

COM368 Intelligent Information Systems

Assignment 1/1

Cross-lingual information retrieval in medical information systems for cancer research

University of Sunderland
Alexander Beisser
99338715S

Introduction to cross-lingual information retrieval

Cross-lingual or multilingual information retrieval (CLIR or MLIR) could be generally defined as information gathering in more than one language. Hull and Grefenstette [Hull 1996] providing five definitions for MLIR, which will be described more detailed in the chapter about the dictionary-based approach to MLIR:

1. Information retrieval (IR) in any language than English
2. IR on a parallel document collection or on a multilingual document collection where the search space is restricted to the query language
3. IR on a monolingual document, which can be queried in multiple languages
4. IR on a multilingual document collection, where queries can retrieve documents in multiple languages
5. IR on multilingual documents, i.e. more than one language can be present in the individual documents

This assignment will describe a dictionary-based and a thesaurus-based approach to MLIR and also a discussion about a self-developed MLIR tool for translating and expanding queries from English to German and vice versa. A chapter will address the problems of multilingual information retrieval.

CLIR and MLIR deal with the same subject. Nevertheless cross-lingual information retrieval is the more common used expression so that the author of this paper decided to use CLIR as the appropriated abbreviation.

Querying across languages: A dictionary-based approach to CLIR

David Hull and Gregory Grefenstette [Hull 1996], both from Rank Xerox Research Center in France, examined a dictionary-based approach to CLIR. They defined CLIR in five different ways, as described in the introduction to this paper.

There research concentrates on definition 3, which assumes that the document collection is monolingual, but the retrieval system can process queries in different languages, which enable document retrieval across language boundaries. Hull and Grefenstette give following reasons for their approach selection:

- To obtain reliable and quantifiable experimental results, a document collection with large numbers of relevance judgements
- The researchers want to address the simpler problem at the beginning, before they generalise their results

The problems of CLIR can be addressed by different approaches. Hull and Grefenstette uses a part of the TIPSTER text collection¹ and the queries 51 – 100 from the third TREC conference in 1995. Before using these queries in their experiments the two researchers had undertaken some pre-testing of the available queries in which they found that the queries could not be used because of their length and content. To address these problems they decided to use shorter queries with an average dimension of seven words. In addition to these the online bilingual French => English Oxford Hachette dictionary of 1994 was utilized, which has to be filtered to reduce the amount of noisy data. Unfortunately it was not possible to filter the dictionary automatically, which left over some disambiguate.

The three steps of the Xerox CLIR approach

The query will be morphologically analysed and its flectional root will replace each term. After this process the flectional root will be translated by a lookup in the bilingual dictionary. In the case in which a term could not be found in the dictionary it will be passed unchanged to the final query. The final query will be sent to the monolingual SMART information retrieval system, where the query is used to gather documents from the document collection. The retrieved documents are ranked in descending order.

Hull and Grefenstette recommend using relevance feedback to improve the quality of the query and to gain a high precision instead of a fast processed set of documents with a high amount of noisy and unusable data. Quality is more important than speed.

The experimental results: The author of this paper spends only a short moment on the optimal performance benchmark for transfer dictionary-based translation.

Hull and Grefenstette found following average precision at 5, 10, 15 and 20 documents retrieved for the original English queries, which are, translate by the dictionary.

¹ Only the news component, which includes article from Wall Street Journal, AP newswire and the San Jose Mercury

Original English	Automatic word-based transfer dictionary	Manual word-based transfer dictionary	Manual multi-word transfer dictionary
0.393	0.235	0.269	0.357

These results show that French queries created with the automatically created translation dictionary have the worst precision. Whereas on the other hand the manual multi-word transfer dictionary produces French queries with a precision nearly as precise as the original English ones.

Future extensions

The more precise and accurate filtering of the used dictionary is one of the most important extensions for the Xerox CLIR approach stated by Hull and Grefenstette. The inclusion of term vectors, in which IBM invested some research effort, could also increase the precision. On the other hand the implementation of a weighted Boolean model might be more useful. The two researchers also mentioned the use of cross-language relationships to help with disambiguation for future system extension. Dagan et al describe this approach in their research article "To languages are more informative than one" in 1991.

Using the SPIDER system for CLIR

Mr Sheridan and Mr Ballerini [Sheridan 1996] from the Swiss Federal Institute of Technology (ETH) in Zurich introduced a new approach to CLIR that is based on query expansion using a thesaurus over a collection of multilingual documents. This approach was implemented into the SPIDER system and tested over a set of Italian documents. The aim of their project was to develop an IR system, which retrieves relevant Italian documents with a German query.

The two Swiss researchers used the findings of a previous research in query expansion for CLIR undertaken by Schaeuble et al. More information about this and other approaches to multilingual retrieval systems can be found in the chapter "Other experiments in CLIR".

Sheridan and Ballerini used a set of 93,229 documents written in German and Italian from the Schweizerische Depeschen Agentur (SDA). Each document has an average length of 112 words. Each document has a description header, which is language independent from a controlled document. This information was used during the process of document alignment and has a size

of 3 to 4 descriptors. This metadata helped to cluster the document selection. The relevance judgement was done by a native Italian to ensure accuracy.

For the reason that the experiments of the CLIR at the ETH relies on query expansion across a given set of multilingual documents correct and careful document segmentation needed to be performed. The documents were cluster by date and also by the previous mentioned descriptors. At the end of this alignment process a collection of 10,293 bi-lingual documents was used for building the similarity thesaurus, which was used for the query expansion throughout the experiments.

