Building Bilingual Dictionaries From Parallel Web Documents

Craig J.A. McEwan[1], Iadh Ounis[1] and Ian Ruthven[2]

[1] Department of Computing Science, University of Glasgow, G12 8QQ
[2]Department of Computer and Information Sciences, University of Strathclyde, G1 1XH

Abstract. In this paper we describe a system for automatically constructing a bilingual dictionary for cross-language information retrieval applications. We describe how we automatically target candidate parallel documents, filter the candidate documents and process them to create parallel sentences. The parallel sentences are then automatically translated using an adaptation of the EMIM technique and a dictionary of translation terms is created. We evaluate our dictionary using human experts. The evaluation showed that the system performs well. In addition the results obtained from automatically-created corpora are comparable to those obtained from manually created corpora of parallel documents. Compared to other available techniques, our approach has the advantage of being simple, uniform, and easy-to-implement while providing encouraging results.

1 Introduction

The content of the Internet is changing from being mainly in the English language to being multi-lingual [11]. At the moment English speakers are the largest group of Internet users, but the number of non-English speaking Internet users is increasing rapidly. For example, it is estimated that by 2005, over 70% of the online population will be non-native English speakers [6].

The Internet is therefore becoming an important source for multi-lingual information, necessitating the development of effective multi-lingual information access tools. This paper describes the development of a system for automatically creating bilingual dictionaries to support these information access tools. The bilingual dictionary can then be put to a variety of uses including Cross-Language Information Retrieval (CLIR) [7]. Furthermore, we examine the potential of using the web as a source of parallel translated documents for the automatic construction of bilingual dictionaries. If the web can be used as a source for parallel documents, then it will allow the development of low-cost, but high quality, translation systems for CLIR.

Our system is composed of three components, comprising three distinct and independent stages. Firstly a collection stage sends a query to a search engine and retrieves the documents from the search engine results links. The second stage uses the HTML tags of the web documents to filter and align the English and Spanish text into parallel sentences. The final stage involves the translation of the words from the parallel blocks. This is achieved by finding word pairs that co-occur in many sentences. The translation stage also incorporates the construction of the dictionary itself. The languages chosen for this implementation are English and Spanish because of the availability of expert evaluators, but the system can be adapted for use with any pair of languages.

Our intention is to provide a system that will automatically cover the whole construction of a dictionary from the initial gathering of parallel documents to the translation of words. However, we must ensure that the documents collected automatically are of sufficient quality. Hence we compared the techniques for creating a dictionary on two sets of data: an automatically collected corpus of parallel web documents of unknown quality and a manually collected corpus of parallel documents that are good quality translations. The evaluation of the two dictionaries created indicates that the automatic corpus produces a dictionary that is of similar quality to the dictionary produced by the manual corpus. This result requires further investigation, but indicates that it may be possible to generate good quality bilingual dictionaries from rapidly collected parallel corpora of unknown quality.

The paper is structured as follows. In section 2 we briefly summarise earlier studies and discuss how our work relates to these. Section 3 discusses the data we collected to construct our dictionary and the means by which we collected the data. In sections 4 and 5 we deal with how we process the documents and in section 6 we discuss how we construct the bilingual dictionary. In section 7 we describe the evaluation of the system and in sections 8 and 9 we conclude with a discussion of our approach and options for future work.

2 Related Work

The idea of building bilingual thesaurus structures using parallel or comparable texts (i.e. comparable on the basis of the similarity between the topics they address [12]) is not new. Comparable texts are usually easier to find or build than parallel texts (i.e. translation equivalent). However, they require appropriate alignment tools to extract cross-language equivalencies. Sheridan and Ballerini [17] created a multilingual similarity thesaurus by aligning news stories from the Swiss news agency (SDA) by topic label and date, and then merging them to build the similarity thesaurus. The alignment process used by Picchi and Peters [13] relies on some contextual information derived from a multilingual machine readable dictionary (MRD). The bilingual MRD is used to establish the links between contexts over languages. The above approaches do not necessarily apply to all pair of languages. Moreover, they are corpus-based techniques and as such they tend to be very application-dependent.

