| | |
|---|---|
| **Project ref. no.** | *IST-1999-11438* |
| **Project acronym** | **MUCHMORE** |
| **Project full title** | Multilingual Concept Hierarchies for Medical InformationOrganization and Retrieval |

| | |
|---|---|
| **Security (distribution level)** | *Public* |
| **Contractual date of delivery** | *Month 6 (Dec 2000)* |
| **Actual date of delivery** | *Month 7 (Jan 2001)* |
| **Deliverable number** | *D3.2* |
| **Deliverable title** | *Performance Testing Plan, including measures, methodology, timetables (Report)* |
| **Type** | *RE* |
| **Status & version** | *Final* |
| **Number of pages** | *13* |
| **WP contributing to the deliverable** | *WP3* |
| **WP / Task responsible** | *EIT* |
| **Author(s)** | *EIT, CMU, DFKI* |
| **EC Project Officer** | *Yves Paternoster* |
| **Keywords** | *Evaluation, cross-language, evaluation methods, test collections* |
| **Abstract (for dissemination)** | *Deliverable 3.2 details the methodologies to be used for the evaluation of the cross-language components that will be developed for MUCHMORE. Well-established measures, such as Precision and Recall, will be used for evaluation of effectiveness. The report lists the available test collections and test data sets to be used for such tests, and gives a timetable with respect to the forthcoming work in workpackage 8.* |

# Table of Contents

# Performance Testing Plan, including measures, methodology, timetables (Report)

## 1 Executive summary

As part of the MUCHMORE project, new components for cross-language information access on medical data will be developed. This deliverable outlines the methodology and the tools that will be used to evaluate the effectiveness of the prototypes that are being built. Both large, so-called "TREC-style" tests for near-final or final prototypes and simpler tests for intermediate prototypes will be employed.

The larger evaluations towards the end of the lifetime of the project will be using both data coming from the public OHSUMED test collection as well as data collected by the MUCHMORE participants. The smaller, intermediary tests that are considered will be using additional data available to the MUCHMORE partners, primarily provided by ZInfo.

It is planned to use well-established and proven measures for effectiveness, such as precision and recall, as well as known-item searches and overlap measures. The meaning of these tools for evaluation is well understood today, thanks to extensive research carried out in the past. Use of these popular measures allows us to maintain comparability with similar evaluations.

# 2 Introduction

This deliverable discusses the planned procedure for the evaluation of the MUCHMORE cross-language components.

The MUCHMORE consortium intends to cover the following steps with regard to evaluation.

1. TREC-style tests on data from the medical domain. Because such tests require a substantial effort, they will be limited to the final or near-final prototype of the software.
2. Tests of the intermediary prototypes. For this, overlap measures and known-item searches are considered.

These methodologies are described below.

This document covers the following questions:

1. Methodologies of the evaluations conducted during the life span of the MUCHMORE project
2. Definition of the evaluation measures to be used
3. Time table for test setup and evaluation experiments

# 3 Background

One of the goals of the MUCHMORE project is to develop a prototype for cross-lingual information access, allowing professionals in the medical field to access information in languages different than their preferred query formulation language.

Since cross-language information retrieval has retrieved considerable interest in recent years, first test beds and forums for the evaluation of cross-lingual retrieval systems have been developed. Specifically, the well-known TREC [http://trec.nist.gov] initiative started including cross-language retrieval evaluation beginning with TREC-6 in 1997. Around the same time, the French Amaryllis [http://www.inist.fr/accueil/profran.htm] forum conducted a limited evaluation in this field. In 1999, the NTCIR [http://research.nii.ac.jp/ntcir/index-en.html] initiative in Japan was founded, introducing a forum for Asian languages. And 2000 saw the start of a new series of evaluation campaigns called CLEF [http://www.clef-campaign.org], which is a spin-off of the earlier cross-language evaluations in TREC that has been moved under new coordination to Europe.