Before constructing the similarity thesaurus word normalisation has to be performed. This process decreases the ambiguity and increases the precision during the expansion process. Because of the different nature of Italian and German, two different procedures were used. For the Italian word normalisation Sheridan and Ballerini developed a rule-based stemming algorithm, which is similar to the Porter stemming algorithm. The algorithm contained 220 normalisation rules in the same notation as Porter's one. The researchers discovered that by using their stemming algorithm the performance of their CLIR system increased by a factor between 57% and 130% better than without stemming.

The German word normalisation demanded a different approach, because of the German word and especially noun construction (most German nouns are build by connecting two or more nouns together) the researcher found out that a stemming algorithm like Porter's or the Italian would fail. They used the German lexicon of CELEX instead to normalise the German documents of the collection by splitting the compound nouns into constituents in case when the compound one was not found in the dictionary. This process brought some problems with itself. Washington for example was divided into was-hing-ton, which destroyed the meaning of the word completely. For this reason some additional measures needed to be considered to protect names etc.

The construction of the similarity thesauri took place after the word normalisation process, which can be found described detailed in [Sheridan 1996].

The ETH CLIR system performed 6 tasks using the SPIDER IR system and ran on different servers, because of the client server architecture of SPIDER:

1. A German query was read

2. The query was submitted to the server with the multilingual documents for query expansion
3. The expanded query with similar Italian and German terms was received back
4. The queries was now filtered for a specified amount of Italian terms
5. The filtered query was than sent to the server with the Italian documents for the purpose of query evaluation
6. A ranked list of Italian documents was received back

The results of the systems performance by using above described systems can be described as following.

German queries on Italian documents performed:

Total relevant documents: 1418		
Query length	# Retrieved	Average precision
10	525	0.212
25	649	0.278
50	638	0.275

These results show that a query length of 25 terms delivered the best precision.

Italian queries on Italian documents performed:

Total relevant documents: 1418		
Description	# Retrieved	Average precision
No stemming	488	0.231
SPIDER stemming	898	0.527

These results show that using the developed SPIDER stemming increases the precision by more than 120 %.

Sheridan and Ballerini implemented the functionality of relevance feedback into their system, which increased the performance by 29% over the automated multilingual retrieval without feedback.

In their discussion the two researchers described their document collection as an ideal experiment environment. They could not find their discovered performance results in the "real" world. Their approach heavily relies on the quality of the available document selection. The

results may also change for different languages. Future research has to be undertaken to verify the discovered result about language borders.

Other experiments in CLIR

Cross-lingual Information Retrieval is not a research subject, which came up just a few years ago. The roots of bi- or multilingual information processing going back to the 1970's where Professor Gerhard Salton from the Cornell University started his experiments in multilingual information retrieval. He used a manual constructed multilingual thesaurus to assign terms into categories [Salton 1970].

Another approach to CLIR is the use of different corpuses to develop thesaurus-based structures, which could be used for query expansion. Schaeuble and Knaus started their research in this subject in 1992 [Schaeuble and Knaus 1992]. Qui and Frei [Qui and Frei 1993] as well as Jing and Croft [Jing and Croft 1994] spent also some effort in this technique. Han et al. [Han et al. 1994] concerned himself with automatic query expansion techniques for Japanese text retrieval in 1994 at the University of Massachusetts.

Mr M. F. Porter [Porter 1980] developed in 1980 his Porter stemming algorithm for suffix stripping, which had become a high impact in modern information retrieval systems. Sheridan and Ballerini [Sheridan 1996] using this stemming algorithm as a resource for developing their own Italian stemming roles.

Oard [Oard et al. 1994] spent in the early 1990's some research effort in addressing the problem of interlingual term correspondence in text translation, term vector translation and Latent Semantic Coindexing.

CLIR uses some research findings from the subject of Machine Translation, in which for example Radwan [Radwan 1994] spent work on it. Different researches and companies like IBM researched in this area, so that Radwan should be understood as one of many other developers.

Another resource for CLIR is examined since the early 1990's. Brown [Brown et al 1993] and other researchers from IBM undertook several examinations in this subject, which is nowadays heavily used in CLIR systems like in the one from Sheridan and Ballerini.

As Sheridan and Ballerini 1996 used this system, the SPIDER system should be also mentioned here. The SPIDER system provides efficient weighted retrieval on dynamic data collections and it is based on signatures and non-inverted item description [Schaeuble 1993]. More information about the SPIDER system and its structure can be found in the reference [Schaeuble 1993].

Davis [Davis w.y.] described another CLIR system for Spanish IR with English query called Recuerdo. This system uses UN parallel corpus and the Collins bi-lingual English-Spanish dictionary combined with fuzzy matching for term translation processing.

CANCIR: CLIR in medical information systems for cancer research

The developed prototype for a cross-lingual information retrieval approach will be explained in this chapter. The developed CLIR system called CANCIR should be understood as a prototype for CLIR in the subject of cancer research.


The medical and also the pharmaceutical research are very expensive. For this reason different universities and research laboratories from all over the world try to work in a research network to minimise the costs and maximise the outcome of their research work. Obviously as soon as you work in a multinational and –lingual environment, language problems arise. The presence of multilingual documents can be identified as one example for these challenges. Therefore an IR system that can retrieve information across language boundaries should be implemented into a multinational researching network.

The prototype of CANCIR addresses this demand. It is programmed in Perl, which is a powerful and the ideal script language for information extraction. As Perl is platform independent, CANCIR runs on UNIX (Linux), Windows and also on Macintosh computers. The system is tested on Windows 2000 and Windows NT. In the School of Computing, Engineering and Technology (SCET) it runs in cell M7 of the David Goldman Informatics Center. It has a dictionary approach to CLIR, so it looks the entered terms up in a dictionary to find similar terms (in the thesaurus), which then will be added to the query. After the query expansion process, it queries two different search engines:

- The online search engine from the German Cancer Research Center and the

- Medline online search engine with the U.S. National Library of Medicine

CANCIR runs in a DOS command window, which makes it faster and more reliable. Windows NT and 2000 separates the 16bits from the 32bits processes, so that the system is less likely to crash. Following screenshot shows the opening screen of the system.



```
Command Prompt - offline2.pl
*****
**          Cross-lingual information retrieval in          **
**  medical information systems for cancer research  **
**          Welcome to CANC-IR the information retrieval    **
**                    tool for cancer research            **
**          *****                                       **
Please input Search term. Enter 0 to stop:
```

Figure 1: Opening screen

In the opening screen the user will be asked to enter a search term. This procedure will be repeated until a "0" is entered.