Parallel texts have been used in several studies on CLIR [2] [5] [8] [18]. In [8], the Latent Semantic Indexing reduction technique has been applied on a relatively small parallel text (i.e. translation equivalent) collections in English with French, Spanish, Greek and Japanese. The effectiveness of this approach has not been demonstrated on large collections of data. In [18], a corpus-based bilingual term-substitution thesaurus, called EBT was proposed. In [2], a thesaurus has been constructed from parallel texts using co-occurrences information. QUILT [5] integrates traditional, glossary-based machine translation technology with IR approaches into a Spanish/English CLIR system. These approaches use parallel collections that are domain-specific and/or costly to obtain. In fact, one of the problems with using parallel texts is the difficulty to find cheap, available, generic, large and reliable parallel texts.

Recently, there have been some attempts to collect cheap parallel texts from the Web. Resnik [14][15] was among the first researchers to investigate methods to collect parallel/translated text from the Web. He uses queries to the AltaVista search engine together with HTML tags to detect potential candidate documents on the Web. His approach can be seen as a filtering process allowing identification of high quality syntactically similar translated documents. Indeed, he did not look into the issue of building a bilingual dictionary from the collected corpus, nor did he investigate the alignment process that would statistically allow such a dictionary to be built. Chen [3], Chen and Nie [4] and Nie et al., [9] all addressed the issue of CLIR using the Web as a source of parallel documents. Their approach was to use a probabilistic translation model based on a training corpus made of parallel documents automatically collected from the Web.

Our approach uses a rather simple but uniform approach for both alignment and translation. We use a simplistic alignment algorithm that only uses the characteristic of the HTML markup in Web documents. For the translation stage, we use: a refinement of the well-established IR EMIM measure for defining the strength of relationships between translated words (instead of using a probabilistic approach [1]). The use of the EMIM technique allows a more accurate interpretation of the co-occurrences information obtained from parallel texts, making it more interesting than the rough co-occurrence technique used in [2]. Moreover, our approach does not need tuning or any other classical pre-operations, as no probabilities have been used. Therefore, like the methodology proposed by Nie et al. [9], it could be seen as a generic methodology for building bilingual dictionaries from the Web, while being cheaper/simpler/ and easier-to-implement. Moreover, it still provides very encouraging results.

[1] EMIM measures are based on a function that is monotonic to a probabilistic measure. This function avoids the need to estimate probabilities directly, instead it uses values based on the absence or presence of terms in sentences.

3 Collection

We collected two corpora of parallel documents. One corpus was collected manually by finding and comparing parallel documents, and a second corpus was collected automatically by sending a query to the AltaVista search engine.

The manual corpus was assembled by searching bilingual websites for appropriate documents. An example of the websites reviewed to collect documents for the manual corpus is the European Union website[2]. Parallel documents in English and Spanish from a variety of websites were assessed by bilingual humans for their suitability for inclusion in the manual corpus. Only the text in the parallel documents was assessed, the HTML code of the documents was not considered.

For the automatic collection we tested several different queries to automatically download candidate pair pages in order to determine which query generated the highest number of good quality candidate pairs for the automatic corpus. These queries look for links or anchors from an initial page to its translation page. A query containing '*anchor:spanish version'* will search for pages containing the text 'Spanish version' within HTML anchor tags (Fig. 1).

Additionally, web page authors often use abbreviations for different languages – *en* is the commonest abbreviation for English and *es* is the typical abbreviation for Spanish. Using a query of the form: *'link:*_es.html'* to search for links which include the abbreviation *es.html* in the URL of the Spanish translation page was therefore tried as another method of finding and downloading parallel pages.

However, web page authors use many other abbreviations to identify Spanish pages and the queries for links that end with *'es.html'* encountered many links which were not related to language differences – for instance *_es.html* was frequently used by Environmental Science departments to identify their pages.

After assessing different possibilities, the automatic corpus was collected using the query '*anchor:spanish version'* and searching English pages because this combination produced the least number of erroneous links together with the highest number of result URL addresses. This query finds and downloads parallel pages asymmetrically (Fig. 1). The query searches for web pages in English that have a link containing the text 'Spanish version'. No check is made on the Spanish page to ensure that it has a corresponding link back to the English page.

Using this approach, a parallel Spanish page is not located for each English document. The reason for the lower number of Spanish pages collected is the variety of different file path possibilities used by web authors to store their Spanish version files which could not be handled by the heuristics employed in the system.

