# Combining Vector Space Model and Multi Word Term Extraction for Semantic Query Expansion

Eric SanJuan[1], Fidelia Ibekwe-SanJuan[2], Juan-Manuel Torres-Moreno[1,2], and Patricia Velázquez-Morales

[1] Laboratoire Informatique d'Avignon UAPV, BP 1228 84911 Avignon, Cedex 9, France {eric.sanjuan,juan-manuel.torres}@univ-avignon.fr
[2] URSIDOC & Université de Lyon 3, 4, cours Albert Thomas, 69008 Lyon Cedex, France ibekwe@univ-lyon3.fr
[3] École Polytechnique/DGI CP 6079 Succ. Centre-ville - H3C3A7 Montréal, Canada

**Abstract.** In this paper, we target document ranking in a highly technical field with the aim to approximate a ranking that is obtained through an existing ontology (knowledge structure). We test and combine symbolic and vector space models (VSM). Our symbolic approach relies on shallow NLP and on internal linguistic relations between Multi-Word Terms (MWTs). Documents are ranked based on different semantic relations they share with the query terms, either directly or indirectly after clustering the MWTs using the identified lexico-semantic relations. The VSM approach consisted in ranking documents with different functions ranging from the classical tf.idf to more elaborate similarity functions. Results shows that the ranking obtained by the symbolic approach performs better on most queries than the vector space model. However, the ranking obtained by combining both approaches outperforms by a wide margin the results obtained by methods from each approach.

## 1 Introduction

Despite the huge amount of studies on query expansion and document ranking, this topic continues to attract a lot of attention. Indeed, earlier studies have established that information seekers rarely use the enhanced search features available on most search engines or in specialised databases. Average query text consists of 1.8 words [1]. This means that query terms are often too imprecise. In technical fields, it can be expected that a unique semantic category can be associated to each domain term (a nound phrase that refers to a unique concept in some specialised field). When an ontology exists, refining by semantic nearest-neighbour term consists in expanding the query terms using terms in the same category as the query. When the query is too imprecise, this process of refinement by adjoining semantically related terms allows to rank documents according to the frequency of such terms in titles or abstracts available in bibliographic databases.

In this paper, we target document ranking in a highly technical field with the aim to approximate a ranking that is obtained through an existing ontology or a knowledge structure. The reference ranking is obtained by refining the query term with terms in the same semantic category in the ontology. In this context, a pre-requisite is that domain terms in the test corpus be previously annotated and assigned a unique semantic category in the ontology. We test two ranking approaches, symbolic methods and the vector-space model which will both try to obtain rankings that come as near as possible to the reference ranking but without using the manually annotated terms nor the semantic category of a term in the ontology.

We explore the two major approaches to query expansion: the vector space model approach for measuring *term – document* similarity and a symbolic approach relying on surface linguistic relations between query terms and documents. The aim is to determine the approach and the methods therein that perform best on multi-word query terms. We implemented the vector space model using the CORTEX system initially designed for automatic summarisation [2]. The symbolic approach is effected via the TermWatch system [3]. This system extracts multi-word terms (MWTs), links them through local morphological, lexical, syntactic and semantic relations, then clusters the MWTs variants based on these relations. Given a query term, these clusters are used to rank documents according to the proportion of shared terms between clusters and documents that also contain the query term. The idea is to refine a query term with its semantic nearest neighbour ($S$-NN) terms. Finally, in a hybrid approach, relations used for ranking in the symbolic approach are combined with different functions from the vector-space model in order to see if this improves the results obtained by each model separately. All these methods are then evaluated against a reference ranking obtained by ranking documents using semantic categories from a hand-built taxonomy associated with the test corpus. As a by product, this experiment also provides a new methodology for comparing the different methods issuing from the two approaches.

The rest of the paper is organised as follows: section 2 describes the corpus used in this experiment and how queries were formulated. Section 3.1 describes the symbolic approach, the following section 3.2 describes the vector space model approach and secyion 3.3 the hybrid approach. Section 4 analyses results while section 5 draws lessons learned from the experiment.

