Data compression approach to monolingual GIRT task: an agnostic point of view

Daniela Alderuccio[1], Luciana Bordoni[1] and Vittorio Loreto[2]
[1]ENEA UDA-Advisor, Centro Ricerche Casaccia
Via Anguillarese 301, 00060 Santa Maria di Galeria (RM), Italy
[2]"La Sapienza" Univ. in Rome, Physics Dept., P.le A. Moro 2, 00185 Rome, Italy
alderuccio@casaccia.enea.it, bordoni@casaccia.enea.it, loreto@roma1.infn.it

Abstract

In this paper we present a data-compression oriented approach to the information retrieval task in the scientific collection of GIRT. For this purpose we use a recently proposed general scheme for context recognition and context classification of strings of characters (in particular texts) or other coded information. Based on data-compression techniques, the key point of the method is the computation of a suitable measure of remoteness of two strings of characters. This measure of remoteness only reflects the distance in information between the two strings, i.e. the difference between the syntactic/structural elements of the sequences. The question we address is whether the informatic measure of remoteness between two sequences could account for their semantic distance. We have focused in particular on the monolingual GIRT tasks for German and English and we present here the results. It is worth stressing the generality and versatility of our information-theoretic method. It applies, in fact, to any kind of corpora of character strings, independent of the type of coding behind them. For texts, it is then language independent since it prescinds from any linguistic knowledge.

1 Introduction

The rapid increase in the amount of electronic information makes the need of effective and efficient accesses to texts and objects associated with texts of paramount importance. The task of an Information Retrieval (IR)[1, 2] system is to estimate the degree to which documents in a collection reflect the information expressed in a user query. IR techniques facilitate this access by developing a representation of user's information need or query, comparing it to representations of information objects, and retrieving those objects most closely matching that need.

Object representations are typically based on the words or vocabulary of a language. This seems to make sense because textual information is conveyed using words. However, language is ambiguous and there are many different ways to express a given idea or concept.

In recent years, there has been tremendous growth in the amount of multilingual, electronic media. The current boom in economic growth has lead more businesses and organizations to expand the boundaries of their organizational structure to include foreign offices and interests. Greater numbers of people are interested in data and information collected in other regions of the world as efforts to address issues of global concern lead to increased multi-national collaboration. The Internet has become a powerful resource for information in areas such as electronic commerce, advertising and marketing, education, research, banking and finance. This growth in availability of multi-lingual data in all areas of the public and private sector is driving an increasing need for systems that facilitate access to multi-lingual resources by people with varying degrees of expertise with foreign languages. Cross-Language Retrieval (CLR) technology is a means of addressing this need.

The Cross-Language Evaluation Forum (CLEF) aims to develop shared experiment designs that will allow research teams to compare their strategies for cross-language search assistance.

To enlarge our knowledge and experience with information retrieval in multilingual document collections, we participated in the Cross-Language Evaluation Forum (CLEF) this year. Our goal for this year was to participate in the monolingual tasks. We submitted monolingual runs for the English and German in the GIRT collection.

The paper is organised as follows. Section II describes the methodology used for doing this research. In section III we describe the monolingual GIRT experiments and the results. In section IV we present some remarks and we draw the conclusions.

2 Method description

Information extraction takes place *via* a two-step process. First comes the so-called *syntactic* step, where one identifies the structures present in the messages without associating any specific meaning to them. It is only in the second (or *semantic*) step that comprehension of meaning takes place, by connecting the syntactic information to previous experience and knowledge. As an example of this two step process, consider how we might identify the language in which a given text is written. In the first step we scan through the text and identify the syntactic structures: articles, verbs, adjectives, etc. But only one who "knows" the language can carry out the second phase, where the incoherent jumble of syntactic data is summarized in the specific meaning of the sentences. Other examples of this two-step process are how we recognize the subject of a given text, its historical background, and eventually its author.