[2] http://europa.eu.int/index_es.htm

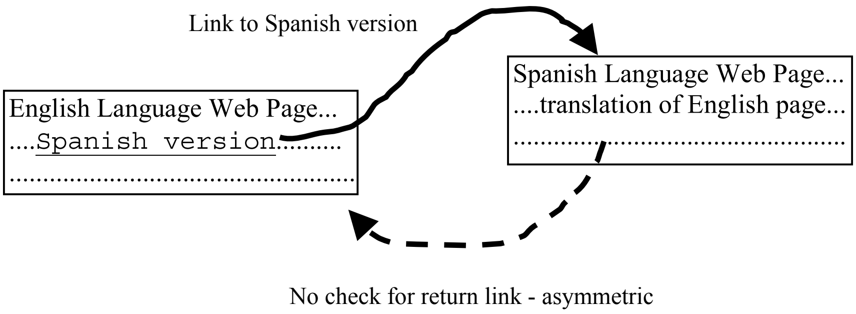

Link to Spanish version
English Language Web Page...
....Spanish version.........
Spanish Language Web Page...
....translation of English page...
No check for return link - asymmetric

Fig. 1. Model of web page links used to collect the automatic corpus

To increase the number of candidate pair pages collected by the system, a different algorithm could be used. There are several different possible ways of doing this. For example, an intelligent crawler could be used to mine through the directory structure of websites where a high concentration of multi-lingual documents occur.

Alternatively, by using a symmetrical approach (Fig. 2), it would be possible to download parallel documents which do not have direct links between them. The query would look for pages with anchors containing the text 'English version' AND 'Spanish version'. The links to the respective versions would be extracted and threads sent to download the candidate pair of pages. Either of these techniques would increase the likelihood of obtaining pairs of documents that were translations of each other [14].

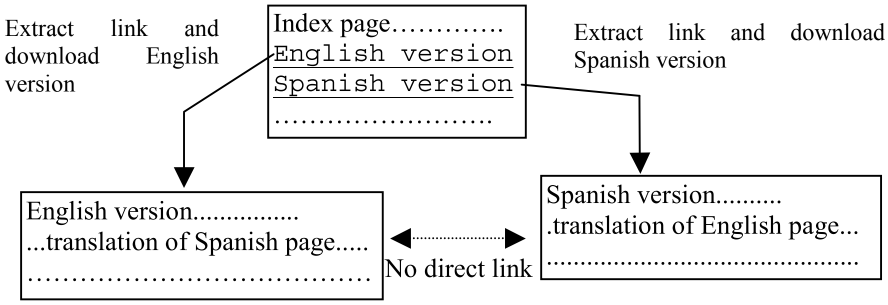

Extract link and download English version
Index page............
English version
Spanish version
....
Extract link and download Spanish version
English version................
...translation of Spanish page.....
No direct link
Spanish version..........
.translation of English page...

Fig. 2. Symmetrical download model of web page links

Once we have targeted documents that are possible translations of each other – the candidate pair documents - we need to process the documents. This involves filtering the documents to eliminate documents that are not likely to be translations, section 4, and then to align the text that is to be used for creating the dictionary, section 5.

4 Filtering

After the collection of candidate pair documents has been completed, the candidate document pairs are filtered to ensure that they have a reasonable chance of being translations of one another.

Several filters are used:

i. *language* filters to prevent documents being classified as belonging to the wrong language, section 4.1,

ii. *length* filters to ensure that parallel documents are of approximately similar length, section 4.2,

iii. *structural* filters to test whether the HTML mark up code of parallel documents are similar, section 4.3.

4.1 Language filtering

The first filter for the candidate pair documents is a language check. The document text is compared against a list of stop words in the language that the document is supposed to contain. For example, English language documents are compared against a list of English stop words, and Spanish documents are compared against a list of Spanish stop words. This stage, then, eliminates documents that have been misclassified as belonging to English or Spanish.

The stop word lists themselves have been checked to ensure that no words with the same spelling occur in both the English and Spanish document. This is done to prevent an English document being recognised as Spanish and vice versa. Examples of the words which were removed are '*he*' – pronoun for a male in English, but also first person conjugation of the verb '*haber*' - to have - in Spanish.

