Application of Variable Length *N*-Gram Vectors to Monolingual and Bilingual Information Retrieval

Daniel Gayo-Avello, Darío Álvarez-Gutiérrez, and José Gayo-Avello

Department of Informatics, University of Oviedo, Calvo Sotelo
s/n 33007 Oviedo Spain
dani@uniovi.es

Abstract. Our group in the Department of Informatics at the University of Oviedo has participated, for the first time, in two tasks at CLEF: monolingual (Russian) and bilingual (Spanish-to-English) information retrieval. Our main goal was to test the application to IR of a modified version of the *n*-gram vector space model (codenamed blindLight). This new approach has been successfully applied to other NLP tasks such as language identification or text summarization and the results achieved at CLEF 2004, although not exceptional, are encouraging. There are two major differences between the blindLight approach and classical techniques: (1) relative frequencies are no longer used as vector weights but are replaced by n-gram significances, and (2) cosine distance is abandoned in favor of a new metric inspired by sequence alignment techniques, not so computationally expensive. In order to perform cross-language IR we have developed a naive n-gram pseudo-translator similar to those described by McNamee and Mayfield or Pirkola *et al*.

1 Introduction

The vector model is a classic approach in text retrieval [1]. In this model any document (or query) can be represented as a vector of terms and, thus, the similarity between text objects can be determined by a distance in the vector space (often, the cosine of the angle between the vectors). This model does not specify how to set vector weights although there are common elements to any term weighting approach: (1) term weight within a particular document, (2) term weight within the document corpus and, (3) document length normalization. Index terms are usually words or word stems, although *n*-grams have been also successfully used (e.g., D'Amore and Mah [2] or Kimbrell [3]).

Although this model is widely used it shows two major drawbacks. First, since documents are represented by *D* dimensional vectors of weights, where *D* is the total amount of different terms in the whole document set, such vectors are not document representations by themselves but representations according to a bigger, potentially growing, "contextual" corpus. Secondly, cosine similarities (the metric most often used) between high dimensional vectors tend to be zero[1], so, to avoid this "curse of

[1] That is, two random documents have a high probability of being orthogonal to each other.

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 73 – 82, 2005.
© Springer-Verlag Berlin Heidelberg 2005

dimensionality" problem it is necessary to reduce the number of features (i.e. terms). When using *n*-grams, this is usually done by setting arbitrary weight thresholds.

blindLight is a new approach differing in two aspects from the classical vector space model: (1) every document is assigned to a unique document vector with no regards to any corpus (so, in fact, there is no vector space!) and, (2), another measure, suitable to compare different length vectors is used.

2 Foundations of the blindLight Approach

blindLight, like other *n*-gram vector space solutions, maps every document to a vector of weights; however, such document vectors are rather different from classical ones. On the one hand, any two document vectors obtained through this technique are not necessarily of equal dimensions, thus, there is no actual "vector space" in this proposal. On the other hand, weights used in these vectors are not relative frequencies but represent the significance of each *n*-gram within the document.

Computing a measure of the relation between elements inside *n*-grams, and thus the importance of the whole *n*-gram, is a problem with a long history of research, however, we will focus on just a few references. In 1993 Dunning described a method based on likelihood ratio tests to detect keywords and domain-specific terms [4]. However, his technique worked only for word bigrams. Later Ferreira da Silva and Pereira Lopes [5] presented a generalization of different statistical measures so that these could be applied to arbitrary length word *n*-grams. In addition to this, they also introduced a new measure, Symmetrical Conditional Probability [6] (equations 1 and 2 where $(w_1 \ldots w_n)$ is an *n*-gram), which overcomes other statistically-based measures. According to Pereira Lopes, their approach obtains better results than those achieved by Dunning.

blindLight implements the technique described by da Silva and Lopes although applied to character *n*-grams rather than word *n*-grams. It measures the relations between characters inside each *n*-gram and, thus, measures the significance of every *n*-gram or, what is the same, the weight for the components in a document vector.

