

510035

CACAO

First validation report (technical adequacy)

Deliverable number	<i>D.6.1</i>
Dissemination level	<i>Public</i>
Delivery date	<i>28 February 2009</i>
Status	<i>Final</i>
Author(s)	<i>Alessio Bosca, Daniele Gobbetti</i>



eContentplus

This project is funded under the *eContentplus* programme¹, a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

¹ OJ L 79, 24.3.2005, p. 1.

Table of Contents

1	GOAL	3
2	METHODOLOGY	3
2.1	BRIEF EXPLANATION OF THE TERMINOLOGY USED	3
2.1.1	<i>Precision</i>	3
2.1.2	<i>Recall</i>	4
2.1.3	<i>F-measure</i>	4
2.1.4	<i>Average precision of precision and recall</i>	4
2.1.5	<i>Multilingual coverage</i>	4
2.2	EVALUATION METHODOLOGIES	5
2.2.1	<i>TEL@CLEF set of queries</i>	5
2.3	CACAO TOPICS PROCESSING	6
3	EVALUATION	6
3.1	TEL@CLEF 2008 RESULTS	6
3.2	EVALUATION OF THE PROTOTYPE	7
3.2.1	<i>Prototype results</i>	7
4	CONCLUSION AND FINAL REMARKS	11

WP6: Assessment and Evaluation

Work package number :	6	Start date:	12	End date:	24				
Work package title:	Assessment and Evaluation								
Applicants involved:	2	3	4	5	8	9			

1 Goal

This deliverable aims to assess the performances of the CLIR after deploying the first prototype¹ of the GUI and thus allowing all the partners to access the CACAO system and test it against data harvested from the partners' libraries. The aspects documented in this deliverable concern the **technical adequacy** of CACAO², namely the increase in F-measure³ and multilingual coverage⁴ provided by the first prototype.

Such measures were performed by using a set of queries and predefined human-selected answers, according to standard international methodologies such as TREC/CLEF⁵.

Although CACAO's participation in last year's TEL@CLEF⁶ task with an early prototype of the CLIR system was rather modest, a considerable improvement of its performance – documented in the present Deliverable – can be stated.

2 Methodology

2.1 *Brief explanation of the terminology used*⁷

There have been several different measures proposed for evaluating the performance of information retrieval systems. These measures require a collection of documents and a query. All common measures described here assume a ground truth notion of relevancy: every document is known to be either relevant or non-relevant to a particular query.

2.1.1 Precision

Precision is the fraction of the documents retrieved that are relevant to the user's information need.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

¹ Cf. D4.1

² The only CACAO engine now in place is the federated one, it is not yet possible to evaluate User Satisfaction. It will be part of D6.2 and D6.3. Cf. Description of Work p. 52

³ Cf. Sec. 2.1.3 of this deliverable

⁴ Cf. Sect. 2.1.5 of this deliverable

⁵ Cf. Sect. 2.2.1 of this deliverable

⁶ <http://www.clef-campaign.org/2008>

⁷ Cf. http://en.wikipedia.org/wiki/Information_retrieval

Precision takes all retrieved documents into account. It can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system.

Note that the meaning and usage of "**precision**" in the field of Information Retrieval differs from the definition of accuracy and precision within other branches of science and technology.

2.1.2 Recall

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

"**Recall**" can be looked at as the probability that a relevant document is retrieved by the query.

It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by computing the precision.

2.1.3 F-measure

The weighted harmonic mean of precision and recall, the traditional¹ F-measure is:

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}$$

This is also known as the F_1 measure, because recall and precision are evenly weighted.

2.1.4 Average precision of precision and recall

The precision and recall are based on the whole list of documents returned by the system. Average precision emphasizes returning more relevant documents earlier. It is average of precisions computed after truncating the list after each of the relevant documents in turn:

$$\text{AveP} = \frac{\sum_{r=1}^N (P(r) \times \text{rel}(r))}{\text{number of relevant documents}}$$

where r is the rank, N the number retrieved, $\text{rel}()$ a binary function on the relevance of a given rank, and $P()$ precision at a given cut-off rank.

In the rest of this deliverable the **Mean value of the Average Precision** (MAP) will be used as a performance indicator.

2.1.5 Multilingual coverage

The TEL@CLEF task was performed on a corpus that consisted of bibliographic metadata divided into 3 collections, extracted from British, French and Austrian national libraries. TEL@CLEF task offers a set of subtasks reflecting the multilinguality of the data,

¹ Cf. http://en.wikipedia.org/wiki/Information_retrieval for further refinements

respectively focusing on monolingual and bilingual information retrieval; 50 topics have been prepared for each of the 3 main collection languages and each topic has 2 fields: a title with 2-4 key terms and a description field, containing a sentence that specifies in more detail the information needs of the user¹. The aforementioned 50 topics were translated into Italian by one of the library partners for evaluation purposes.