## 2 The test corpus

In order to have a reference ranking, we needed a corpus with an associated knowledge structure (taxonomy or ontology) but where each term in the corpus can be traced back to the ontology. This is because our QR systems extract terms automatically from the corpus and the associated knowledge structure will be used to build the reference ranking. The GENIA corpus[4] satisfied our requirements in that it comes with a hand-built ontology where terms from the abstracts have been manually annotated and assigned to categories in the

[4] *http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/*

ontology by domain specialists. This corpus consists of 2000 bibliographic records drawn from the MEDLINE database using the keywords: *Human, Blood Cells,* and *Transcription Factors*. We shall refer to the titles and abstracts of these records as documents henceforth. The annotations in XML format indicate the terms together with their semantic category, defined as the leaves of a small hand-built ontology, the GENIA ontology. There are 36 such categories at the leaf nodes and a total of 31, 398 terms. The largest category, called "*other name*" has 10, 505 terms followed by the "*protein molecule*" category with 3, 899 terms and the "*dna domain or region*" category with 3, 677 terms. The distribution of terms in the categories follow a zipfian curve. In this context, each annotated term can be viewed as a potential query that will extract all documents in the GENIA corpus containing this term or semantically close terms in the same GENIA category (in the ontology). The extracted documents can therefore be ranked according to the number of annotated terms in the same GENIA category as the term query. The ranking obtained for each query using the manually annotated terms and the GENIA categories constitutes the reference ranking. The QR experiment thus consists in testing the ability of different methods from the two approaches to produce a ranking as similar as possible to the reference ranking. Of course, none of the QR methods tested used the manually annotated terms nor had prior knowledge of their semantic category in the GENIA ontology.

The query terms used in this experiment were manually annotated terms in the GENIA corpus that occurred in at least 50 documents and which were associated with a category other than "*other name*". We excluded one word terms like "*cell*". In the GENIA corpus, this term will select practically all the documents. Sixteen MWTs matched these criteria. Table 1 shows the query terms together with their GENIA category, the number of elements in this category and the number of documents containing each term. We now describe the two approaches to the QR task.

**Table 1.** Queries used in the experiment

| **Query** | **GENIA Category** | **Nb** | **Docs** |
|---|---|---|---|
| activated T cell | cell_type | 1723 | 51 |
| B cell | cell_type | 1723 | 120 |
| Epstein-Barr virus | virus | 352 | 66 |
| glucocorticoid receptor | protein_family_or_group | 2452 | 96 |
| human immunodeficiency virus type 1 | virus | 352 | 52 |
| human monocyte | cell_type | 1723 | 69 |
| Jurkat cell | cell_line | 1992 | 66 |
| Jurkat T cell | cell_line | 1992 | 58 |
| NF-kappa B | protein_molecule | 3885 | 271 |
| nuclear extract | cell_component | 205 | 74 |
| nuclear factor | protein_family_or_group | 2452 | 54 |
| nuclear factor of activated T cells | protein_family_or_group | 2452 | 51 |
| protein kinase C | protein_molecule | 3885 | 83 |
| T cell | cell_type | 1723 | 339 |
| T lymphocyte | cell_type | 1723 | 115 |
| transcription factor | protein_family_or_group | 2452 | 487 |

# 3 Methodology

## 3.1 Symbolic approach

The symbolic approach to QR is implemented via the TermWatch system [3] comprising three modules: a term extractor, a relation identifier which yields the terminological network and a clustering module. Clustering is based on general linguistic relations that are not dependent on a particular domain and do not require specific work for every text collection.

Different linguistic relations for expanding query terms into their $S$-NN terms were tested, ranging from coarse-grained ones like identity of grammatical head words to fine-grained ones.Thus, any query term is mapped onto the set of automatically extracted $S$-NN terms. Since these $S$-NN terms have been grouped into clusters, the query term can be represented by the cluster vector with as many dimensions as there are clusters and, whose values are the number of variants that the query has in each cluster. Since every document can also be represented by a similar vector that gives for each cluster, the number of its terms in the document, the relevance of the document against the query can be evaluated as the scalar product between the two vectors (cluster and document). We describe in more details the relations tested.