If both documents in the pair contain a word from the stop word list of their respective languages, they are assumed to be in the correct language and progress to the length check filter.

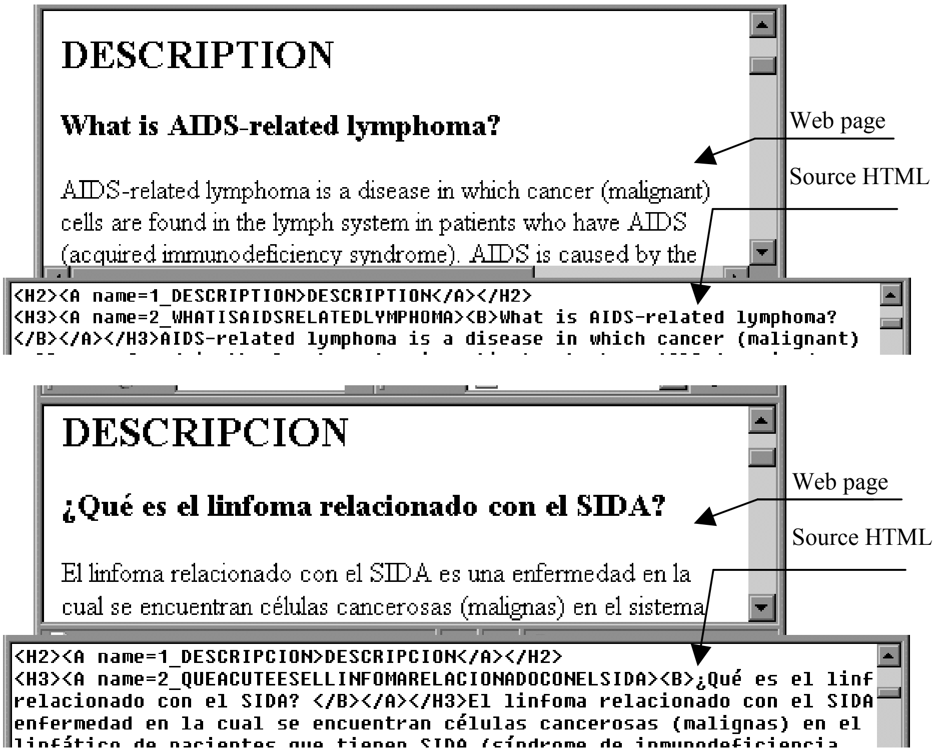
4.2 Length filtering

A length filter is used since it is assumed that very long documents will not be translations of very short documents and vice versa [10]. To determine quantitative parameters for the length filter, 10 pairs of parallel documents of varying lengths were selected at random from the manual corpus. These documents were stripped of their HTML code and the number of words counted. The word counts of these documents showed that the Spanish versions of the documents varied between 1.02 and 1.42 times the length of the English versions.

For the initial runs of the length filter, the system uses 0.9 as the minimum length factor and 1.5 as the maximum length factor. That is, to be considered as a translation, a Spanish document cannot have less than 0.9 times the number of words in its English pair document, nor more than 1.5 times the number of words in its English pair. This is an approximation to filter out candidate pairs of documents that have widely differing lengths, a further length check is done at the sentence alignment stage.

4.3 Structural filtering

The main advantages of using web documents to build a parallel corpus is that they are part of a large and continually growing source of translated documents which contain HTML mark-up code. The filtering and alignment, section 5, processes assume that parallel documents will have very similar HTML mark up code around the translated text (Fig. 3).



DESCRIPTION

What is AIDS-related lymphoma?

AIDS-related lymphoma is a disease in which cancer (malignant) cells are found in the lymph system in patients who have AIDS (acquired immunodeficiency syndrome). AIDS is caused by the

```
<H2><A name=1_DESCRIPTION>DESCRIPTION</A></H2>
<H3><A name=2_WHATISAIDSRELATEDLYMPHOMA><B>What is AIDS-related lymphoma?
</B></A></H3>AIDS-related lymphoma is a disease in which cancer (malignant)
```

DESCRIPCION

¿Qué es el linfoma relacionado con el SIDA?