CANCIR looks now whether it can find the entered word in the prototype's dictionary or not. If the term was found it would be added automatically to the final query and the similar words from the thesaurus will expand the query.

In case that the term was not found in the dictionary following screen will appear.

```

C:\> Command Prompt - offline2.pl
**          tool for cancer research          **
**                                          **
*****
Please input Search term. Enter 0 to stop: cancer
Please input Search term: Enter 0 to stop: drugs
Please input Search term: Enter 0 to stop: perl
Please input Search term: Enter 0 to stop: 0

I am extracting the searchterms.

I am now translating your terms you searched for.
Processing term: cancer
Processing term: drugs
Processing term: perl

Your searchterm perl was not found in the dictionary.
Do you want to add perl to your search? [input yes or no]:

```

Figure 2: User dialog, if term is not in the dictionary

The user can now decide whether the term should be added to the query or not (by enter “yes” or “no”).

```

C:\> Command Prompt - offline2.pl
*****
**          Cross-lingual information retrieval in medical          **
**          information systems for cancer research          **
**                                          **
*****
** Your translated query is: **
** Krebs OR Tumor OR Melanom OR Karzinom OR Neoplasia OR cancer OR Medikamente OR **
** R Allopurinol OR Azaserie OR Mercaptopurine OR Ifosfamid OR Cytarabine OR drugs **
**                                          **
*****

```

Figure 3: The expanded and translated query

The final query will be generated after processing the last enter term and then be displayed to the user, who can then decide to query the available search engines. The query results are stored in the file called “results.html” in the folder “H:\netware\com368”.

Another version of CANCEIR has a graphical user interface (GUI), which works on a similar way as the command-line CANCEIR. This version has the disadvantage of the missing function of querying the two search engines, which is caused by the firewall set-up for the UNIX servers at the SCET.

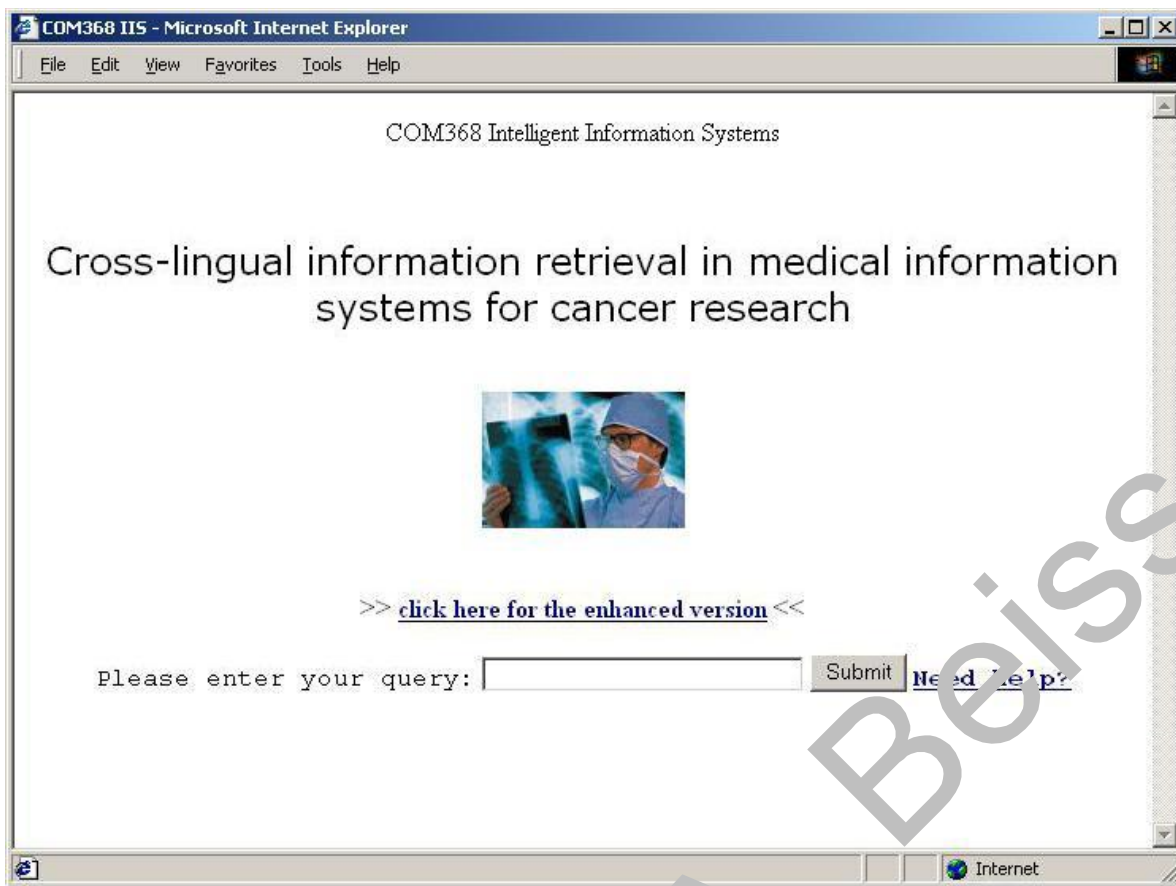


Figure 4: The graphical user interface of CANCEIR

As CANCEIR is a prototype system, it has some other disadvantages or missing functions. These are described in the chapter “Conclusion and discussion”.

