A Domain Specific Lexicon Acquisition Tool for Cross-Language Information Retrieval

Djoerd Hiemstra
University of Twente/CTIT
hiemstra@cs.utwente.nl

Franciska de Jong
University of Twente/CTIT
fdejong@cs.utwente.nl

Wessel Kraaij
Netherlands Organization for Applied Scientific Research (TNO)
kraaij@tpd.tno.nl

Abstract

With the recent enormous increase of information dissemination via the web as incentive there is a growing interest in supporting tools for cross-language retrieval. In this paper we describe a disclosure and retrieval approach that fulfills the needs of both information providers and users by offering fast and cheap access to a large amounts of documents from various language domains. Relevant information can be retrieved irrespective of the language used for the specification of a query. In order to realize this type of multilingual functionality the availability of several translation tools is needed, both of a generic and a domain specific nature. Domain specific tools are often not available or only against large costs. In this paper we will therefore focus on a way to reduce these costs, namely the automatic derivation of multilingual resources from so-called parallel text corpora. The benefits of this approach will be illustrated for an example system, i.e. the demonstrator developed within the project Twenty-One, which is tuned to information from the area of sustainable development.

Keywords: Full Text Retrieval, IR tools, Lexicon Acquisition, Parallel Corpora, Statistical Natural Language Processing, Cross-Language Information Retrieval.

1 Introduction

The recent enormous increase in the use of information from Internet and CD-ROM has led to databases being available in many languages. Often the relevance of the documents in these databases goes beyond the scope of a region or country. In cases where the documents are only available in a foreign language, cross-language retrieval functionality is needed to provide access to the documents for users who are non-native speakers of the foreign language or not a speaker of the language at all. Cross-language information retrieval (CLIR) minimally requires translations tools that are capable of translating document indexes and/or queries to help the user to identify documents that are relevant to his information need. In order to

1

realize this type of multilingual functionality the availability state-of the-art IR-tools are not sufficient. It could be argued that plain substitution of words occurring in a query by the corresponding words in one or more other languages can do the job. Our judgment however is that this is a very poor solution to a problem that deserves a more ambitious approach. By not relying on the poor quality of existing translation software, but by carefully coupling several natural language processing (NLP) techniques, such as sound monolingual morphological and syntatctic parsing, and various search modes, more adequate support for the multilingual information searchers will become available. The required NLP also includes translation knowledge, both of a generic and a domain specific nature. Domain specific resources are often not available or only against large costs. The automatic derivation of multilingual resources from so-called parallel text corpora is a way to prohibit large investments and still be able to fulfill the requirements for CLIR. The approach of looking for the optimal combination of NLP and search techniques is being applied within Twenty-One, a project aiming at the development of a system for the disclosure and retrieval of information on sustainable development. This paper explains how we envisage to realize CLIR in this domain. Both the translation technique(s), as well as the acquisition of suitable resources will be discussed. In particular it will focus on a tool for the acquisition of domain-specific translation knowledge, such as preferred word meaning and corresponding translations, and the translation of multi-word expressions from the domain of ecology and sustainable development. With the application domain of Twenty-One in mind this statistics based tool was applied to Agenda 21 to derive a domain specific probabilistic bilingual lexicon for English and Dutch. Agenda 21 is the document that contains the results from the 1992 UNCED conference (Rio de Janeiro) on sustainable development. It is available in numerous languages. Therefore it is a suitable document for the development and evaluation of the lexicon acquisition tool to be described below. An example of the kind of dictionary entries it generates for English-Dutch word pairs is given in Table 1.

sustainable	
duurzame	0.80
duurzaam	0.20

Table 1: an example entry

Section 2 will present project Twenty-One in more detail. In section 3 the will be focus on the envisaged multilingual functionality of the Twenty-One demonstrator, and section 4 will address the acquisition of a bilingual lexicon from a so-called parallel corpus (Agenda 21) and the way it can be put to use. Finally, some discussion and concluding remarks will be presented in section 5

2 Project Twenty-One

Twenty-One[1] is a project funded within the EU Telematics Applications Pogramme, sector Information Engineering. Project partners include academic partners (Universities of Twente and Tübingen), companies (Getronics, Xerox, Highland Software), contract research organizations (TNO and DFKI) and a number of non-profit environmental organizations like Friends of the Earth Europe.

2.0.1 general characteristics

The project can be characterized by the following keywords:

Multimedia The Twenty-One 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 with textual annotations[2].

