevance clustering benefits largely from knowledge on collocations.

1. Introduction

The performance of an Information Retrieval engine significantly depends on the quality of its query expansion technique. The most popular expansion techniques are based on term co-occurrence: document terms that are found to co-occur significantly often with query terms are considered suitable expansion terms.[1] Depending on the strategy adopted, the query terms are processed in isolation (Sparck Jones, 1971; Schütze and Pedersen, 1994), as a set of terms (Qiu and Frei, 1993; Jing and Croft, 1994; Xu and Croft, 2000), as elements of a phrase (Mitra et al., 1997; De Lima and Pedersen, 1999), or as elements of a compound (Jacquemin and Tzoukermann, 1999; Peñas et al., 2002). Linguistically oriented strategies are largely outnumbered by statistical co-occurrence strategies. However, while the usefulness of linguistic information in IR is still questioned by some scholars, evidence is available that the performance of (especially web-based) IR can be improved by using NLP. As suggested above, so far, mainly two types of linguistic information have been used: phrase structures and lexical compounds. The goal of our work is to explore the use of lexically restricted co-occurrence, i.e., *collocation*, information for *local* query expansion. Local query expansion techniques search for suitable expansion terms the first n top-ranked documents retrieved as response to the original user query.

[1] Some other techniques involve, e.g., the exploitation of terms in the syntactic context of the query terms found in the document collection (Grefenstette, 1992; Ruge, 1992), or the use of most common terms in the n-top ranked documents obtained for the original query. The use of thesauri and lexica as source of expansion terms (e.g., hyperonyms and synonyms of query terms in the case of a thesaurus, and morphological derivates in the case of lexica) has also been suggested; cf. among others, (Hersh et al., 2000; Woods et al., 2001).

Following (Xu and Croft, 2000), we hypothesize that top-ranked documents tend to form two clusters—a cluster of documents that are relevant to the query of the user and a cluster of documents that are irrelevant to this query. Our work is thus divided into two stages: (i) relevance clustering of the top-ranked document set; (ii) query expansion using suitable expansion terms from the relevant document cluster. In this paper, we focus on the first stage. We investigate to what extent *collocations* that occur in the original user query can be used for the relevance clustering task. The use of collocation information for query expansion is described elsewhere.

Our working document collection is the Spanish part of the document collection of the CLEF 2002 competition (Peters, 2002). For our experiments, we use top-ranked document sets retrieved within the CLEF 2002 competition by the COLE IR-system (Vilares et al., 2002).

The remainder of the paper is organized as follows. In Section 2., we introduce the phenomenon of collocation underlying our work. Section 3. provides some evidence for the significant co-occurrence of collocations in both queries as used within the CLEF competition and the document collection. Section 4. describes the preliminary experiments we carried out so far and their evaluation. Section 5. presents the conclusions.

2. The Phenomenon of Collocation

A collocation is a term combination $t_1 + t_2$ that expresses a concept configuration $c_1 \oplus c_2$ such that t_1 (the *base* of the collocation) is a standard "context free" option for the expression of c_1, while the choice of t_2 (the *collocate* of the collocation) for the expression of c_2 depends on the availability of t_1. Cf. the concept combinations 'sanctions'⊕'installation', 'state of emergency'⊕'installation', and 'customs duty'⊕'installation'. In all three of them, c_2 is 'installation'. However, in connection with *sanctions*, it is expressed by the term *imposition*, in the case of *state of*

N_{Co_Q} being the number of different LFs in Q, N_{Co_D} the number of different LFs from Q that occur D, $f(co)$ the frequency of the collocation co or a similar instance of the same LF in the document D, $f(B)$ the frequency of the base B of co in D, and N_{Co^-} the total number of the LF-instances with B in D that are opposed to the LF of which co is an instance.

The advantage of an LF-metric over a term co-occurrence metric is its generalization potential: it covers all semantically similar term sequencies rather than only one single term sequence.