We should stress that the syntactic and semantic levels are not always strictly related; the correlation between them may depend a lot on the specific source of information. In other words, suppose we could extract efficiently the syntactic information of a given sequence (e.g. a text). Could we obtain from this measure the information (in the semantic sense) we were trying to extract from the sequence? The answer to this question is far from being trivial.

Having this plan in mind, the first logical step is to provide ourselves with tools that can measure the amount of syntactic or structural information contained in a given string.

We shall follow an Information Theoretic approach [3, 4]. In Information Theory (IT) the word "information" acquires a very precise meaning, namely the "entropy" of the string. In a sense, entropy measures the *surprise* the source emitting the sequences can give us. Suppose the surprise one feels upon learning that an event E has occurred depends only on the probability of E. If the event occurred with probability 1 (certainty) our surprise at its occurrence would be zero. On the other hand if the probability of occurrence of the event E were quite small our surprise would be proportionally larger.

It is possible to extend the definition of the entropy for a generic string of characters, without any reference to its source. This is what we need to analyze a text whose source and/or statistical properties are *a priori* unknown. Among the many equivalent definitions of entropy the best for this case is the so-called Chaitin-Kolmogorov complexity, or Algorithmic Complexity (see for instance [5]): **the algorithmic complexity of a string of characters is given by the length (in bits) of the smallest program which produces the string as output**. A string is called complex if its complexity is proportional to its length. This definition is really abstract: for one thing, it is impossible, even in principle, to find such a program [5]. Since the definition is not based on the time the best program should take to reproduce the sequence, one can never be sure that there might not exist a shorter program that could eventually output the string in a larger (eventually infinite) time.

To overcome the intrinsic difficulty of measuring the entropy of a generic string of characters, we employ the relation between the entropy, h, and the maximum compression rate of a sequence $X_1, X_2, X_3, \ldots)$, expressed in an alphabet with M symbols. If the length T of the sequence is large enough, then it is not possible to compress it into another sequence (with an alphabet with M symbols) whose size is smaller than $Th/\ln M$. Therefore, noting that the number of bits needed for a symbol in an alphabet with M symbol is $\ln M$, one has that the maximum allowed compression rate is $h/\ln M$.

We now recall that there exist algorithms explicitly conceived to approach the theoretical limit of optimal coding. These are the file compressors or zippers. It is then intuitive that a typical zipper, besides reducing the space a file occupies on a memory storage device, can be considered an entropy meter. The better the compression algorithm, the closer the length of the zipped file to the minimal entropic limit, and hence the better will be the estimate of the entropy provided by the zipper. It is indeed well known that compression algorithms provide a powerful tool for the measure of entropy and more generally for the estimation of more sophisticated measures of complexity [3, 6].

Any algorithm that tries to estimate the entropy (or better the Algorithmic Complexity) of a string of characters (with arbitrary statistical properties) only carries out the syntactic step. To proceed to the semantic level we need to add other ingredients that could bridge the gap between the syntactic properties of a sequence (e.g. a text) and its semantic aspects. With this precise aim recently it has been proposed a general method [7] for context recognition and context classification of strings of characters or other coded information. Based on data-compression techniques, the key point of the method is the computation of a suitable measure of remoteness of two bodies of knowledge. This idea has been used for authorship attribution and, defining a suitable distance between sequences, for languages phylogenesis.

The key point of the new method is simple. Suppose you want to estimate the "distance" (or similarity) between texts A and B in terms of their informatic content. For instance, for two texts written in different languages (e.g. English and Italian), their "distance" is a measure of the difficulty experienced by a typical speaker of mother tongue A in understanding the text written in language B. More generally one can imagine to measure the remoteness between two languages by measuring the distance between two *typical* texts representative of the two languages. A possible way to estimate the remoteness between two texts has been proposed in [7] in terms of the so-called the relative entropy [8]. At the zero-th order of approximation we can describe the procedure as follows. We take a long English text and we append to it an Italian text, then we zip the resulting text. The zipper begins reading the file starting from the English text and after a while it is able to encode (almost) optimally the English file (this is the aim of the compression procedure). When the Italian part begins, the zipper starts encoding it in a way which is optimal for the English. So the first part of the Italian file is encoded with the English code. After a while the zipper "learns" Italian and changes its rules. Therefore if the length of the Italian file is "small enough", the difference between the length (in bits) of the zipped English+Italian text and the length (in bits) of the English text zipped alone will give a measure of the "distance" between the two texts. Figure 1 illustrates this *learning* process when using the LZ77 scheme [11], one of the best known compression algorithm. We do not enter in too many details here and we refer the reader to [7, 9, 10].

