The Workshop Programme

9.00 – 9.15	**Opening remarks – René Schneider, University of Hildesheim, Germany**
9.15 – 9.45	**“Experiences in Evaluation in CLIR” - Christa Womser-Hacker, University of Hildesheim, Germany**
9.45 – 10.15	**“The Forgotten Key Point for Assuring the Knowledge Consistency in CLIR Systems” - Jesus Cardeñosa, Universidad Politécnica de Madrid, Madrid, Spain**
10.15 – 10.45	**“Assessing Relevance in CLEF” - Michael Kluck, Social Science Information Centre (IZ) , Bonn, Germany**
10.45 – 11.15	*Coffee Break*
11.15 – 11.45	**“Analysis of Topic Features in Cross-Language Information Retrieval Evaluation” - Thomas Mandl, University of Hildesheim, Germany**
11.45 – 12.15	**“Quality Gates: a New Device for Development and Evaluation in Cross-Language Information Retrieval?“ - René Schneider, University of Hildesheim, Germany**
12.15 – 13.00	**Discussion and closing remarks**

Workshop Organisers

Thomas Mandl	**University of Hildesheim, Germany**
René Schneider	**University of Hildesheim, Germany**
Christa Womser-Hacker	**University of Hildesheim, Germany**

Workshop Programme Committee

Marcello Federico	**I.R.S.T, Italy**
Norbert Fuhr	**University of Duisburg, Germany**
Gregory Grefenstette	**Clairvoyance Corp., Pittsburgh, USA**
Donna Harman	**NIST, USA**
Paul Heisterkamp	**DaimlerChrysler AG, Germany**
Thomas Mandl	**University of Hildesheim, Germany**
Dagobert Soergel	**University of Maryland, USA**
Christa Womser-Hacker	**University of Hildesheim, Germany**

Table of Contents

Author Index

Workshop LECLIQ: Lessons Learned from Evaluation: Towards Integration and Transparency in Cross-Lingual Information Retrieval with a special Focus on Quality Gates

Thomas Mandl, René Schneider, Christa Womser-Hacker

University of Hildesheim
Information Science
Marienburger Platz 22- D-31141 Hildesheim, Germany
{mandl, rschneid, womser}@uni-hildesheim.de

Abstract

In this paper we give an overview of the workshop on "Lessons Learned from Evaluation: Towards Integration and Transparency in Cross-Lingual Information Retrieval" that was held in Lisbon, Portugal on May, 30, 2004 in conjunction with the 4th International Conference on Language Resources and Evaluation LREC-04.

1 Introduction

Recent work in Cross-Lingual Information Retrieval (CLIR) has shown that systems perform differently with respect to queries, topics, data sets and the corresponding query or result language. Since the state-of-the-art in CLIR is far from finding an ideal method for this variety of parameters, a better performance might be achieved through the integration of several retrieval devices resp. functionalities. The process of integrating these different methodologies might be controlled by so called quality gates. Quality gates - having their origin in car manufacturing and being used in IT-project management as well - are characterized as preventive and process accompanying control mechanisms to check critical parameters and ensure quality standards during the design, development, and deployment of software tools. They generally consist of checklists combined with appropriate rules to guarantee that work procedure failure is recognized in time to prevent repetition.

The objective of this workshop was to bring together researchers from the field of Information Retrieval and Language Engineering with three primary goals:

1. to discuss, evaluate and judge the different methodologies used in CLIR and to relate the results to the languages being retrieved,
2. to find out appropriate measuring points to install a dynamic network of quality gates for on-the-run tuning of retrieval tasks,
3. to define desired learn effects caused by the interaction of quality gates.

Topics of interest include:

- Lessons learned in evaluation initiatives such as CLEF, TREC, INEX, NTCIR, etc.
- Criteria, checklists, parameters and metrics for (on-the-run) evaluation of search and retrieval methods
- Quality criteria and evaluation for resources and tools in CLIR (thesauri stemmer, translation tools and services, etc.)
- Semaphore logic for the integration of distributed retrieval systems
- Explanation and verification of success stories and failure analysis in CLIR
- Data and Text Mining on CLIR evaluation results
- Current evaluation issues: shortcomings, gaps, reliability of results, etc.
- Transfer of evaluation results to system design and product development
- Design for (coupled or networked) quality gates in CLIR (for systems and building blocks)
- Crossing-over techniques for quality assurance
- Hybrid systems for CLIR

Contributions discussing the themes above were invited from participants in evaluation initiatives as well as from experienced researches in the areas of (Cross-Language) Information Retrieval or related fields.

2 Lessons Learned from Evaluation

TREC (Text REtrieval Conference[1]) has set new standards within information retrieval evaluation and has led to a great improvement of retrieval algorithms (Harman & Voorhees, 1997; Buckland & Voorhees, 2003). TREC set up a environment for comparative evaluation including document collections, topics and relevance assessments.

The Cross Language Evaluation Forum (CLEF) is a large European evaluation initiative which is dedicated to cross-language retrieval for European languages. CLEF (Cross Language Evaluation Forum[2]) (Braschler et al., 2003, 2004) has been implemented as a consequence to the rising need for cross- and multi-lingual retrieval research and applications. CLEF provides a multi-lingual testbed for retrieval experiments. The evaluation campaign of CLEF comprises several components: the evaluation methodology, the evaluation software packages, the data collections, the topics, the overall results of the participants, the assessed results of the participants, and the calculated statistical results. CLEF uses the evaluation methodology developed at the TREC. Multilingual retrieval for Asian languages are evaluated within NTCIR (Eguchi et al., 2002).

[1] http://trec.nist.gov

[2] http://www.clef-campaign.org

Retrieval systems improved considerable during TREC (and CLEF). Appropriate methods for many tasks have been developed and many system components have been evaluated. However, many questions remain unanswered. How well do the evaluation results mirror real world tasks and how can the results be transferred into operating retrieval systems? For the implementation of quality gates, the validity of results is of special interest. Therefore, we review some of the work done in this area (see also Sparck Jones, 1995).

The validity of large-scale information retrieval experiments has been the subject of a considerable amount of research. Zobel (1998) concluded that the TREC (Text REtrieval Conference[3]) experiments are reliable as far as the ranking of the systems is concerned. Buckley & Voorhees (2002) have analyzed the reliability of experiments as a function of the size of the topic set. They concluded that the typical size of the topic set in TREC is sufficient for a satisfactory level of reliability.

Further research is dedicated toward the question whether expensive human relevance judgments are necessary or whether the constructed document pool of the most highly ranked documents from all runs may serve as a valid approximation of the human judgments. According to a study by Soboroff et al. (2001) the ranking of the systems in TREC correlates positively to a ranking based on the document pool without further human. However, there are considerable differences in the ranking which are especially significant for the highest ranks. Human judgments are therefore necessary to achieve the highest reliability of the system ranking. Still, relevance assessment is a very subjective task. Consequently, assessments by different jurors result in different sets of relevant documents. However, these different sets of relevant documents do not lead to different system rankings according to an empirical analysis by Voorhees (Voorhees, 2000). Thus, the subjectivity of the jurors does not call into question the validity of the evaluation results. Another important aspect is pooling. Not all submitted runs can be judged manually by jurors and relevant documents may remain undiscovered. Therefore, a pool of documents is built to which the systems are contributing differently. In order to measure the potential effect of pooling, a study was conducted which calculated the final rankings of the systems by leaving out one run at a time (Braschler et al., 2003). It shows that the effect is negligible and that the rankings remain stable. However, our analysis shows that leaving out one topic during the result calculation changes the system ranking in most cases. It has also been noted that the differences between topics are larger than the differences between systems. This effect has been noted in many evaluations and also in CLEF (Braschler et al., 2004). As a consequence, topics are an important part of the design in an evaluation initiative and need to be created very carefully.

Voorhees & Harman measured the difficulty of TREC topics from two perspectives (Braschler et al., 2003). One was the estimation of experts and the second was the actual outcome of the systems measured as the average precision which systems achieved for that topic. They found no correlation between the two measures. This result was confirmed in a study of the topics of the Asian languages retrieval evaluation NTCIR[4] (Mandl & Womser-Hacker, 2004a). Furthermore, Eguchi et al. tried to find whether the system rankings change within the NTCIR evaluation campaign when different difficulty levels of topics were considered. They conclude, that changes in the system ranking occur, however, the Kendall correlation coefficient between the overall rankings does not drop dramatically. For that analysis, the actual difficulty measured by the precision of the runs was used. The overall rankings remain stable, however, top ranks could be affected (Eguchi et al., 2002). According to the results from the study by Buckley & Voorhees, which analyzed the reliability of experiments as a function of the size of the topic set, such a small set does not lead to fully reliable results (Buckley & Voorhees, 2002).

The difficulty of topics is a notion worth further exploration within the context of quality gates. What makes a topic difficult? The identification of linguistic or statistical phenomena which make topics more difficult for systems would be desirable. For retrieval evaluation it is important to be aware of influencing factors within the topics. Named entities seem to play an important role especially in multilingual information retrieval (Mandl & Womser-Hacker, 2004b). This assumption is backed by experimental results. The influence of named entities on the retrieval performance is considerable.

3 Quality Gates

3. 1 Definitions

Generally, a Quality Gate (QG) is a checkpoint consisting of a set of predefined quality criteria that a project must meet in order to proceed from one stage of its lifecycle to the next.

Quality gates thus serve as amendments to milestones and deliverables to

- support planning,
- improve status visibility,
- measure the current project status, and
- control necessary changes or improvements.

Each quality gate is characterized by its own entry and exit criteria. A typical entry criteria is the completion and baseline of deliverables while an exit criteria can be the removal of the identified defects.

By including metrics at every stage of the development process, projects are monitored against their stated goals. By these means, QG point out new strategies for the integration and validation of different methods or routines.

3. 2 Benefits of Quality Gates in (CL)IR

Since the successful implementation of a retrieval system and the corresponding participation at an evaluation initiative (such as CLEF or TREC) depends considerably on a large number of quality criteria, quality gates ensure that the project deliverables meet the criteria necessary to carry out subsequent project activities. Similar to complex manufacturing processes or product

[3] http://trec.nist.gov

[4] http://research.nii.ac.jp/ntcir/

development, the release version of a retrieval system consists of many separate components, which may be developed at different times and are based on concurrent, sequential, or recursive applications of a standard development pattern. This is esp. true in cross-lingual information retrieval, where the number of critical parameters is multiplied by the different languages and their implications for effective retrieval.

During development and testing of complex systems, many requests for changes are generated - by developers as they realize something can be done in a different way, by result evaluation, by users when they try out the product, etc.

Sometimes these changes are small and a decision can easily be made whether to implement the change: but the changes to specification should be noted. Some requests may be kept open depending on the project's progress against timetable, some will be deferred as taking too long to implement. All of these specifications should be kept appropriately and probably be converted into quality criteria for further development circles.

3. 3 Goals

One of the first steps for the introduction of quality gates in Information Retrieval would be the collection and validation of experiences made in various evaluation initiatives. This primary goal may be formulated through the following question: How can we transfer our different experiences to objective quality criteria that improve the development, testing, and deployment of retrieval systems and avoid making the same mistakes again and again?

3. 4 Future Perspectives

Our vision is that of using quality gates as a concrete methodology in implementation in case QGs prove to be a successful and enriching method for quality assurance and evaluation. This means that after a first period of intellectual specification and testing as specified above, the idea will become part of the software by connecting the different components of a retrieval system via a network of coupled quality gates to

- control system parameters,
- steer information flow, and
- document learning effects.

4 Lessons to be Learned from Evaluation: Workshop Overview

The key question of information retrieval evaluation is how to integrate knowledge from evaluation results into working systems. The contributions to the workshop take very different approaches to this issue.

Cardeñosa et al. (2004) review the architecture of CLIR systems and identify ontologies as a key for quality assurance in multi-lingual information systems. Their formal language UNL is capable of providing a framework for knowledge representation. Unified knowledge representation is a key issue for quality in information retrieval (Cardeñosa et al., 2004).