CANCEIR is aimed as an input device for a CLIR, which uses thesaurus-based query expansion. It was not intended to develop a fully functional medical CLIR system. It should illustrate the idea of a multilingual query expansion system, where English queries are translated into German ones and vice versa.

The problems with CLIR

CLIR has not reached a status where a system could be declared as perfect. There is still a need for improvement. Gregory Grefenstette [Grefenstette w.y.] described in his article “The problems of cross-lingual information retrieval three main problems, which arose more rapidly since the success of the World Wide Web in the late 1990’s:

1. The first problem is the need of knowledge how a term might be written in different language.
2. The second problem is the selection of the accurate translation of the term.

3. The third problem of CLIR is the properly weighting of the translated alternatives

In CLIR the system has still the possibility to find relevant documents even when the wrong translation is used unlikely to machine translation techniques, where the accurate translation is the primary requirement. This fact illustrates the problem of term and translation weighting, which needs be solved.

Grefenstette described different approaches to find translations of a search terms. Those four solutions will be briefly described in this chapter. More detailed information can be found in the reference [Grefenstette w.y.]. The first discussed key for finding accurate translations are the use of dictionaries, which is the easiest way to get acceptable results. A couple of important dictionaries are available online or in form of databases, which can be implemented easier into a CLIR system. Filtering especially of large dictionaries is need. Hull and Grefenstette described one dictionary-based approach in their article [Hull 1996]. Word normalisation processes like stemming might need to be also performed [Sheridan 1996].

The second key to successful term translation in accordance to Grefenstette is the use of parallel corpora. It is possible to get bi-lingual term equivalents with a document selection of the proper size. The SPIDER system is one of IR system that can use corpuses to perform translation, which is described in the articles of Schaeuble [Schaeuble 1993] and of Sheridan and Ballerini [Sheridan 1996].

Grefenstette describes pruning of translation alternatives as another key factor and referred to the EMIR system from Fluhr. In this system a database of known compounds from the target language is used to filter out translations from compound words translated word-to-word [unknown reference].

As the last key factor Grefenstette mentioned the weighting process of alternatives, which was also investigated by Salton and Buckley in 1988 [Salton and Buckley 1988]. Their approach consists of an indexing method and a retrieval function, where the indexing method determines the descriptions of the stored data items and of the user given query items. The retrieval function matches the descriptions of the query items against the descriptions of the stored data items. More about this method can be found in the reference [Schaeuble 1993]

Conclusion and discussion

CLIR is a discipline where machine translation and information retrieval come together [Grefenstette w.y.].

It was difficult to implement a CLIR system into the field of cancer research. Medical terms are still mostly written in Latin, so that it difficult to find English or German translations. The CANSEARCH expert system [POLLITT 1987] helped to find some usable translations.

As one of finding of this assignment it is recommended to use professional expert online dictionary for the translation process. A stemming algorithm would increase the precision significantly, but the scope of the given task was the implementation of a CLIR prototype system. Therefore the author of the assignment decided to create a short dictionary with five different terms manually.

English	German
Drugs	Medikamente
Therapy	Therapie
Types	Arten
Cancer	Krebs
Systems (Organ systems)	Systeme (Organsysteme)

Table1: The word translation matrix.

Each of these terms had some similar terms stored in an array.

The implementation of a graphical user interface started to be another challenge. The Internet server here at the University of Sunderland is a UNIX server with its own firewall, so that the querying of search engines could not be performed by a graphical user interface. To overcome this problem a command-line oriented system was developed and deployed for cell M7. An example of a graphical CLIR that uses query expansion is available over the Internet under following URL: <http://osiris.sunderland.ac.uk/~ca9abe/com368/index.htm>

The last problem, which the author was faced wit, was to find on medicine specialist search engines. Free available engines do not fulfil the requirements for a CLIR. The used search engines should be seen for that reason as an example, of what was intended to perform with the

translated and expanded query. Specialist search engine for the World Wide Web are only accessible for the institutes own researches. The author had not the possibility to integrate the planned machines, which should be taken into account in rating and marking the given prototype.

As mentioned before CANCEIR should demonstrate the possibility to translate and expand a multilingual query.

The implementation of an online dictionary with stemming algorithms and also adding specialist search engines would improve the prototype. A graphical user interface would help the user to work better with the system than in the present version.

Appendix:

- System overview
- Flowchart
- Pseudo code
- Code of CANCEIR
- User manual
- References

COM368 Intelligent Information Systems

Assignment 1/1

Cross-lingual information retrieval in medical information systems for cancer research

University of Sunderland
Alexander Beisser
99338715S

Introduction to cross-lingual information retrieval

Cross-lingual or multilingual information retrieval (CLIR or MLIR) could be generally defined as information gathering in more than one language. Hull and Grefenstette [Hull 1996] providing five definitions for MLIR, which will be described more detailed in the chapter about the dictionary-based approach to MLIR:

1. Information retrieval (IR) in any language than English
2. IR on a parallel document collection or on a multilingual document collection where the search space is restricted to the query language
3. IR on a monolingual document, which can be queried in multiple languages
4. IR on a multilingual document collection, where queries can retrieve documents in multiple languages
5. IR on multilingual documents, i.e. more than one language can be present in the individual documents

This assignment will describe a dictionary-based and a thesaurus-based approach to MLIR and also a discussion about a self-developed MLIR tool for translating and expanding queries from English to German and vice versa. A chapter will address the problems of multilingual information retrieval.

CLIR and MLIR deal with the same subject. Nevertheless cross-lingual information retrieval is the more common used expression so that the author of this paper decided to use CLIR as the appropriated abbreviation.