$$Avp = \frac{1}{n-1} \sum_{i=1}^{i=n-1} p(w_1 \ldots w_i) \cdot p(w_{i+1} \ldots w_n) \tag{1}$$

$$SCP\_f((w_1 \ldots w_n)) = \frac{p(w_1 \ldots w_n)^2}{Avp} \tag{2}$$

With regard to comparisons between vectors, a simple similarity measure such as the cosine distance cannot be straightforwardly applied when using vectors of different dimensions. Of course, it could be considered as a temporary vector space of dimension d_1+d_2, with d_1 and d_2 the respective dimensions of the document vectors to be compared, assigning a null weight to the *n*-grams of one vector that are not present in the other and vice versa. However, we consider the absence of a particular *n*-gram within a document as distinct from its presence with null significance.

Eventually, comparing two vectors with different dimensions can be seen as a pairwise alignment problem. There are two sequences with different lengths and some

(or none) elements in common that must be aligned, that is, the highest number of columns of identical pairs must be obtained by only inserting gaps, changing or deleting elements in both sequences.

One of the simplest models of distance for pairwise alignment is the so-called Levenshtein or edit distance [7] which can be defined as the smallest number of insertions, deletions, and substitutions required to change one string into another (e.g. the distance between "accommodate" and "aconmodate" is 2).

However, there are two noticeable differences between pairwise-alignning text strings and comparing different length vectors, no matter that the previous ones can be seen as vectors of characters. The first difference is important, namely, the order of components is central in pairwise alignment (e.g., DNA analysis or spell checking) while unsuitable within a vector-space model. The second is also highly significant: although not taking into account the order of the components, "weights" in pairwise alignment are integer values while in vector-space models they are real.

Thus, distance functions for pairwise alignment, although inspiring, cannot be applied to the problem under examination. Instead, a new distance measure is needed and, in fact, two are provided. Classical vector-space based approaches assume that the distance, and so the similarity, between two document vectors is commutative (e.g., cosine distance). blindLight, however, proposes two similarity measures when comparing document vectors. For the sake of clarity, we will call them the query (Q) and target (T) documents although these similarity functions can be equally applied to any pair of documents, not only for information retrieval purposes.

Let Q and T be two blindLight document vectors with dimensions m and n:

$$Q = \{(k_{1Q}, w_{1Q}) \quad (k_{2Q}, w_{2Q}) \quad \dots \quad (k_{mQ}, w_{mQ})\} \tag{3}$$

$$T = \{(k_{1T}, w_{1T}) \quad (k_{2T}, w_{2T}) \quad \dots \quad (k_{nT}, w_{nT})\} \tag{4}$$

k_{ij} is the i-th n-gram in document j while w_{ij} is the significance (computed using SCP [6]) of the n-gram k_{ij} within the same document j.

We define the total significance for document vectors Q and T, S_Q and S_T respectively, as:

$$S_Q = \sum_{i=1}^{m} w_{iQ} \tag{5}$$

$$S_T = \sum_{i=1}^{n} w_{iT} \tag{6}$$

Then, the pseudo-alignment operator, Ω, is defined as follows:

$$Q\Omega T = \left\{ (k_x, w_x) \middle/ \begin{array}{l} (k_x = k_{iQ} = k_{jT}) \wedge (w_x = \min(w_{iQ}, w_{jT})), \\ (k_{iQ}, w_{iQ}) \in Q, 0 \le i < m, \\ (k_{jT}, w_{jT}) \in T, 0 \le j < n \end{array} \right\} \tag{7}$$

Similarly to equations 5 and 6 we can define the total significance for $Q\Omega T$:

$$S_{Q\Omega T} = \sum w_{iQ\Omega T} \tag{8}$$

Finally, we can define two similarity measures, one to compare Q vs. T, Π (uppercase Pi), and a second one to compare T vs. Q, P (uppercase Rho), which can be seen as analogous to precision and recall measures:

$$\Pi = S_{Q\Omega T} / S_Q \tag{9}$$

$$\mathrm{P} = S_{Q\Omega T} / S_T \tag{10}$$

To clarify these concepts we will show a simple example based on (one of) the shortest stories ever written. We will compare the original version of Monterroso's Dinosaur with a Portuguese translation; the first one will play the query role and the second one the target, the n-grams will be quad-grams.

