



How Might CLIR be Accomplished?

Elizabeth D. Liddy

**Professor & Director, Center for NLP
School of Information Studies
Syracuse University**

November 13, 2000



CLIR

- **Cross-Language Information Retrieval**



CLIR

- **Cross-Language Information Retrieval**

- **Definition:**

*Users enter their query in one language
and the system retrieves relevant
documents in other languages.*



Why Cross Language Retrieval matters:



Why Cross Language Retrieval matters:

- **Globalization of the economy**



Why Cross Language Retrieval matters:

- **Globalization of the economy**
 - International corporations with divisions located in various countries



Why Cross Language Retrieval matters:

- **Globalization of the economy**
 - International corporations with divisions located in various countries
 - Vital documents requiring dynamic access exist in multiple languages



Why Cross Language Retrieval matters:

- **Globalization of the economy**
 - International corporations with divisions located in various countries
 - Vital documents requiring dynamic access exist in multiple languages
 - **ISO standards apply to all documentation**



Why Cross Language Retrieval matters:

- **Globalization of the economy**
 - International corporations with divisions located in various countries
 - Vital documents requiring dynamic access exist in multiple languages
 - ISO standards apply to all documentation
 - **Multilingual litigation**



Why Cross Language Retrieval matters:

- **Globalization of the economy**
 - International corporations with divisions located in various countries
 - Vital documents requiring dynamic access exist in multiple languages
 - ISO standards apply to all documentation
 - Multilingual litigation
 - **Intellectual property coverage worldwide**



Why Cross Language Retrieval matters:

- **Globalization of the economy**
 - International corporations with divisions located in various countries
 - Vital documents requiring dynamic access exist in multiple languages
 - ISO standards apply to all documentation
 - Multilingual litigation
 - Intellectual property coverage worldwide
 - **Customers speak multiple languages**



Why Cross Language Retrieval matters:

- **Globalization of the economy**
 - International corporations with divisions located in various countries
 - Vital documents requiring dynamic access exist in multiple languages
 - ISO standards apply to all documentation
 - Multilingual litigation
 - Intellectual property coverage worldwide
 - Customers speak multiple languages
 - **Employees speak multiple languages**



Why Cross Language Retrieval matters: (cont'd)