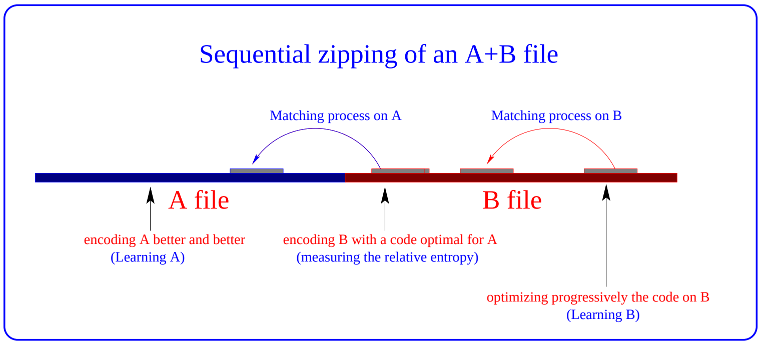
Sequential zipping of an A+B file
Matching process on A
Matching process on B
A file
B file
encoding A better and better
(Learning A)
encoding B with a code optimal for A
(measuring the relative entropy)
optimizing progressively the code on B
(Learning B)

Figure 1: Schematic description of how a sequential zipper optimizes its features at the interface between two different sequences A and B while zipping the sequence $A + B$ obtained by simply appending B after A.

Once one has defined the entropy-based remoteness between two texts, or more generally between two strings of characters one has in his hands a powerful tool to implement suitable algorithms for recognizing the “context” of a given sequence (for example for a sequence of characters representing a text, we might like to recognize the language in which it is written, the subject treated, and its author).

In the framework of CLEF 2003 we have applied our method to the GIRT task both in English and in German. From the point of view of our method the GIRT task could be summarized as follows. We have a a corpus of known documents D_i $(i = 1, 151319)$ and a corpus of topics T_j $(j = 76, 100)$. For each run and for each topic T_j we compare an artificial topic text AT_j with a corpus of artificial reference documents AD_i $(i = 1, 151319)$. The construction of the artificial topics as well as of the artificial reference documents depends on the specific run and it will be discussed in the next section. The idea of constructing artificial texts follows the need of selecting, for each run, only few information contained in the topics and in the reference documents. In this sense an artificial text is supposed to possess the features of the original documents neeeded for the specific run under consideration.

It is important to stress that, in the implementation of our method, a text is considered as a unique sequence of characters (ASCII format) where all the carriage returns and punctuation marks have been removed. Additionally all the capital letters (even in German) have been converted to small letters.

For each topic, one is interested in detecting the artificial reference texts AD_i closest (according to some rule) to the AT_j one and in providing a ranking of them. In our case for each artificial topic text AT_j we rank all the other documents (restricting the list to the first 1000) according to the value of the so-called cross-entropy per character between the artificial topic text AT_j and each artificial reference text. We recall that the cross-entropy between a text A and a text B is given by the minimal number of bits necessary to codify text A with a code optimal for text B (see [9, 10] for details).

3 Monolingual GIRT experiments

The German Indexing and Retrieval Data Base (GIRT) is a domain-specific text collection which covers the domain of social sciences.