Kluck (2004) presents the assessment of CLEF in detail, showing the rules and procedures, illustrating the assessment organization and processes. It also discusses the validity of the assessments in the context of the pooling method. Finally the specific issues of the assessment in the multilingual context are examined and comparable activities are analyzed (Kluck, 2004).

Mandl & Womser-Hacker (2004b) show how much retrieval evaluation results may depend on the test design. An analysis of CLEF topics showed a medium correlation between proper names and retrieval performance. This results needs to be considered in the design of retrieval experiments. It can also serve as input for systems and quality gates which identify topic difficulty and treat different topics appropriately (Mandl & Womser-Hacker, 2004a, 2004b).

Schneider (2004) shows how the notion of quality gates can be integrated into information retrieval systems. This paper outlines and discusses the perspectives quality gates offer in cross-lingual information retrieval to ensure that the development process benefits from evaluation. Adequate evaluation in this context is possible through a combination of modular quality gates at the inch-pebble level, their linear connection in networks and re-organisation during different development cycles. As a consequence, the strict separation between development and evaluation disappears (Schneider, 2004).

References

Buckley, C. & Voorhees, E. (2002). The Effect of Topic Set Size on Retrieval Experiment Error. In Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '02). (Tampere, Finland, Aug. 11-15, 2002). ACM Press, pp. 316-323.

Cardeñosa, J., Gallardo, C. & Iraola, L. (2004). The Forgotten Key Point for Assuring Knowledge Consistency in CLIR Systems. In Proceedings of the LREC-04 Workshop on "Lessons Learned from Evaluation: Towards Transparency and Integration in CLIR", Lisbon, Portugal 30.05.04, to appear.

Eguchi, K., Kando, N. & Kuriyama, K. (2002). Sensitivity of IR Systems Evaluation to Topic Difficulty. In C. P. S. Araujo & M. G. Rodríguez (Eds.). Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002) (Las Palmas de Gran Canaria, Spain, May 29-31, 2002). ELRA, Paris, 2002, pp. 585-589.

Harman, D. & Voorhees, E. (1997). Overview of the Sixth Text REtrieval Conference. In D. Harman and E. Voorhees (Eds.). The Sixth Text REtrieval Conference (TREC-6). NIST Special Publication, National Institute of Standards and Technology, Gaithersburg, Maryland, 1997, http://trec.nist.gov/pubs/.

Kluck, M.(2004). Assessing Relevance in CLEF. In Proceedings of the LREC-04 Workshop on "Lessons Learned from Evaluation: Towards Transparency and Integration in CLIR", Lisbon, Portugal 30.05.04., to appear.

Mandl, T. & Womser-Hacker, C. (2004a). Proper Names in the Multilingual CLEF Topic Set. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2004)

Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2003, Trondheim. Berlin et al.: Springer [Lecture Notes in Computer Science] 2004. to appear. Preprint http://www.clef-campaign.org.

Mandl, T. & Womser-Hacker, C. (2004b). Analysis of Topic Features in Cross-Language Information Retrieval Evaluation, in: Proceedings of the LREC-04 Workshop on "Lessons Learned from Evaluation: Towards Transparency and Integration in CLIR", Lisbon, Portugal 30.05.04., to appear.

Braschler, M., Gonzalo, J., Kluck, M. & Peters, C. (Eds.) (2003) Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2002, Rome. Berlin et al.: Springer [Lecture Notes in Computer Science 2785].

Braschler, M., Gonzalo, J., Kluck, M. & Peters, C. (Eds.) (2004) Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2003, Trondheim. Berlin et al.: Springer [Lecture Notes in Computer Science], to appear. Preprint http://www.clef-campaign.org.

Schneider, R. (2004). Quality Gates: a New Device for Development and Evaluation in Cross-Language Information Retrieval ? in: Proceedings of the LREC-04 Workshop on "Lessons Learned from Evaluation: Towards Transparency and Integration in CLIR", Lisbon, Portugal 30.05.04., to appear.

Soboroff, I., Nicholas, C. & Cahan, P. (2001). Ranking retrieval systems without relevance judgements. In Proceedings of the 24th Annual International ACM SIGIR Conference of Research and Development in Information Retrieval, 2001.

Sparck Jones, K. (1995). Reflections on TREC. Information Processing & Management, 31, 3 (May/June 1995), pp. 291-314.

Voorhees, E. (2000). Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness. Information Processing & Management, 36, 5 (2000), pp. 679-716.

Buckland, L. & Voorhees, E. (Eds.) (2003). The Eleventh Text Retrieval Conference (TREC 2002), NIST Special Publication: SP 500-251. http://trec.nist.gov/pubs/trec11/t11_proceedings.html.

Zobel, J. (1998). How Reliable are the Results of Large-Scale Information Retrieval Experiments? In Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '98), Melbourne, Australia, Aug. 24-28, 1998. ACM Press, New York, pp. 307-314.

The Forgotten Key Point for Assuring Knowledge Consistency in CLIR Systems

Jesús Cardeñosa, Carolina Gallardo, Luis Iraola

Facultad de Informática – Universidad Politécnica de Madrid
Campus de Montegancedo, 28660 Madrid – SPAIN
{carde,carolina,luis}@opera.dia.fi.upm.es

Abstract

CLIR is the acronym of a great variety of techniques, systems and technologies that associate information retrieval (normally from texts) in a multilingual environments. Many of these systems are based on a double architecture composed by systems in charge of extracting information with a great dependency on the language together with classical machine translation systems. In the early 90's, machine translation systems fell from grace due to the failure of big machine translations projects in Europe, Japan and USA. Due to this reason some approaches, particularly those of linguistic knowledge representation were undeservedly forgotten, and above all the so called "interlinguas". Recently, the re-emergence of these models under the generic name of "ontologies" are supporting most of knowledge representation initiatives, even in an language independent way. However consistency problems are not well solved yet. UNL, initially conceived as a contents representation and multilingual generation system, can also be applied to the CLIR.. This paper aims to show how to create and apply domain specific ontologies using the UNL apparatus, particularly the UNL language as a way of ensuring a consistent representation mechanisms.

1 Introduction

Cross-Language Information Retrieval (CLIR) deals with the problem of issuing a query in one language and retrieving relevant information in other languages. It aims to help the user in finding relevant information without being limited by linguistic barriers.

In order to overcome the language barrier, three major approaches exist:

- to translate the query into the documents' languages
- to translate the documents into the query's language
- to translate both into an intermediate representation through the use of domain-specific interlinguas.

1. 1 Query Translation

Online translation can be applied to the query entered by the user. Online query translation will help the user to formulate his/her query in a language other than his/her own. If the user either has at least some reading skills in the target language, it may be possible for him/her to reformulate, elaborate or narrow down the translation proposed.

Because of its simplicity, query translation via machine-readable bilingual or multilingual dictionaries is a very most common approach (Grefenstette, 1996; Ballesteros & Croft, 1997; Davis & Ogden, 1997). Compared to translating an entire document collection, translating a query by dictionary look-up is far more efficient. However, it is unreliable since short queries do not provide enough context for disambiguation in choosing proper translations of query words, and also because it does not exploit domain-specific semantic constraints and corpus statistics in solving translation ambiguities.

A wide array of resources is used in CLIR (Radwan & Fluhr, 1995; Oard, 1997), ranging from multilingual glossaries or dictionaries to multilingual collections of texts and sophisticated taggers and parsers (e.g., Mulinex and MIETTA projects).

1. 2 Document Translation

Full document translation can be applied offline to produce translations of an entire document. The translations provide the basis for constructing an index for information retrieval and also offer the user the possibility to access the content in his/her own language. Machine or (large scale) human translation, however, is not always available as a realistic option for every language pair. Typically machine translation systems only translate between language pairs which involve one of the major languages, such as English, German or Spanish, and often English plays a pivotal role.

1. 3 Domain-specific Ontologies for CLIR

Recent CLIR projects (MuchMore, LIQUID) employ a domain-specific ontology that contains the knowledge of the application domain and serves as an interlingual backbone for a multilingual thesaurus. Relevant terms contained in a query are translated into several languages using the term-to-concept links established in the multilingual thesaurus. Domain knowledge represented in the conceptual layer is exploited for expanding the initial query (see figure 1).

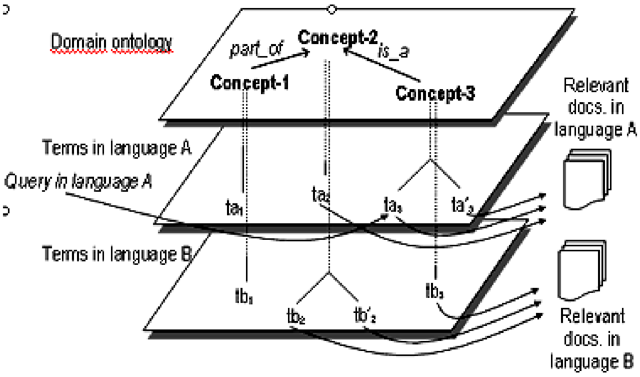

Figure 1: Linking documents and queries through a multilingually mapped ontology

2 Ontologies and Support Languages

Like in many other cases, the definition of an ontology is not completely fixed and agreed on. There are several definitions of ontologies, but for our purpose we will cling to Gruber's one: *an ontology is an explicit specification of a conceptualisation* (Gruber, 1993).

There are two main issues in this definition:

- Explicit specification
- Conceptualization

The "explicit specification" of an ontology leads us to the formalization of ontologies and used languages. In this section, we will deal with ontologies support languages as the main way for attaining such explicitness and machine readability.

A conceptualization is related to the creation of a model of a given domain pointing out the relevant concepts, their relations and functions that made up a complete domain.

In order to support an ontology and inference mechanisms, the question of the language support is crucial. There are two main factors that determine the evolution of ontology languages. These are the *knowledge representation formalism* and *web orientation*.

Regarding the knowledge representation formalism, there appear to be two clear periods that we will refer to as *First Generation Languages* and *Second Generation Languages*. First generation ontology languages are basically frame-based and correspond to the first attempts to build ontologies and establish the ontology engineering discipline (beginning of 90ies). As the most representative frame-based languages are Loom (MacGregor, 1987), Ontolingua (Faquhar et al., 1996) or KIF (Genesereth et al., 1992).

In its beginnings, ontology engineering was highly oriented towards knowledge reuse and share (Neches et al., 1991). All of these languages can be considered as languages for knowledge representation, being KIF (*Knowledge Interchange Format*) the most oriented towards knowledge reuse, since it conforms a sort of "interlingua" of knowledge representation languages.

The common feature of these languages is its frame-based nature. Thus, they are endowed with the usual expressiveness of frames. Basically, they allow for:

- Representing classes and subclasses
- Distinguishing between classes and instances
- Establishing relations between classes.
- Establishing default values.

In a way we could say that these languages are oriented toward a hierarchical conceptualisation of a domain. Needless to say, the Semantic Web wasn't the main goal in this period. So there is no web integration of this ontologies.

The second generation of ontology languages shows a more logical flavour (although some retain the frame flavour). We are referring to RDF (Lassila et al., 1999), OIL (Horrocks et al., 2000), SHOE (Luke et al., 2000), DAML-OIL (Horrocks et al., 2001) or even XML (Yergeau et al., 2004). Let's mention some of the properties of these languages:

- They are based on first order logic (with some possible extensions).
- Use of logic (formal semantics for deduction processes)
- The distinction between class and instance is supported.
- The establishment of taxonomies (class – subclass) is normally supported.
- Representation and inclusion of axioms are supported in some of them.
- Normally no default values are allowed.
- Relations (of different arity) are more or less covered.

Some of them are oriented towards the Semantic Web (developed by the W3C consortium or either compatible with XML).

These ontology languages show the second parameter: web orientation, they extends the traditional definition of an ontology and try to conceptualise the whole web, that is, the target is no more reuse of knowledge but to achieve the so-called Semantic Web. Thus many of them are based on web languages and technologies (such as XML and RDF developed by the W3C consortium).

