The EU project ’Twenty-One’ and cross-language IR [*]

Wessel Kraaij
Netherelands Organization for Applied Scientific Research (TNO)
Institute of Applied Physics
PO Box 155 2600 AD Delft
The Netherlands
kraaij@tpd.tno.nl or kraaijw@acm.org

April 21, 1997

Abstract

TwentyOne is a EU funded project which aims at developing advanced indexing and retrieval techniques for multimedia document bases. The document base consists of documents in four languages: Dutch, English, French and German. This paper focusses on the multilingual aspects of the project: cross-language retrieval, partial document translation techniques and automatic hyperlinking between source text and translations.

1 Introduction

TwentyOne[1 2] is a project funded by the EU Telematics pogramme (IE-2108). Project partners include academic partners like the Universities of Twente and Tübingen, companies like Getronics and Xerox, contract research organistations like TNO and DFKI and non-profit environmental organisations like Friends of the Earth. The project can be characterised by the following keywords:

Document conversion The TwentyOne system aims at the disclosure of documents of different media types and / or data formats e.g. paper documents, WEB documents, word processor documents, text annotated images, audio or video material.

Knowledge based disclosure The TwentyOne Multimedia document base will be disclosed using several advanced techniques like fuzzy matching, NLP-based phrase indexing, relevance ranking and automatic hyperlinking.

Multilinguality The TwentyOne database consists of documents in different languages, initially Dutch, English, French and German but extensions to other European languages are envisaged.

Sustainable Development The name of the project refers to the UN conference on this topic in Rio de Janeiro 1992. The aim of the project is to build a system that supports and improves dissemination of information about ’local agenda 21’ initiatives.

Dissemination Model The environmental partners develop an information transaction model which works like a perpetuum mobile. Both information providers and seekers profit from the model, the former by increasing the number of potential customers, the latter because more information becomes available. The project tries to stimulate interaction and raise awareness of local agenda 21 initiatives in Europe.

Application oriented The most important deliverable of the project is the profiling system which produces an index on the multilingual multimedia document base. This index will be available via CDROM and accessible via a WEB server.

[*] A variant of this paper will appear in the proceedings of the AAAI spring symposium 1997 on Cross-Language Text and Speech Retrieval

[1] This paper describes joint work with colleagues at TNO-TPD, University of Twente, DFKI, Xerox and University of Tuebingen

[2] The TwentyOne homepage can be found at: `http://twentyone.tpd.tno.nl/`

Because TwentyOne is funded by an application oriented programme, the project has only limited resources for fundamental research, Some aspects of the system touch open research problems though. The consortium has assessed the state of the art of technology in these areas (e.g. cross language information retrieval). Because a first version of the system has to be finished by the end of 1997, we take a pragmatic approach, by integrating available tools and resources and developing solutions for missing links.

This paper focusses on the multilingual functions of TwentyOne. They are threefold:

1. retrieval of documents in another language than the query language (CLIR) , supported languages are Dutch, English, French and German
2. (partial) translation of documents to enable content judgement by the user
3. Automatic hyperlinks between index terms and their translations (aligned multilingual documents)

From a research perspective, attacking four languages at once complicates things considerably. Scalability of the system and separation of language dependent from language independent resources becomes more important than in the two-language case which has been investigated in detail, especially in the last few years. A related European project is the ESPRIT II project EMIR [2] which covers a subset of the TwentyOne languages: English, French and German. EMIR is based on the SPIRIT ranked boolean engine combined with a multilingual thesaurus as front-end. EMIR is currently being extended to Russian.

It's clear that development of (partial) document translation software which can also be used for aligned hyper-multi-language documents is quite ambitious. We claim however that a rudimentary translation of a document can greatly enhance the usability of the system in the case of languages which are completely unknown to the user.

The automatic hyperlinking function attaches typed hyperlinks between terms, phrases or images etc. These links can be either static (generated off-line) or dynamic, in which case a link is evaluated by a CGI program. We have planned to generate hyperlinks for all translated Noun-Phrases, which makes it easy for the user to jump between translated and original text.

In this paper we will concentrate on CLIR and partial document translation, because these functions can be combined in several aspects. We will first present some results from CLIR experiments which have inspired the design. Subsequently we will discuss the TwentyOne approach to CLIR and Document translation which is also influenced by the availability of linguistic resources like bilingual dictionaries.