2.2 Evaluation methodologies

2.2.1 TEL@CLEF set of queries

Each query in the TEL@CLEF (known as **topic**) is made up by a **title** and a **description**, as well as a unique identifier (**identifier**) and a **language** element. The queries are organised in an XML document as illustrated below:

<pre><topic lang="en"> <identifier>10.2452/451-AH</identifier> <title>Roman Military in Britain</title> <description> Find books or publications on the Roman invasion or military occupation of Britain. </description> </topic></pre>
<pre><topic lang="fr"> <identifier>10.2452/451-AH</identifier> <title>L'armée romaine en Grande-Bretagne</title> <description> Trouver des livres ou des publications sur l'invasion et l'occupation de la Grande-Bretagne par les Romains. </description> </topic></pre>
<pre><topic lang="de"> <identifier>10.2452/451-AH</identifier> <title>Römisches Militär in Britannien</title> <description> Finden Sie Bücher oder Publikationen über die römische Invasion oder das Militär in Britannien. </description> </topic></pre>
<pre><topic lang="it">² <identifier>10.2452/451-AH</identifier> <title>L'occupazione romana della Britannia</title> <description> Trova libri o pubblicazioni sull'invasione e l'occupazione della Britannia da parte dei Romani. </description> </topic></pre>

¹ Cf. http://www.clef-campaign.org/2008/working_notes/bosca-paperCLEF2008.pdf

² NB: As stated in Sect. 2.1.5 the Italian topics translation is not official and was prepared by one of the library partners for evaluation purposes

For each topic a list of the correct results is built by experts and kept secret: it represents the baseline against all the CLEF participants are evaluated. The disclosure of this list yields to a **gold standard** allowing each participant to test further improvements in own algorithms. The aforementioned **gold standard** covers the official languages of the TEL@CLEF task (English, German and French).

2.3 CACAO Topics processing¹

The approach adopted by the CACAO system **in the TEL@CLEF task** for dealing with user queries is based on the free keywords search; therefore while the title field of TEL topics already fits this model, the description field has been processed in order to extract a set of relevant keywords from the sentence.

For this purpose a simple keyword extractor module has been used for each of the main languages present in the corpus (English, French and German).

Each description sentence has been analysed in order to extract two different kinds of data, one representing the content type of the items to be retrieved (as novels, poetry or photo collections) and the other conveying additional detail on user interests.

The keywords retrieved in this process have been lemmatised² and different weights have been assigned to them by the system according to their frequency in both of the topic fields (title and description). During the lemmatisation process named entities³ have also been identified, to avoid their translation into the target languages. According to the subtasks (monolingual or bilingual) the keywords were either translated into the target language or directly submitted to the Lucene search engine.

The translation process has exploited internal resources (inter-lingual indexes or bilingual dictionaries) and online dictionaries; the so-obtained translation candidates have been disambiguated using the corpus based semantic vectors⁴, computed by the CACAO system on the collections metadata.

Experiments involving query expansion have enriched the keywords groups (either in the original or in the target language) through the exploitation of the corpus based semantic vectors.

3 Evaluation

3.1 TEL@CLEF 2008 results

The following table presents the evaluation results from CACAO participation in TEL@CLEF 2008 competition (see <http://www.clef-campaign.org/> for further details). Such performances are therefore relative to the early implementation of the CACAO system, at month 5 of project lifespan.

¹ Cf. http://www.clef-campaign.org/2008/working_notes/bosca-paperCLEF2008.pdf for a more detailed explanation

² Cf. <http://en.wikipedia.org/wiki/Lemmatization>

³ Cf. http://en.wikipedia.org/wiki/Named_entity_recognition

⁴ Cf. <http://research.celi.it/Wiki.jsp?page=Semantic%20Vectors>

Lang	Index	#Relevant	#Retrieved	MAP	R-Prec	P@5	P@10	P@20
en	BL	2533	1625	0.17	0.21	0.42	0.32	0.28
fr	BL	2533	699	0.06	0.08	0.15	0.12	0.1
de	BL	2533	627	0.05	0.06	0.14	0.12	0.09
en	BNF	1339	451	0.07	0.09	0.11	0.11	0.09
fr	BNF	1339	742	0.17	0.19	0.31	0.28	0.23
de	BNF	1277	86	0.02	0.02	0.03	0.03	0.03
en	ONB	1637	489	0.04	0.06	0.1	0.08	0.08
fr	ONB	1637	96	0.02	0.02	0.03	0.03	0.03
de	ONB	1637	740	0.09	0.1	0.15	0.1	0.8