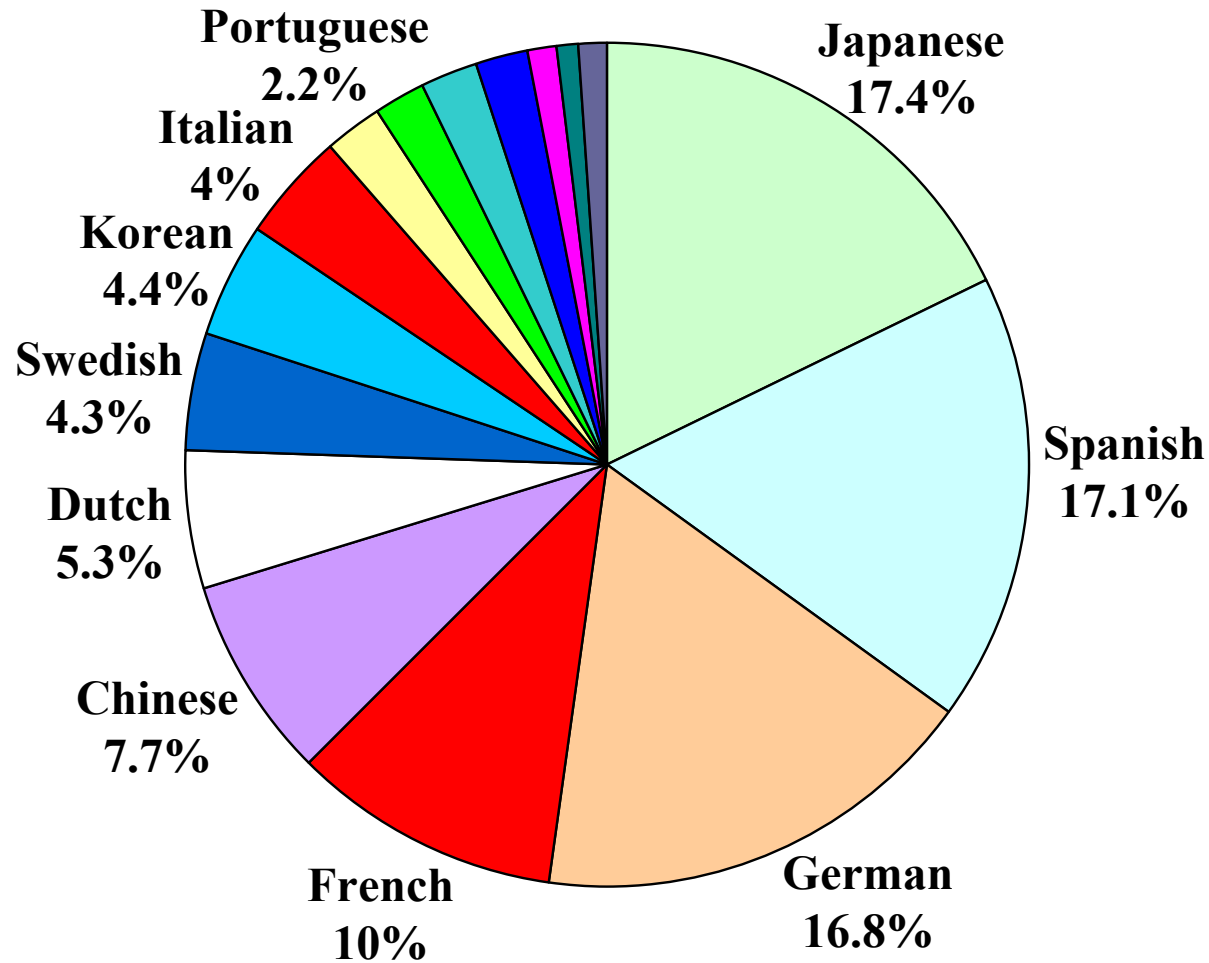
- **Internationalization of the Internet**



Why Cross Language Retrieval matters: (cont'd)

- **Internationalization of the Internet**
 - Non-English speakers represent the fastest growing group of new Internet users

83 Million People Online Who Access the Internet in Other Languages





Why Cross Language Retrieval matters: (cont'd)

- **Internationalization of the Internet**
 - Non-English speakers represent the fastest growing group of new Internet users
 - In 1997, 8.1 million Spanish speakers



Why Cross Language Retrieval matters: (cont'd)

- **Internationalization of the Internet**
 - Non-English speakers represent the fastest growing group of new Internet users
 - In 1997, 8.1 million Spanish speakers
 - **By 2001, there will be 37 million**



Why Cross Language Retrieval matters: (cont'd)

- **Government Agencies**



Why Cross Language Retrieval matters: (cont'd)

- **Government Agencies**
 - Require ability to monitor events in many languages



Why Cross Language Retrieval matters: (cont'd)

- **Government Agencies**
 - Require ability to monitor events in many languages
 - Analysts with knowledge of foreign languages are becoming more scarce



Why Cross Language Retrieval matters: (cont'd)

- **Government Agencies**
 - Require ability to monitor events in many languages
 - Analysts with knowledge of foreign languages are becoming more scarce
 - Languages of interest change quickly



Cross-Language Information Retrieval



Cross-Language Information Retrieval

- **Many different approaches to CLIR have been implemented and much progress has been made in recent years**



Cross-Language Information Retrieval

- **Many different approaches to CLIR have been implemented and much progress has been made in recent years**
- **How a system “crosses the language barrier”**



Cross-Language Information Retrieval

- **Many different approaches to CLIR have been implemented and much progress has been made in recent years**
- **How a system “crosses the language barrier”**
 - **Machine Translation**