Cuando despertó, el dinosaurio todavía estaba allí. (Query)

Quando acordou, o dinossauro ainda estava lá. (Target)

Fig. 1. "El dinosaurio" by Augusto Monterroso, Spanish original and Portuguese translation

Q vector (45 elements)		**T vector (39 elements)**		**QΩT (10 elements)**	
Cuan	2.489	va_l	2.545	saur	2.244
l_di	2.392	rdou	2.323	inos	2.177
stab	2.392	stav	2.323	uand	2.119
...		...		_est	2.091
saur	2.313	saur	2.244	dino	2.022
desp	2.313	noss	2.177	_din	2.022
...		...		esta	2.012
ndo_	2.137	a_lá	2.022	ndo_	1.981
nosa	2.137	o_ac	2.022	a_es	1.943
...		...		ando	1.876
ando	2.012	auro	1.908		
avía	1.945	ando	1.876		
_all	1.915	do_a	1.767		
				Π: 0.209 P: 0.253	

Fig. 2. blindLight document vectors for both documents in Fig.1 (truncated to show ten elements, blanks have been replaced by underscores). QΩT intersection vector is shown plus Π and P values indicating the similarities between both documents

So, the blindLight technique, although vector-based, does not need a predefined document collection and thus, it can perform IR over ever-growing document sets. Relative frequencies are abandoned as vector weights in favor of a measure of the importance of each n-gram. In addition to this, similarity measures are analogous to those used in pairwise-alignment although computationally inexpensive and, also, non commutative which allows us to "tune" both measures, Π and P, into any linear combination.

3 Information Retrieval Using blindLight

blindLight has been used to extract key phrases and summaries from single documents [8] and to perform language identification and classification of natural languages [9]. At this moment we are interested in the evaluation of this technique applied to information retrieval; this is the reason why we developed a "quick and dirty" prototype to take part in CLEF 2004.

As with any other application of blindLight, a similarity measure to compare queries and documents is needed. At this moment just two have been tested: Π and a more complex one (see equation 11) which provides rather satisfactory results.

$$\frac{\Pi + norm(\Pi \mathrm{P})}{2} \tag{11}$$

The goal of the *norm* function shown in previous equation is just to translate the range of $\Pi \cdot P$ values into the range of Π values, thus making possible a comprehensive combination of both (otherwise, P, and thus $\Pi \cdot$P values, are negligible when compared to Π).

The operation of the blindLight IR system is really simple:

- For each document in the database an *n*-gram vector is obtained and stored.
- When a query is submitted to the system this computes an *n*-gram vector and compares it with every document obtaining Π and P values.
- From these values a ranking measure is worked out, and a reverse ordered list of documents is returned as a response to the query.

This way of operation has both advantages and disadvantages: documents may be added to the database at any moment because there is no indexing process; however, comparing a query with every document in the database can be rather time consuming and not feasible with very large datasets. In order to reduce the number of document-to-query comparisons a clustering phase may be done in advance, in a similar way to the language tree used within the language identifier. Of course, by doing this working over the ever-growing datasets is no longer possible because the system must be shut down periodically to perform indexing. Thorough performance analysis is needed to determine what database size requires this previous clustering.

Before performing CLEF experiments, we tested the blindLight IR prototype on two very small standard collections with encouraging results. These collections were CACM (3204 documents and 64 queries) and CISI (1460 documents and 112 queries). Figure 3 shows the interpolated precision-recall graphs for both collections and ranking measures (namely, pi and piro).