It is interesting to see the influence of an standard entity such as W3C as an standardizing body. It is quite obvious the convergence of all these languages towards a unique standard one: OWL (Bechofer et al., 2004).

All these languages seems to have derived in OWL, which is an extension of XML, RDF, DAML and OML. According to the authors, it provides "*greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics".* It was in February, 2004 when it was proposed by W3C to become the standard language for ontology representations in the web.

3 Knowledge Representation vs. Cross-Linguality

Ontologies and knowledge representation are two close concepts. At the end, conceptualisation and formalization of a model or domain are two quite well known issues of Knowledge representation. Ontology engineering does not begin from scratch, many of its theoretical foundations are borrowed form Knowledge Engineering, being formalisms and representation languages no exception.

Historically, semantic nets was the first formalism suitable to represent knowledge, as it extended the expressiveness of pure logical models. The semantic nets were proposed in 1968 by Quillian and he was also who study the knowledge extraction from texts some years later (Quillian, 1968). Even today the degree of conceptual advance in comparison with those years is not high. Possibly the real advance is coming from the capacity of managing great knowledge bases based in an increase of computing power and an increase of the interoperability between heterogeneous systems through standardized formalisms. In those years the main problem was the lack of standardisation of the possible number of relations and also the necessities to expand the amount of information associated to the concepts of a net. This was also the convergence between the conceptual definition of "frame" proposed by Minsky (Minsky, 1974) and the necessities to expand the capacities to encapsulate information in the so called "frame nets" that were the combination of semantic nets with the expansion of the concepts into frames.

Wood in (Brachman et al., 1985) stated two issues that prevents semantic nets from being a good candidate for knowledge representation:

Ambiguities in its representation (no specific account of the distinction between class and instance)
Lack of a common understanding of the semantic labels, that eventually Wood defines as the "asemanticity" of semantic nets.

For these two reasons, ontology languages turn to frame and logic based formalisms, disregarding the adequacy of semantic nets for the specification of non-hierarchical relations (that is, functions and roles between concepts). Curiously, current ontologies do not fully exploit the most expressive characteristics of semantic nets, resulting in a massive use of relation IS-A. Bearing in mind the features of ontology languages, we could state that there is coverage for vertical relations (class, subclass, instance, plus other) but not for horizontal relations (roles and links between concepts). Horizontal relations enrich the domain representation, as shown in Burg (1997) and Shamsfard et al. (2004) as attempts to build ontologies from natural language texts. Even if we accept Wood's objection to semantic nets, there is still a wide amount of information that semantic nets offers and ontologies do not exploit, being this the capacity of semantic nets to express horizontal relations, that could be easily integrated into ontology support languages in principle.
Thus relations would not be only limited to a is-a or a-kind-of types, but richer relations will have to be included. A hint of what sort of horizontal relation should be included in domain models is given by natural languages (languages are the main vehicle of expressing knowledge), this is the approach followed in the GUM, following the theoretical positions that Functional Grammar established (Bateman et al, 1990), or as we will see later in the Universal Networking Language (UNL).
A major problem for knowledge based approaches is the creation of the necessary resources: in addition to a multilingual thesaurus such as MeSH (Medical Subject Index), SNOMED (Systematized Nomenclature of Medicine) and ICD (International Classification of Diseases) for the medical domain, these systems require a domain-specific ontology. In order to extract relevant knowledge from technical documentation containing the domain knowledge, several person-years of highly qualified work are required (Gonzalo et al., 1998).
By knowledge bases in our context we understand the set of concepts belonging to a specific domain and the relations between these concepts that also belong to this domain. But when we turn to ontologies, the richness of a domain becomes relegated to a mere enumeration of concepts and a taxonomic organization of them. That is, there is danger of identifying ontologies as mere thesauri.

4 Some Advances: New Approaches

Our group is the Spanish Language Centre (www.unl.fi.upm.es) of the UNL Programme of the United Nations (www.undl.org). UNL is basically an artificial language for knowledge representation designed for representing contents written in any language and for generating such contents in any natural language. Borrowing the term from the Machine Translation literature, UNL is an interlingua since it plays the role of an intermediate representation of the text meaning in a language independent way. The next section will depict UNL in more detail.

4. 1 UNL as an Interlingua

Formally speaking, UNL follows the schema of semantic nets (that is, UNL expresses binary relations between concepts, labelled by a number of semantic tags). The specifications of the language (UNL Center, 2003) formally define the set of relations, concepts and the so-called attributes. Let's have a look at them in more detail.

Universal words. They conform the vocabulary of the language, i.e., they can be considered the lexical items of UNL. To be able to express any concept occurring in a natural language, the UNL proposes the use of English words modified by a series of semantic restrictions that eliminate the innate ambiguity of the vocabulary in natural languages. If there isn't any English word suitable to express the concept, the UNL allows the use of words from other languages, if the semantic restrictions describe the meaning of the base word with precision. In this way, the language gets an expressive richness from the natural languages but without their ambiguity. Take, for example, the English word "construction" meaning "the action of constructing" and the "final product". Thus, the word "construction" will be paired with two different universal words:

construction$_1$ → construction(icl>action)
construction$_2$ →construction(icl>concrete thing)

where "icl" is the abbreviation for "included". The set of UWs is included in the UNL dictionary.

Relations. These are a group of 41 relations that define the semantic relations among concepts. They include argumentative (agent, object, goal), circumstantial (purpose, time, place), logic (conjunction, and disjunction) relations, etc. For example, in a sentence like "The boy eats potatoes in the kitchen", there is a main predicate ("eats") and three arguments, two of them are instances of argumentative relations ("boy" is the *agent* of the predicate "*eats",* whereas "potatoes" is the *object*) and one circumstantial relation ("kitchen" is the *physical place* where the action described in the sentence takes place). The specifications provides a definition in natural language of the intended meaning of semantic relations and establishes the contexts where relations may apply, like the nature of the origin and final concept of the relation.
For example, an agent relation can link an action (as opposed to an event or process) and an volitional agent (as opposed to a property or a substance). This characterization of concepts implies a top "ontology" or "taxonomy" similar to the Wordnet (Fellbaum, 1998), whose main purpose is validating the correct application of conceptual relations.

Attributes. They express several types of semantic information that usually modifies the predication described by the net of uws linked through the relations. This information includes time and aspect of the event, polarity and modality of the predication, type of reference of the entities described by the UWs, number and/or gender, etc. In the previous sentence, attributes are needed to express plurality in the object ("potatoes"), definite reference in the both the agent ("boy") and the place ("kitchen") and finally and special attribute denoting which UW is the head of the whole expression (the *entry* node).

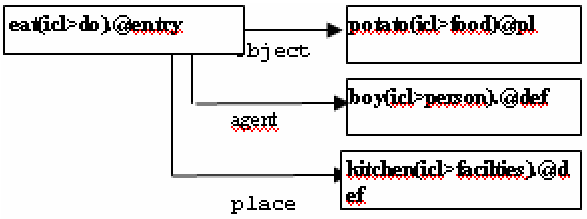

Figure 2: Representation of a UNL expression.

The textual representation in this UNL graph is the following:

agt(eat(icl>do).*@entry*, boy(icl>person).*@def*)
obj(eat(icl>do).*@entry*, potato(icl>food).*@pl*)
plc(eat(icl>do).*@entry*, kitchen(icl>facilities).*@def*)

One of the main objections done to GUM (impossibility of representing information about hearer and speaker) is resolved in UNL by means of attributes: the subset of the language UW + Conceptual Relations defines the propositional part of a given text, the addition of attributes adds contextual meaning like epistemic and deontic modality, speaker's intention, speaker's attitudes, informative structure etc.

4. 2 UNL as Language for Knowledge Representation

UNL is mainly used as a support language for multilingual generation of contents coming from different languages. However, its design allows for non language centred applications, that is, UNL could serve as a support for knowledge representation in generic domains. When there is a need to construct domain-independent ontologies, researches turn back to natural language (such as Wordnet, GUM or even CyC[1]) to explore the "semantic atoms" that knowledge expressed in natural languages is composed of. UNL follows this philosophy, since it provides an interlingual analysis of natural language semantics. The reasons why UNL could be backed as a firm knowledge representation language are:

- The set of necessary relations existing between concepts is already standardized and well defined (overcoming the objection posed by Woods about the asemanticity of semantic networks).
- It is the product of intensive research on the thematic roles existing in natural languages by a number of experts in the area of MT and IA, guaranteeing wide coverage of all contents expressed in any natural language.
- Similarly, the set of necessary attributes that modify concepts and relations is fixed and well-defined, guaranteeing a precise definition of contextual information.
- UNL syntax and semantics are formally defined.

But to really serve as a language for knowledge representation, it must support deduction mechanisms and must specify how a knowledge base could be build up in the UNL language. We will explore this idea by looking closer at the UWs part of the UNL system and how to link them in knowledge base.

4. 2. 1 The UNL Dictionary and its Companion KB

The UW dictionary is a repository of UWs and as such does not organise its contents in any way. It is just a (big) set of UWs, each element having no relation with any other. The necessity of establishing certain relations between UWs arises when considering several desirable features of the UNL system:

- Setting the combinatory possibilities of each UW with respect to any other UW regarding the conceptual relations that may link them and the attributes they may accept.
- Enabling a "fall-back" generation mechanism for those UWs that are not linked with HWs in a given language at a given time. Those UWs would be replaced with semantically close, linked UWs so allowing generation to continue.

If word sense disambiguation were the *only* reason for introducing semantic restrictions into UNL, any of the previous approaches could be adopted. However, semantic restrictions have been also used for a different though related purpose: providing a semantic structure to the otherwise "flat" UW dictionary. However, and in order to support these features, the devised solution consists in creating a *network* with the set of UWs as nodes and *semantic relations* as arcs. In such a network, we use the same information both for disambiguating and for building the KB. The semantic restrictions attached to the UWs for disambiguation purposes *also* express knowledge stored in the KB and conversely; the semantic knowledge serves for disambiguation. Such network is called the UNL KB.
From an *extensional* point of view, the UNL KB can be viewed as a finite set of tuples of the form:
<semantic relation, uw_1, uw_2>

which can be graphically displayed as:

uw_1 —semantic relation→ uw_2

The following are examples of tuple, being "icl" and "agt" abbreviations for "included" and "agent" respectively:

helicopter —icl→ concrete thing
ameliorate —icl→ do
do —agt→ thing

Given the huge amount of tuples that it may contain, the UNL KB is best viewed from an *intensional* point of view as a first order logical theory composed of a finite set of axioms and inference rules[2].
Most of the axioms state plain semantic relations among UWs, now viewed as atomic formulas:

relation(uw_1, uw_2)
Examples:
icl(helicopter, concrete thing)

[1] http://www.cyc.com

[2] This idea is fully developed in the document "The UNL Knowledge Base, a formal description", Luis Iraola. Internal Report, Spanish Language Center. January 1999.

icl(ameliorate, do)
agt(do, thing)

Besides atomic formulas, the theory contains complex formulas, like the one stating the transitivity of the "icl" relation:

$$\forall w_1 \forall w_2 \forall w_3 (\ icl(w_1, w_2) \wedge icl(w_2, w_3) \rightarrow icl(w_1, w_3)\ )$$

As for the inference rules, a subset of the standard rules present in first order theories may suffice for defining the relation of syntactic consequence among formulas. The UNL KB is then formally defined as the closure of the set of axioms under the consequence relation.

We can now turn to the tasks the UNL KB is intended to be used for, and get a clearer picture of its concrete contents according to those tasks. The first task we have mentioned is setting the combinatory possibilities of every UW with respect to the rest of UWs and to the set of conceptual relations (and attributes) included in UNL. For any two UWs w_1, w_2 and any conceptual relation r, the UNL KB should be able to determine whether linking w_1, w_2 with r is allowed (makes sense in principle) or if it is against the intended use of w_1, w_2 and r. If we view the KB as a theory, the question is then if the formula $r(w_1, w_2)$ is a consequence (a theorem) of the set of axioms that form the KB or it is not. The axioms needed for answering such questions are mostly derived from the intended usage of the UNL conceptual relations and the broad semantic classes each UW belongs to.