Cross-Language Information Retrieval

- **Many different approaches to CLIR have been implemented and much progress has been made in recent years**
- **How a system “crosses the language barrier”**
 - **Machine Translation**
 - **Machine-Readable Dictionary/Thesaurus**



Cross-Language Information Retrieval

- **Many different approaches to CLIR have been implemented and much progress has been made in recent years**
- **How a system “crosses the language barrier”**
 - **Machine Translation**
 - **Machine-Readable Dictionary/Thesaurus**
 - **Corpus Based**



Cross-Language Information Retrieval

- **Many different approaches to CLIR have been implemented and much progress has been made in recent years**
- **How a system “crosses the language barrier”**
 - **Machine Translation**
 - **Machine-Readable Dictionary/Thesaurus**
 - **Corpus Based**
 - **Parallel**
 - **Comparable**
 - **GVSM**
 - **Latent Semantic Indexing**



Cross-Language Information Retrieval

- **Many different approaches to CLIR have been implemented and much progress has been made in recent years**
- **How a system “crosses the language barrier”**
 - **Machine Translation**
 - **Machine-Readable Dictionary/Thesaurus**
 - **Corpus Based**
 - **Parallel**
 - **Comparable**
 - **GVSM**
 - **Latent Semantic Indexing**
 - **Conceptual Interlingua**



Machine Translation Approach



Machine Translation Approach

- **Either, an MT system is used to translate documents into each of the possible querying languages**



Machine Translation Approach

- **Either, an MT system is used to translate documents into each of the possible querying languages**
 - not viable on large collections



Machine Translation Approach

- **Either, an MT system is used to translate documents into each of the possible querying languages**
 - not viable on large collections
 - too many languages into which to translate the documents



Machine Translation Approach

- **Either, an MT system is used to translate documents into each of the possible querying languages**
 - not viable on large collections
 - too many languages into which to translate the documents
- **Or, the query is translated into all the languages of the document collection**



Machine Translation Approach

- **Either, an MT system is used to translate documents into each of the possible querying languages**
 - not viable on large collections
 - too many languages into which to translate the documents
- **Or, the query is translated into all the languages of the document collection**
 - queries often do not contain enough context for disambiguation



Machine Readable Dictionary



Machine Readable Dictionary

- **Query terms are looked up in a bilingual dictionary and replaced with a/some translations from the language of the documents being searched**



Machine Readable Dictionary

- **Query terms are looked up in a bilingual dictionary and replaced with a/some translations from the language of the documents being searched**
 - **problems with selecting correct entry for ambiguous terms**



Machine Readable Dictionary

- **Query terms are looked up in a bilingual dictionary and replaced with a/some translations from the language of the documents being searched**
 - problems with selecting correct entry for ambiguous terms
 - **requires a bilingual MRD for every query & document language pair**



Corpus-based Approaches



Corpus-based Approaches

- **Parallel corpora**



Corpus-based Approaches

- **Parallel corpora**
 - used to identify lexical equivalencies across languages



Corpus-based Approaches

- **Parallel corpora**
 - used to identify lexical equivalencies across languages
 - UN corpus in French, Spanish & English



Corpus-based Approaches

- **Parallel corpora**
 - used to identify lexical equivalencies across languages
 - UN corpus in French, Spanish & English
- **Comparable corpora**



Corpus-based Approaches

- **Parallel corpora**
 - used to identify lexical equivalencies across languages
 - UN corpus in French, Spanish & English
- **Comparable corpora**
 - Swiss news reports in German, French & Italian



Corpus-based Approaches

- **Parallel corpora**
 - used to identify lexical equivalencies across languages
 - UN corpus in French, Spanish & English
- **Comparable corpora**
 - Swiss news reports in German, French & Italian
- **Must have corpus resources in each language of interest**



Corpus-based Approaches (Cont'd)