**Ranking by head word occurrence (Head)** This consists in ranking documents based on an occurrence count of the head word of the query term in the documents that contain that head word but in any grammatical position. The justification for using this coarse relation is the well-known role of head nouns in noun phrases: they depict the subject of phrase and thus also of the query. Thus documents in which the head word has a high frequency could correspond to documents with the highest number of terms in the same GENIA category. Document ranking with this relation is performed outside TermWatch as it relies simply on an occurrence count of a head word in documents.

**Ranking by Basic TermWatch's clusters (TW)** The most coarse-grained clustering relation in TermWatch consists in merging all terms sharing the same head word into the same cluster. This relation generated clusters of identical heads and on this corpus produced $3,670$ clusters involving all the extracted multiword terms $(36,702)$. Given a query term, documents are ranked according to the number of their terms which had the head word of the query term also in their head position.

For instance, given the query term *T cell* where *cell* is the head word, the topmost ranked document by this relation had the most number of terms with "cell" in its head position: *B cell, cell, blood cell, differentiated cell, hematopoietic cell, HL60 cell, L cell, lympoid cell, macrophage cell, monocyte-macrophage cell, nucleated cell, peripheral blood cell, S cell, T cell.*

**Ranking by tight semantic clusters (Comp)** This consists in ranking using terms in the connected components formed by spelling variants, substitutions of synonymous variants acquired via WordNet and expansions relations (where only one word was added to a term). The idea is to restrict the $S$-NN of a query term to only those terms which do not involve a topical shift and are its closest $S$-NN in terms of all the variation relations used in TermWatch. In this experiment, $2,382$ were found involving only $8,019$ terms.

**Ranking by looser semantic clusters (Var)** Relations are added to *Comp* ones in order to form bigger clusters involving weaker expansion variants (addition of more than one modifier word) and substitution of modifier words. The idea here is to expand the $S$-NN of a query term to farther semantic neighbours where the link with the original subject of the query term may be weaker. Clustering in this case produced $3,637$ clusters involving $14,551$ terms. For instance, for the same "*T cell*" query, the topmost document ranked by *Var* clusters had six terms bearing the word *cell* in their head position some of which were also modifier substitutions of the query term (*cell, jurkat cell, naive cell, responding cell, stimulated cell, T cell*). In contrast, the topmost document ranked by the reference ranking obtained through the GENIA ontology contained more variants of the query term (*jurkat T cell, L cell, T cell, activated T cell, cell, endothelial cell, human T, human T cell,...*). This document was ranked 10th by *Var* relations.

### 3.2 Vector-based model approach

We tested two ways of ranking documents based on the vector model. The first method supposes that word frequency can be estimated on the whole set of documents represented as an inverted file. The second method works on the restricted set of documents containing at least one occurrence of the query term.

Let $\Delta$ be the set of all abstracts in the bibliographic database and let $\Omega$ bet the set of uniterms (terms with only one word). For any abstract $d$, we shall denote by $\Omega_d$ the set of uniterms occurring at least once in $d$ and by $\Delta_w$ the set of documents in which $w$ occurred.

We assume the existence of an inverted file which for any word $w$ and abstract $d$ in the bibliographic database gives the frequency $f_{d,w}$ of $w$ in $d$. Based on such inverted file, documents can be ranked following the *tf.idf* score of query terms in the document with or without query expansion mechanism *QE*. It consists in first computing the *tf.idf* function and then replacing the query term vector by the sum of the top ranked document vectors. This expanded query is then used to perform another ranking.

Now, we do not more assume the existence of an inverted file. Given a query sequence $T$ in the form of a MWT the following measures are computed on the restricted set of documents $\Delta(T)$ where the string $T$ occurred. These documents are represented in a vector space [4, 5] using the CORTEX [2] system that includes a set of independent metrics combined by a Decision Algorithm. This vector space representation takes into accounts nouns, compound words, conjugated verbs numbers (numeric and/or textual) and symbols. Other grammatical categories like articles, prepositions, adjectives and adverbs are eliminated using a stop list. Lemmatisation and stemming [6, 7] are performed thus yielding higher word frequencies. Compound words are identified, then transformed into a unique lemmatised/stemmed uniterm using a dictionary.