4. 2. 2 Example

The *instrument relation* ("ins") holds between an event and the concrete thing involved as instrument used for completing the action. In the UNL specifications this is expressed very much like one of our previous formulas:

ins(do, concrete thing)

That is, there is an "ins" arc between UWs "do" and "concrete thing":

do —ins→ concrete thing

On the other hand, the method relation ("met") holds between an event and the mean or method applied for doing the action. This is expressed in the specifications with the formula:

met(do, abstract thing)

Graphically:

do —met→ abstract thing

The differences between "ins" and "met" impose a semantic difference between concrete and abstract things. In order to set the combinatory possibilities of nominal concepts as destination of these relations, nominal UWs must be included under the "concrete thing" or "abstract thing" respectively. Verbal concepts included under "do" qualify as origin of both relations. These inclusions plus the axioms governing "ins" and "met" are all that is needed in the KB for setting the combinatory possibilities regarding "ins" and "met".

5 Considerations for Building the UNL Ontology

The UNL ontology has been developed with several considerations in mind:

- Linguistic relevance. The main goal of the ontology is to aid to the tasks carried out by Analyzers and Generators of UNL, and more generally to any task related with the processing of natural language.
- Language independence. The divisions made in the uppermost levels of the ontology (which are presented in this document) try to be based on very general semantic distinctions present in most of the natural languages.
- Exhaustive and disjoint classifications. The ontology should cover the whole range of concepts (universal words) and, at least in its up-most levels, its divisions should be disjoint.
- Clear membership criteria. Though there always be concepts difficult to situate in the ontology, the goal of giving clear criteria for applying the classification is considered central.

One of the main characteristics of UNL is its flexibility both formally and linguistically. From a linguistic point of view, the UNL ontology serves to a wide variety of natural languages. From the formal point of view, its integration with other support languages (HTML, XML, OWL) could be easily achieved. UNL and OWL could be considered as complementary, integrating thus the formal rigour and machine readability of OWL and the expressiveness and language and domain independence of UNL.

Essentially, UNL has the capability of representing knowledge. However the classical problem emerges. It is the semantic validation process, that is, the set of mechanisms able to deduce coherent domain knowledge from existing one. This is still an open problem, that so far has only attained some partial solutions based on the application of the logic verification rules. However verification rules are not enough to establish a model with sufficient semantic coherence.

6 An Illustration of our Approach

Possibly, one of the best examples of the utility of UNL is its capability to build knowledge bases from texts in an automatic way. In the following example, we have three sentences from a Spanish document about Heritage policies, more specifically it shows some procedures of how to catalogue existing heritage.

The text is:

1. *To integrate the catalogues of all the Spanish museums in the General Catalogue of Historical Heritage*
2. *To establish the necessary mechanisms to integrate all the information from the Autonomous Communities.*
3. *The registry campaigns to make the General Inventory of Moveable Assets of the Church.*

Figure 3, 4 and 5 show the UNL representation of first, second and third sentences respectively.

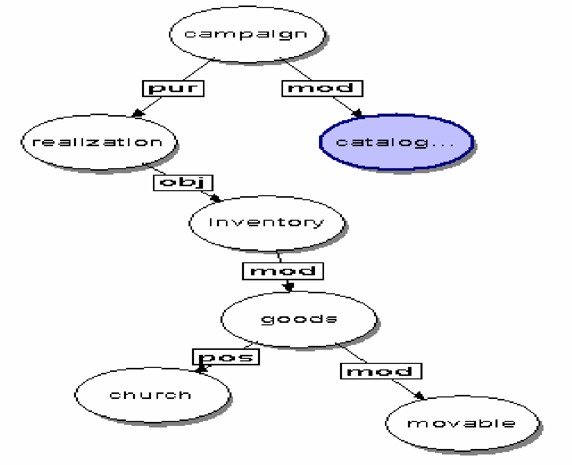

Figure 3: UNL representation of sentence 1

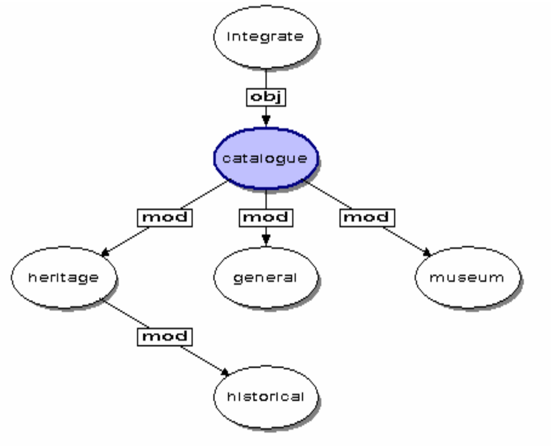

Figure 4: UNL representation of sentence 2

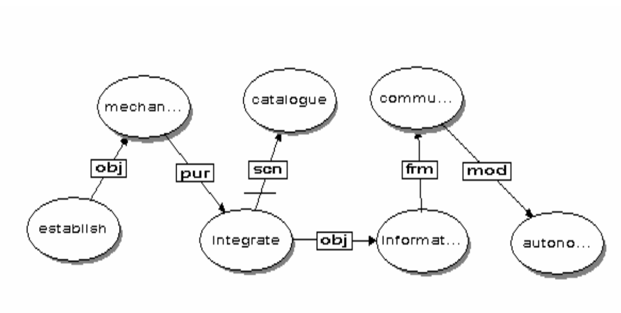

Figure 5: UNL representation of sentence 3

All these graphs could be merged. Here we are aiming at the capacity of UNL to represent and organize contents, and as a side effect, to generate such contents into different natural languages. By adding all sentence information into a single representation, we can deduct more information about the concepts present in each sentence. Figure 6 represent the merged graph.

The joined representation offers new relations among concepts than those presented in the sentence representations. For example, the term "catalogue" is related to several concepts, some relations may be true whereas others not. For example, here a "catalogue" is describes as an entity that could be generic, could belong to museums, to historical heritages, can be a virtual place where some actions are carried out or can be the effective object of a "carrying out" activity. This can be applied to the problem of query expansion in a knowledge based manner, which can complement well known linguistic query expansion (morphological and syntactical variations).

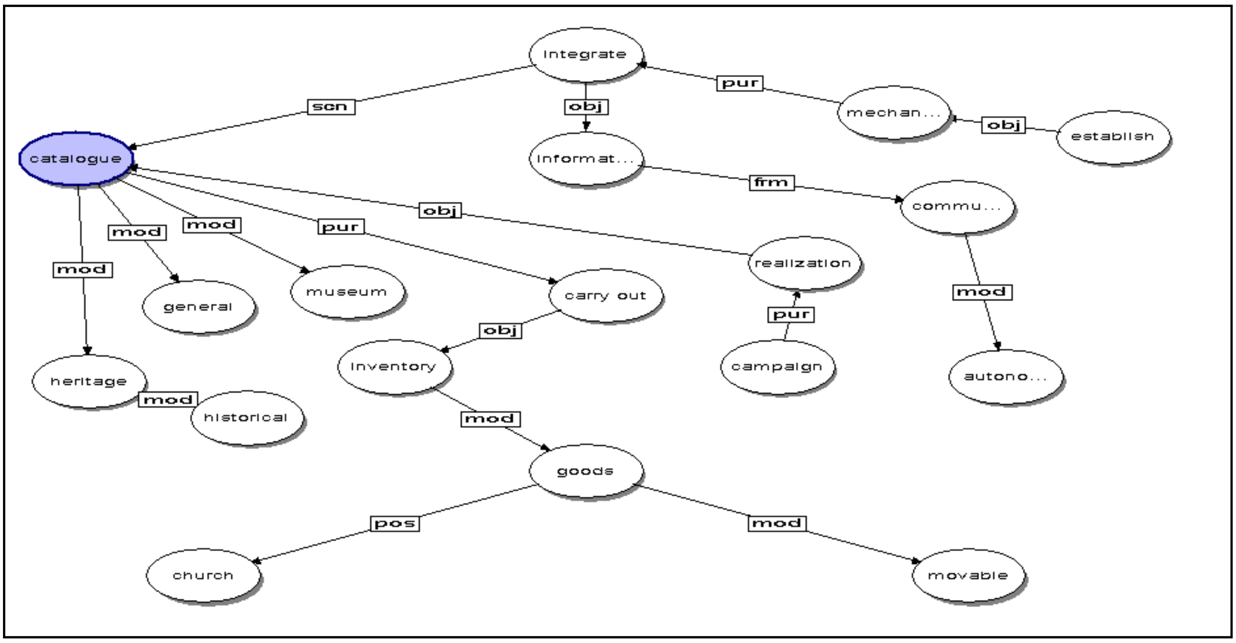

Figure 6: UNL merged graph

7 Conclusions

UNL is a language capable of being the formal basis of knowledge representation and also a language capable of representing information coming from textual sources. In order to propose this language as a possible standard for these purposes, we want to bring the attention to the fact that it is maintained by an open world-wide organisation which provides the necessary institutional support.

UNL researchers face the same problems than any others in this field, that is, to define mechanisms that guarantee the coherence of inferred knowledge. We, of course, assume that knowledge represented into UNL can transform implicit knowledge into explicit one. As in many other problems, inferred knowledge needs to be validated, and for that it is necessary to design the domain ontology where the valid combination of relations are defined, including the restrictions that cannot be violated.

A possible way for building such domain ontology is to painstakingly encode all the concepts and relations for the application domain. Our approach relies on statistical analysis of UNL representations of domain specific texts. By exploiting these representations, we hope to be able to build such a domain ontology. At the moment we have obtained encouraging initial results in the cultural heritage domain as a side effort of the Spanish Language Center in the Herein project[3]. We hope to have reliable results during this year.

References

Ballesteros, L. & Croft, W.B. (1996). Dictionary Methods for Cross-Lingual Information Retrieval. In Proceedings of the 7th International DEXA Conference on Database and Expert System, 791-801.

Ballesteros, L. & Croft, W.B. (1997). Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In Proceedings of ACM SIGIR Conference, 20, 84-91.

Bateman, J.A., Henschel, R. & Rinaldi, F. (1995). The Generalized Upper Model 2.0. http://www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html.

Bechhofer. S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P.F. & Stein, L. I. (2004). OWL: Web Ontology Language Reference. W3C Recommendation 10 February 2004 http://www.w3.org/TR/owl-ref/.

Buitelaar, P., Netter, K. & Xu, F. (1998). Integrating Different Strategies in Cross-Language Information Retrieval in the MIETTA Project. In Proceedings of the 14th Twente Workshop on Language Technology (TWLT 14). Language Technology in Multimedia Information Retrieval, December 1998. Enschede, The Netherlands.

Burg, J. F. M. & van de Riet, R. P. (1997). The impact of linguistics on conceptual models: consistency and understandability. In Data & Knowledge Engineering, Vol. 21, 131 – 146.

Farquhar, A., Fikes, R. & Rice, J. (1996). The Ontolingua Server: a Tool for Collaborative Ontology Construction. In Proceedings of the 10th knowledge acquisition for knowledge-based systems workshop, Canada.

Fellbaum, C. (Ed.) (1998). WordNet: An Electronic Lexical Database. Language, Speech, and Communication Series. MIT Press.

Fensel, D., Horrocks I, van Harmelen, F., Decker S., Erdmann, M. & Klein, M. (2000). OIL in a nutshell. In Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling, and Management.

Genesereth, M. R. & Fikes, R. E. (1992). Knowledge Interchange Format. Version 3.0. Reference Manual. Computer Science Department. Standford University, California.