Table 1: CLEF 2008 evaluation results

3.2 Evaluation of the prototype

This section presents the results of the TEL@CLEF evaluation procedure performed against the current implementation of CACAO, month 15 of project lifespan. Different tests have been set up for the purpose:

- **Test1:** using as **input** both the title and description of the topics and **searching** both source¹ language and destination/target language.
 - The destination language is considered to be the official one of the collection: English for British Library (BL), French for Bibliothèque nationale de France (BNF), German for Österreichische Nationalbibliothek (ONB).
- **Test2:** using as **input** only the title of the topics (emulating a CACAO-like query) and **searching** both source language and destination/target language.
- **Test3:** using as **input** only the title of the topics (emulating a CACAO-like query) and **searching** only the destination/target language.
- **Test4:** using as **input** both the title and description of the topics, **searching** only the destination/target language.
- **Test5:** like test 4 but using SYSTRAN@² for translating textual information.
- **Test6:** like test 3 but using SYSTRAN@ for translating textual information.

The first test (Test 1) will be used as comparison with the TEL@CLEF performances while Test 2 and 3 are more close to the expected performances in real-world usage of the CACAO system. Test 5 and 6 finally allow evaluating the performances of “internal” multilingual resources (available from CACAO partners) against commercially available machine translation systems.

3.2.1 Prototype results

¹ “Source” language being the language of the query and “destination/target” language being the language of desired documents

² SYSTRAN@ is a registered trademark of SYSTRAN Software Inc. Cf. <http://www.systransoft.com>

Test 1

Lang	Index	#Relevant	#Retrieved	MAP	R-Prec	P@5	P@10	P@20
de	BL	2533	1212	0.13	0.17	0.32	0.26	0.21
en	BL	2533	1632	0.22	0.25	0.51	0.42	0.35
fr	BL	2533	1085	0.09	0.13	0.22	0.2	0.17
it	BL	2533	1194	0.12	0.16	0.3	0.25	0.21
de	BNF	1339	438	0.08	0.09	0.14	0.13	0.11
en	BNF	1339	595	0.17	0.19	0.29	0.22	0.16
fr	BNF	1339	1004	0.21	0.22	0.34	0.28	0.22
it	BNF	1339	606	0.12	0.13	0.2	0.17	0.16
de	ONB	1637	742	0.11	0.15	0.29	0.22	0.2
en	ONB	1637	669	0.1	0.13	0.23	0.18	0.14
fr	ONB	1637	436	0.05	0.07	0.14	0.12	0.09
it	ONB	1637	411	0.07	0.09	0.16	0.14	0.09

*Table 2: Test 1***Test 2**

Lang	Index	#Relevant	#Retrieved	MAP	R-Prec	P@5	P@10	P@20
de	BL	2339	836	0.11	0.14	0.24	0.2	0.17
en	BL	2533	1390	0.15	0.19	0.36	0.3	0.25
fr	BL	2533	1087	0.1	0.13	0.23	0.2	0.16
it	BL	2533	1027	0.08	0.11	0.19	0.16	0.13
de	BNF	1244	292	0.06	0.06	0.08	0.09	0.07
en	BNF	1339	613	0.12	0.13	0.24	0.17	0.13
fr	BNF	1299	846	0.16	0.18	0.28	0.24	0.19
it	BNF	1304	599	0.11	0.13	0.17	0.16	0.13
de	ONB	1605	532	0.08	0.11	0.25	0.2	0.15
en	ONB	1637	517	0.08	0.11	0.19	0.15	0.14
fr	ONB	1637	434	0.06	0.08	0.12	0.11	0.09
it	ONB	1577	399	0.06	0.08	0.12	0.11	0.08

*Table 3: Test 2***Test 3**

Lang	Index	#Relevant	#Retrieved	MAP	R-Prec	P@5	P@10	P@20
de	BL	2339	837	0.12	0.14	0.25	0.2	0.17
en	BL	2533	1390	0.15	0.19	0.36	0.3	0.25
fr	BL	2426	1122	0.11	0.13	0.23	0.19	0.16
it	BL	2533	997	0.08	0.11	0.2	0.16	0.12
de	BNF	1201	302	0.06	0.07	0.1	0.09	0.08
en	BNF	1339	601	0.11	0.12	0.19	0.16	0.13
fr	BNF	1299	846	0.16	0.18	0.28	0.24	0.19
it	BNF	1304	611	0.12	0.13	0.2	0.17	0.14
de	ONB	1605	532	0.08	0.11	0.25	0.2	0.15
en	ONB	1576	439	0.08	0.11	0.21	0.18	0.13
fr	ONB	1637	423	0.08	0.1	0.16	0.14	0.11
it	ONB	1577	407	0.06	0.07	0.11	0.1	0.08