To describe the selected metrics we used for QR, we shall use the following notations for any $w \in \Omega$ and $d \in \Delta(T)$:

$$\Delta(T)_w = \Delta_w \cap \Delta(T) \quad f_{d,.} = \textstyle\sum_{\omega \in \Omega_d} f_{d,\omega} \quad f_{.,w} = \sum_{\delta \in \Delta(T), w \in \Omega_\delta} f_{\delta,w}$$

$$\Omega(T) = \{\omega \in \Omega : f_{.,w} > 1\} \;\; f_{.,.} = \textstyle\sum_{\omega \in \Omega(T)} f_{.,\omega} \;\; \Omega(T)_d = \Omega_d \cap \Omega(T)$$

We tested the metrics described above as well as combinations of them: the angle (noted A), three different measures of query overlapping (D, L, O) and the frequency of informative words (F). We also considered the following combinations of sets of metrics $\{A, D, O\}$, $\{A, L, O\}$, $\{A, D, L, O\}$, $\{F, L, A, D, O\}$ based on CORTEX's decision algorithm.

**A** is the angle between $T$ and $d$. Although not all words in $T$ have the same informative value since words closed to the term head have an higher probability to be correlated to the term's category. Thus, we have represented the query term $T = t_1 ... t_n h$ by a vector $\boldsymbol{T} = (x_w)_{w \in \Omega(T)}$ where:

$$x_w = \begin{cases} 15 & \text{if } w = h \\ j & \text{if } w = t_i \text{ for some } i \in [1..n] \\ 0 & \text{otherwise} \end{cases}$$

**D** is the sum of the word frequencies in abstract $d$ multiplied by its probability of occurrence in $\Delta(T)$ as follows: $D(d) = \sum_{w \in \Omega(T)_d} \left( \frac{f_{.,w}}{f_{.,.}} \times f_{d,w} \right)$

**O** focus on documents involving terms that occurred in almost all documents: $O(d) = \sum_{w \in \Omega(T)_d} (|\Delta(T)_w| \times f_{d,w})$

**L** reveals documents that overlap with query words but with a larger vocabulary: $L(d) = |\Omega(T)_d| \times \sum_{w \in \Omega(T)_d} (|\Delta(T)_w|)$

**F** is the term frequency sum $F = f(., w)$ It favours documents with a small vocabulary on tha contrary of metrics D,O,L.

The Decision Algorithm (DA) relies on all the normalised metrics $\hat{\mu}(d)$ combined in a sophisticated way. Here is the decision algorithm that allows to include the vote of each metrics:

$$\alpha = \sum_{\hat{\mu} \in \{X_1,...,X_k\}, \hat{\mu}(d) > 0.5} (\hat{\mu}(d) - 0.5) \; ; \; \beta = \sum_{\hat{\mu} \in \{X_1,...,X_k\}, \hat{\mu}(d) < 0.5} (0.5 - \hat{\mu}(d))(1)$$

The value $\Lambda$ attributed to every sentence is then calculated:

$$\text{If } \alpha > \beta \text{ then } \Lambda = 0.5 + \tfrac{\alpha}{k} \text{ else } \Lambda = 0.5 - \frac{\beta}{k}$$

### 3.3 Hybrid approach

Clusters built by TermWatch target a high degree of semantic homogeneity. They rely on the existence of a restricted family of linguistic variation relations among terms and thus are generally small. As a consequence, when mapping a query term $T$ onto its $S$-NN terms in clusters, this often grasps only a few clusters. Thus, ranking documents according to their overlap with these clusters produces a substantial proportion of ties. We then tried to use CORTEX's normalised metrics to break these ties. Indeed as pointed out in the preceding section, high scores of selected CORTEX metrics are obtained for documents containing the query words in $T$ and words frequently associated to them, i.e, their co-occurrence contexts. Since document scores based on cluster overlapping are integers, tails can be simply broken by adding to this integer score, CORTEX's decision score which is a real number in $[0, 1]$. This leads to a new document ranking system (summarised in figure 1) where documents are:

1. extracted in full text Boolean mode based on a sentence expressed in natural language,
2. ranked according to the linguistical relations they share with the multiword terms in the query,
3. re-ranked by breaking ties based on vector similarities with the query.