These results are similar to those obtained by several systems but not as good as those achieved by others; for instance, 11-pt. average precision was 16.73% and 13.41% for CACM and CISI, respectively, while the SMART IR system achieves 37.78% and 19.45% for the same collections. However, it must be said that these experiments were performed over the documents and the queries just as they are, that is, common techniques such as stop-word removal, stemming, or query term

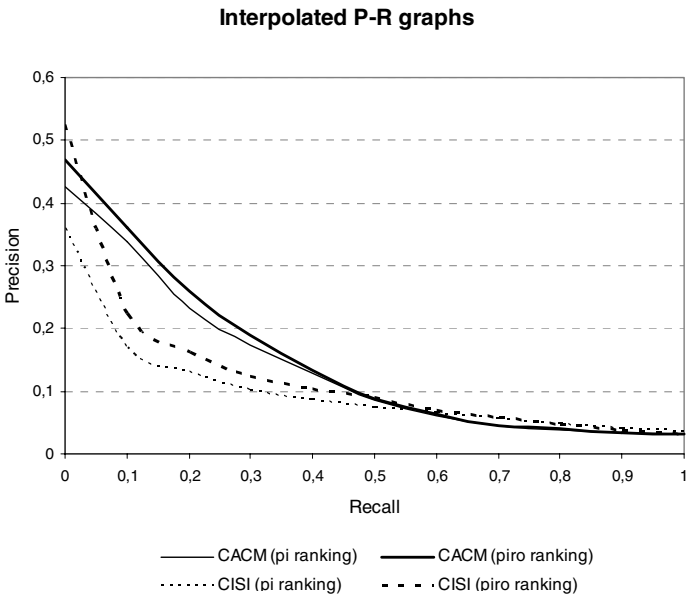

Fig. 3. Interpolated precision-recall graphs for the blindLight IR system applied to CACM and CISI test collections. Top-10 average precision for CACM and CISI was 19.8% and 19.6% respectively, in both cases using `piro` ranking

weighting were not applied to the document set and the queries were provided to the system in a literal fashion[2], as if they were actually submitted by real users. By avoiding such techniques, the system is totally language independent, at least for non ideographic languages, although performance must be improved. One obvious area for future work is represented by the similarity measures; we are planning to use genetic programming in order to test new measures.

4 CLEF 2004 Tasks

4.1 Information Retrieval Method

We applied our prototype to two "ad hoc" tasks [10] from CLEF 2004 [11]: monolingual and bilingual IR. Specifically, we queried the Russian collection with the Russian topics and the English collection with the Spanish topics. All the queries were automatically built from the topics using both title and description fields.

The method employed to obtain the results was the following one:

1. Every SGML file from a collection was parsed to extract individual pieces of news.

[2] An example query from the CACM collection: `#64 List all articles on EL1 and ECL (EL1 may be given as EL/1; I don't remember how they did it)`. The blindLight IR prototype processes queries like this one in an "as is" manner.

2. For each piece of news a quad-gram vector was computed, as described above, from the permitted fields (typically, TEXT and TITLE or HEADLINE) and stored.
3. Once the entire collection was processed the topics to query it were also parsed, computing for every topic another quad-gram vector from title and description fields.
4. After parsing the topics file, queries (i.e., their corresponding vectors) were submitted to the prototype in batch mode obtaining ranked lists of one thousand documents. The similarity measure employed to rank the results was the so-called piro since this was the one that performed the best when applied to CACM and CISI collections; however, as has been explained this measure is far from being good and this area needs to be studied in much more depth.

4.2 Pseudo-Translation of Queries

For bilingual information retrieval, the above method needs minor changes with respect to how the query vectors are obtained. This was done without performing actual machine translation using a sentence aligned corpus of source (S) and target (T) languages.

A query written in the source language, *QS*, is split into word chunks (from one word to the whole query). The *S* corpus is searched looking for sentences containing any of these chunks. Every sentence (up to ten) found in *S* is replaced by its counterpart in the *T* corpus. For every sentence found in *T* an *n*-gram vector is computed and then all these vectors are Ω-intersected. Since such *T* sentences contain, allegedly, the translation of some words from language *S* into language *T*, it can be supposed that the Ω intersection of their vectors would contain a kind of "translated" *n*-grams (see Figure 4). Those word chunks that do not appear in the S corpus are incorporated without "translation". Thus, we obtain a vector which is similar, in theory, to that which could be computed from a real translation from the original query.