# 4 Evaluation methodology

The MUCHMORE project intends to conduct some tests using the methodology that is usually applied at the above mentioned evaluation forums. Unfortunately, it does not seem possible to directly use the test collections coming out of these forums, however, since there is a significant discrepancy in the domain of the test data (these forums all use some form of newspaper or academic reports/papers, whereas MUCHMORE is concerned with medical data).

The goal is therefore to follow the well-established methodology, but using suitable documents from the medical domain. All these forums offer "automatic tests", modeled by providing participants with test data and a set of corresponding "expressions of information need". It is then the task of the participants to use the topics to construct queries appropriate for their systems and run these against the test data. The results of this process are then submitted to the coordinator of the evaluation campaign for analysis.

The test data and the test queries collectively make up the "test collection". In summary, automatic evaluation works by executing the following steps:

1. A suitable test collection is acquired.
2. The test queries are run using the system to be evaluated.
3. The results are analyzed, usually in terms of recall and precision.

# 5 Evaluation measures

The two measures most commonly computed for evaluation are "precision" and "recall", defined as follows:

$$\mathrm{Pr}\,ecision = \frac{number\_of\_relevant\_documents\_retrieved}{total\_number\_of\_relevant\_documents}$$

This is the ratio of the relevant ("good") documents returned by the system compared to the overall number of retrieved documents. In a database scenario, where the search is conducted on structured data with controlled content fields, precision is usually 100%, i.e. items are retrieved only if they perfectly match the search criteria. In full text search, precision usually is substantially below the theoretically optimal level, due to the ambiguities of natural language and interpretation problems of the searcher's intent.

$$\mathrm{Re}\,call = \frac{number\_of\_relevant\_documents\_retrieved}{total\_number\_of\_relevant\_documents\_in\_collection}$$

This is the ratio of the relevant documents returned by the system compared to the actual number of relevant documents that is in the entire collection. In a database scenario, recall is usually 100%, for the same reasons as stated above in the case of precision. A

full text search could achieve 100% recall by returning the entire collection as a search result – but this is clearly a theoretical solution since the user would be overwhelmed by an excessive amount of irrelevant information, i.e. the precision would be very low. Consequently, as in the case of precision, a full text retrieval system usually achieves a recall that is substantially lower than 100%.

These two measures model the following assumptions:

1. the user wants to see as few non-relevant documents as possible
2. the user wants to retrieve as many relevant documents as possible.

Assumption #1 generally holds in any conceivable scenario. Assumption #2 may not be as important in some scenarios where the user is looking for specific answers to questions that can be derived from relatively few relevant documents, with more documents not giving additional information (e.g. "how tall is the Eiffel tower?"). On the other hand, for some applications, such as patent research, good recall is essential. For an in-depth definition of these measures, consider e.g. [Schäuble, 1997].

There is usually a trade-off between precision and recall. If a system is tuned towards higher recall, it is likely that along with the additional relevant documents, more non-relevant documents are also retrieved. If, on the other hand, the system is optimized for high precision, it usually filters documents very strictly so that some relevant information is missed.

This means that the two measures are usually investigated for a variety of combination levels. The most common procedure is to determine the precision at various levels of recall, i.e. to use a function that assigns each recall value a precision value. This function is then visualized in a precision/recall graph.

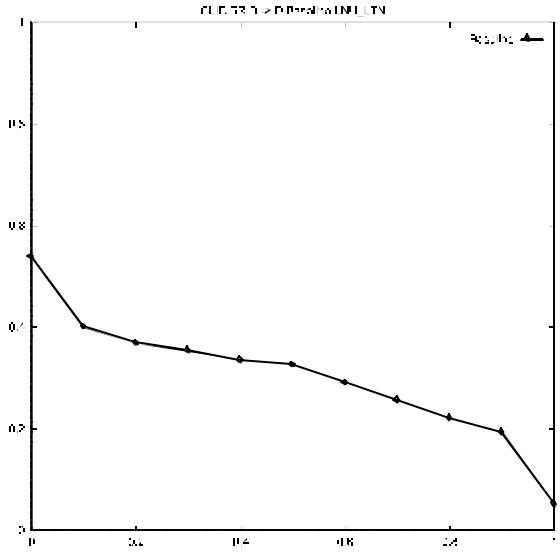

**Figure 1: A sample precision/recall graph. Initial precision, at low recall levels, is high, due to the ordering of the retrieved items in order of their estimated relevance. Once the user inspects more documents, the precision continuously drops as the recall increases.**

A popular measure to sum up the overall performance is the use of n-point average precision. For this, the average precision at various levels of recall is calculated:

$$Avg\,\Pr ec = \frac{1}{n}\sum_{i=1}^{n} prec(recall_i)$$

Popular choices for n are 3 and 11.