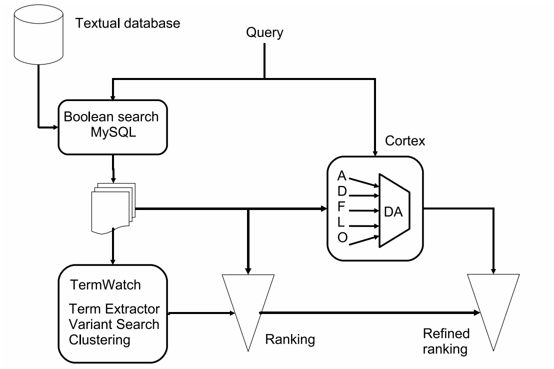

**Fig. 1.** Hybrid ranking system

## 4 Results

We now analyse results from the three approaches : vector space, symbolic and hybrid. Given a query term, we evaluate the methods described in sections 3.1 and 3.2 according to their capacity in ranking documents with regard to an existing ontology, i.e., top ranked documents should contain terms from the semantic category in the GENIA ontology as the query term.

For each query, we compared the ranking of documents produced by the different methods to the reference ranking by calculating the Kendall's W coefficient of concordance [8]. This coefficient stems from the family of robust non-parametric tests which do not make any assumption on the Gaussian distribution of the data. Kendall's W coefficient is 1 in the case of complete agreement between two rankings and 0 for total disagreement. As in all statistical tests, to interpret the intermediary values, it is necessary to verify if the score obtained by a method is significantly different from that of a random ranking on the same data. We computed Kendall's W coefficient and its "$p$-value" using R software for statistical computing with the Concord package[5]. We did not use precision-recall as evaluation metric because all the ranking methods work from the same list of documents, i.e., they are all based on the selection of documents containing the initial query term. What differed was the way in which they ranked these documents. Hence, calculating recall does not make sense here.

### 4.1 Global comparison of methods

Figure 2 gives the boxplots of Kendall's W coefficient of concordance on all queries for each method. According to these boxplots, refining TW's ranking by CORTEX's metrics ($X_1...X_k$-tw where $X_1, ..., X_k$ is any combination of {A, D, F, O, L}) outperformed single TW which in turn outperformed the Head method, any one of CORTEX metrics (A, D, F, O, L) taken separately or any of their combinations and MySQL rankings (tf.idf and QE). We now check if these differences are statistically significant. For that, we apply the non parametric paired Wilcoxon signed rank test and Friedman's rank sum test both available in the standard R software package. These two tests are used to compare the median Kendall's W scores obtained by each method.

We first analysed the combinations of CORTEX's metrics to see if any one performs better than the others. Friedman's test showed with a confidence of 99% that there exists significative differences. However, running the same test only on combination of at least two CORTEX measures among {A, D, O, L} shows that there is no statistical evidence of differences among members of this group ($p$-value $> 0.8$). This shows that combining CORTEX metrics based on its decision algorithm 3.2 significantly improves the results.

Now observing the group of methods based on a single CORTEX metric significantly differs among themselves as found by Friedman's test with a confidence of 99%. Indeed, based on Wilcoxon test we found out that O and D are not statistically different ($p$-value=0.86), neither are F and L ($p$-value=0.82). The first two appear to be more adapted to this experiment than F and L (see their Kendall's W values on Figure 2). Metrics O and D top-rank documents in which the frequent words correspond to the query words or are strongly associated to them, whereas metrics L and F focus on the vocabulary coverage of documents irrespective of the query words. L is very sensitive to documents with a wide vocabulary coverage and F does the reverse. Thus these two rank