The European Parliament Proceedings Parallel Corpus 1996-2003 [12] has been used as the sentence aligned corpus and the results obtained have been really interesting. In average terms, 38.59% of the *n*-grams from pseudo-translated query vectors are present within the vectors from actual translated queries and, in turn, 28.31% of the *n*-grams from the actual translated query vectors correspond to n-grams within the pseudo translated ones. In order to check this we have compared vectors obtained through pseudo translation of Spanish queries into English with the vectors computed from actual English topics. This constitutes another area for future work employing different parallel corpora (e.g., OPUS, `http://logos.uio.no/opus`) and improving the "translation" method.

This technique is related to those described by Pirkola *et al* [13] to find cross-lingual spelling variants or by McNamee and Mayfield [14] to "translate" individual *n*-grams. The difference between such techniques and ours is that we do not attempt to obtain word translations nor individual *n*-gram translations but a pseudo-translation for a whole *n*-gram vector containing *n*-grams from the target language that would likely appear in actual query translations. Such a vector can then be straightforwardly submitted to the IR system.

Topic 206 written in language S (Spanish)

Encontrar documentos en los que se habla de las discusiones sobre la reforma de instituciones financieras y, en particular, del Banco Mundial y del FMI durante la cumbre de los G7 que se celebró en Halifax en 1995.

Some sentences from corpus S (Europarl Spanish)

(1315) …mantiene excelentes relaciones con las instituciones financieras internacionales.
(5865) …el fortalecimiento de las instituciones financieras internacionales...
(6145) La Comisión deberá estudiar un mecanismo transparente para que las instituciones financieras europeas...

Counterpart sentences from corpus T (Europarl English)

(1315) …has excellent relationships with the international financial institutions..
(5865) …strengthening international financial institutions...
(6145) The Commission will have to look at a transparent mechanism so that the European financial institutions...

Pseudo-translated query vector (Ω-intersection of previous T sentences)

(al_i, anci, atio, cial, _fin, fina, ial_, inan, _ins, inst, ions,
itut, l_in, nanc, ncia, nsti, stit, tion, titu, tuti, utio)

Fig. 4. Procedure to pseudo translate a query written originally in a source language (in this case Spanish) onto a vector containing appropriate n grams from the target language (English in this example). Blanks have been replaced by underscores, just one chunk from the query has been pseudo translated (shown underlined)

5 Results Obtained by blindLight IR

As we said before our group submitted results for just two tasks: monolingual retrieval on the Russian collection and bilingual retrieval querying the English

Table 1. Top-5 and bottom-5 performing topics for monolingual and bilingual tasks. Top 5 are those with highest precision at 5 documents. Bottom-5 topics are those which do not provide any relevant result; the more relevant documents available within the collection, the worse the query performs. As can be seen, focused topics related to people, places and/or particular events are the best performers within blindLight IR prototype while broad queries are poorly managed by our system

Top-5 performing topics (ES-EN)	Top-5 performing topics (RU)
218 Andreotti and the Mafia	230 Atlantis-Mir Docking
248 Macedonia Name Dispute	209 Tour de France Winner
202 Nick Leeson's Arrest	210 Nobel Peace Prize Candidates
224 Woman solos Everest	211 Peru-Ecuador Border Conflict
205 Tamil Suicide Attacks	202 Nick Leeson's Arrest
Bottom-5 performing topics (ES-EN)	**Bottom-5 performing topics (RU)**
212 Sportswomen and Doping	227 Altai Ice Maiden
235 Seal-hunting	203 East Timor Guerrillas
241 New political parties	207 Fireworks Injuries
214 Multi-billionaires	228 Prehistorical Art
216 Glue-sniffing Youngsters	250 Rabies in Humans

collection using Spanish as query language. For the Russian task our prototype returned 72 of the 123 relevant documents with an average precision of 0.1433. With regards to the bilingual task we obtained 145 of the 375 relevant documents showing an average precision of 0.0644.