Document conversion The system incorporates a component for the conversion of the various document formats into standard representation (SGML/HTML), including a tool for the conversion of paper documents into electronic format, on the basis of layout semantics analysis and OCR.

Advanced disclosure techniques The Twenty-One Multimedia document base will be disclosed using several advanced techniques like fuzzy matching, rule-based NLP-for phrase indexing, relevance ranking and automatic hyperlinking.

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

Domain-tuning: 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. This requires a special effort in the acquisition of linguistic resources that are tuned to the language and vocabulary in this domain. Still the technology to be developed is supposed to be generic.

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 supports the objectives of the users involved in the project by trying to stimulate interaction and raise awareness of local agenda 21 initiatives in Europe.

---

[1] The Twenty-One homepage can be found at: `http://www.tno.nl/twentyone/21-home.html`

[2] Within Pop-Eye, a EU-project in the sector Language Engineering, the automatic disclosure of video-fragments is pursued by using subtitles as a basis for indexing and retrieval. Cf. `http://www2.echo.lu/langeng/en/pop-eye/pop-eye.html`

Application oriented The most important deliverable of the project is the disclosure system which produces an index on the multilingual multimedia document base. This index will be available via CD-ROM and accessible via a web-server.

2.0.2   Multilingual characteristics

The description of Twenty-One will focus on the multilingual functionality of the system. Three aspects are crucial:

Cross-language retrieval Retrieval of documents in another language than the query language; the languages presently covered in the project are Dutch, English, French and German; extension with other languages is considered. 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 is more important than in the two-language case, which has been investigated in detail, especially in the last few years[3].

(Partial) translation of documents To enable content judgement by the user, translations of documents that match his query but that are written a language the user is not familiar are very useful. As explained below, this functionality can be realized in various ways.

Automatic hyperlinking 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 a link is evaluated by a CGI-program. Hyperlinks will be generated for all translated noun phrases, which should enable the user to easily jump between translated and original text.

The next section will explain the multilingual functionality of the Twenty-One demonstrator in more detail. First some results from CLIR experiments will be presented which have inspired the design. Subsequently we will discuss the Twenty-One approach to CLIR and its relation to the monolingual NLP components. As mentioned above, both aspects heavily rely on the availability of linguistic resources like bilingual dictionaries. This paper focuses on the tuning to the domain specific language and vocabulary.

3   Multilingual Fuctionality

Before presenting the approach to followed in the design of the Twenty-One system we will present some possibilities for CLIR by discussing a few parameters. The taxonomy behind is implied is slightly different from the one used in the overview article by Oard & Dorr [13]. CLIR systems can be classified according to three features:

---

[3]For example within the related ESPRIT II project EMIR [6] which covers a subset of the Twenty-One languages, namely 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.

1. The stage in the disclosure process at which the language transfer takes place. Translation can be performed either during indexing time (off-line) or as a pre-processing step in the retrieval process (on-line).

2. The translation can apply either to the objects in the document base or to the queries.

3. The translation process can be based on three sources of translation knowledge (also referred to as transfer knowledge):

   (a) MT systems
   (b) Bilingual dictionaries or thesauri
   (c) Parallel corpora

Below a series of possible combinations of approaches and resources will be presented.

3.1 On-line query translation

3.1.1 Dictionary based approach

Simple word by word translation of query terms has been evaluated by Hull [10]. It is the most simple approach to CLIR ambiguity turned out to be left unresolved: each (lemmatized) word is substituted by all its possible translations. There are two prominent problems with this approach:

1. Polysemy:
   Translation of query concepts is likely to decrease precision when word sense cannot be disambiguated. For 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 very well for short (1-3 word) queries, because a query term has multiple translations, documents that match only one query concept have a high probability of being totally off topic.

2. Multi word expressions(MWE's)
   Idiomatic expressions, terminology and 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.

3.1.2 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 query translation. The average precision of the latter system turned out to be much better.

3.1.3 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. Cf. section 4.

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 "multi-lingual semantic space" on which the monolingual documents are mapped. Positive results are reported in [5]. An advantage of this approach is that alignment of the parallel corpora is only necessary on the document level.

3.2 Off line document translation

3.2.1 MT based full translation

If we translate all documents to the query language, than CLIR is reduced to a monolingual IR case. 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 [5] reported favourable results of document translation by SYSTRAN in combination with monolingual LSI.

3.2.2 Partial translation techniques

Because most indexing models are based on lemmatized 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 knowledge source is 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 [11] 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.