In a series of preliminary experiments, we clustered a number of 100 top-ranked documents sets returned by the IR-system of the COLE-group of the University of La Coruña. Table 3 shows the *fallout* f, *precision* p and *recall* r of five clustering runs on five different 100 top-ranked sets.

f(allout) %	p(recision) %	r(ecall) %
97.77	50.00	28.57
61.40	56.86	67.44
51.35	67.92	65.45
80.64	10.00	50.00
97.80	66.67	50.00

$f = \frac{NR_Q}{NR_d}$, with NR_Q as the number of non-relevant documents recognized by the metric, NR_d as the number of non-relevant documents in the corresponding top-100 set; $p = \frac{R_Q}{R_m}$, with R_Q as the number of relevant documents recognized by the metric and R_m as the number of documents classified by the metric as relevant; $r = \frac{R_Q}{R_d}$, with R_d as the number of relevant documents in the corresponding top-100 set.

Table 3: Quality figures of five clustering runs.

The table reveals that f is consistently high, which means that the metric functions well for filtering out irrelevant documents from the top-ranked sets. p and r may vary significantly—depending on the LFs involved in the query: instances of certain LFs are better discriminators than instances of other LFs.

5. Conclusions

We argued that collocation information is an important type of linguistic information that must be taken into account for local document relevance clustering and for query expansion. We presented some initial evidence for the importance of collocations for local document relevance clustering. More work is needed to obtain reliable figures that reveal to what extent query expansion improves when collocations in queries are being taken into account. A topic we still did not explore so far at all is the role of collocations in indexing.

6. References

E. De Lima and J. Pedersen. 1999. Phrase Recognition and Expansion for Short, Precision-Biased Queries Based on a Query Log. In *Proceedings of the 22nd SIGIR Conference*, pages 145–152. ACM Press, NY.

G. Grefenstette. 1992. Use of Syntactic Context to Produce Term Association Lists for Retrieval. In *Proceedings of the 15h SIGIR Conference*, pages 89–97. ACM Press, NY.

W. Hersh, S. Price, and L. Donohoe. 2000. Assessing Thesaurus-Based Query Expansion Using the UMLS Metathesaurus. *Journal of Americian Medical Informatics Association*, Symposium Supplement.

C. Jacquemin and E. Tzoukermann. 1999. Nlp for term variant extraction: Synergy between morphology, lexicon, and syntax. In T. Strzalkowski, editor, *Natural Language Processing Information Retrieval*, pages 25–74. Kluwer, Boston.

Y. Jing and W.B. Croft. 1994. An association thesaurus for information retrieval. In *Proceedings of the RIAO Conference*, pages 146–160.

I.A. Mel'čuk. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–102. Benjamins Academic Publishers, Amsterdam.

M Mitra, C. Buckley, A. Singhal, and C Cardie. 1997. An analysis of statistical and syntactic phrases. In *Proceedings of the Fifth RIAO Conference*.

A. Peñas, J. Gonzalo, and F. Verdejo. 2002. Distinción semántica de compuestos léxicos en recuperación de información. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 26.

C. Peters. 2002. *Results of the CLEF 2002 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2002 Workshop*. Rome, Italia.

Y. Qiu and H.P. Frei. 1993. Concept based query expansion. In *Proceedings of the 16th SIGIR Conference*, pages 160–169. ACM Press, NY.

G. Ruge. 1992. Experiments on linguistically-based term associations. *Information Processing and Management*, 28(3):317–332.

H. Schütze and J. Pedersen. 1994. A cooccurrence-based thesaurus and two applications to Information Retrieval. In *Proceedings of the RIAO Conference*, pages 266–274.

K. Sparck Jones. 1971. *Automatic KeyWord Classification for Information Retrieval*. Butterworths, London.

J. Vilares, M.A. Alonso, F.J. Ribadas, and M.Vilares. 2002. COLE Experiments at CLEF 2002 Spanish Monolingual Track. In C. Peters, editor, *Results of the CLEF 2002 Evaluation Campaign*, pages 153–160.

L. Wanner, B. Bohnet, and M. Giereth. in print. Making Sense of Collocations. *Computer Speech and Language*.

L. Wanner. 2004. Towards Automatic Fine-Grained Semantic Classification of Verb-Noun Collocations. *Natural Language Engineering Journal*, 10(2):95–143.

W. Woods, S. Green, P. Martin, and A. Houston. 2001. Aggressive morphology and lexical relations for query expansion. In *The Tenth Text REtrieval Conference (TREC 2001)*.

J. Xu and W.B. Croft. 2000. ACM Transactions on Information Systems. *Improving the Effectiveness of Information Retrieval with Local Context Analysis*, 18(1):79–112.