Gonzalo, J., Verdejo, F., Peters, C. & Calzolari, N. (1998). Applying EuroWordNet to Cross-Language Text Retrieval. Computers and the Humanities, Special Issue on EuroWordNet.

Gonzalo, J., Verdejo, F. & Chugur, I. (1999). Using EuroWordNet in a Concept-Based Approach to Cross-Language Text Retrieval. Applied Artificial Intelligence Special Issue on Multilinguality in the Software Industry: the AI contribution.

Gruber, T. A. (1993). A translation Approach to portable ontology specifications. Knowledge Acquisition. Vol. 5, 199-220.

Horrocks I., van Harmelen, F. (2001). Reference description of the DAML+OIL ontology markup language. http://www.daml.org/2001/03/reference.html.

Hovy, E. H., Ide, N., Frederking, R. E., Mariani, J. & Zampolli, A. (Eds.) (2000). Multilingual Information Management. Available at http://www.cs.cmu.edu/~ref/mlim - 20-12-00.

Hull, D. & Grefenstette, G. (1996). Experiments in Multilingual Information Retrieval. In Proceedings of ACM, SIGIR'96. Zurich.

Lassila, O and Swick, R. R (1999). Resource Description Framework (RDF) Model and Syntax Specification. W3C recommendation, 1999. www.w3.org/TR/PR-rdf-syntax.

LIQUID project: http://liquid.sema.es/.

Luke, S. & Heflin, J. (2000). SHOE 1.01. Proposed specification. http://www.cs.umd.edu/projects/plus/SHOE/spec.html.

MacGregor, R. & Bates R. (1987). The Loom Knowledge Representation Language. Technical Report ISI-RS-87-188. USC Information Sciences Institute, Marina del Rey, CA.

MIETTA project: http://www.mietta.info/.

Minsky, M. (1974). A Framework for Representing Knowledge. MIT Press.

MuchMore project: http://muchmore.dfki.de/.

MULINEX: Multilingual Indexing, Navigation and Editing Extensions for the World-Wide Web. In Proceedings of the AAAI Spring 1997 Symposium on Cross-Language Text and Speech Retrieval. Menlo Park, CA.

Neches, R, Fikes, R. E., Finin, T., Gruber, T. R., Senator, T. & Swartout, W. (1991). Enabling technology for knowledge sharing. AI Magazine, 12(3), 36-56.

Oard, D. (1997). Alternative Approaches for Cross-Language Text Retrieval. In Proceedings of the AAAI Spring 1997 Symposium on Cross-Language Text and Speech Retrieval. Menlo Park, CA.

Quillian, M.R. (1968). Semantic Memory. In M. Minsky (Ed.), Semantic Information Processing.Cambridge: MIT Press.

[3] http://www.european-heritage.net

Sánchez V. & Belskis, I. (2002). Multilingual terminology extraction and validation. In Proceedings of the LREC 2002, Las Palmas de Gran Canaria.

Shamsfard, M. & Abdollahzadeh, A. (2004). Learning Ontologies from Natural Language Texts. International Journal of Human Computer studies. Vol. 60, 17-63.

UNL Center. UNL specifications v.3.2. July, 2003. http://www.undl.org/unlsys/unl/UNL%20Specifications .htm.

Vintar, S., Buitelaar, P. & Volk, M. (2003). Semantic Relations in Concept-Based Cross-Language Medical Information Retrieval. In Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM), September 22nd , 2003, Cavtat-Dubrovnik, Croatia.

Volk, M., Vintar, S. & Buitelaar, P. (2003). Ontologies in Cross-Language Information Retrieval. In Proceedings of WOW2003 (Workshop Ontologie-basiertes Wissensmanagement), Luzern, Switzerland, April 2003.

Wood, J. (1985). What's in a link ? In R. Brachman and H. Levesque (Eds.), Readings in Knowledge Representation. Morgan Kaufmann.

www.undl.org

Yang, Y., Carbonell, J., Brown, R. & Frederking, R., (1998). Translingual Information Retrieval: Learning from Bilingual Corpora. AI Journal Special Issue: Best of IJCAI'97.

Yergeau, F, Bray, T, Paoli, J, Sperberg-McQueen, C. M & Maler E. (2004). Extensible Markup Language (XML) 1.0 (Third Edition). W3C Recommendation. http://www.w3.org/TR/REC-xml/.

Assessing Relevance in CLEF

Michael Kluck

Informationszentrum Sozialwissenschaften (IZ)
Lennéstr. 30, 53113 Bonn, Germany
kluck@bonn.iz-soz.de

Abstract

The paper presents the assessment of CLEF in detail, showing the rules and procedures, illustrating the assessment organization und processes. It also discusses the validity of the assessments in the context of the pooling method. Finally the specific issues of the assessment in the multilingual context are examined and comparable activities are analyzed.

1 Introduction

The series of campaigns organized by the Cross-language Evaluation Forum (CLEF) aim on system evaluation of Cross-Language Information Retrieval (CLIR) systems and the promotion of research and development in this area by providing a test and evaluation infrastructure[1]. The evaluation is based on two resources presented to the participants: the topics expressing user information needs and the corpora to be analyzed whether they contain relevant documents fulfilling these information needs. Then, the results delivered by the participants are intellectually judged whether they are relevant answers to the topic in question. The paper presents the assessment of CLEF in detail, showing the rules and procedures, illustrating the assessment organization und processes. It also discusses the validity of the assessments in the context of the pooling method. Finally the specific issues of the assessment in the multilingual context are examined and comparable activities are analyzed.

2 Organization of the Assessment in CLEF

The relevance assessment is based on the pooling method that has been developed by the TREC-Initiative (Harman & Vorhees, 2001; Vorhees, 2002). The procedure and the metrics of TREC are used in CLEF, too. This means mainly the pooling method, which is applied to large test collections as they are also used in CLEF. The pooling method creates a subset of documents out of the whole collection to be judged for a specific topic, as it is not possible to judge all documents of a large collection for each single topic. Un-judged documents (those not included in the pool) are assumed to be not relevant. The rules for the assessment are also closely related to those used in TREC.

Precondition for a sufficient coherence of the assessments is a full understanding of the topics and their coverage. This is especially important in a multilingual context. Thus, already the topic creation phase establishes the basics for the further steps of assessment. The topic creation phase consists of the following steps: “invention” of topic ideas by each language group, first test of these topics against the collections from different languages, proposal of a set of topics from each language group, face-to-face discussion of all language groups on the topics, refinement of the topics especially with respect to the common understanding and comprehensibility in all languages, final decision on the topics of a campaign, translation of the topics into the other official and unofficial languages, re-check of all translations into the official languages (Kluck & Womser-Hacker, 2002; Womser-Hacker, 2002; Mandl & Womser-Hacker, 2003).

The participants run all the given topics against the data collections belonging to the chosen task. This may be done in different types of tasks: 1. as monolingual task (topics and data have the same language), 2. as bilingual task (topics from one language against data from another language) or 3. as multilingual task (topics from one language against adapt from several languages). Then for each topic, the participating research groups merge the results form all data collections (and languages) into one result set of documents[2], which are considered relevant.

In the next step these result sets, which are separated by topics and coming from the participating groups, are pooled (by the CLEF team) by using a cut-off for each results set. In the current CLEF campaigns, this cut-off is set at the top 60 documents. This means all 60 top-ranked document numbers from the runs[3] of each group are merged into one list for the respective topic. Those 60 documents are seen to be most likely relevant to that topic, since the retrieved results are delivered by the participating systems in a ranked list with decreasing order of relevance. By processing this pool all identical numbers are eliminated. Because of some overlap between the result sets from the different participating systems, the resulting list consists of about a quarter of the possible maximum amount of documents (which would be 60 times the number of delivering groups). Thus the pooled result list of a specific topic includes all as relevant declared hits of any participating group up to the 60th hit. In contrast to TREC and INEX there is no cut-off (like top 1000) for the whole pooled result list of each topic, which means that result hits are only included in the pool, until a sum of 1000 hits per topic is reached.

Afterwards every result list is divided per topic into sub-lists for any concerned language. The documents in these

[1] See also www.clef-campaign.org . CLEF also aims at creating test-suites of reusable data which be employed by system developers for benchmarking purposes and further research. See also Harman et al. 2001; Kluck/Mandl/Womser-Hacker 2002; Kluck 2002)

[2] The documents are represented by their document identifiers (numbers), which allow a one-to-one identification.

[3] Run means each set of results for all topics which has been treated with a different retrieval methodology and/or a different retrieval software and has been delivered by a participating group within a specific track or task of CLEF.

sub-lists are sorted by their numbers neither regarding any degree of relevance nor the producer of these results. This is to avoid any influence of these factors on the assessors.
These language and topic related lists are given to the assessors of the respective language group. Then the results are judged whether they really represent a relevant answer to the respective topic (information need). A binary decision is requested: relevant or not relevant. Thus, no ambiguity is allowed in the decision. The assessment is supported by the Assess software from NIST, which allows the highlighting of search terms during the assessment process[4]. Then the relevance measures are computed and the recall-precision curves and other figures produced and sent out to the participants. At the same time the single judgments are spread out to the participants. The calculations include statistics per participant and overall comparisons.

3 The Assessment Process in CLEF

The assessment process itself is integrated in the workflow of the pooling as shown below (Figure 1). The assessment is based on the clear definition of the scope of each topic. The definition is given in the information elements of a given topic (title, description and narrative). The topic creation group provides sometimes additional information for all assessors. This additional information may give hints like the names searched for, the time-span of an expected result, alternative spellings of the searched person, institution or event, exclusion reasons or examples.
As orientation for the assessors general assessing rules have been provided. The essence of these rules is that the assessors should judge a document as relevant regardless of other documents even if they are containing the same information or are occurring more than once. The coordinators of the language groups know the topic creation discussion in detail to be able to answer questions of the assessors concerning the topic meanings. If any uncertainty occurs during the assessment process an e-mail discussion is executed between the language groups.
One single assessor for any topic carries out the assessment itself. The assessors are advised to execute the assessment for each single topic as one procedure without break to avoid shifts in the judgments. During the assessment a two-stage approach is used. In the first run decisions are made on clearly relevant documents, in the second run unclear cases are re-examined by the assessors. Remaining problems of understanding are discussed with the other assessors of the language group and/or the coordinator. It is very important to make sure that there is no shift of criteria during the time of assessing even if there are a lot of documents delivered for one topic.

4 Discussion on Validity of Assessments

In the context of the use of the pooling method the discussion on the validity of the assessments arises quite often. But the comparative evaluation method has been proved as reliable by several studies on this problem (Braschler, 2001, 2002, 2003, 2004; Hiemstra, 2001). The assumption is that a sufficient number of included runs will turn up the most of the relevant documents. But if a system did not contribute to the pool of judged documents, it might be unfairly penalised by the evaluation statistics. These fears have been rejected by those studies examining the effect of adding or discarding the contribution of one system. On the other hand it was demonstrated that the variations of judgements by different assessors and for the same assessor over time do not affect the comparative evaluation. The conclusion has been that the comparative approach of the pooling methodology is sufficiently backed how the assessments are done (Vorhees, 2000, 2002; Zobel, 1998).