This provides a "one-figure" performance measure for a variety of different applications, from high precision/low recall to high recall/low precision. Care must be taken however when interpreting average precision and other similar measures, since a range of different characteristics can get obscured in the averaging process. Average precision can therefore not replace a more detailed analysis as outlined above.

A key measure for cross-language retrieval is the relative performance of the system in cross-language mode with respect to monolingual mode.

$$Ratio = \frac{Avg\,\Pr ec(Cross - Language)}{Avg\,\Pr ec(Monolingual)}$$

In order to compute either of these measures, assessments of the relevance of the retrieved documents (for precision) or of all documents (for recall) are needed. Determining the relevance of these documents is usually a manual, costly process. Especially the need to know the overall number of relevant documents in the collection makes such evaluations impractical in isolated settings.

# 6 Test Collections

## *6.1 TREC/CLEF test collections*

As mentioned, the most widely used test collections in recent years are made available through the US National Institute of Standards and Technology (NIST). NIST organizes TREC, the Text REtrieval Conference, a yearly conference bringing together major research groups and companies interested in information retrieval. These participants use the TREC test collections to submit sample results to NIST, which are then evaluated at NIST by relevance assessors. The conference itself provides a forum for comparisons of results and reports about the methods employed by different participants. An overview of TREC can be found in [Harman, 1995]. TREC introduced a cross-language track beginning with TREC-6 [Braschler et al., 2000].

The considerable success of this cross-language track prompted the spin-off of this activity into an independent forum called CLEF (Cross-Language Evaluation Forum). CLEF is funded by the fifth framework programme of the European Commission.

The advantage of these evaluation forums is in using the synergy of having multiple participants. This justifies the effort necessary to produce the relevance assessments. Furthermore, provided sufficient participation, it is possible to calculate approximate figures for recall by using a "pooling technique". The idea is that the number of relevant documents that go undetected (i.e. that are not retrieved by any of the participants) is likely to be small for a sufficient number of different result sets. Therefore, the total number of relevant documents retrieved by the participants is a useful approximation of the total number of relevant documents in the collection. This assumption has been validated in several research studies, notable is e.g. [Voorhees, 1998].

The main problem when applying the test collections produced by these evaluation forums is the style of the documents that are contained in the test data. These are usually mostly newspaper or newswire articles. With MUCHMORE concentrating on retrieval of medical information, we cannot test our system exclusively with documents from a different domain.

## *6.2 OHSUMED collection*

An alternative to using the cross-language test collections from TREC or CLEF is the use of the OHSUMED collection. This collection has been created originally as part of a study by William Hersh at the Oregon Health Sciences University around 1994 [Hersh, 1994]. It uses documents derived from a commercial CD-ROM-based search system which allows access to MEDLINE journal references. This means that the document base is very close to what we intend to use in MUCHMORE. The test queries were developed by physicians and librarians that were used to conduct searches on this kind of data. A first analysis by ZInfo showed that they are an appropriate representation of the kind of search requests that we can expect MUCHMORE users to issue.

In total, OHSUMED contains 348,566 documents, a total of around 400 MB of data. There are two sets of queries, the "OHSU" set and the "MeSH" set. The OHSU set contains a total of 106 queries, while the MeSH set contains 14321 queries (categories) (for use in TREC – see below – a set of 5023 (4904+119) queries was used).

OHSUMED has been used in various experiments, notably also in the filtering track of the latest TREC conference, TREC-9. This means the collection is well accepted in terms of its quality with respect to queries and relevance assessments. Therefore, we intend to use this collection for testing of the near-final and final prototypes of the systems that are to be implemented in MUCHMORE.