The GIRT collection used for CLEF2003 campaign contains 151319 documents, available as two parallel corpora, containing the same documents in German (GIRT4-DE) and English (GIRT4-EN). It is composed by two databases: FORIS and SOLIS. FORIS (Social Science Research Information System) offers information on theoretical and empirical research projects in Austria, Germany, Switzerland. Project information comes from the higher education sector. Sources of information are surveys and institutions doing social science research. SOLIS (Social Science Literature Information System) contains information on German social science literature (journal articles, monographs, grey literature, etc. In addition to monographs and compilations, around 550 journals are examined for relevant articles). Both databases are in German. There are also English translations of the titles for nearly all documents and about 17% of the documents have an English abstract.

Reference documents contain bibliographical information (author, language of the document, publication year), controlled-terms, free-terms, classification texts and abstracts. Controlled terms are indexing terms assigned by experts; on average each document has 10 indexing terms. Classification terms assign to each document one or more specifications of topic. In our experiments we have used only abstracts and controlled-terms.

The topics contain three main parts: a short title, a one sentence description and a more complex narrative, specifying the relevance assessment criteria. In our experiments we have used only the titles and the description parts.

RUN ID	ENEASAPDTC	ENEASAPDTDC	ENEASAPDTA	ENEASAPETC	ENEASAPETDC
Recall	0.6528	0.6145	0.5904	0.7087	0.6996

Table 1: Average recall (averaged over all the topics) for our five runs. What is reported is, for each run, the ratio between the total number of relevant documents retrieved (independently of the position in the ranking) and the total number of existing relevant documents.

3.1 Experiment with the German and English collections

For the German Monolingual task we have performed the following three runs:

ENEASAPDTC In this run the artificial topic texts are composed only by the titles. On the other hand the artificial reference texts are obtained by appending, for each original reference text, all the controlled-terms.

ENEASAPDTDC In this mandatory run the artificial topic texts are composed only by appending the titles and the desciptions of each topic. The artificial reference texts are obtained by appending, as before, for each original reference text, all the controlled-terms.

ENEASAPDTA In this run the artificial topic texts are composed only by the titles. The artificial reference texts are composed by the abstracts of the original reference texts.

For the English Monolingual task we have performed the following runs:

ENEASAPETC In this run the artificial topic texts are composed only by the titles. On the other hand the artificial reference texts are obtained by appending, for each original reference text, all the controlled-terms.

ENEASAPETDC In this mandatory run the artificial topic texts are composed only by appending the titles and the desciptions of each topic. The artificial reference texts are obtained by appending, as before, for each original reference text, all the controlled-terms.

In this case the equivalent of the ENEASAPDTA run is not present since only about 17% of the English reference documents contained an abstract. The retrieval would not be exhaustive because the documents without abstract would have been ignored.

Table 1 reports, for the different runs, the average recall (averaged over all the topics), i.e. the ratio between the number of relevant documents retrieved (independently of the position in the ranking) and the total number of existing relevant documents.

Fig. 2 and fig. 3 report the results for the recall obtained for each single topic and for each run, for English and German, respectively.

We report in Fig. 4 the results for the average precision obtained for each single topic and for each run, compared with the average performances.

4 Remarks and Conclusions

A few remarks are in order to comment our results. It is worth to recall how our method, being completely agnostic, i.e. not at all informed from the linguistic point of view, presents lights and shadows. On the one hand we consider our method very powerful for its generality: we can apply it to whatever language without any kind of modifications or tuning. Moreover the fact of not being informed linguistically makes our method almost free from biases, except those introduced by the choice of the procedure to construct the artificial texts to be compared. We stress how we consider the agnostic character very important in order to have algorithms suitable for a wide spectrum of applications. On the other hand it is clear how the agnostic character could become