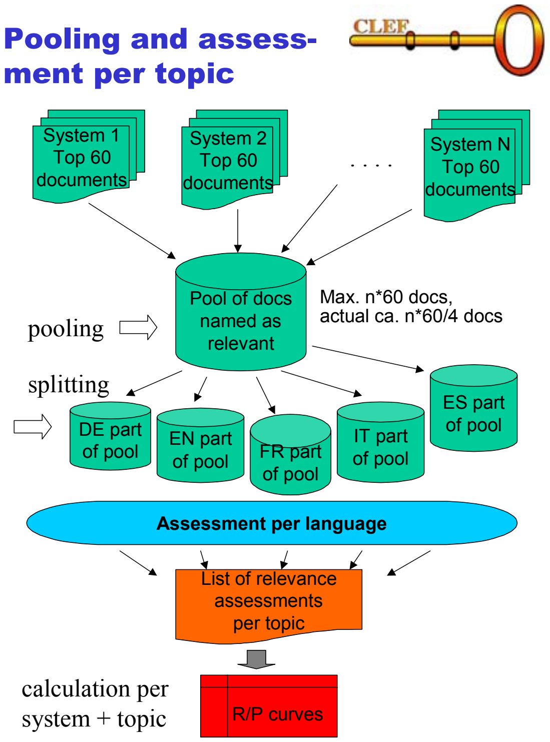

Figure 1: Pooling and Assessment in CLEF

5 Specific Problems of Cross-Language Assessment

Some specific problems arise from the fact that multiple language corpora and result sets have to be handled by distributed language groups. Thus, establishing coherence of the judgments across the language groups is a major problem. Precondition is a translation of the topics from their original language (the language of that language group which developed the respective topic) into the other languages. These translations must be done without loss of information and without shift of meaning, but in a way, which is adequate to the respective language. An example is given below in figure 2.

[4] Reidsma et al. (2003) report on an experiment that was carried out during the assessment of the Dutch results for the CLEF 2002 campaign. The goal of the experiment was to examine possible influences on the assessments caused by the use of highlighting in the assessment program.

```
<top>
<num> C064 </num>
<NL-title> Muisarm </NL-title>
<NL-desc> Zoek documenten waarin melding wordt gemaakt van een muisarm. </NL-desc>
<NL-narr> Relevante documenten melden klachten die veroorzaakt worden door het langdurig gebruik van een computermuis. Documenten waarin maatregelen worden genoemd voor het voorkomen van een muisarm tijdens het werken achter de computer zijn ook relevant. </NL-narr>
</top>
<top>
<num> C064 </num>
<ES-title> Síndrome RSI y ratones de ordenador </ES-title>
<ES-desc> Encontrar documentos que informen sobre RSI ("repetitive strain injuries" o "enfermedad del periodista") producidas por el uso del ratón del ordenador. </ES-desc>
<ES-narr> Los documentos relevantes informan sobre daños causados por el uso continuado de un ratón de ordenador. Los documentos que proponen formas de evitar el RSI cuando se usa el ordenador también son relevantes. </ES-narr>
</top>
```

Figure 2: Dutch Topic example with Spanish translation: RSI

Some further problems are handling of spelling variants of acronyms and proper names in different languages, of differences in language sub-areas like American versus British English, or German in Germany versus German in Switzerland/Austria (Womser-Hacker, 2002; Kluck & Womser-Hacker, 2002; Mandl & Womser-Hacker, 2002).
Nevertheless for each topic there are as much different assessors doing judgments as language collections are involved. For each language several different native assessors assess every topic. In the CLEF 2001 campaign there have been between 2 and 10 assessors per language group (Hiemstra, 2001).
For specific problems of the assessment of the domain-specific GIRT data see Kluck (2004).

6 Comparison with Other Evaluation Campaigns

In comparison to TREC, NTCIR and INEX there is a lot of similarities, but also some differences. Most of the rules and metrics have been developed in the TREC context and been taken over for CLEF, NTCIR and INEX. Depending on the specific tasks of each evaluation campaign amendments had to be developed. The big overlap allows comparable views on the assessment procedures in each campaign (Fuhr et al., 2002, 2003; Kazai et al., 2003; Kando, 2001, 2002, 2003; Harman & Vorhees, 2001; Braschler & Peters 2001, 2002, 2003, 2004). The assessments in all campaigns are done with the pooled result sets. But different types of judgement are applied. TREC and CLEF use one-dimensional binary decisions whether a document is considered to really be relevant or not (relevant, irrelevant). As INEX treats XML-documents, which provide also structured subsections of documents, INEX uses a two-dimensional scheme, which has four values for each dimension. One dimension represents the topicality of a document with four grades (irrelevant, marginally relevant, fairly relevant, highly relevant) and the other dimension the document coverage represented by the retrieved (sub)-section of the document with four grades (no coverage, too large, too small, exact coverage). NTICIR uses a one-dimensional scheme with 4 grades (highly relevant, fairly relevant, partially relevant, irrelevant).
In our experience the usage of four-value scales does not improve the stability and reliability of the assessors' judgements or the results. In the end the important difference lies between relevant and irrelevant, and "in most cases, however even though users and/or judges were able to conceptualize a meaningful difference between relevant and partially relevant documents, the experimental results were combined, collapsed, or grouped into a single category of "relevant" for the purpose of analysis." (Greisdorf, 2000) For instance, the NTCIR campaign groups the result into three separate ranked list of relevance: the first only including "highly relevant" documents, the second including "highly relevant" and "fairly relevant" documents, the third including "highly relevant", "fairly relevant", and "partially relevant" documents as well.

References

Braschler, M. (2001). CLEF 2000 - Overview of the Results. In C. Peters (2001) , pp. 89- 101.

Braschler, M. (2002). CLEF 2001 - Overview of Results. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2002), pp. 9-26.

Braschler, M. & Peters, C. (2002). CLEF Methodology and Metrics. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2002).

Braschler, M. & Peters, C. (2002). CLEF 2002 Methodology and Metrics. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2003) pp. 512-528.

Braschler, M. (2003). CLEF 2002 - Overview of Results. In Braschler, M., Peters, C. (2002): CLEF Methodology and Metrics. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2003) pp. 9-27.

Braschler, M. & Peters, C. (2002). CLEF Methodology and Metrics. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2002) pp. 394-404.

Braschler, M. (2004). CLEF 2003 - Overview of Results.

Braschler, M. & Peters, C. (2002). CLEF Methodology and Metrics. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2004) forthcoming.

Fuhr, N., Gövert, N., Kazai, G. & Lalmas, M. (2002). INEX: Initiative for the Evaluation of XML Retrieval. In R. Baeza-Yates, N . Fuhr, Maarek, S. Yoelle (Eds.), Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval. see http://www.is.informatik.uni-duisburg.de/bib/fulltext/ir/Fuhr_etal:02a.pdf .

Fuhr, N., Gövert, N., Kazai, G. & Lalmas, M. (Eds.) (2003). Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX), Schloss Dagstuhl, Germany, December 9-11, 2002. Sophia Antipolis: ERCIM see http://www.ercim.org/publication/ws-proceedings/ INEX2002.pdf .

Greisdorf, H. (2000). Relevance: An Interdisciplinary and Information Science Perspective. In Informing Science 2 (3), pp. 67-71.

Harman, D., Braschler, M., Hess, M., Kluck, M., Peters, C., Schäuble, P. & Sheridan, P. (2001). CLIR Evaluation at TREC. In Peters, C. (E.) (2001) pp. 7-23.

Hiemstra, D. (2001). The CLEF Relevance Assessment in Practice. Talk at the CLEF 2001 Workshop, September

3, Darmstadt, Germany, see http://videoserver. iei.pi. cnr.it:2002/DELOS/CLEF/hiemstra-wkshp01.pdf .

Kazai, G., Gövert, N., Lalmas, M. & Fuhr, N. (2003). The INEX evaluation initiative. In H. Blanken, T. Grabs, H.-J. Schek, R. Schenkel, & G. Weikum (Eds.), Intelligent Search on XML Data - Applications, Languages, Models, Implementations, and Benchmarks. pp. 279-293), see http://www.is.informatik.uni-duisburg.de/bib/ fulltext/ir/Kazai_etal:03b.pdf

Kando, N. (2001). NTCIR Workshop: Japanese- and Chinese-English Cross-Lingual Information Retrieval and Multi-Grade Relevance Judgments. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2001), pp. 24 – 35.

Kando, N. (2002). CLIR System Evaluation at Second NTCIR Workshop. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2002), pp. 371-388.

Kando, N. (2003). CLIR at NTCIR Workshop 3: Cross-Language and Cross-Genre Retrieval. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2003), pp. 485-504.

Kluck, M., Mandl, T. & Womser-Hacker, C. (2002). Cross-Language Evaluation Forum (CLEF). Europäische Initiative zur Bewertung sprachübergreifender Retrievalverfahren. In nfd – Information Wissenschaft und Praxis 2 (53), pp. 82-89.

Kluck, M. & Womser-Hacker, C. (2002). Inside the Evaluation Process of the Cross-Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment. In M. G. Rodríguez & C. P. S. Araujo (Eds.). Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, Las Palmas de Gran Canaria 29-31 May 2002, pp. 573-576. Paris: ELRA.

Kluck, M. (2002). Das Cross-Language Evaluation Forum (CLEF) - Evaluationsumgebung und Forschungskontext für mehrsprachiges Information Retrieval (mit einer Skizze der Ergebnisse von CLEF 2002). In R. Hammwöhner, C. Wolff, & C. Womser-Hacker (Eds.), Information und Mobilität. Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissen schaft (ISI 2002), pp. 225-237. Konstanz: UVK.

Kluck, M. (2004). Evaluation of Cross-Language Information Retrieval Using the Domain-Specific GIRT Data as Parallel German-English Corpus. (in the Proceedings volume of LREC 2004).

Mandl, T. & Womser-Hacker, C. (2003). Linguistic and Statistical Analysis of the CLEF topics. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2003), pp. 505-511.

Peters, C. (Ed.) (2001). Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Information Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised Papers. Berlin et al.: Springer (Lecture Notes in Computer Science, 2069).

Braschler, M., Gonzalo, J., Kluck, M. & Peters, C. (Eds.) (2002). Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001. Revised Papers. Berlin et al.: Springer (Lecture Notes in Computer Science, 2406).

Braschler, M., Gonzalo, J., Kluck, M. & Peters, C. (Eds.) (2003). Advances in Cross-Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002, Rome, Italy, September 19 - 20, 2002 ; Revised Papers. Berlin et al.: Springer (Lecture Notes in Computer Science; 2785).

Braschler, M., Gonzalo, J., Kluck, M. & Peters, C. (Eds.) (2004). Advances in Cross-Language Information Retrieval: Results of the Cross-Language Evaluation Forum - CLEF 2003 Trondheim, Norway, August 2003 Revised Papers. Berlin et al.: Springer (Lecture Notes in Computer Science; forthcoming).

Hiemstra, D., de Jong, F., Kraaij, W. & Reidsma, D. (2003). Cross-Language Retrieval at the University of Twente and TNO. In C. Peters, et al (2003), pp. 197-206.

Vorhees, E. M. (2002). The Philosophy of Information Retrieval Evaluation. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2002), pp. 355-370.

Voorhees, E. (2000). Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness. In Information Processing & Management (36), pp. 679-716.

Voorhees, E. & Harman, D. (2001). Overview of the Ninth Text Retrieval Conference (TREC-9). In E. Voorhees & D. Harman (Eds.) The Ninth Text REtrieval Conference (TREC 9) pp. 1-14. Gaithersburg: NIST see http:// trec.nist.gov/pubs/trec9/t9_proceedings.html.

Womser-Hacker, C. (2002). Multilingual Topic Generation within the CLEF 2001 Experiments. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2002), pp. 389-393.

Zobel, J. (1998). How Reliable Are the Results of Large-Scale Information Retrieval Experiments? In W. Bruce, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (eds.). Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 307-314. New York: ACM Press.

Analysis of Topic Features in Cross-Language Information Retrieval Evaluation

Thomas Mandl, Christa Womser-Hacker

University of Hildesheim
Information Science
Marienburger Platz 22- D-31141 Hildesheim, Germany
{mandl,womser}@uni-hildesheim.de

Abstract

Lessons learned from large scale information retrieval evaluation go beyond the optimization of systems. Data mining on evaluation results allows insights into the relationships between system, user and query features. The following study provides an example for such a meta-analysis for named entities in retrieval tasks. The existence of named entities within topics has a significant influence on the performance of retrieval systems participating in the Cross Language Evaluation Forum (CLEF). Named entities in topics lead to better retrieval quality in general and for most systems. The performance of the individual retrieval systems varies for topics with no, few and more named entities. As a consequence, a fusion approach is envisioned which directs a topic toward an appropriate system. Based on the number of named entities the topic contains, a system is chosen which performs well for topics of this category. We assume that such an approach has great potential for optimizing system results.