Querying across languages: A dictionary-based approach to CLIR

David Hull and Gregory Grefenstette [Hull 1996], both from Rank Xerox Research Center in France, examined a dictionary-based approach to CLIR. They defined CLIR in five different ways, as described in the introduction to this paper.

There research concentrates on definition 3, which assumes that the document collection is monolingual, but the retrieval system can process queries in different languages, which enable document retrieval across language boundaries. Hull and Grefenstette give following reasons for their approach selection:

- To obtain reliable and quantifiable experimental results, a document collection with large numbers of relevance judgements
- The researchers want to address the simple problem at the beginning, before they generalise their results

The problems of CLIR can be addressed by different approaches. Hull and Grefenstette use a part of the TIPSTER text collection¹ and the queries 51 – 100 from the third TREC conference in 1995. Before using these queries in their experiments the two researchers had undertaken some pre-testing of the available queries in which they found that the queries could not be used because of their length and content. To address these problems they decided to use shorter queries with an average dimension of seven words. In addition to these the online bilingual French => English Oxford Hachette dictionary of 1994 was utilized, which has to be filtered to reduce the amount of noisy data. Unfortunately it was not possible to filter the dictionary automatically, which left over some disambiguate.

The three steps of the Xerox CLIR approach

The query will be morphologically analysed and its flectional root will replace each term. After this process the flectional root will be translated by a lookup in the bilingual dictionary. In the case in which a term could not be found in the dictionary it will be passed unchanged to the final query. The final query will be sent to the monolingual SMART information retrieval system, where the query is used to gather documents from the document collection. The retrieved documents are ranked in descending order.

Hull and Grefenstette recommend using relevance feedback to improve the quality of the query and to gain a high precision instead of a fast processed set of documents with a high amount of noisy and unusable data. Quality is more important than speed.

The experimental results: The author of this paper spends only a short moment on the optimal performance benchmark for transfer dictionary-based translation.

Hull and Grefenstette found following average precision at 5, 10, 15 and 20 documents retrieved for the original English queries, which are, translate by the dictionary.

¹ Only the news component, which includes article from Wall Street Journal, AP newswire and the San Jose Mercury

Original English	Automatic word-based transfer dictionary	Manual word-based transfer dictionary	Manual multi-word transfer dictionary
0.393	0.235	0.269	0.357

These results show that French queries created with the automatically created translation dictionary have the worst precision. Whereas on the other hand the manual multi-word transfer dictionary produces French queries with a precision nearly as precise as the original English ones.

Future extensions

The more precise and accurate filtering of the used dictionary is one of the most important extensions for the Xerox CLIR approach stated by Hull and Grefenstette. The inclusion of term vectors, in which IBM invested some research effort, could also increase the precision. On the other hand the implementation of a weighted Boolean model might be more useful. The two researchers also mentioned the use of cross-language relationships to help with disambiguation for future system extension. Dagan et al describe this approach in their research article "To languages are more informative than one" in 1991.

Using the SPIDER system for CLIR

Mr Sheridan and Mr Ballerini [Sheridan 1996] from the Swiss Federal Institute of Technology (ETH) in Zurich introduced a new approach to CLIR that is based on query expansion using a thesaurus over a collection of multilingual documents. This approach was implemented into the SPIDER system and tested over a set of Italian documents. The aim of their project was to develop an IR system, which retrieves relevant Italian documents with a German query.

The two Swiss researchers used the findings of a previous research in query expansion for CLIR undertaken by Schaeuble et al. More information about this and other approaches to multilingual retrieval systems can be found in the chapter "Other experiments in CLIR".

Sheridan and Ballerini used a set of 93,229 documents written in German and Italian from the Schweizerische Depeschen Agentur (SDA). Each document has an average length of 112 words. Each document has a description header, which is language independent from a controlled document. This information was used during the process of document alignment and has a size

of 3 to 4 descriptors. This metadata helped to cluster the document selection. The relevance judgement was done by a native Italian to ensure accuracy.

For the reason that the experiments of the CLIR at the ETH relies on query expansion across a given set of multilingual documents correct and careful document segmentation needed to be performed. The documents were cluster by date and also by the previous mentioned descriptors. At the end of this alignment process a collection of 10,293 bi-lingual documents was used for building the similarity thesaurus, which was used for the query expansion throughout the experiments.

Before constructing the similarity thesaurus word normalisation has to be performed. This process decreases the ambiguity and increases the precision during the expansion process. Because of the different nature of Italian and German, two different procedures were used. For the Italian word normalisation Sheridan and Ballerini developed a rule-based stemming algorithm, which is similar to the Porter stemming algorithm. The algorithm contained 220 normalisation rules in the same notation as Porter's one. The researchers discovered that by using their stemming algorithm the performance of their CLIR system increased by a factor between 57% and 130% better than without stemming.

The German word normalisation demanded a different approach, because of the German word and especially noun construction (most German nouns are build by connecting two or more nouns together) the researcher found out that a stemming algorithm like Porter's or the Italian would fail. They used the German lexicon of CELEX instead to normalise the German documents of the collection by splitting the compound nouns into constituents in case when the compound one was not found in the dictionary. This process brought some problems with itself. Washington for example was divided into was-hing-ton, which destroyed the meaning of the word completely. For this reason some additional measures needed to be considered to protect names etc.

The construction of the similarity thesauri took place after the word normalisation process, which can be found described detailed in [Sheridan 1996].

The ETH CLIR system performed 6 tasks using the SPIDER IR system and ran on different servers, because of the client server architecture of SPIDER:

1. A German query was read

2. The query was submitted to the server with the multilingual documents for query expansion
3. The expanded query with similar Italian and German terms was received back
4. The queries was now filtered for a specified amount of Italian terms
5. The filtered query was than sent to the server with the Italian documents for the purpose of query evaluation
6. A ranked list of Italian documents was received back

The results of the systems performance by using above described systems can be described as following.