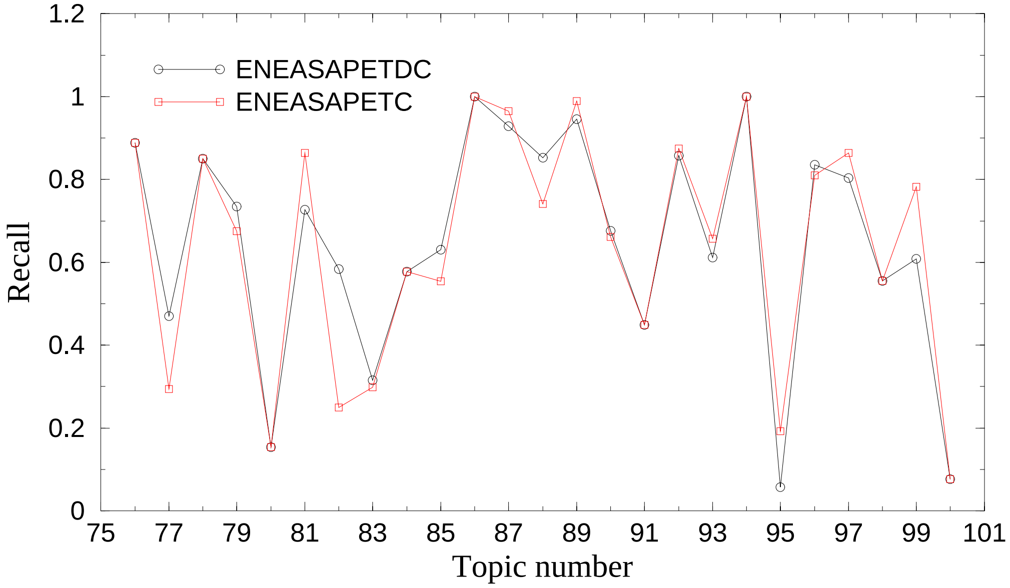

Figure 2: English: recall obtained for each single topic and for each run.

a difficulty whenever a disambiguation procedure would be important. This point can become crucial when the length of the texts to be compared is very small (a few words), as it is the case for the CLEF topics.

We think that the properties of our algorithms, with their lights and shadows, can explain the quality of our results and in particular the fact that we have obtained a quite high overall recall with a not corrispondently high precision. It is precisely in the direction of improving precision we should work. We can imagine two possible paths. On the one hand the possibility to perform some expansion procedures for the queries and/or the reference text, in order to reduce the strong bias due to the extreme shortness of the documents to be compared. On the other hand one could imagine some hybrid procedure that could make an effective balance between the need of keeping the method as uninformed as possible and the insertion of linguistically relevant tips.

References

[1] Rijsbergen CJ (1979) Information Retrieval. Second Edition Butterworths, London.

[2] Croft B (ed.) (2003) Advances in Information Retrieval - Recent research from the Centre for Intelligent Information Retrieval. Kluwer Academic Publishers.

[3] Shannon CE (1948) A mathematical theory of communication. The Bell System Technical Journal, 27:379-423 and 623–656.

[4] Zurek WH (editor) (1990) Complexity, Entropy and Physics of Information. Addison-Wesley, Redwood City.

[5] Li M and Vitányi P (1997) An introduction to Kolmogorov complexity and its applications. Springer, 2nd ed..

[6] Khinchin AI (1957) Mathematical Foundations of Information Theory. Dover, New York.

[7] Benedetto D, Caglioti E and Loreto V (2002) Language trees and zipping. Physical Review Letters 88:048702-048705.