El linfoma relacionado con el SIDA es una enfermedad en la cual se encuentran células cancerosas (malignas) en el sistema

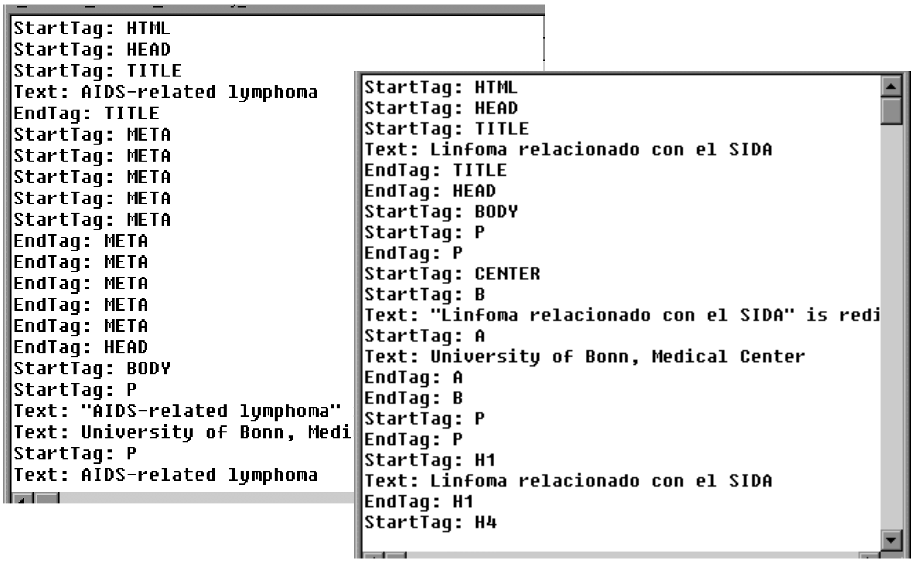
```
<H2><A name=1_DESCRIPCION>DESCRIPCION</A></H2>
<H3><A name=2_QUEACUTEESELLINFOMARELACIONADOCONELSIDA><B>¿Qué es el linf
relacionado con el SIDA? </B></A></H3>El linfoma relacionado con el SIDA
enfermedad en la cual se encuentran células cancerosas (malignas) en el
linfático de pacientes que tienen SIDA (síndrome de inmunodeficiencia
```

Fig. 3. Web documents and the source HTML code for two parallel translated texts. Note the similar appearance of the web pages and the similarity of the HTML source code for both pages. The text contained in each page is a high quality translation of the other.

Once the system has completed the length filtering it applies a structural filter. Structural filtering uses the HTML tags around the text of the candidate pair documents to test whether the documents are *sufficiently* similar to be considered as parallel translated documents. This approach has been successfully applied to align English: French, English: Spanish and English: Chinese bilingual corpora [3][4][14].

This process is called ‘linearisation’ [14]. Examples of linearised English and Spanish documents are shown below (Fig. 4).

Once we have the linear sequences of tags and text the system can align the text contained within the tags. We discuss this in the next section.



```
StartTag: HTML
StartTag: HEAD
StartTag: TITLE
Text: AIDS-related lymphoma
EndTag: TITLE
StartTag: META
StartTag: META
StartTag: META
StartTag: META
StartTag: META
EndTag: META
EndTag: META
EndTag: META
EndTag: META
EndTag: META
EndTag: HEAD
StartTag: BODY
StartTag: P
Text: "AIDS-related lymphoma"
Text: University of Bonn, Medi
StartTag: P
Text: AIDS-related lymphoma
```

```
StartTag: HTML
StartTag: HEAD
StartTag: TITLE
Text: Linfoma relacionado con el SIDA
EndTag: TITLE
EndTag: HEAD
StartTag: BODY
StartTag: P
EndTag: P
StartTag: CENTER
StartTag: B
Text: "Linfoma relacionado con el SIDA" is redi
StartTag: A
Text: University of Bonn, Medical Center
EndTag: A
EndTag: B
StartTag: P
EndTag: P
StartTag: H1
Text: Linfoma relacionado con el SIDA
EndTag: H1
StartTag: H4
```

Fig. 4. Linear sequence of tags and text for an English and Spanish parallel document pair from the manual corpus. Note that although the pattern of tags and text is similar, it is not identical. In this example, the English language page (Fig. 4, left hand side) has a number of META tags which do not appear on the Spanish language page (Fig. 4, right hand side).