German queries on Italian documents performed:

Total relevant documents: 1418		
Query length	# Retrieved	Average precision
10	525	0.212
25	649	0.278
50	638	0.275

These results show that a query length of 25 terms delivered the best precision.

Italian queries on Italian documents performed:

Total relevant documents: 1418		
Description	# Retrieved	Average precision
No stemming	488	0.231
SPIDER stemming	898	0.527

These results show that using the developed SPIDER stemming increases the precision by more than 120 %.

Sheridan and Ballerini implemented the functionality of relevance feedback into their system, which increased the performance by 29% over the automated multilingual retrieval without feedback.

In their discussion the two researchers described their document collection as an ideal experiment environment. They could not find their discovered performance results in the "real" world. Their approach heavily relies on the quality of the available document selection. The

results may also change for different languages. Future research has to be undertaken to verify the discovered result about language borders.

Other experiments in CLIR

Cross-lingual Information Retrieval is not a research subject, which came up just a few years ago. The roots of bi- or multilingual information processing going back to the 1970's where Professor Gerhard Salton from the Cornell University started his experiments in multilingual information retrieval. He used a manual constructed multilingual thesaurus to assign terms into categories [Salton 1970].

Another approach to CLIR is the use of different corpuses to develop thesaurus-based structures, which could be used for query expansion. Schaeuble and Knaus started their research in this subject in 1992 [Schaeuble and Knaus 1992]. Qui and Frei [Qui and Frei 1993] as well as Jing and Croft [Jing and Croft 1994] spent also some effort in this technique. Han et al. [Han et al. 1994] concerned himself with automatic query expansion techniques for Japanese text retrieval in 1994 at the University of Massachusetts.

Mr M. F. Porter [Porter 1980] developed in 1980 his Porter stemming algorithm for suffix stripping, which had become a high impact in modern information retrieval systems. Sheridan and Ballerini [Sheridan 1996] using this stemming algorithm as a resource for developing their own Italian stemming roles.

Oard [Oard et al. 1994] spent in the early 1990's some research effort in addressing the problem of interlingual term correspondence in text translation, term vector translation and Latent Semantic Coindexing.

CLIR uses some research findings from the subject of Machine Translation, in which for example Radwan [Radwan 1994] spent work on it. Different researches and companies like IBM researched in this area, so that Radwan should be understood as one of many other developers.

Another resource for CLIR is examined since the early 1990's. Brown [Brown et al 1993] and other researchers from IBM undertook several examinations in this subject, which is nowadays heavily used in CLIR systems like in the one from Sheridan and Ballerini.

As Sheridan and Ballerini 1996 used this system, the SPIDER system should be also mentioned here. The SPIDER system provides efficient weighted retrieval on dynamic data collections and it is based on signatures and non-inverted item description [Schaeuble 1993]. More information about the SPIDER system and its structure can be found in the reference [Schaeuble 1993].

Davis [Davis w.y.] described another CLIR system for Spanish IR with English query called Recuerdo. This system uses UN parallel corpus and the Collins bi-lingual English-Spanish dictionary combined with fuzzy matching for term translation processing.

CANCIR: CLIR in medical information systems for cancer research

The developed prototype for a cross-lingual information retrieval approach will be explained in this chapter. The developed CLIR system called CANCIR should be understood as a prototype for CLIR in the subject of cancer research.

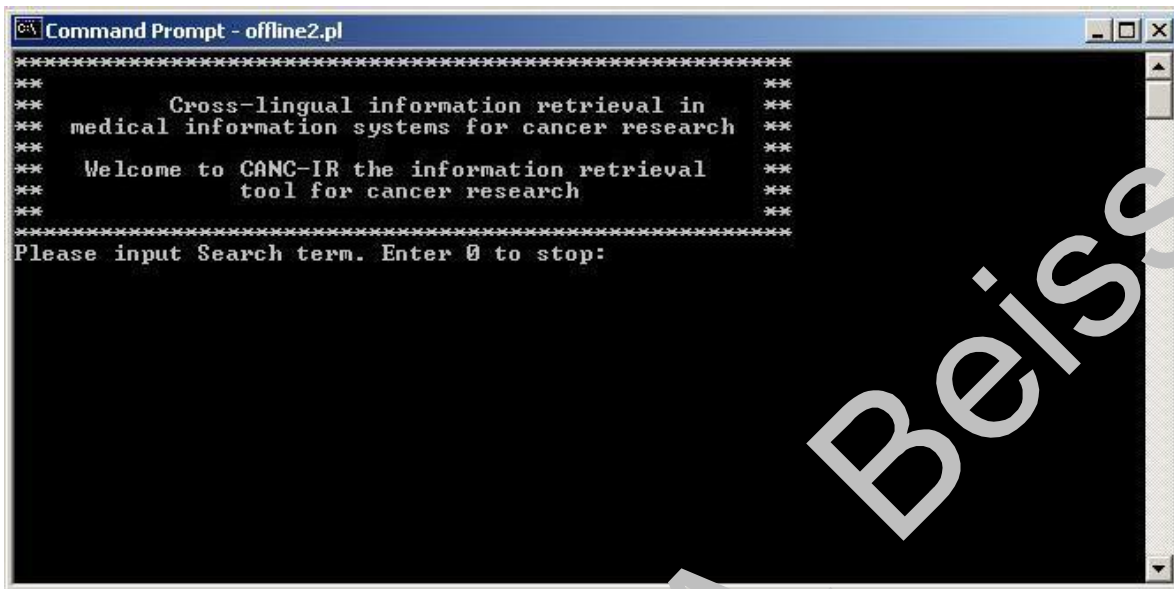
The medical and also the pharmaceutical research are very expensive. For this reason different universities and research laboratories from all over the world try to work in a research network to minimise the costs and maximise the outcome of their research work. Obviously as soon as you work in a multinational and –lingual environment, language problems arise. The presence of multilingual documents can be identified as one example for these challenges. Therefore an IR system that can retrieve information across language boundaries should be implemented into a multinational researching network.