- **Generalized Vector Space Model (GVSM)**



Corpus-based Approaches (Cont'd)

- **Generalized Vector Space Model (GVSM)**
 - uses a bilingual training corpus to build matrices of documents & term weights in each language



Corpus-based Approaches (Cont'd)

- **Generalized Vector Space Model (GVSM)**
 - uses a bilingual training corpus to build matrices of documents & term weights in each language
 - requires parallel corpora for each language pair



Corpus-based Approaches (Cont'd)

- **Generalized Vector Space Model (GVSM)**
 - uses a bilingual training corpus to build matrices of documents & term weights in each language
 - requires parallel corpora for each language pair
- **Latent Semantic Indexing**



Corpus-based Approaches (Cont'd)

- **Generalized Vector Space Model (GVSM)**
 - uses a bilingual training corpus to build matrices of documents & term weights in each language
 - requires parallel corpora for each language pair
- **Latent Semantic Indexing**
 - reduces the GVSM further



Corpus-based Approaches (Cont'd)

- **Generalized Vector Space Model (GVSM)**
 - uses a bilingual training corpus to build matrices of documents & term weights in each language
 - requires parallel corpora for each language pair
- **Latent Semantic Indexing**
 - reduces the GVSM further
 - requires parallel or comparable corpora



Corpus-based Approaches (Cont'd)

- **Generalized Vector Space Model (GVSM)**
 - uses a bilingual training corpus to build matrices of documents & term weights in each language
 - requires parallel corpora for each language pair
- **Latent Semantic Indexing**
 - reduces the GVSM further
 - requires parallel or comparable corpora
 - **computationally expensive**



Conceptual Interlingua Approach



Conceptual Interlingua Approach

- **A conceptual space in which the terms / phrases from multiple languages which refer to the same concept are mapped into a language independent schema**



Conceptual Interlingua Approach

- **A conceptual space in which the terms / phrases from multiple languages which refer to the same concept are mapped into a language independent schema**
- **Both documents and queries are mapped into the Conceptual Interlingua**



Conceptual Interlingua Approach

- **A conceptual space in which the terms / phrases from multiple languages which refer to the same concept are mapped into a language independent schema**
- **Both documents and queries are mapped into the Conceptual Interlingua**
- **Permits matching and retrieval based on any combination of languages involved, rather than relying on pairwise translations**



Conceptual Interlingua Approach

- **Enables language-independent CLIR based on natural language concepts**



Conceptual Interlingua Approach

- **Enables language-independent CLIR based on natural language concepts**
- **E.g. EuroWordNet**



Conceptual Interlingua Approach

- **Enables language-independent CLIR based on natural language concepts**
- **E.g. EuroWordNet**
 - **a hierarchically organized concept lexicon**



Conceptual Interlingua Approach

- **Enables language-independent CLIR based on natural language concepts**
- **E.g. EuroWordNet**
 - a hierarchically organized concept lexicon
 - **with language-neutral concept groups**