The only drawback in using OHSUMED is the monolingual nature of this test collection. Both queries and documents are in English. We can easily test German->English cross-language access by manually translating the queries into German, but the other direction, English->German, cannot be evaluated using this collection.

We intend to additionally conduct a limited TREC-style English->German evaluation. CMU is preparing to produce relevance assessments for such a collection, and we hope that we can also shift manpower in a way that would allow ZInfo to join in this task. Such an experiment would be carried out on the parallel medical abstracts that we are collecting (see below). This additional, smaller, collection gives us the advantage of having a real parallel test set, and therefore being able to investigate the relation between English->German and German->English experiments.

<top>
<num> Number: OHSU1
<title> 60 year old menopausal woman without hormone replacement therapy
<desc> Description:
Are there adverse effects on lipids when progesterone is given with estrogen rep
lacement therapy
</top>

<top>
<num> Number: MSH1
<title> Calcimycin
<desc> Description:
An ionophorous, polyether antibiotic from Streptomyces chartreusensis. It binds
and transports cations across membranes and uncouples oxidative phosphorylation
while inhibiting ATPase of rat liver mitochondria. The substance is used mostly
as a biochemical tool to study the role of divalent cations in various biologica
l systems.
</top>

**Figure 2: Sample of OHSUMED queries (one query from the OHSU and MeSH sets each)**

.I 1
.U
87049087
.S
Am J Emerg Med 8703; 4(6):491-5
.M
Allied Health Personnel/*; Electric Countershock/*; Emergencies; Emergency Medic
al Technicians/*; Human; Prognosis; Recurrence; Support, U.S. Gov't, P.H.S.; Tim
e Factors; Transportation of Patients; Ventricular Fibrillation/*TH.
.T
Refibrillation managed by EMT-Ds: incidence and outcome without paramedic back-u
p.
.P
JOURNAL ARTICLE.
.W
Some patients converted from ventricular fibrillation to organized rhythms by defibrillation-trained ambulance technicians (EMT-Ds) will refibrillate before hospital arrival. The authors analyzed 271 cases of ventricular fibrillation managed by EMT-Ds

working without paramedic back-up. Of 111 patients initially converted to organized rhythms, 19 (17%) refibrillated, 11 (58%) of whom were reconverted to perfusing rhythms, including nine of 11 (82%) who had spontaneous pulses prior to refibrillation. Among patients initially converted to organized rhythms, hospital admission rates were lower for patients who refibrillated than for patients who did not (53% versus 76%, P = NS), although discharge rates were virtually identical (37% and 35%, respectively). Scene-to-hospital transport times were not predictively associated with either the frequency of refibrillation or patient outcome. Defibrillation-trained EMTs can effectively manage refibrillation with additional shocks and are not at a significant disadvantage when paramedic back-up is not available.
.A
Stults KR; Brown DD.

**Figure 3: Sample of an OHSUMED document.**

# 7 Other evaluations

In order to evaluate the intermediary stages of development of the prototype, we intend to use some evaluation methods that are less costly as a supporting measure in case the more comprehensive tests are inconclusive. Their simpler nature has potential to allow us a repeated use, in sync with development progress.

Two measures are considered for use: known-item searches, and overlap measures. While generally agreed to be inferior to full precision/recall evaluation, they provide us with a workable compromise between obtaining meaningful results and feasibility in terms of cost.