3.3 Monolingual components

In the previous section the options for realizing CLIR were presented. The next section will explain the choices made for Twenty-One. This section introduces some of the crucial design

choices that are not affected by the multilingual functionality. For the part of the functionality that is independent of the cross-language retrieval the following elements are crucial.

Search kernel In CLIR, translation functionality is of course an add-on to monolingual retrieval functionality. For Twenty-One the monolingual Full Text Retrieval kernel developed at TNO-TPD is used. It supports various search modes.

- Vector Space retrieval
- Boolean retrieval
- Fuzzy matching

Monolingual NLP tools For morphological processing and part-of-speech tagging, Xerox finite state tools are used. For syntactic analysis a fast parser based on a phrase structure grammar for the extraction of NPs is available. This parser has been developed at TNO-TPD. These tools will be made availble for all the languages covered by the project. The extracted NPs are the basis for the indexing module.

Automatic indexing The NPs extracted from the texts and their frequencies are the basis for the construction of a term-based index.

3.4 CLIR in TwentyOne

At various stages NPs can be submitted to *term translation* TT. With *term* we refer to the main indexing units within Twenty-One: noun-phrases. 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 monolingual NLP modules which include of morphological analysis, POS disambiguation and parsing. Identification and translation of multi-word expressions is a tough problem, but by combining corpus based approaches and bilingual dictionaries this problem can be tackled up to a level that is adequate for the purposes of CLIR. Term translation can fullfil three roles:

1. It is the basis for the generation of a series of monolingual indexes (one for each project language). The monolingual NLP-modules identify the NPs in a document as the indexing units. By off line TT these source language index terms get three target language equivalents. These index terms are stored in the four monolingual indexes. During retrieval queries are matched on these monolingual indexes.

2. If monolingual query handling does not lead to any hits, TT can also be applied on-line to the query terms. Query translation can partly alleviate the effects of poor quality MT in the following ways:

    (a) A document with a relevant term which contains an OCR error can be found via fuzzy matching with the translated query concept.

(b) 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

(c) A word based translation approach followed by a ranked Boolean query (cf. [10]) can act as a disambiguating filter.

(d) Interactive disambiguation by the user

3. TT is the basis for the establishment of hyperlinks between terms and their translations. The result is a (part of 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).

In addition to TT Twenty-One will use off line Document Translation (DT) for the purpose of enabling users to judge the relevance of retrieved material. Experiments have been performed with word based translation and full text translation by the on-line software made available by SYSTRAN. These experiments have shown an enormous difference in quality between these approaches. Therefore for presentation purposes we favour the storage of translated documents at the Twenty-One site . We know already, however, that not all language pairs are covered by commercial MT tools, so partial translation of documents by applying TT is needed as a fall back option here[4]. Document translation could also be used as a basis for monolingual indexing in the three target language versions of a document. This could even obviate query translation. But as the quality of this translation would be poor on average the more reliable TT for on-line translation of the NPs from a query is presumably a more adequate basis for CLIR.

Both NLP and TT require lexical resources. Machine readable dictionaries as owned by commercial lexicon publishers could be useful for the generic lexical knowledge required by the monolingual NLP-component and the translation modules. Such lexical databases usually do not only contain information on single words but even contain idioms and collocations plus their translation, which can be extremely valuable. As the acquisition of machine readable dictionaries for our purposes is complicated by the fact that coverage of all the four project languages is rare. Therefore tools that can automate the acquisition of lexical resources is not only important for the domain specific vocabulary, but could also be of value for the generic part. In addition to general purpose dictionaries, special terminology banks might be useful, e.g. the EUROVOC thesaurus, collection, of commonly used terminology in EU documents. The dictionaries envisaged for Twenty-One will be a merge of these various lexical resources.

4 Lexicon acquisition from parallel corpora

Parallel text corpora are large amounts of texts that are available in two or more languages in such a way that they can be considered to be translations of each other. Parallel corpora can

---

[4]Partial translation (noun phrases only) 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. Therefore it is only a fall back option.

be viewed as the implicit storage of all the knowledge about translation relation between words and complex expression that has been put into the translation by the human translator(s). Part of this kind of knowledge is of course also available in a more explicit way, but not never formalized in way that facilitates automation of the translation task. This is due to the fact that there are at least two sources of knowledge required to do translation, namely linguistic knowledge (knowledge about translation relations between words and linguistic constructions) and knowledge about the context of use. For the former type of knowledge a formal representation is possible up to a certain level. It is however a widely acknowledged fact that contextual information is very hard to formalize in a rule-based manner up to a level that allows automatic disambiguation. And without adequate disambiguation, translation results will be poor. So given the lack of explicit translation knowledge various researchers have focused on ways to somehow reconstruct the knowledge that is implicitly available via parallel corpora. This field of research heavily relies on the possibility to apply statistical methods to the analysis of corpora.