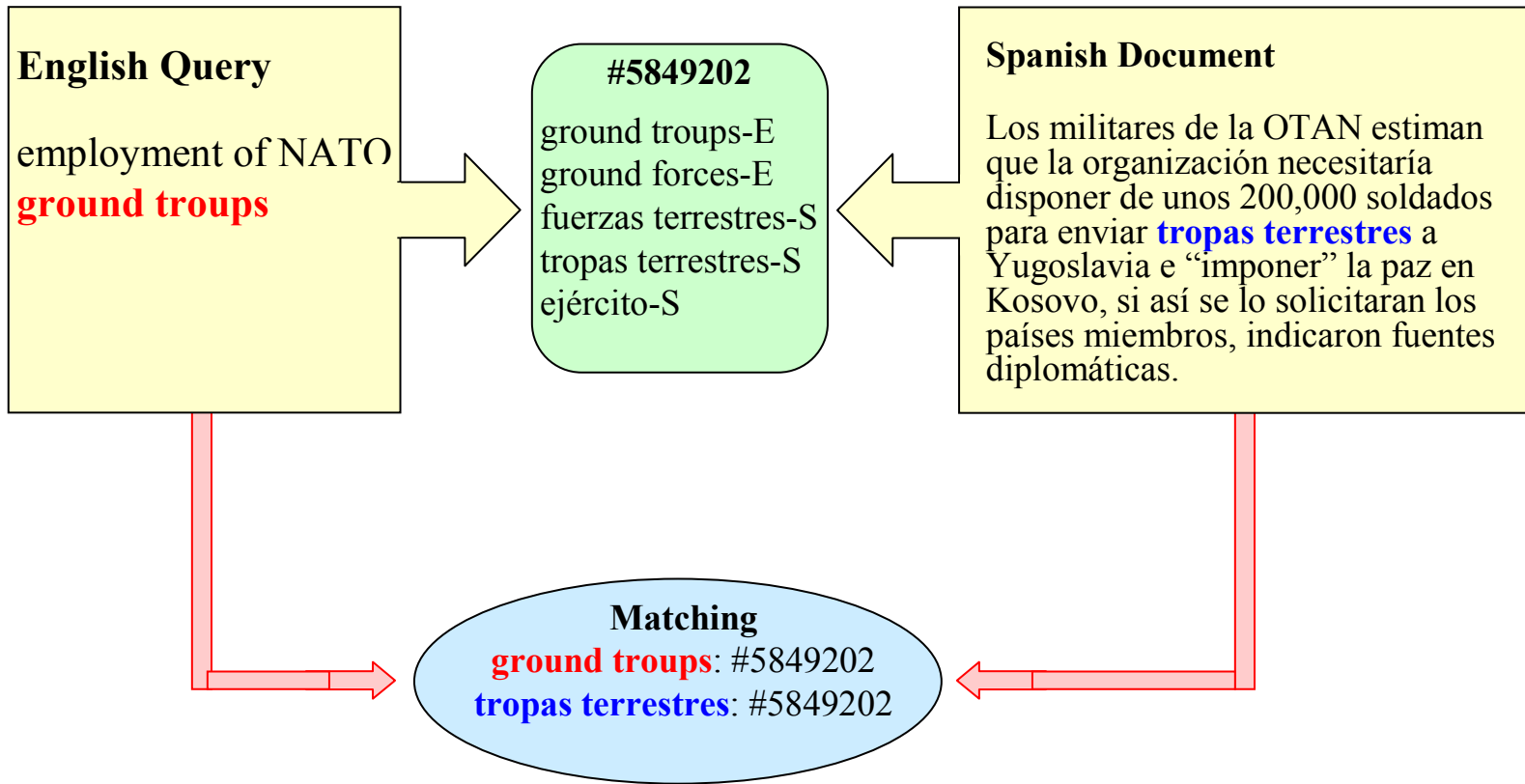
Conceptual Interlingua Approach

- **Enables language-independent CLIR based on natural language concepts**
- **E.g. EuroWordNet**
 - a hierarchically organized concept lexicon
 - with language-neutral concept groups
 - **which link to their terminological instantiations in various languages**



Conceptual Interlingua Approach

- **Enables language-independent CLIR based on natural language concepts**
- **E.g. EuroWordNet**
 - a hierarchically organized concept lexicon
 - with language-neutral concept groups
 - which link to their terminological instantiations in various languages
- **Enables matching to synonyms in all languages**





Conceptual Interlingua Performance

- **Conceptual interlingua matching in one language shown to produce up to 30% improvement over word-form matching when disambiguation is done (Gonzalo et al, 1998)**



Conceptual Interlingua Performance

- **Conceptual interlingua matching in one language shown to produce up to 30% improvement over word-form matching when disambiguation is done (Gonzalo et al, 1998)**
 - early work using WordNet for retrieval suggests caution