Such results are far from being good but we found them kind of encouraging. Firstly, it is our first participation in CLEF. Secondly, although average results are rather poor we can clearly separate classes of topics that obtain good results from other types which perform poorly (e.g., broad queries) showing us a future line of work.

6 Conclusions and Future Work

blindLight is a new technique related to classical *n*-gram vector space models and developed to perform several natural language processing tasks. We have shown that it is well-suited to extract keyphrases and automatic summaries from single documents [8] in addition to performing language identification and classification of natural languages [9]. At this moment we are testing its applicability to information retrieval since we totally agree with McNamee and Mayfield when they say that "knowledge-light methods can be quite effective" [14]. With regards to this goal, it must be said that partial results are not outstanding but we feel optimistic about this issue since poor performance is mostly constrained to broad topics and focused queries usually achieved reasonable precision.

Three areas require further work: (1) Similarity measures between queries and documents must be improved, perhaps with genetic programming. (2) Different parallel corpora should be used to enhance the n-gram pseudo-translator employed to perform bilingual IR. And (3) thorough research is needed to improve precision when broad topics are submitted to the system.

References

1. Salton, G., Wong, A. and Yang, C.S.: A vector space model for information retrieval. Communications of the ACM, 18(11), pp. 613-620 (1975)
2. D'Amore, R., Mah, C.P.: One-time complete indexing of text: Theory and practice. Proc. of SIGIR 1985, pp. 155-164 (1985)
3. Kimbrell, R.E.: Searching for text? Send an n-gram! Byte, 13(5), pp. 297-312 (1988)
4. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), pp. 61-74 (1993)
5. Ferreira da Silva, J., Pereira Lopes, G.: A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora. In Proc. of MOL6 (1999)
6. Ferreira da Silva, J., Pereira Lopes, G.: Extracting Multiword Terms from Document Collections. Proc. of VExTAL, Venice, Italy (1999)
7. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals, (English translation from Russian), Soviet Physics Doklady, 10(8), pp. 707-710 (1966).

8. Gayo-Avello, D., Álvarez-Gutiérrez, D., Gayo-Avello, J.: Naive Algorithms for Key phrase Extraction and Text Summarization from a Single Document inspired by the Protein Biosynthesis Process, in Biologically Inspired Approaches to Advanced Information Technology: 1st International Workshop, BioADIT 2004, A.J. Ijspeert, M. Masayuki, and N. Wakamiya (Eds), LNCS 3141, pp. 440-455, (2004)
9. Gayo-Avello, D., Álvarez-Gutiérrez, D., Gayo-Avello, J.: One Size Fits All? A Simple Technique to Perform Several NLP Tasks, in 4th International Conference, EsTAL 2004, J.L. Vicedo *et al* (Eds), LNAI 3230, pp. 267-278, (2004)
10. Peters, C. and Braschler, M. and Di Nunzio, G., and Ferro, N.: CLEF 2004: Ad Hoc Track Overview and Results Analysis, in 5th Workshop of the Cross-Language Evaluation Forum (CLEF 2004), Peters, C. *et al* (Eds), LNCS (in print).
11. Peters, C.; What happened in CLEF 2004, in 5th Workshop of the Cross-Language Evaluation Forum (CLEF 2004), Peters, C. *et al* (Eds), LNCS (in print).
12. Koehn, P.: Europarl: A Multilingual Corpus for Evaluation of Machine Translation, Draft, Unpublished, `http://www.isi.edu/~koehn/publications/europarl.ps`
13. Pirkola, A, Keskustalo, H., Leppänen, E., Känsälä, A. and Järvelin, K. (2002) Targeted *s*-gram matching: a novel *n*-gram matching technique for cross- and monolingual word form variants. Information Research, 7(2) (2002)
14. McNamee, P., Mayfield, J.: JHU/APL Experiments in Tokenization and Non-Word Translation. Working Notes for the CLEF 2003 Workshop. 21-22 August, Trondheim, Norway