[5] 'http://www.r-project.org/

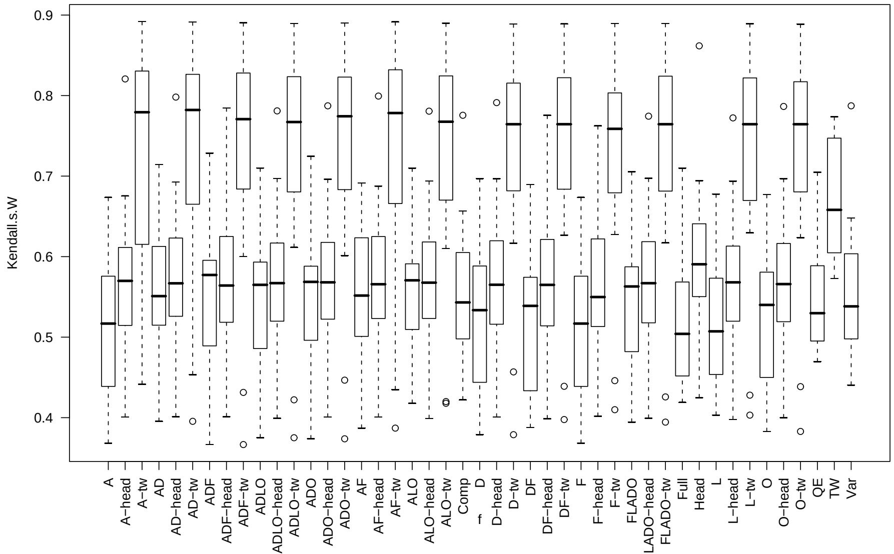

**Fig. 2.** Boxplots showing median Kendall's W scores and extreme values for each method. Symbols A, D, F, L, O and their combinations in upper case refer to CORTEX metrics (e.g. FLADO); "Head", "TW" and "Var" refer to the rankings based on the respective TermWatch's clusters. Symbols representing CORTEX's metrics followed by lower case "tw" or "head" refer to hybrid approaches. "QE" stands for *tf.idf with QE.*

documents based on criteria intrinsic to the documentsbut not to the query. Metric A that takes into account the position of each word in the query remains apart. Finally, we take a look at performances amongst symbolic methods to see if there is any statistical difference among their rankings. Wilcoxon's test enabled us to ascertain that the hypothesis of equal medians between *basic TW* and Head's rankings can indeed be rejected with a risk lower than 5%. The same test also showed with a confidence of 90% that *Head* method outperformed *Var*, but that the observed differences between *Head* and *COMP*'s rankings were not statistically significant ($p$-value=0.23).

Let us now compare the rankings obtained by the hybrid approach. We have already observed that there is no statistical difference between median scores of combinations of at least two CORTEX's metrics. We have the same phenomena between any TermWatch's ranking refined by any CORTEX's metric. Indeed the $p$-value resulting from the Friedman test on this family of methods is higher than 0.54. Since we have already ascertained the effectiveness of CORTEX's decision algorithm, we shall only need to consider *FLADO-tw* which is the refinement of TW ranking based on the combination of all selected CORTEX metrics among all possible combinations. In the same way, we found out that there is no statistical evidence of differences between refinements of Head's rankings with any CORTEX's metrics. Thus we shall only consider the *FLADO-Head* combination. We then obtain, based on Wilcoxon's test, that *FLADO-tw* outperforms *TW* with a confidence of 95%, and that *Head* outperforms *FLADO-head* with a confidence of 99%. Since we have previously shown that *TW* outperforms *Head*, we deduce that *FLADO-tw* clearly outperforms *FLADO-head* and *FLADO*. This turned out to be the case with a confidence level higher than 99.98%.

Following these statistical tests, it appears using that the combination of CORTEX's metrics (FLADO) chosen by its decision algorithm to refine TermWatch's TW's semantic rankings produces the best hybrid approach. Contrarily, refining the ranking produced by the *Head* method with CORTEX's metrics degrades results considerably.