Table 4: Test 3

Test 4

Lang	Index	#Relevant	#Retrieved	MAP	R-Prec	P@5	P@10	P@20
de	BL	2533	1271	0.14	0.17	0.36	0.29	0.23
en	BL	2533	1632	0.22	0.25	0.51	0.42	0.35
fr	BL	2533	1257	0.11	0.14	0.3	0.24	0.2
it	BL	2533	1265	0.12	0.15	0.3	0.25	0.2
de	BNF	1339	471	0.08	0.09	0.16	0.12	0.1
en	BNF	1339	612	0.13	0.14	0.22	0.17	0.15
fr	BNF	1339	1004	0.21	0.22	0.34	0.28	0.22
it	BNF	1339	640	0.12	0.14	0.21	0.18	0.15
de	ONB	1637	742	0.11	0.15	0.29	0.22	0.2
en	ONB	1637	594	0.08	0.12	0.28	0.21	0.14
fr	ONB	1637	455	0.05	0.08	0.18	0.15	0.11
it	ONB	1637	420	0.06	0.08	0.18	0.13	0.09

Table 5: Test 4

Test 5

Lang	Index	#Relevant	#Retrieved	MAP	R-Prec	P@5	P@10	P@20
de	BL	2533	1416	0.17	0.21	0.42	0.33	0.26
en	BL	2533	1632	0.22	0.25	0.51	0.42	0.35
fr	BL	2533	1476	0.19	0.22	0.46	0.37	0.3
it	BL	2533	1305	0.14	0.16	0.36	0.29	0.24
de	BNF	1339	772	0.13	0.14	0.26	0.2	0.16
en	BNF	1339	866	0.15	0.16	0.3	0.24	0.18
fr	BNF	1339	1004	0.21	0.22	0.34	0.28	0.22
it	BNF	1339	649	0.14	0.15	0.21	0.18	0.14
de	ONB	1637	742	0.11	0.15	0.29	0.22	0.2
en	ONB	1633	589	0.09	0.11	0.3	0.21	0.16
fr	ONB	1637	596	0.11	0.13	0.34	0.27	0.19
it	ONB	1637	416	0.07	0.09	0.2	0.14	0.1

Table 6: Test 5

Test 6

Lang	Index	#Relevant	#Retrieved	MAP	R-Prec	P@5	P@10	P@20
de	BL	2365	1027	0.1	0.13	0.25	0.21	0.18
en	BL	2533	1390	0.15	0.19	0.36	0.3	0.25
fr	BL	2533	1277	0.13	0.18	0.34	0.29	0.24
it	BL	2258	1007	0.09	0.12	0.22	0.17	0.15
de	BNF	1339	671	0.1	0.11	0.16	0.15	0.12
en	BNF	1339	757	0.11	0.12	0.22	0.19	0.15
fr	BNF	1299	846	0.16	0.18	0.28	0.24	0.19
it	BNF	1264	458	0.1	0.1	0.16	0.14	0.11
de	ONB	1605	532	0.08	0.11	0.25	0.2	0.15
en	ONB	1377	410	0.06	0.08	0.19	0.15	0.12
fr	ONB	1582	545	0.1	0.13	0.32	0.24	0.18
it	ONB	1489	366	0.06	0.07	0.16	0.11	0.08

Table 7: Test 6

4 Conclusion and final remarks

The following table presents a performance comparison in MAP (Mean Average Precision) and R-Precision between results from [TEL@CLEF 2008](#) and test 1.

Lang	Index	#Relevant	#Retrieved	R-Prec	#Relev	#Retrie	MAP	R-Prec	R-prec gain (%)
de	BL	2533	627	0.06	2533	1212	0.13	0.17	171.13
en	BL	2533	1625	0.21	2533	1632	0.22	0.25	17.7
fr	BL	2533	699	0.08	2533	1085	0.09	0.13	58.54
de	BNF	1277	86	0.02	1339	438	0.08	0.09	376.19
en	BNF	1339	451	0.09	1339	595	0.17	0.19	100.85
fr	BNF	1339	742	0.19	1339	1004	0.21	0.22	16.03
de	ONB	1637	740	0.1	1637	742	0.11	0.15	50
en	ONB	1637	489	0.06	1637	669	0.1	0.13	110.7
fr	ONB	1637	96	0.02	1637	436	0.05	0.07	250

Table 8: Performance Comparison (clef08 and test1)

As the comparison shows, the increase in performances since TEL@CLEF is significant (the ratio going from 16% for French documents in the BNF collection up to 376% for German documents in the same collection). The rows for language Italian are missing from table 8 because no comparison with TEL@CLEF results is possible: the performances in this language are however comparable (in terms of MAP) with other languages (ranging from 0.07 to 0.12)¹.

¹ Cf. Table “Test 1” in section 3.2.1 of this document.