The prototype of CANCIR addresses this demand. It is programmed in Perl, which is a powerful and the ideal script language for information extraction. As Perl is platform independent, CANCIR runs on UNIX (Linux), Windows and also on Macintosh computers. The system is tested on Windows 2000 and Windows NT. In the School of Computing, Engineering and Technology (SCET) it runs in cell M7 of the David Goldman Informatics Center. It has a dictionary approach to CLIR, so it looks the entered terms up in a dictionary to find similar terms (in the thesaurus), which then will be added to the query. After the query expansion process, it queries two different search engines:

- The online search engine from the German Cancer Research Center and the

- Medline online search engine with the U.S. National Library of Medicine

CANCIR runs in a DOS command window, which makes it faster and more reliable. Windows NT and 2000 separates the 16bits from the 32bits processes, so that the system is less likely to crash. Following screenshot shows the opening screen of the system.



```
Command Prompt - offline2.pl
*****
**          Cross-lingual information retrieval in          **
**  medical information systems for cancer research  **
**          Welcome to CANC-IR the information retrieval    **
**                   tool for cancer research            **
**          *****          **
Please input Search term. Enter 0 to stop:
```

Figure 1: Opening screen

In the opening screen the user will be asked to enter a search term. This procedure will be repeated until a "0" is entered.

CANCIR looks now whether it can find the entered word in the prototype's dictionary or not. If the term was found it would be added automatically to the final query and the similar words from the thesaurus will expand the query.

In case that the term was not found in the dictionary following screen will appear.

```

C:\> Command Prompt - offline2.pl
**          tool for cancer research          **
**                                          **
*****
Please input Search term. Enter 0 to stop: cancer
Please input Search term: Enter 0 to stop: drugs
Please input Search term: Enter 0 to stop: perl
Please input Search term: Enter 0 to stop: 0

I am extracting the searchterms.

I am now translating your terms you searched for.
Processing term: cancer
Processing term: drugs
Processing term: perl

Your searchterm perl was not found in the dictionary.
Do you want to add perl to your search? [input yes or no]:

```

Figure 2: User dialog, if term is not in the dictionary

The user can now decide whether the term should be added to the query or not (by enter “yes” or “no”).

```

C:\> Command Prompt - offline2.pl
*****
**          Cross-lingual information retrieval in medical          **
**          information systems for cancer research          **
**                                          **
*****
** Your translated query is: **
** Krebs OR Tumor OR Melanom OR Karzinom OR Neoplasia OR cancer OR Medikamente OR **
** R Allopurinol OR Azaserie OR Mercaptopurine OR Ifosfamid OR Cytarabine OR drugs **
**                                          **
*****

```

Figure 3: The expanded and translated query

The final query will be generated after processing the last enter term and then be displayed to the user, who can then decide to query the available search engines. The query results are stored in the file called “results.html” in the folder “H:\netware\com368”.

Another version of CANCIR has a graphical user interface (GUI), which works on a similar way as the command-line CANCIR. This version has the disadvantage of the missing function of querying the two search engines, which is caused by the firewall set-up for the UNIX servers at the SCET.