1 Introduction

The difficulty of a topic has been an issue in information retrieval research for some time. The identification of difficult topics and their proper treatment seems to be one of the remaining research tasks with great potential for system optimization. We approached this challenge by analyzing the linguistic structure of a topic. A primary study (Mandl & Womser-Hacker, 2004) revealed that mainly named entities seem to be a promising factor. The current study is therefore dedicated to named entities and their effect on retrieval performance.

2 Named Entities in the Multilingual Topic Set

The data for this study was extracted from the Cross Language Evaluation Forum (CLEF) (Braschler et al., 2003; Braschler et al., 2004). CLEF is a large evaluation initiative which is dedicated to cross-language retrieval for European languages. The setup is similar to the Text Retrieval Conference (TREC) (Harman & Voorhees, 1997; Buckland & Voorhees, 2003). The main tasks for multilingual retrieval are:

- The core and most important track is the **multilingual task**. The participants choose one topic language and need to retrieve documents in all main languages. The final result set needs to integrate documents from all languages ordered according to relevance regardless of their language.
- The **bilingual task** requires the retrieval of documents different from the chosen topic language.

The topic creation for CLEF needs to assure that each topic is translated to all languages without modifying the content and providing equal chances for systems which start with different topic languages. Therefore, a thorough translation check of all translated topics in CLEF was performed to check if the translations to all languages resulted in the same meaning (Womser-Hacker, 2002). Nevertheless, the topic generation process follows a natural way and avoids artificial construction (Kluck & Womser-Hacker, 2002).

The topic language is the language which the system designers chose to construct their queries. The retrieval performance of the runs for the topics was extracted from the appendix of the CLEF proceedings (Braschler et al., 2003; Braschler et al., 2004).

An intellectual analysis of the results and the properties of the topics had identified named entities as a potential indicator for good retrieval performance. Because of that, named entities in the CLEF topic set were analyzed in more detail.

The analysis included all topics from the campaigns in the years 2001 through 2003. The number of named entities in the topics were assessed intellectually. We focused on the number of types of named entities in the topics. Table 1 shows the overall number of named entities found in the topic sets.

CLEF year	Number of topics	Total number of named entities	Average number of named entities in topics	Standard deviation of named entities in topics
2001	50	60	1.20	1.06
2002	50	86	1.72	1.54
2003	60	97	1.62	1.18

Table 1: Number of named entities in the CLEF topics

For our further analysis, only tasks with more than eight runs were considered.

3 Named Entities and General Retrieval Performance

Our first goal was to measure whether named entities had any influence on the overall quality of the retrieval results. In order to measure this effect we first calculated the correlation between the overall retrieval quality achieved for a topic and the number of named entities encountered in this topic. In the second section, this analysis is refined to single tasks and specific topic languages.

3.1 Correlation between Average Precision and Number of Proper Names

First, we show the overall performance in relation to the number of named entities in a topic. The 160 analyzed topics contain between zero and six named entities. For each number n of named entities, we determine the overall performance by two methods: (a) take the best run for each topic (b) take the average of all runs for a topic. For both methods, we obtain a set of values for n named entities. Within each set we can determine the maximum, the average and the minimum. For example, we determine for method (a) the following values: best run for n named entities, average of all best runs for n named entities and worst run among all best runs for n named entities. The last value gives the performance for the most difficult topic within the set of topics containing n named entities. The maximum of the best runs is in most cases 1.0 and is therefore omitted. The following table 2 and figure 1 show these values.

Number of named entities	0	1	2	3	4	5
Number of Topics	42	43	40	20	9	4
Average of Best System per Topic	0.62	0.67	0.76	0.83	0.79	0.73
Minimum of Best System per Topic	0.09	0.12	0.04	0.28	0.48	0.40
Standard Deviation of Best System per Topic	0.24	0.24	0.24	0.18	0.19	0.29

Table 2: Method a: Best run for each topic in relation to the number of named entities in the topic (topics 41 to 200)

The CLEF campaign contains relatively few topics with four or more named entities. The results for these values are therefore not significant.

It can be observed that topics with more named entities are generally solved better by the systems. This impression can be confirmed by a statistical analysis. The average performance correlates to the number of named entities with a value of 0.43 and the best performance with a value of 0.26. Disregarding the topics from the campaign in 2003 leads to a correlation coefficient of 0.35. These relations are statistically significant at a level of 95%.

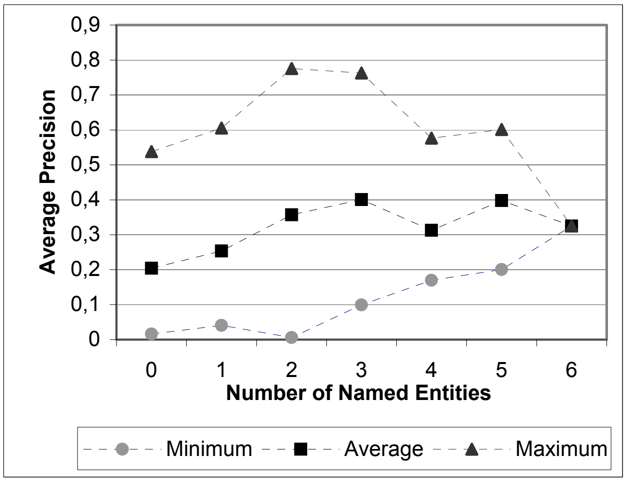

Figure 1: Method b: Relation between system performance and the number of named entities

3.2 Correlation for Individual Tasks and Topic Languages

The correlation analysis was also carried out for the individual retrieval tasks or tracks. This can be done by calculating the average precision for each topic achieved within a task, by taking the maximum performance for each topic (taking the maximum average precision that one run achieved for that topic) and by calculating the correlation between named entities and average precision for each run individually and taking the average for all runs within a task. Except for one task (multilingual with topic language English in 2001), all correlations are positive. Thus, the effect which was before observed overall, occurs within most tasks and even within most single runs.

There is no difference in the average strength of the correlation for German (0.27) and English (0.28) as topic language. The average for each language in the last column shows a more significant difference. The correlation is stronger for German (0.19) than for English (0.15) as topic language. Furthermore, there is a considerable difference between the average correlation for the bi-lingual (0.35) and multi-lingual run types (0.22). This could be a hint, that the observed positive effect of named entities on retrieval quality is smaller for multi-lingual retrieval.

4 Performance Variation of Systems for Named Entities

In this chapter, we show that the systems tested at CLEF perform differently for topics with different numbers of named entities. Although proper names make topics easier in general and for almost all runs, the performance of systems varies within the three classes of topics based on the number of named entities. As already mentioned, we distinguished three classes of topics, (a) the first class with no proper names called *none*, (b) the second class with one and two named entities called *few* and (c) one class with three or more named entities called *lots*. This categorization is similar to the one proposed in other experiments where topics were grouped according to their difficulty (Braschler et al., 2003). However, our approach is suited for an implementation and allows the categorization before the experiments and the relevance assessment. It requires no intellectual intervention but solely a named entity recognition system.

4. 1 Variation of System Performance

As we can see in table 2, the three categories are well balanced for the CLEF campaign in 2002. For 2003, there are only few topics in the first and second category. Therefore, the average ranking is extremely similar to the ranking for the second class *few*.

A look a the individual runs shows large differences between the three categories. Sometimes even the best runs perform quite differently for the three categories. Other runs perform similarly for all three categories.

4.2 Correlation of System Rankings

The performance variation within the classes leads to different system rankings for the classes. An evaluation campaign including, for example, only topics without named entities may lead to different rankings. To analyze this effect, we determined the rankings for all runs within each named entity class, *none*, *few* and *lots*. Table 5 shows that the system rankings can be quite different for the three classes. The difference is measured with the Pearson rank correlation coefficient.

For most tracks, the original average system ranking is most similar to the ranking based only on the topics with one or two named entities. For the first and second category, the rankings are more dissimilar. The ranking for the top ten systems in the classes usually differs more from the original ranking. This is due to the minor performance differences between top runs.

5 Outlook: Further Analysis of Topics

The promising results for named entities are encouraging for further analysis of topic features. We intend to explore the relationship between named entities and further languages. In addition, our study needs to be extended to include named entities and the corpus language as well as the frequency of occurrence of the named entity in the corpus. In this case, the occurrences to named entities need to be assessed in various languages because the number of named entities is sometimes different in the languages.

The recognition of named entities was carried out intellectually in our study. In a working retrieval engine applying our proposed fusion, this task needs to be delegated to software. therefore, the correlation between human and machine for named entity recognition needs to be examined.

It may also be useful to include the type of named entity. Maybe, some categories of named which lead to better retrieval performance than others.
We also plan to conduct a POS analysis of the topics and search for relationships to the system performance.

References

Allan, J. & Raghavan, H. (2002). Using part-of-speech Patterns to Reduce Query Ambiguity. In Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR ´02). Tampere, Finland, Aug. 11-15, 2002. ACM Press, pp. 307-314.

Braschler, M. (2004). CLEF 2002 - Overview of Results. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2004) Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2003, Trondheim. Berlin et al.: Springer [Lecture Notes in Computer Science] 2004, to appear. Preprint. http://www.clef-campaign.org.

Braschler, M., Gonzalo, J., Kluck, M. & Peters, C. (Eds.) (2003) Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2002, Rome. Berlin et al.: Springer [Lecture Notes in Computer Science 2785].

Braschler, M., Gonzalo, J., Kluck, M. & Peters, C. (Eds.) (2004) Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2003, Trondheim. Berlin et al.: Springer [Lecture Notes in Computer Science], to appear. Preprint http://www.clef-campaign.org.

Buckland, L. & Voorhees, E. (Eds.) (2003). The Eleventh Text Retrieval Conference (TREC 2002), NIST Special Publication: SP 500-251. http://trec.nist.gov/pubs/trec11/t11_proceedings.html.

Harman, D. & Voorhees, E. (1997). Overview of the Sixth Text REtrieval Conference. In D. Harman and E. Voorhees (Eds.). The Sixth Text REtrieval Conference (TREC-6). NIST Special Publication, National Institute of Standards and Technology, Gaithersburg, Maryland, 1997, http://trec.nist.gov/pubs/.

Kluck, M. & Womser-Hacker, C. (2002). Inside the Evaluation Process of the Cross-Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment. In C. P. S. Araujo & M. G. Rodríguez, (Eds.), Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain, May 29-31, 2002. Paris: ELRA, pp. 573-576.

Mandl, T. & Womser-Hacker, C. (2004). Proper Names in the Multilingual CLEF Topic Set. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.) (2004) Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2003, Trondheim. Berlin et al.: Springer [Lecture Notes in Computer Science] 2004. to appear. Preprint http://www.clef-campaign.org.

Womser-Hacker, C. (2002). Multilingual Topic Generation within the CLEF 2001 Experiments. In M. Braschler, J. Gonzalo, M. Kluck & C. Peters (Eds.). (2002) Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum (CLEF 2001), Darmstadt, Germany, September 3-4, 2001. Berlin et al.: Springer, Lecture Notes in Computer Science 2406, pp. 389-393.

Quality Gates: a New Device for Development and Evaluation in Cross-Language Information Retrieval ?

René Schneider

University of Hildesheim
Information Science
Marienburger Platz 22- D-31141 Hildesheim, Germany
rschneid @ uni-hildesheim.de

Abstract

This paper outlines and discusses the perspectives quality gates offer in cross-lingual information retrieval to ensure that the development process benefits from evaluation. Adequate evaluation in this context is possible through a combination of modular quality gates at the inch-pebble level, their linear connection in networks and re-organisation during different development cycles. As a consequence, the strict separation between development and evaluation disappears.