2 Concise overview of approaches to CLIR

We will present the possibilities for CLIR in a slightly different taxonomy[3] than the one used in the overview article by Oard[6]. CLIR systems can be classified in two ways:

1. The stage in the disclosure process at which the language transfer takes place. Translation can be done either during indexing time (off-line) or as a pre-processing step in the retrieval process (on-line).
2. The translation process can be based on three sources of transfer knowledge:
   (a) MT systems
   (b) Bilingual dictionaries or thesauri
   (c) Parallel corpora

We will discuss all possible combinations of the approaches and resources.

2.1 Query translation (on-line translation)

2.1.1 Dictionary based approach

Simple word by word translation of the query terms has been evaluated in e.g. [4]. It is the most simple approach to CLIR as ambiguity is left unresolved: each (lemmatised) word is substituted by all its possible translations. Two problems are prominent:

[3] This taxonomy highlights the first dilemma in CLIR systems design: either on-line query translation or off-line document translation

1. Polysemy:
   Translation of query concepts is likely to decrease precision when word sense cannot be disambiguated. Example: the Dutch word "slag" can be translated to both "battle" or "stroke".
   On the other hand, if more than one equivalent translation is available, translation could increase recall, because synonyms are added to the query. Hull proposes to use a ranked Boolean query model as a possible way to cope with this problem. In this model documents are ordered on the number of (translations of) query concepts that are matched. This model will probably not work so well for short (1-3 term) queries. Because documents that match only one query concept have a high probability of being totally off topic when that query term has multiple translations.

2. Multi word expressions(MWE's)
   Idiomatic expressions, terminology, collocations are a notorious problem in CLIR. Word based translation fails here because often the meaning of the MWE is not compositional, e.g. *yellow pages*. A terminology or idiomatic dictionary can only partly leverage the problem because most of the MWE's are highly domain specific.

2.1.2 MT based approach

Typical queries in current popular IR systems like "Web search engines" tend to be very short. Therefore the advantage of MT systems (which in principle can exploit syntactic and semantic aspects of context to improve translation) with respect to dictionary based approaches is questionable. On the other hand, for longer queries (Query by example, search similar documents) MT could yield good results. The EMIR project has compared SYSTRAN query translation with thesaurus based translation, average precision of the latter system turned out to be much better.

2.1.3 Corpus based approach

Parallel corpora implicitly encode a lot of transfer knowledge. This knowledge can be exploited in different ways:

1. Deriving bilingual dictionaries from aligned corpora. Especially domain specific aligned corpora are of great value to infer translations of or at least identify MWE's. These are of key value for CLIR but can't be dealt with by simple word-based translation. In fact this is also a dictionary based approach.

2. Store dual-language documents in a dual-language vector space, Perform Latent Semantic indexing on the dual language documents before folding in the monolingual documents . The LSI space captures a "multilingual semantic space" on which the monolingual documents are mapped. Positive results are reported in [1]. An advantage of this approach is that alignment of the parallel copora is only necessary on the document level.

2.2 Document translation (off-line)

2.2.1 MT based full translation

If we translate all documents to the query language, than CLIR is reduced to a monolingual IR case. A disadvantage of the approach is the dependency on imperfect MT systems which are often closed monolithic systems with (probably) limited coverage of domain terminology. Another disadvantage is that MT system deliver only one translation in case of synonyms. An advantage however is that the translated documents can also be used for presentation to the user, which makes sense when translating from languages of which the user even has no passive knowledge. Machine translation of complete documents is obviously more worthwhile than translating short queries, because the MT system can use the whole document as context. Dumais [1] reported favourable results of document translation by SYSTRAN in combination with monolingual LSI.

2.2.2 Partial translation techniques

Because most Indexing models are based on lemmatised content words, a CLIR system could be based on lemma based translation of non-stopwords as a front end for a monolingual system. However this transfer step is hampered by the same problems as dictionary based query translation. The main difference with query translation is the availability of context. The question is how to use this context to improve the translation. A possible knowledgesource are word association statistics like the expected mutual information measure (EMIM). Such statistics can also be used to identify multi word terminology (sometimes referred to as “statistical phrases” in IR literature). Johansson [5] reports that highly associated bigrams are not always good index terms, but this could be remedied by removing stop words before or after the bigram finding process.

Partial translation of noun phrases for presentation purposes has to meet higher requirements than the query translation case: getting the word senses right is not enough because word order and inflection have to be correct in order to make the translation readable. This step requires syntactic and morphological knowledge.

