Verbal associations as a natural language semantic parameter: beyond WordNet and BableNet

Abstract: Working out an algorithm for QE we have analyzed some parameters of initial queries and target documents. We have been looking for parameters to decrease amount of documents for the global analysis after QE. Intuitively a researcher considers that processing a huge text collection corresponds to the best result, relevant to the query. The more you process to satisfy the query, more possibilities you get to catch relevant documents, the less a chance to miss something important. Even if it does not match the reality, experiments with a dump of Wikipedia represent a model for the Internet searching procedure. The diversity of texts, videos, images in the Internet with a term from the query overwhelms algorithms of the model. Thus you need to systematize the diversity, to find a parameter in the initial query, which predicts during the global analysis processing a subcorpus with a term relevant to its sense and usage area instead of a whole document collection.

Terms behave as words from a natural language, therefore they are characterized by grammar markers, semantics, lexical compatibility and pragmatics. Some features of a term predetermine a direction for QE of the initial query. QE requires frequent words from texts with a term as a frequent word or a key word.

Actually, text attributes depend on its genre and topic. Let's clarify the thesis on the basis of differences between academic writing and literature, or fiction. For academic writing any restrictions limit repeatability of words, because an author has to use special terminology. Academic communication demands a plain language to escape ambiguity. So important lexemes and keywords are frequent in the academic texts. On the contrary, for fiction a writer is expected to provide readers with bright descriptions and remarkable characters, so he / she must diversify a language by epithets and synonyms. The same trend works for media. Thus, frequent words in fiction are less significant for the content and information, than ones in science or humanities.

Thereby texts differ in the word frequency distribution. For scientific texts a proportion and a value of frequent words would be relevant to the content, while for fiction a proportion of frequent words has not essential value. However, after you apply content analysis to the words from a fiction narrative, you will get essential results. A new word is needed to explicit a concept; with this new word as a lexical unit we receive a new set of linguistic features: semantic links, context restrictions, valency. For media texts a list of frequent words appears to be more informative, than for fiction. Therefore, frequent words for QE from texts of different types and genres possess divergent values for information extraction. In case with fiction and media the frequent words, and even concept nominations, cause to increase semantic distance between the initial query and the information requested by a user.
So the problem is not just to widen a field for the global analysis, but to restrict the field according to the semantics and pragmatics of the term.

Bio: Irina G. Ovchinnikov, graduated from Saint-Petersburg State University in 1986 (PhD in Psycholinguistics) and have been working in Perm State University, Russia, during two decades. She have been in charge of education in Computational Linguistics as the Head of Speech Communication Department. Since 2008 Irina lives and works in Haifa. Prof. Ovchinnikova published 5 books and more than 40 papers, including chapters in monographs and textbooks for students. The most recent project is targeted at those developing and using translation tools and parallel corpuses. The specific « Israeli » project is a database of verbal associations of Russian Israelis: from the point of computational linguistics, the database is considered as a pilot sample for association research in Hebrew and English. Irina is an active blogger of The Jerusalem Post and Chanal 9 TV.