### 4.2 Query by query comparison of ranking methods

Global results can mask important differences as suggested by the length of the boxes in figure 2 and by the existence of extreme values. The detailed view of the performances for the main methods is shown in Table 2. This table shows the Kendall's W score for each method per query. For each query, only the relative position of the score between methods can be directly interpreted. Thus, Table 2 can only be read vertically, column by column. Indeed, Kendall's score depends on the number of ranked documents and on the number of tails. The absolute Kendall's W value cannot be interpreted without considering the probability of finding this value in non correlated rankings. The confidence level is the complement of this probability. Table 2 only shows figures with a confidence level of at least 90%. It evaluates the expectation of the correlation between the ranking produced by the methods and the reference ranking.

Table 2 shows that *FLADO-tw* is the only method that produced 14 rankings out of 16 with more than 90% probability of being correlated with the reference ranking. The two non correlated ranking were produced for the longest queries "*nuclear factor of activated T cells*" involving a preposition and "*human immunodeficiency virus type 1*". We will comment on this later.

It also appears clearly that *FLADO-tw* improves *TW* on all queries, thus showing that CORTEX is adapted to resolving ties in *TW*'s rankings. Conversely, a similar combination of metrics degrades Head's ranking, whereas the two methods *TW* and *Head* considered separately obtain similiar Kendall's W scores on several queries where the category is mainly determined by the head word. If we look at CORTEX's metrics in isolation, we obtain weaker results than for *Head* and *TW* methods. However it is interesting to observe that the three measures A, D and O are required in order to cover the whole set of queries where the FLADO combination is significant. It is also interesting to notice that *Comp* method based on tight semantic relations performed well mainly on queries where no CORTEX metric obtained good scores like "*nuclear factor, T lymphocyte, activated T cell*". This points to the fact that a hybrid approach is indeed desirable for query expansion and the two systems TermWatch and CORTEX are indeed complementary for this task.

We now take a look at queries where the hybrid approach did not perform as well as expected, i.e., where independent methods obtained better rankings.

The *Head* method significantly outperformed all other methods on the "*Epstein-Barr virus*" query due to the fact that the head word "*virus*" characterises the terms in this GENIA category, i.e., almost all terms in this category include the word "*virus*". Thus counting the occurrences of this head word in documents is equivalent to counting occurrences of terms in this category. There is however a difference between the ranking produced by *Head* and the reference ranking because the latter records the single presence of a term in a document even if the term has multiple occurrences.

*Tf.idf* function is the only one that obtained a significantly correlated ranking on the query "*human immunodeficiency virus type 1*" notwithstanding the ambiguity of the subject of this query, which is not the last token *1* but the entire phrase *virus type 1.* One query was not included in the table ("*nuclear factor of activated T cells*") because no method attained the confidence level of 90% on it. This query had the particularity of containing a preposition. Permutation variants are amongst those identified by TermWatch and could be used in future work to efficiently process queries with prepositions.

| Queries: | B cell | protein kinase C | T cell | NF-kappa B | Jurkat cell | transcription factor | T lymphocyte | Epstein-Barr virus | nuclear extract | glucocorticoid receptor | human monocyte | nuclear factor | Jurkat T cell | activated T cell | human immunodeficiency virus type 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Head | **0.61** | **0.69** | **0.58** | | 0.68 | **0.60** | | **0.86** | 0.65 | | *0.59* | 0.58 | *0.62* | | |
| FLADO-head | 0.58 | *0.62* | **0.60** | | 0.69 | **0.58** | | **0.77** | *0.61* | | | | *0.63* | | |
| A | | 0.63 | | **0.67** | 0.65 | | | | | *0.59* | | | | | |
| D | 0.58 | 0.63 | **0.57** | **0.62** | 0.69 | | | | | | | | | | |
| F | | 0.63 | | **0.67** | 0.65 | | | | | *0.59* | | | | | |
| L | | *0.62* | | **0.67** | 0.65 | 0.56 | | | | | | | | | |
| O | *0.56* | 0.63 | 0.55 | **0.65** | 0.67 | 0.55 | | | | | | | | | |
| FLADO | *0.57* | *0.61* | **0.57** | **0.63** | **0.70** | **0.57** | | | | | | | | | |
| Comp | | | 0.57 | | | | *0.61* | | | | | *0.65* | | **0.77** | |
| Var | | | **0.60** | | **0.78** | | | | | | | *0.63* | *0.64* | | |
| TW | **0.74** | | **0.72** | 0.58 | **0.77** | **0.60** | 0.65 | **0.75** | *0.63* | *0.60* | **0.75** | | 0.67 | 0.75 | |
| FLADO-tw | **0.88** | 0.67 | **0.88** | **0.61** | **0.88** | **0.73** | **0.80** | **0.73** | **0.84** | **0.73** | 0.68 | **0.75** | **0.80** | **0.77** | |
| QE | | 0.65 | | **0.64** | **0.70** | 0.54 | | | | | | | *0.61* | | |
| tf.idf | | | | | | | | 0.68 | | | | | **0.70** | | 0.70 |