Conceptual Interlingua Performance

- **Conceptual interlingua matching in one language shown to produce up to 30% improvement over word-form matching when disambiguation is done (Gonzalo et al, 1998)**
 - early work using WordNet for retrieval suggests caution
 - **CLIR is even more sensitive to ambiguity than monolingual retrieval**



Cross-Language Concerns

- **Disambiguation is essential**



Cross-Language Concerns

- **Disambiguation is essential**
 - Syntactic via:



Cross-Language Concerns

- **Disambiguation is essential**
 - Syntactic via:
 - **part-of-speech tagging**



Cross-Language Concerns

- **Disambiguation is essential**
 - Syntactic via:
 - part-of-speech tagging
 - **Semantic via:**



Cross-Language Concerns

- **Disambiguation is essential**
 - Syntactic via:
 - part-of-speech tagging
 - Semantic via:
 - dealing with ‘false cognates’



Cross-Language Concerns

- **Disambiguation is essential**
 - Syntactic via:
 - part-of-speech tagging
 - Semantic via:
 - dealing with ‘false cognates’
 - words which are spelled identical in different languages but refer to different concepts



Cross-Language Concerns

- **Disambiguation is essential**
 - Syntactic via:
 - part-of-speech tagging
 - Semantic via:
 - dealing with ‘false cognates’
 - words which are spelled identical in different languages but refer to different concepts
 - **disambiguation among synsets**



Cross-Language Concerns

- **Disambiguation is essential**
 - Syntactic via:
 - part-of-speech tagging
 - Semantic via:
 - dealing with ‘false cognates’
 - words which are spelled identical in different languages but refer to different concepts
 - disambiguation among synsets
 - **Student Papers & Awards - 1:30 Today**



Disambiguation among synsets:

- **With uncontrolled synset expansion, the highest ranked documents will be those containing the query terms with the most senses**



Disambiguation among synsets:

- **With uncontrolled synset expansion, the highest ranked documents will be those containing the query terms with the most senses**
 - e.g., in a 5 word query, if 1 word has 9 senses and you don't disambiguate, the preponderance of synonyms from these 9 synsets will swamp the other terms



Disambiguation among synsets: (Cont'd)

- **Caution is needed as earlier results showed that incorrect sense resolution has a more serious impact on retrieval than spurious matches (Voorhees, 1993)**



Disambiguation among synsets: (Cont'd)

- **Caution is needed as earlier results showed that incorrect sense resolution has a more serious impact on retrieval than spurious matches (Voorhees, 1993)**
- **CLIR needs heuristics from psycholinguistics to select which sense is intended in a particular context and weight that synset's members accordingly**



Potential for CLIR will be determined by:

- **The amount of effort required of users to locate, retrieve, and put foreign-language information to use**



Potential for CLIR will be determined by:

- **The amount of effort required of users to locate, retrieve, and put foreign-language information to use**
- **The quality of the retrieval**



Potential for CLIR will be determined by:

- **The amount of effort required of users to locate, retrieve, and put foreign-language information to use**
- **The quality of the retrieval**
 - Precision



Potential for CLIR will be determined by:

- **The amount of effort required of users to locate, retrieve, and put foreign-language information to use**
- **The quality of the retrieval**
 - Precision
 - Recall



Potential for CLIR will be determined by:

- **The amount of effort required of users to locate, retrieve, and put foreign-language information to use**
- **The quality of the retrieval**
 - Precision
 - Recall
- **The availability & quality of translation facilities**



Potential for CLIR will be determined by:

- **The amount of effort required of users to locate, retrieve, and put foreign-language information to use**
- **The quality of the retrieval**
 - Precision
 - Recall
- **The availability & quality of translation facilities**
- **Impact of conferences such as TREC, CLEF, and NTCIR**



Further Resources

- **ARIST Chapter on CLIR – 1998**
Oard, Doug & Diekema, Anne
- **TREC 8 Proceedings CLIR Track**
http://trec.nist.gov/pubs/trec8/t8_proceedings.html