1 Introduction

Cross-lingual information retrieval (CLIR) has attracted increasing interest during the last decade not only through national and international evaluation initiatives such as TREC[1], CLEF[2] and NTCIR[3] but also through the rising amount of multilinguality in the world wide web and a stronger user interest on information from non-english and non-indoeuropean languages. Simultaneously, evaluation techniques have become “a must” for any research in natural language processing and development of applications in human language technology.

However, the situation is by no means satisfying. Evaluation in CLIR has only partly gone beyond the scope of precision, recall and f-measure or remains too modular for such complex systems where a multitude of resources, processes and methods merge in applications that differ considerably from each other. These differences emerge not only from the fundamental concepts resulting in different system architectures but also from the changes and adaptations of a single system during the different development cycles, keeping in mind that the “evaluation method is intimately connected to the software life cycle of the emerging technologies.” (Hirschman & Mani, 2003, p. 415)

This paper discusses the perspectives that quality gates offer for the area of cross-lingual information retrieval. After a brief overview of the different approaches used in evaluation in general and their application to CLIR, we will give an introduction to the concept of quality gates before outlining their concrete installation within the context of CLIR.

2 Evaluation in Cross-Lingual Information Retrieval

Evaluation in cross-lingual information retrieval offers a wide field of application due to the complexity and the large number of factors that have an implication for the retrieval results. For the sake of illustration we will name a few following the distinction of (Hirschman & Mani, 2003) and show their potential as well as their limitations in CLIR.

- *Gold standard based measures*: This method of defining, training and evaluating a gold standard has proved to be of great benefit in different fields, esp. for the task of named entity recognition. As recent work shows, high precision in named entity recognition has an equivalent impact on the overall quality of retrieval results. Nevertheless, the distribution, relevance and meaning of named entities varies considerably from genre to genre so that the definition of a generic gold standard is far from being found so far.

- *Feature based metrics* consist of checklists that record “critical features for different functional properties of components to be evaluated” (Hirschman & Mani, 2003, p. 418). Although they are usually set in opposition to corpus based methods, the gathering, annotation and evaluation of language corpora or any other data collection can easily be supported by this means to ensure representativeness and quality in test sets. As will be seen later (see section 3), feature based metrics are a basic element and starting point for the settlement of quality gates.

- *Embedded component evaluation* plays a vital role in the area of information retrieval. With the help of this evaluation device it is possible to track different scenarios of a system and to compare their corresponding results. Furthermore this strategy is strongly connected with the concept of user relevance feedback to either adapt a system according to the specific needs of a user or to measure general acceptance and usability respectively. Thus embedding component evaluation may collect experiences from the life cycle partly as a reaction to first results from evaluation metrics (such as the difference between the systems output and a gold standard) or from the interaction progress between the user and a system.

[1] trec.nist.gov

[2] www.clef-campaign.org

[3] www.ntcir.org

- Due to the fact that machine translation (MT) is a "conditio sine qua non" in CLIR the varying strategies that have been developed for *output evaluation* are of great importance for any evaluation of a cross-lingual retrieval engine. Translation occurs at different points of the system with either the queries and/or the retrieval results (or parts of them) being translated. Different translation strategies may be appropriate at different points of time and for different input.

- As already mentioned earlier, *user relevance feedback* is a very powerful means to discover the interactivity and usability a retrieval system has – and in case of being tracked appropriately – it will affect any further development of a retrieval system considerably. As a result it offers the combination with machine learning strategies to enable fast adaptation to the specific needs of a user. Unfortunately their implementation remains difficult so far.

As the enumeration listed above shows, CLIR offers a large number of interfaces towards evaluation, whose number and complexity changes with any further language that is integrated into an existing system. In some cases a method used so far may remain useful for a language pair (such as n-gram based methods that have proved successful for cross lingual retrieval of indo-european languages), but sometimes this method will lose its power and new components will have to be applied (e.g. in hamito-semitic languages with root-inflection n-gram based feature extraction becomes less valid). This will necessarily have an impact on the evaluation method used so far.

As a consequence the need for a framework arises that captures the dynamic complexity of CLIR in a synergetic system without being to complex but rather basic and feasible itself. The following section will propose a solution to this specific problem.

3 The Concept of Quality Gates

Quality gates had their origin in car manufacturing before being used metaphorically in quality assurance and project management. Generally, a Quality Gate (QG) is a checkpoint consisting of a set of predefined quality criteria that a project must meet in order to proceed from one stage of its life cycle to the next. Quality gates thus serve as amendments to milestones and deliverables which meet predefined quality benchmarks to

- support planning,
- improve status visibility,
- measure the current project status and
- control necessary changes or improvements.

Each quality gate is characterized by its own entry and exit criteria. A typical entry criteria is the completion and baseline of deliverables while an exit criteria can be the removal of the identified defects. By including metrics at every stage of the development process projects are monitored against their stated goals. Another important feature of quality gates is that they can be installed at any point during the life cycle of a project. The appropriate linking and enlargement of their simple structure allows project planning, control and measurement, whereas three different levels of complexity may be differentiated:

- *Binarity:* A very simple but extremely powerful realisation consists in "binary quality gates at the inch-pebble level" (Suzuki, 2003), where "binary" refers to meeting a requirement with no partial credit being given in order to avoid any variance between the planned and the actual performance and the "inch-pebble level" refers to a detailed tasks of short duration to prevent long term periods without control. As mentioned earlier, binary quality gates can be compared with feature based metrics.

- *Interconnectedness:* Growing complexity of a system leads to a connection of sequential or parallel quality gates resulting in a network with semaphores at the intersections to direct further activities and to highlighten the status of a system depending on the fulfilment or missing of a task. This idea corresponds strongly to that of "embedded components" as described earlier, whereas interconnections enable activation of a component or vice versa. Different results will be compared and transposed into appropriate conditionals for further use in equal settings.

- *Recursion:* Finally, since every project is far away from being terminated with a first yield, quality gates show a big part of their potential in keeping record of the whole project life cycle to prevent the repetition of failures. Monitoring of "lessons learned" from previous development cycles enables adaptive and re-active control mechanisms for succeeding activities, e.g. follow-ups, re-implementation or up-dates of a system. These gates – located at the end of a test suite or a life cycle – are used for output evaluation and become input for any refinement to occur.

As can be seen from this short introduction, quality gates have a local aspect (i.e. their distribution and interconnectedness over a system) as well as a temporal aspect (modification in form and content over time). Nevertheless, they are characterized through formal simplicity consisting in binary features in combination with semaphore logic. Thus they can easily be visualized to serve developers, project managers and users for the creation of different plug-and-play settings, the design of different test suites or the creation of user-specific search engines.

4 Integration and Transparency in CLIR

4. 1 Lessons Learned from Evaluation Campaigns

Similar to complex manufacturing processes or product development the release version of a retrieval system consists of many separate components, which may be developed at different times and are based on concurrent, sequential or recursive applications of several development patterns. This is esp. true in cross-lingual information retrieval, where the number of critical parameters and test suites is multiplied by the number of different languages a system is designed for and the implications that these languages have for retrieval strategies.

Thus the successful implementation of a retrieval system and the corresponding participation in an evaluation initiative (such as CLEF, TREC or NTCIR) depends considerably on a large number of quality criteria. Quality gates ensure that the project deliverables meet the criteria necessary to carry out subsequent project activities. In this context it should be noted explicitly, that quality gates are not considered for evaluation and comparison of several participants during an evaluation campaign, but have to be installed before and after the participation in an evaluation campaign.

The results that a system generates during participation will lead to many requests for changes: by developers as they realize something can be done in a different way, by comparison to strategies that other participants applied, etc. The collection and validation of these experiences will be discussed and transferred into appropriate alternations of the system. Sometimes these changes are small and a decision can easily be made whether to implement the change: but the changes to specification should be noted. Some requests may be kept open depending on the projects progress against timetable and/or some will be deferred as taking too long to implement. All of these specifications should be kept appropriately and probably be converted into quality criteria for further development circles.

The remaining question is then: How can we transfer our different experiences to objective quality criteria, that improve the development, testing and deployment of retrieval systems and avoid making the same mistakes again?

4. 2 First steps

Our vision is that of using quality gates as a concrete method not only for project management but also for development and implementation, i.e. that – after a first period of intellectual and manual specification - of using their potential for effective planning, control and measurement in CLIR. After the definition of desired quality criteria, different components of a retrieval system will be connected via a network of coupled quality gates to control system parameters, to steer information flow and to document learning effects. To illustrate this vision, we dedicate the following paragraph to a first outline of quality gates in CLIR according to the differentiation in section 3. Due to space limitations we will restrict ourselves to three examples already mentioned, namely data collections, fusion of strategies and user relevance feedback.

- CLIR is a heavily data-oriented approach. Consequently, results in CLIR depend to a big part on the data collections used for development and testing. Therefore, the quality of the corpora used for test suites and system development have to be described in terms of binary criteria concerning quantity, heterogeneity, data format, conversion, compression etc. A growing number of fulfilled criteria reflects growing validity of retrieval results and maturity of the system. Concrete realisation of this task might be achieved through simple templates that report on the adequacy of the data collection.

- Secondly, the overall system has to have knowledge concerning the components being used and coupled. To attain this, system components will be linked via gates that have information about the use and purpose of the specific components within a given context (e.g. the languages were n-gram based feature extraction has proved of great benefit) and allow steering and retaining of the information flow. Information flow and fusion of strategies might be controlled via conditionals and their correspondence to semaphores.

- Concurrent, sequential or recursive application of different system components will necessarily lead to different results. Combined with user relevance feedback (where users e.g. show their satisfaction by clicking on respective buttons to label documents as relevant or irrelevant) this information will serve as input to the whole system and has to be stored adequately and will lead to a reorganisation of the system and a change of the criteria and connections of lower-level quality gates.

5 Conclusion

The paper presents some reflections on the integration of quality gates into the process of developing and evaluation in cross-lingual information retrieval. This methodology is certainly not limited to this area, but promises to be helpful: on the one hand due to the high complexity of CLIR, on the other hand due to the fact that those systems – in the context of evaluation - initiatives have to be redesigned at least in a yearly interval.

While many of the techniques described can be found in the literature concerning evaluation, the ambition of the concept here is to bundle experiences and methodologies within a single framework (based on the metaphor of quality gates) to ensure adaptive and re-active project management.

References

Braschler, M., Harman, D., Hess, M., Kluck, M., Peters, C. & Schäuble, P.(2000). The Evaluation of Systems for Cross-Language Information Retrieval. In Proceedings of the Second Conference on Language Ressources and Evaluation (LREC-2000).

Carbonell, J., Yang, Y., Frederking, R., Brown, R. D., Geng, Y. & Lee, D. (1997). Translingual information retrieval: A comparative evaluation. In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence.

Charvat, J. How to use quality gates to guide IT projects. ZDNet Australia. Builder. http://web.zdnet.com.au/builder/manage/project/story/0,2000035082,20271712,00.

EAGLES. 1996. *EAGLES: evaluation of natural language processing systems.* Final Report EAGLES Document EAG-EWG-PR.2. http://issco.www.unige.ch/projects/ewg96/ewg96.html

Hirschman, L. et al. (1997). Evaluation. In Ron Cole et al. (Eds.), Survey of the State of the Art in Human Language Technology, Cambridge University Press and Giardini.

Hirschman, L. & Mani, I. (2003). Evaluation. In Mitkov, R. (Ed.), The Oxford Handbook of Computational Linguistics. Oxford University Press, pp. 414-429.

Kando, N. (2000). NTCIR-Workshop: an Evaluation of Cross-Lingual Information Retrieval. In Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, Aug. 30-Sept.1, 1999, Tokyo.

Quality Gates (2003). http://www.compulink.co.uk/~querrid/STANDARD/quality_gates.htm

Saracevic, T. (1995). Evaluation of Evaluation in Information Retrieval. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 9-13, 1995, pp. 138-146.

Suzuki, J. (2003). Best practices. Software Consulting. http://members.cox.net/johnsuzuki/best.htm.