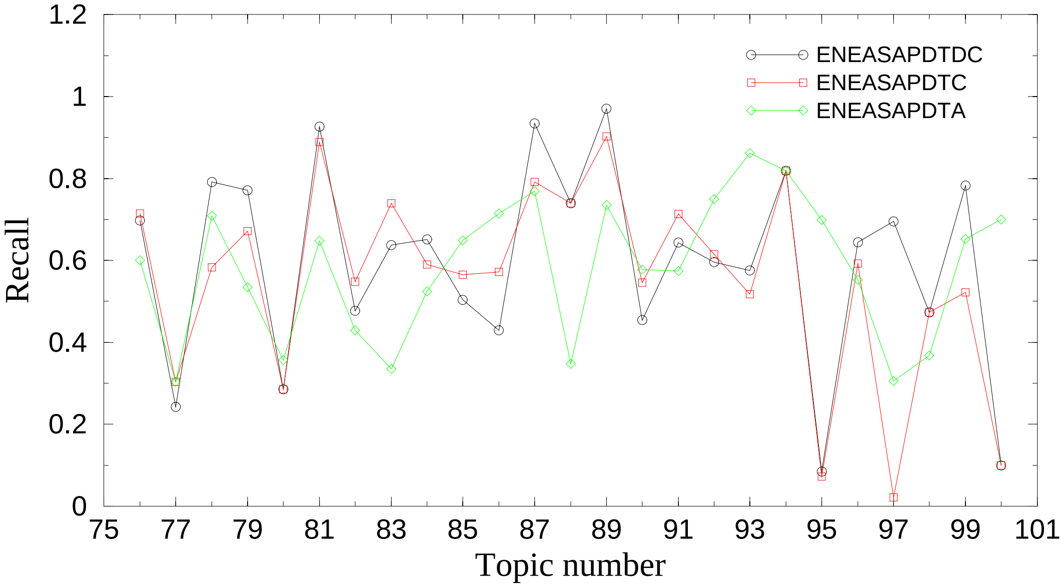

Figure 3: German: recall obtained for each single topic and for each run.

[8] Ziv J and Merhav N (1993) A measure of relative entropy between individual sequences with applications to universal classification. IEEE Transactions on Information Theory, 39:1280-1292.

[9] Puglisi A, Benedetto D, Caglioti E, Loreto V and Vulpiani A (2003) Data compression and learning in time sequences analysis. Physica D, 180:92-107.

[10] Benedetto D, Caglioti E and Loreto V (2003) Zipping out relevant information. Invited column ”Computing Prescriptions” in the AIP/IEEE journal *Computing in Science and Engineering* January-February issue.

[11] Ziv J and Lempel A (1977) A Universal Algorithm for Sequential Data Compression. IEEE Trans. Inf. Th., 23: 337-343.

[12] Braschler M and Ripplinger B (2003) Stemming and Decompounding for German Text Retrieval. Advances in Information Retrieval - Proceedings of the 25th European Conference on IR Research, ECIR 2003, Pisa, Italy.

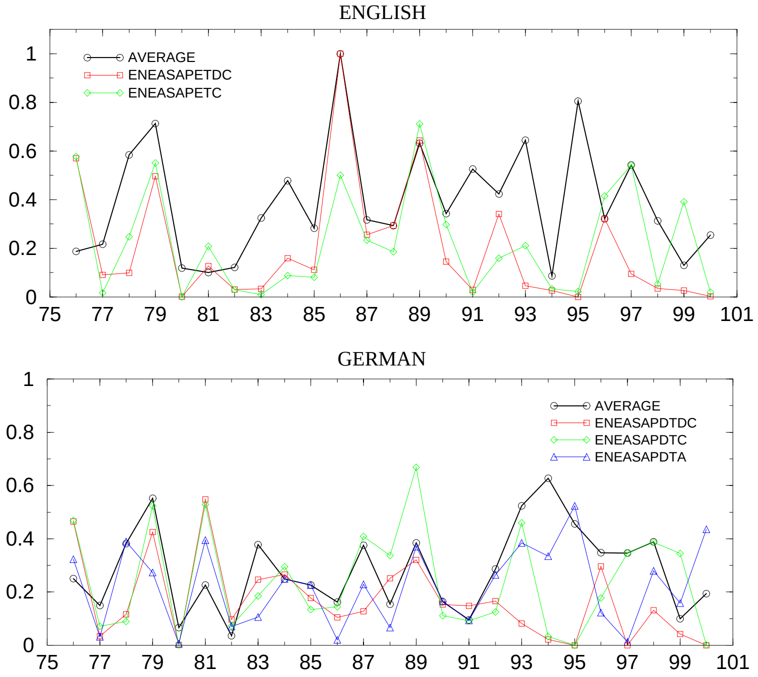

Figure 4: Average precisions obtained for each single topic and for each run, compared with the average performances. Upper graph: English. Lower graph: German.