4.1 Two kinds of alignment

For the purpose of analyzing Agenda 21 in order to automatically derive the translation knowledge that was used during the creation of the translation, two steps can be distinguished: (i) sentence alignment and (ii) word alignment. The objective of doing sentence alignment is achieving a one-to-one correspondence between the sentences from the corpus. If two sentences can be considered to be translations, than there is probably also a correspondence between the words of these sentences.

Recently much research was done into aligning bilingual corpora at the sentence level [2, 7, 12]. For the development of our tool we used the program published by Gale and Church [7]. The program makes use of the fact that longer sentences tend to be translated into longer sentences, and shorter sentences tend to be translated into shorter ones. Throughout the rest of this section we will use the fact that we know the translation of each sentence in the corpus, but not the translation of the words. The lexicon compilation tool is based on a statistical algorithm called the Expectation Maximisation algorithm (EM-algorithm). The EM-algorithm was proofed to be correct by Dempster, Laird and Rubin in 1977 [4] and was first used to analyze bilingual corpora at IBM in 1990 [1, 3]. The IBM article inspired many research centers over the world to use statistical methods for automatic translation purposes. Our approach [9, 8] contributes to this research area in two ways:

Bi-directionality A version of the EM-algorithm was developed that is able to compile a **bi-directional** lexicon, i.e. a lexicon that can be used to translate from for example English to Dutch as well as from Dutch to English). We believe that there are two good reasons to conduct a bi-directional approach. Firstly, a bi-directional lexicon will need less space than two uni-directional lexicons. Secondly, we believe that a bi-directional approach will lead to better estimates of the translation probabilities than the uni-directional approach.

Application during retrieval The lexicons compiled with the EM-algorithm have been

applied within a document retrieval environment as the basic tool for query translation. In an experiment recall and precision of a monolingual (Dutch) retrieval engine were compared to recall and precision of a bilingual (Dutch-to-English) retrieval engine. The experiment was conducted with the help of eight naive users who formulated the queries and judged the relevance of the retrieved documents

This section is organized as follows. In section 4.2 we will give an informal description of the probability model and the estimation algorithm. Section 4.3 gives a brief description of the conducted experiments in which the bilingual dictionary was the basis for query translation.

4.2   Assigning probabilities to translations

By applying the EM algorithm to a parallel corpus with aligned sentences a *probabilistic* bilingual lexicon is derived. A probabilistic bilingual lexicon is a lexicon with a probability assigned to each possible translation of an entry (see table 1). Such a lexicon can be used both directly as a statistical translation tool or as a information to enhance an existing general purpose MT system with domain dependent translations.

The result of the our tool is a *probabilistic* bilingual lexicon. A probabilistic bilingual lexicon assigns a probability to each possible translation of an entry (see table 1). Suppose we want to derive a probabilistic bilingual lexicon from the parallel corpus of table 2. The corpus consists of four pairs of Dutch and four pairs of English sentences which are each others translation.

he waits.	*hij wacht.*
you wait.	*jij wacht.*
he can.	*hij kan.*
you can.	*jij kunt.*

Table 2: an example corpus

Under a statistical approach the corpus can be viewed as consisting of randomly drawn samples of English-Dutch sentence pairs. Each sentence pair will be called an observation. In the example corpus, there are five different English words and also five different Dutch words. This makes a total of twenty-five possible translations. that can be formalized by a so-called contingency table. Each sentence pair of the corpus of table 2 can be displayed by contingency table of table 3. The cell frequencies n_{ij} in the table represent the number of times the English word i and the Dutch word j are each others translation in the corpus. The marginal totals $n_{i.}$ represent the number of times the English word i appears in the corpus. The marginal totals $n_{.j}$ represent the number of times the Dutch word j appears in the corpus. In terms of cell frequencies n_{ij} the marginal totals are given by:

$$n_{i.} = \sum_{j=1}^{5} n_{ij}, \;\; n_{.j} = \sum_{i=1}^{5} n_{ij} \tag{1}$$

	he	$waits$	you	$wait$	can	
hij	n_{11}	n_{12}	$\cdots$		n_{1c}	$n_{1.}$
$wacht$	n_{21}				$\vdots$	$n_{2.}$
jij	$\vdots$					$\vdots$
kan						
$kunt$	n_{r1}	$\cdots$			n_{rc}	$n_{r.}$
	$n_{.1}$	$n_{.2}$	$\cdots$		$n_{.c}$	$n_{..}$