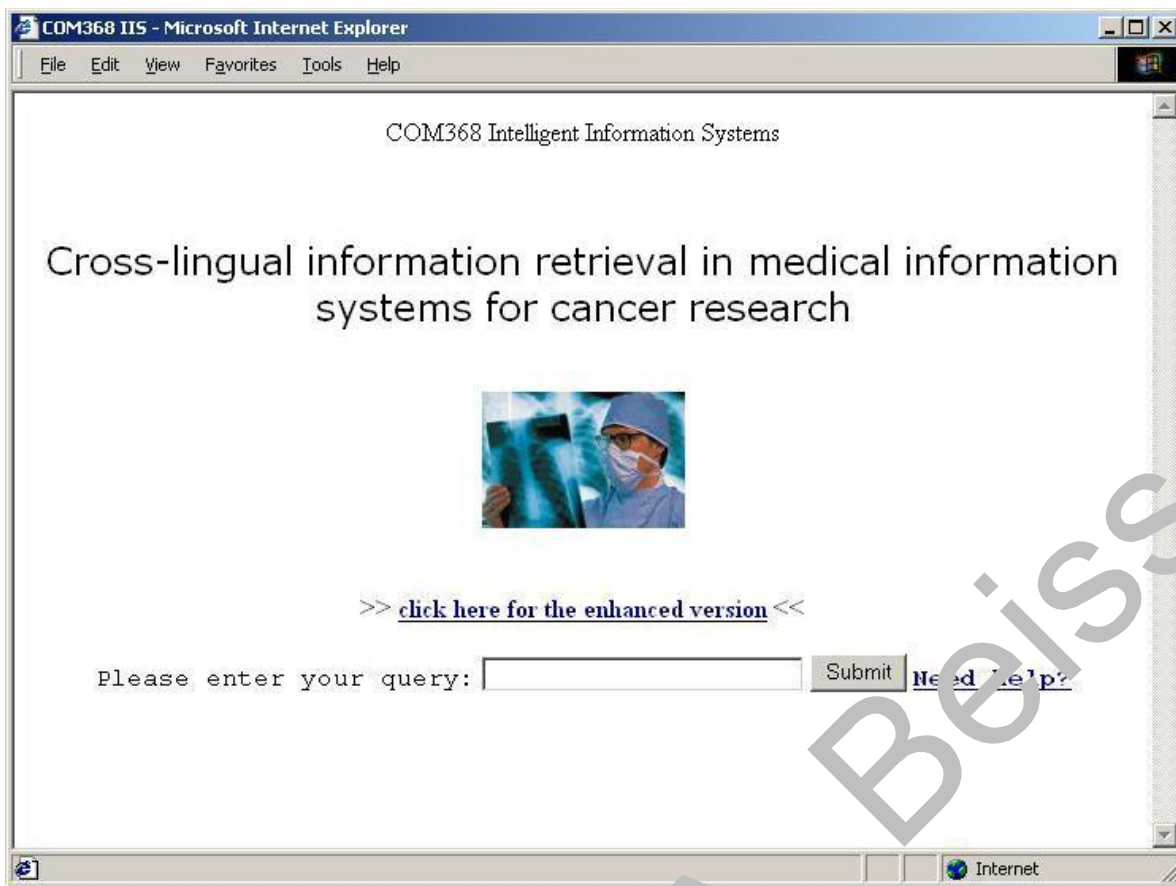


Figure 4: The graphical user interface of CANCEIR

As CANCEIR is a prototype system, it has some other disadvantages or missing functions. These are described in the chapter “Conclusion and discussion”.

CANCEIR is aimed as an input device for a CLIR, which uses thesaurus-based query expansion. It was not intended to develop a fully functional medical CLIR system. It should illustrate the idea of a multilingual query expansion system, where English queries are translated into German ones and vice versa.

The problems with CLIR

CLIR has not reached a status where a system could be declared as perfect. There is still a need for improvement. Gregory Grefenstette [Grefenstette w.y.] described in his article “The problems of cross-lingual information retrieval three main problems, which arose more rapidly since the success of the World Wide Web in the late 1990’s:

1. The first problem is the need of knowledge how a term might be written in different language.
2. The second problem is the selection of the accurate translation of the term.

3. The third problem of CLIR is the properly weighting of the translated alternatives

In CLIR the system has still the possibility to find relevant documents even when the wrong translation is used unlikely to machine translation techniques, where the accurate translation is the primary requirement. This fact illustrates the problem of term and translation weighting, which needs be solved.

Grefenstette described different approaches to find translations of a search terms. Those four solutions will be briefly described in this chapter. More detailed information can be found in the reference [Grefenstette w.y.]. The first discussed key for finding accurate translations are the use of dictionaries, which is the easiest way to get acceptable results. A couple of important dictionaries are available online or in form of databases, which can be implemented easier into a CLIR system. Filtering especially of large dictionaries is need. Hull and Grefenstette described one dictionary-based approach in their article [Hull 1996]. Word normalisation processes like stemming might need to be also performed [Sheridan 1996].

The second key to successful term translation in accordance to Grefenstette is the use of parallel corpora. It is possible to get bi-lingual term equivalents with a document selection of the proper size. The SPIDER system is one of IR system that can use corpuses to perform translation, which is described in the articles of Schaeuble [Schaeuble 1993] and of Sheridan and Ballerini [Sheridan 1996].

Grefenstette describes pruning of translation alternatives as another key factor and referred to the EMIR system from Fluhr. In this system a database of known compounds from the target language is used to filter out translations from compound words translated word-to-word [unknown reference].

As the last key factor Grefenstette mentioned the weighting process of alternatives, which was also investigated by Salton and Buckley in 1988 [Salton and Buckley 1988]. Their approach consists of an indexing method and a retrieval function, where the indexing method determines the descriptions of the stored data items and of the user given query items. The retrieval function matches the descriptions of the query items against the descriptions of the stored data items. More about this method can be found in the reference [Schaeuble 1993]

Conclusion and discussion

CLIR is a discipline where machine translation and information retrieval come together [Grefenstette w.y.].

It was difficult to implement a CLIR system into the field of cancer research. Medical terms are still mostly written in Latin, so that it difficult to find English or German translations. The CANSEARCH expert system [POLLITT 1987] helped to find some usable translations.

As one of finding of this assignment it is recommended to use professional expert online dictionary for the translation process. A stemming algorithm would increase the precision significantly, but the scope of the given task was the implementation of a CLIR prototype system. Therefore the author of the assignment decided to create a short dictionary with five different terms manually.

English	German
Drugs	Medikamente
Therapy	Therapie
Types	Arten
Cancer	Krebs
Systems (Organ systems)	Systeme (Organsysteme)

Table1: The word translation matrix.

Each of these terms had some similar terms stored in an array.

The implementation of a graphical user interface started to be another challenge. The Internet server here at the University of Sunderland is a UNIX server with its own firewall, so that the querying of search engines could not be performed by a graphical user interface. To overcome this problem a command-line oriented system was developed and deployed for cell M7. An example of a graphical CLIR that uses query expansion is available over the Internet under following URL: <http://osiris.sunderland.ac.uk/~ca9abe/com368/index.htm>

The last problem, which the author was faced wit, was to find on medicine specialist search engines. Free available engines do not fulfil the requirements for a CLIR. The used search engines should be seen for that reason as an example, of what was intended to perform with the

translated and expanded query. Specialist search engine for the World Wide Web are only accessible for the institutes own researches. The author had not the possibility to integrate the planned machines, which should be taken into account in rating and marking the given prototype.

As mentioned before CANCEIR should demonstrate the possibility to translate and expand a multilingual query.

The implementation of an online dictionary with stemming algorithms and also adding specialist search engines would improve the prototype. A graphical user interface would help the user to work better with the system than in the present version.

Appendix:

- System overview
- Flowchart
- Pseudo code
- Code of CANCEIR
- User manual
- References