GALATEAS

Generalised Analysis of Logs for Automatic Translation and Episodic Analysis of Searches: the challenge

Every day, millions of search queries are issued to content providers. These range from all-purpose web information sites (e.g. Google and Yahoo!) to digital library sites and merchant sites (e.g. Kelkoo and PriceGrabber). These queries are a precious resource in understanding user behaviour. From careful analysis of these queries, content providers can understand what information users are really looking for, the strategies they use to retrieve digital objects and the match between user needs and the content offered by the web site. Yet no provider has attempted to go beyond the standard log analysis service such as that provided Google analytics, to name but one. These services provide tools to segment user queries into words. They provide statistics on the occurrences of single words, but this is far from content providers needs:

- Existing analysis tools only consider words as chains of characters. This means that any generalization on the searches for "sea" and "ocean" is simply lost. Moreover, the intrinsic ambiguities in language are not resolved, so that a word such as "bank" will be ranked irrespective of its meaning.

- Current tools do not perform any match between user searches and the informative backbone of the content aggregation, be it a standard classification system, subject heading, product type, or plain list of categories. They do not provide hints on search 'episodes'. Each search is seen as an isolated event with no attempt at determining sequential patterns of search activities.

Besides being a precious resource for understanding user behaviour queries, as recorded in log files, could also become a key resource in cross-lingual access to information if we apply appropriate alignment algorithms. If one could dispose of a sufficiently large amount of pairs of queries which are translated equivalents, it would be relatively easy to derive a machine translation system especially adapted to deal with queries which could be used to provide a service for plain query translation. Such a service could be accessed by any kind of monolingual search engine to acquire cross-language functionalities.

Goal:

GALATEAS will create two web services:

LangLog will analyze transaction logs of queries to search engines for a given content provider. By applying statistical technologies coupled with language oriented services, it will produce reports concerning the informational needs of the users. In a way similar to that of standard log analysis systems whichprovide generalizations of paths of users inside a web site, LangLog will provide generalizations of the actions that information seekers perform in order to find content inside a searchable collection of digital objects.

QueryTrans will translate queries from an external search engine into several target languages. The external search engine will use these translations to return to the user, results in languages different from the one in which the query was formulated. QueryTrans will not be a cross-language information retrieval system, performing indexing and search, but just a *query translation service*.

These two web services are tightly linked and will access the same range of NLP based web services. LangLog is essential to allow the continuous acquisition of large quantities of queries in different languages. It is on the basis of such acquisition that the machine translation systems that constitute the backbone QueryTrans can be trained and thus provide the translation service.

GALATEAS partners:

GALATEAS
project timeframe:
1/04/2010
-
1/04/2013

Funded by the ICT Policy Support Programme (or ICT PSP)

Where will you find GALATEAS?

The project will be presented as an umber of scientific conferences in the fields of natural language processing and digital libraries.

Presence in major fair trades in the field of IT support for (digital) libraries.

CeBIT and SMAU 2012/2013

Scientific innovation:

Mainstream services in the field of web logs are click rate, visited pages and user paths inside the document tree. The GALATEAS services will analyse the information contained in queries from the point of view of language interpretation, not data.

Making sense of short queries and translating them into conceptual units will allow administrators and managers to answer questions such as: "What are the topics most commonly searched in my collection, in a certain language?"; "How do these topics relate to my catalogue?"; "Which named entities (people, places) are most popular among my users?".

In machine translation, GALATEAS will investigate technologies for statistical machine translation systems to provide meaningful results for short, syntactically and contextually poor texts such as queries to search engines.

The result:

Galateas will offer two types of services to content providers:

LangLog will extract meaning out of lists of queries for library, content aggregator and site managers.

QueryTrans has the ambitious and innovative goal to providing the first web translation service specially tailored to query translation.

Languages addressed by LangLog and QueryTrans are Italian, French, English, German, Dutch, Modern Arabic and Polish.

Impact:

Indirect "users" of the service are information seekers, who can benefit from improved, possibly cross lingual, search services. The GALATEAS services are not however provided directly to end users, but to administrators and managers of content aggregators and search engine installations.

GALATEAS will therefore target a high end B2B market where customers will be mostly represented by organizations running middle and large sized aggregated content. It will primarily address the following customer requirements:

- Need to understand what users are looking for, irrespective of the content they actually access.
- Need to understand how collections should be extended.
- Need to understand which categories in the catalogue fit more or less the desiderata of final users.
- Need to understand user behaviours
- Provide cross language information retrieval in a seamless way, without changing how documents are indexed and managed.

Contact:

Xerox Research Centre Europe

Frédérique Segond
6 chemin de Maupertuis

frederique.segond@xrce.xerox.com

http://www.galateas.eu