Table 3: contingency table for the example corpus

Each cell frequency n_{ij} will be assigned an unknown probability parameter p_{ij} which is the probability that the English word i and the Dutch word j appear in the corpus as a translation pair. The unknown parameters p_{ij} form the probabilistic bilingual lexicon we are looking for. Three assumptions must be made in order to finish the translation model. Firstly, it is assumed that the word translation pairs in a sentence pair 'appear' independently of each other. Furthermore a sentence is modelled as a collection of words, i.e. there is no sequence between words or translation pairs of words. Finally we assume that each word in one language is alligned to only one word in the other language, and vice versa. These assumptions lead to the definition of a probability measure P, which is a function of the observations n_{ij} and the parameters p_{ij}

$$P(N = n_{11} \cdots n_{rc}) = \frac{n_{..}!}{n_{11}! \cdots n_{rc}!} \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{n_{ij}} \qquad (2)$$

Equation 2 is the well known multinomial distribution. The estimate $\hat{p}_{ij}$ of p_{ij} that makes the observations as likely as possible is given by

$$\hat{p}_{ij} = \frac{n_{ij}}{n_{..}} \qquad (3)$$

which is the maximum likelihood estimate of the unknown parameters.

Every observation in the parallel corpus must be represented by Table 3. However, the information needed to fill table 3 is not explicitly present in the observations. The observations are *incomplete*, i.e. the marginal totals $n_{i.}$ and $n_{.j}$ of the cell frequencies n_{ij} are known, but the cell frequencies themselves are unknown. Table 4 shows the incomplete observation of the first sentence in the example corpus. For convenience, cell frequencies that are 0 are not displayed.

From the definition of the EM-algorithm [4] the following iterative solution can be constructed.

(i) Take an initial estimate of the probability parameters.

(ii) *Expectation-step*: For each sentence, calculate $E(N|n_{1.} \cdots n_{r.}, n_{.1} \cdots n_{.j}, p_{11} \cdots p_{rc})$, the expected cell frequencies given the marginal totals and the probability parameters.

	he	waits	you	wait	can	
hij	?	?	-	-	-	1
wacht	?	?	-	-	-	1
jij	-	-	-	-	-	0
kan	-	-	-	-	-	0
kunt	-	-	-	-	-	0
	1	1	0	0	0	2

Table 4: incomplete observation of (*he waits*, *hij wacht*)

(*iii*) *Maximisation-step*: Add the expected observations and calculate the maximum likelihood estimate as defined by equation 3.

(*iv*) Repeat (*ii*) and (*iii*) until the probability parameters do not change significantly anymore.

If no linguistic knowledge is used, initially every word pair is equally likely as a translation. For the example corpus of table 2 the initial estimate then must be $p_{ij} = \frac{1}{25}$ for each possible i and j. Table 5 and 6 give an impression of the way the algorithm behaves on the simple example

	he	waits	you	wait	can	
hij	1.0	0.5	-	-	0.5	2
wacht	0.5	0.5	0.5	0.5	-	2
jij	-	-	1.0	0.5	0.5	2
kan	0.5	-	-	-	0.5	1
kunt		-	0.5	-	0.5	1
	2	1	2	1	2	8

Table 5: expected complete observation of the corpus in the first iteration

corpus of table 2. After five iterations of the algorithm the parameters of the model do not change significantly anymore. The number of possible complete observations that matches an incomplete observation increases exponentially with the maximum length of both sentences. To be able to calculate the expected complete observation we used an approximation algorithm called *iterative proportional fitting* [9].