3 The TwentyOne approach

In this section we will discuss the design choices we have made in order to build a system with the three multilingual aspects which were introduced earlier. We will start by listing the relevant resources.

3.1 Availability of Resources

Bi- or multilingual dictionaries We have contacts with two Dutch publishers. The material is either a collection of bilingual dictionaries from Dutch to the other languages or a multilingual thesaurus, including morphological information. The lexical database even contains translations of idioms and collocations, which might be extremely valuable. We don’t know about MRD’s of publishers in other European countries which offer translation to and from Dutch

EU materials The EU has published the EUROVOC thesaurus, a collection of commonly used terminology in EU documents. The thesaurus is electronically available

Parallel texts We have the official “Agenda 21” conference document in all the four languages. We are still trying to find parallel texts at EU or UN institutions.

Commercial MT software Recently a survey of these tools (examples can be found on the WEB) has been started at DFKI. We are not aware of a commercially available MT system which supports the four languages supported by TwentyOne

Monolingual IR system TwentyOne will use the monolingual IR kernel of TNO-TPD which supports:

- Vector Space retrieval
- Boolean retrieval
- Fuzzy matching

NLP tools Xerox provides their finite state tools for morphological analysis and POS disambigation. A Fast PSG parser developed at TNO-TPD will be used for NP-extraction.

3.2 Document translation in TwentyOne

Experiments with word based translation and translation by Systran via the WWW have shown the enormous difference in quality between these approaches. Therefore we will store translations of the documents at the TwentyOne site for the purpose of presentation. We know already, however, that not all language pairs are covered by commercial MT tools so a fall-back option is needed.

The fall-back option is called *term translation*. With *term* we refer to the main indexing units of the TwentyOne system: noun-phrases. This means that in most cases, a term is complex i.e. consists of more than one concept. The challenge is to develop robust term translation techniques which can preserve the morphosyntactic information of the NP structure. This structure is available because every document is processed by the NLP module which consists of morphological analysis, POS disambiguation and noun-phrase extraction.

Identification and translation of MWE’s is a tough problem, but we will combine corpus based approaches with bilingual dictionaries.

3.3 Translation Hyperlinks

A second reason why we want to develop our own term translation methodology is that we want to establish hyperlinks between terms and their translations. The result is a document aligned with its three translations. The alignment between terms will be implemented by hyperlinks. MT systems are file oriented and thus would require post translation alignment (reverse engineering).

3.4 CLIR in TwentyOne

Figure 1 shows that the TwentyOne system will include both Document translation *and* Query Translation because we expect that both approaches can improve the performance of the system in their own way. The main approach to CLIR is Document translation, because DT can fully exploit context for disambiguation. But we expect the following problems:

1. OCR errors will not be translated
2. Part of the domain specific terminology is not covered by the available transfer resources
3. Some language pairs might be stuck by poor DT functionality

Query translation can partly alleviate the effects of these problems in the following ways:

1. A document with a relevant term which contains an OCR error can be found via fuzzy matching with the translated query concept.
2. The user can perform relevance feedback in the target language, once a relevant document is found in the particular foreign language. This technique is also useful to overcome the effects of translation ambiguity
3. A word based translation approach followed by a ranked boolean query (cf. [4]) can act as a disambiguating filter.
4. Interactive disambiguation by the user

Query translation in TwentyOne will use a multilingual lexicon which comprises both lemmas(including syntactic category) and multi-word-expressions. This lexicon will be based on the merge of existing multilingual thesauri, bilingual machine readable dictionaries and dictionaries derived from parallel corpora [3] and probably also some hand-coded translations for automatically indentified MWE's.

4 Outlook

In the project there is some time available for evaluation. The evaluation will be both based on feedback from "real" users because the system will be operational on the WEB during the project, but also a small scale test with the usual measures like average precision is foreseen. We face the problem of the unavailability of test corpora, though we probably will not test all of the 12 possible language transfer directions..

References

[1] Susan T. Dumais, Thomas K. Landauer, and Michael L. Littman. Automatic cross-linguistic information retrieval using latent semantic indexing. In *Workshop on Cross-Linguistic Information Retrieval (SIGIR'96)*, pages 16–24, 1996.

[2] C. Fluhr and Kh. Radwan. Fulltext databases as lexical semantic knowledge for multilingual interrogation and machine translation. In *EWAIC'93*, 1993.