## *7.1 Known-Item searches*

By using test queries that are constructed to only have one answer (i.e. there is only one relevant document in the collection), evaluation of the search results is simplified considerably: the precision measure is replaced by the rank of the relevant document (the higher the document is ranked, the better, usually referred to as "mean reciprocal rank") and recall is replaced by the percentage of test queries that have an answer returned.

$$"Precision" = Mean\_\text{Re}ciprocal\_Rank = \frac{\sum \frac{1}{rank\_of\_answer}}{nuimber\_of\_queries}$$

(for queries that do not return the answer, a "0" is substituted in the sum)

$$\text{Re}call = \frac{number\_of\_queries\_correctly\_answered}{total\_number\_of\_queries}$$

### 7.2 Overlap measures

These measures are based on the assumption that a good baseline is available. The system is then evaluated by comparing the results with the results provided by the baseline system. If the overlap is high, it is assumed that the results of the experimental system are appropriate. In terms of cross-language retrieval, we can use a monolingual search result as the baseline: first, a search within the language is carried out (e.g. English queries run against English documents), which is then compared to the cross-language search (e.g. German translations of the English queries run against the English documents). This procedure is questionable if the quality of the baseline cannot be guaranteed.

## 8 Test Data

The following test data sets are in consideration for known-item searches and overlap measure evaluations:

Monolingual

- Medical abstracts (from Medline)
- Web documents (from Dr. Antonius)
- Medical abstracts & Web documents combined

Bilingual

- Autopsy reports (comparable corpus: German from Zinfo, English from associated partner Johns Hopkins University Medical School)
- Medical abstracts (parallel corpus: German/English from "Springer LINK" Website)

The availability of the autopsy report comparable corpus will depend on the cooperation with Johns Hopkins University.
In order to test various components of the prototype system (i.e. document classification, semantic annotation of terms and relations), parts of these data sets need to be additionally annotated with medical standardized terminologies (i.e. MeSH, ICD). To get

a meaningful test collection, these annotations have to be done either completely by hand or are to be corrected manually in a semi-automatic way.

# 9 Timetable

Evaluation of both intermediate and prefinal/final components will be conducted as part of workpackage 8. This workpackage is due to start month 7, and will run through month 24. The results will be reported in deliverable 8.1, "Performance comparisons of different methodologies within the context of concept-based CLIR. (Report)", deliverable 8.2, "Platform architecture for combination of methods and resources (Specification/Report)" and deliverable 8.3, "Information Access system prototype for user testing and demonstration (Software)", all due in month 24. In the meantime, preliminary test results will be actively disseminated between participants as an integral part of decision making with regard to development of the affected components.
The set of test collections to be used in the tests (constituting deliverable 3.1, "Test Collection (Data)" (the OHSUMED collection, the Springer LINK corpus, the German autopsy reports) has been acquired by the partners. It will be extended according to section 8 of this document should need arise.

# 10 Summary

Two sorts of evaluations are planned for the MUCHMORE cross-language components:

Intermediate prototypes will be tested using known-item searches and overlap measures. We will use test data that is available from the MUCHMORE partners (primarily ZInfo). Parts of this test data need to be annotated (manually corrected) with suitable medical terminologies (if not already present).

Near-final and final prototypes will be tested using TREC-style evaluation and the precision and recall measures. These tests will make use of the OHSUMED test collection, which has a well-developed set of relevance judgments. We will also conduct some smaller evaluations using our own data, which will require to make the relevance assessments ourselves.

Combining both the simpler and more comprehensive evaluations gives us the advantage of being able to conduct continuous evaluation, and therefore assuring effective quality control. The more extensive final tests, on the other hand, give us excellent comparability to other studies into cross-language retrieval performance.

# References

[Braschler et al., 2000] Braschler, M., Harman, D., Hess, M., Kluck, M., Peters, C., and Schäuble, P. 2000. The Evaluation of Systems for Cross-Language Information Retrieval. In *Proceedings of the 2nd International Conference on Language Resources & Evaluation (LREC 2000)*.

[Harman, 1995] Harman, D. 1995. The TREC Conferences. In *Proceedings of HIM '95*. Reprint in Sparck-Jones, K., and Willett, P. (eds.): *Readings in Information Retrieval.* Morgan Kaufmann Publishers.

[Hersh, 1994] Hersh, W., Buckley, C., Leone, T. J., Hickam, D. 1994. An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1994)*, pages 192-201.

[Schäuble, 1997] Schäuble, P. 1997. Content-Based Information Retrieval from Large Text and Audio Databases. Kluwer Academic Publishers.

[Voorhees, 1998] Voorhees, E. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*