4.3 Experimental results

To test the performance of the algorithm, we compiled a bilingual probabilistic lexicon from the parallel corpus consisting of the English and Dutch version of Agenda 21. Only half the corpus was used to derive the lexicon. The other part of the corpus has been kept aise for testing later on. The training corpus consisted of 4664 parallel sentences. With the training

	he	waits	you	wait	can	
hij	2.0	-	-	-	-	2
wacht	-	1.0	-	1.0	-	2
jij	-	-	2.0	-	-	2
kan	-	-	-	-	1.0	1
kunt		-	-	-	1.0	1
	2	1	2	1	2	8

Table 6: expected complete observation of the corpus in the fifth iteration

corpus, a bilingual lexicon was compiled consisting of 3854 English words and 5462 Dutch words. More than 21 million unknown parameters were estimated. Preliminary experiments with cross-language retrieval (in a much less elaborate retrieval environment than the Twenty-One system) show that even simple word-by-word translation via a corpus-based bilingual lexicon is useful. Comparison of mono-lingual Dutch retrieval with cross-language Dutch-to-English retrieval showed an decrease of average precision form 78% to 67%, but a an increase of average recall from 51% to 82%. For more details cf. [8]. To explain the unexpected high recall of cross-language retrieval a closer look ate the bilingual lexicon should be taken Tables 7 and 8 give some examples of the results of the algorithm after six training steps.

local		duurzame	
plaatselijke	0.51	sustainable	0.93
lokale	0.24	unsustainable	0.02
lokaal	0.15	renewable	0.02
plaatselijk	0.09	consumption	0.01
maken	0.01	sustainability	0.01

Table 7: example entries of morphologically related words and synonyms

The six most probable translations of the entry are displayed, together with the probability of each possible translation. The (*null*) token represents the fact that the word was not translated at all in the corpus. Table 7a and 7b show examples of how the algorithm handles morphologically related words and synonyms. Morphology and synonyms often get special attention in information retrieval systems. The richer Dutch morphology and the relatively frequent use of synonyms in the Dutch part of the corpus will lead to an increase of the average recall, but will not effect the average precision of cross-language retrieval.

Table 8a and 8b show example entries of translations that cannot be modeled very well by the approach taken in this paper. In Dutch, nouns can be compounded to form new words. For example the Dutch word *volksgezondheid* is a compound noun and should be translated as *people's health*. Because nouns are usually not compounded in English, the algorithm will

volksgezondheid		health	
health	1.00	gezondheid	0.28
		gezondheidszorg	0.20
		volksgezondheid	0.11
		gezondheidsprobl.	0.05
		gezondheids	0.04
		te	0.02

Table 8: example entries of compound nouns

find only a partial translation. Partial translations will lead to an increase of average recall, but will lead to a decrease of average precision. Because of the limitation of our domain the effect on the precision is less severe than the effect on recall

4.4 Plans

Currently the possibility is investigated to automatically compile dictionaries for multi-word expressions, or in other words, to take context into account. Other improvements are expected from the incorporation of the morphological processing tools from Xerox that are used within the Twenty-One disclosure modules as well. In particular the compound splitter is expected to have positive effects. The expectation is that the approach described here can contribute substantially to the quality of the term translation tool described in the previous section.

5 Discussion and Concluding remarks

The research presented in this paper proofs that, as long as a parallel corpus is available on the application domain, it is relatively simple to automatically compile bilingual dictionaries. This is a promising result in view of the problems developers of retrieval systems with CLIR functionality often encounter when looking for proper multilingual resources. The alert reader will probably notice that relying on lexicon compilation tools may solve the lexicon acquisition problem, but that at the same time introduces the problem of getting parallel corpora available. Indeed here is a problem that deserves some attention. But as stated at the beginning, there is this enormous increase of information that is made available in electronic form and it is to be expected that more and more organizations will operate with multilingual scope. If making their parallel documents available will bring them the profit of having them disclosed at lower costs, this problem may be solved soon enough.

References

[1] P.F. Brown, J.C. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, and P.S. Roossin R.L. Mercer. A statistical approach to machine translation. *Computational*

Linguistics, 16(2):79–85, 1990.

[2] P.F. Brown, J.C. Lai, and R.L. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting for the Association for Computational Linguistics*, pages 169–176, Berkeley CA, 1991.

[3] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. The mathematcis of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):169–176, 1993.

[4] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em-algorithm plus "discussions on the paper". *Journal of the Royal Statistical Society*, 39(B):1–38, 1977.

[5] 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.

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

[7] W.A. Gale and K.W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102, 1993.

[8] Djoerd Hiemstra. Deriving a bilingual lexicon for cross language retrieval. In *Proceedings of Gronics'97*.

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

[10] 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.

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

[12] Kay and M. Röscheisen. Text-translation alignment. *Computational Linguistics*, 19(1):121–142, 1993.

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