[3] Djoerd Hiemstra. Using statistical methods to create a bilingual dictionary. Master's thesis, University of Twente, 1996.

[4] David Hull and Gregory Grefenstette. A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

[5] Christer Johansson. Good bigrams. In *Proceedings of COLING 1996*, pages 592–597, 1996.

[6] Douglas W. Oard and Bonnie J. Dorr. A survey of multilingual text retrieval. Technical report, University of Maryland, 1996.

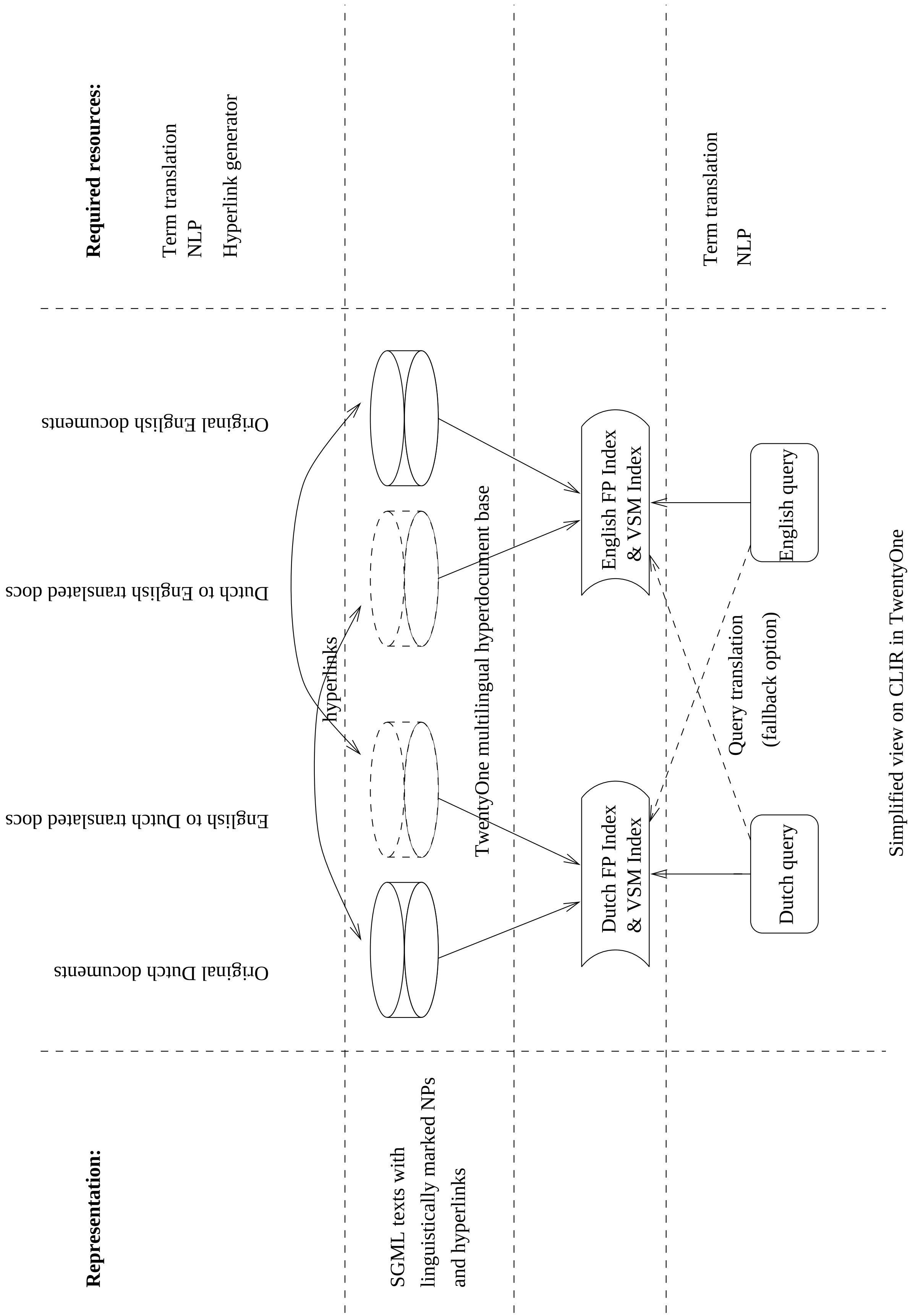

Figure 1: Overview of CLIR in TwentyOne