**Table 2.** Kendall's W scores per query. Only scores with a confidence level of at least 90% appear. Figures with confidence between 90% and 95% are in italic. Figures in Bold have a confidence greater than 99%.

## 5 Conclusion

The task introduced in this paper that we have termed *Semantic Query Expansion oriented Document Ranking* (SQEDR) is quite novel and has not been dealt with in the TREC's campaigns [9]. The results we obtained show on the GENIA corpus that such rankings can be approximated combining MWT term extraction and bag-of-word text representation.

In the recent TREC2005 Robust track, [10] used WSD (word sense disambiguation) and semantic query term expansion in the document retrieval task. WSD is first applied to multi-word query terms in order to determine the exact sense of the consitituent words in the context of the query. This is done using all the available information in WordNet. When this fails, the authors resort to a Web search for the WSD process. After WSD is performed, semantically-associated terms to the chosen sense (synset) from WordNet are used to expand the query term. As we can see, query expansion here is heavily reliant on WordNet's coverage of words in the document collection.

Work in progress is carried out in testing if SQEDR could be usefull in this TREC's standard task.

We are also working in drawing records from general MEDLINE corpus. SQR can be carried out on this corpus using Mesh thesaurus [6] and the UMLS[7]. However, these two contain only terms from a controlled vocabulary (humanly fabricated terms) which are not necessarily present in MEDLINE's abstracts. Our approach of SQR could handle this gap between real terms from texts and a controled vocabulary.

## References

1. Ray, E.J., Seltzer, R., Ray, D.S.: The AltaVista Search Revolution. Osborne-McGraw Hill (1997)
2. Torres-Moreno, J.M., Velázquez-Morales, P., Meunier, J.G.: Condensés de textes par des méthodes numériques. In: JADT 2002, France (2002) 723–734
3. SanJuan, E., Ibekwe-SanJuan, F.: Text mining without document context. Information Processing and Management **42** (2006) 1532–1552
4. Salton, G.: The SMART Retrieval System - Experiments un Automatic Document Processing. Englewood Cliffs (1971)
5. Morris, A., Kasper, G., Adams, D.: The effects and limitations of automated text condensing on reading comprehension performance. In: Advances in automatic text summarization, U.S.A., The MIT Press (1999) 305–323
6. Paice, C.D.: Another stemmer. SIGIR Forum **24**(3) (1990) 56–61
7. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3) (1980) 130–137
8. Siegel, S., Castellan, N.: Nonparametric statistics for the behavioral sciences. McGraw Hill (1988)
9. Buckley, C.: Looking at limits and tradeoffs: Sabir research at trec 2005. In: Proc. of the 14th Text REtrieval Conference (TREC 2005), Gaithersburg, Maryland, U.S.A. (2005) 13
10. Liu, S., Yu, C.: University of Illinois Chicago at TREC 2005. In: Proc. of the 14th Text REtrieval Conference (TREC 2005), Gaithersburg, Maryland, U.S.A. (2005) 7

[6] Medical Subject Headings, the thesaurus associated to MEDLINE descriptors.

[7] Unified Medical Language System.