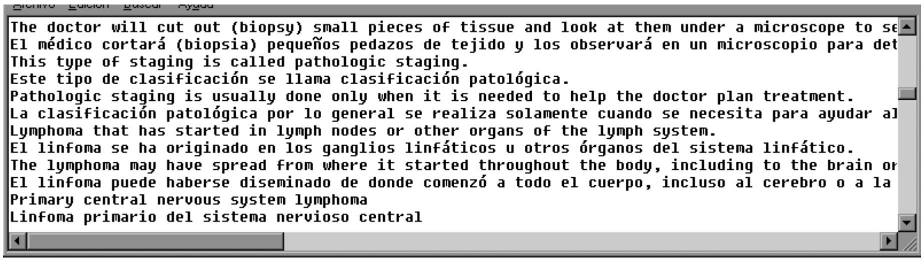
5 Alignment process

After filtering, the sentences contained within one document are aligned with their translations in the parallel document. In section 5.1, we describe how text is aligned and, in section 5.2, we describe the results of the alignment process on our corpora.

5.1 Aligning text blocks

The linear sequence of tags and text for the English language document is compared with the linear sequence of tags and text from the Spanish language document. Web authors may use identical HTML code around the text in parallel translated documents, but this is uncommon even in sites of governmental organisations. It is much more common to have HTML code which is broadly similar but not identical around the parallel texts.

The alignment process relies on matching the HTML tags of the text in the two languages. To quantify the alignment, matching <Start>, <End> and <Text> tags in both languages are counted. In addition, since longer sentences in one language will translate to longer sentences in another language [10], a sentence level word count ensures that short sentences are not aligned against long ones (Fig. 5). Where a <Text> tag in the English document of a pair does not align with a <Text> tag in the Spanish document of the pair, the system searches for the next <Text> tag in the Spanish document.



```
The doctor will cut out (biopsy) small pieces of tissue and look at them under a microscope to se
El médico cortará (biopsia) pequeños pedazos de tejido y los observará en un microscopio para det
This type of staging is called pathologic staging.
Este tipo de clasificación se llama clasificación patológica.
Pathologic staging is usually done only when it is needed to help the doctor plan treatment.
La clasificación patológica por lo general se realiza solamente cuando se necesita para ayudar al
Lymphoma that has started in lymph nodes or other organs of the lymph system.
El linfoma se ha originado en los ganglios linfáticos u otros órganos del sistema linfático.
The lymphoma may have spread from where it started throughout the body, including to the brain or
El linfoma puede haberse diseminado de donde comenzó a todo el cuerpo, incluso al cerebro o a la
Primary central nervous system lymphoma
Linfoma primario del sistema nervioso central
```

Fig. 5. An example of an aligned text file. English and Spanish sentences alternate. Long English sentences align with long Spanish sentences.

If the aligned text strings are similar in length, sentences within the text blocks are identified by searching for full stop characters '.'. One English sentence is then aligned against one Spanish sentence. In this system, it is assumed that one sentence will be translated into one sentence since this occurs in about 90% of sentences in parallel documents [10]. Untranslated sentences, or one sentence translating to 2 sentences account for the remaining 10% of sentences in the parallel documents.

5.2 Results of filtering and alignment

In this section we discuss the results of the filtering and alignment process on our two corpora; the automatically retrieved and manually created sets of parallel documents.

Of the 423 candidate pairs collected by the automatic system, 105 pairs passed the three filtering steps described in section 3.

Candidate English and Spanish pairs which do not have a high level of HTML tag matching are discarded by the system. Currently, the threshold for matching tags is set to 60%. That is, 6 out of every 10 English and Spanish lines must have identical HTML tags to be considered translations else the candidate pairs are discarded.

Of the 105 files which passed the language and length filters, 33 were discarded because they fell below the alignment threshold. This leaves 72 aligned text files from the original 423 pairs collected by the automatic collection system.

A corpus of 41 parallel pairs of web pages was collected manually – that is by reading and reviewing both the English and Spanish versions of the documents. If the translation was a good one, the document was included in the manual corpus.

The manual corpus was also filtered and aligned. Of the 41 pairs, 37 pairs passed the language and length filtering stage and of these only 2 were discarded because they fell below the alignment threshold (Fig. 6).

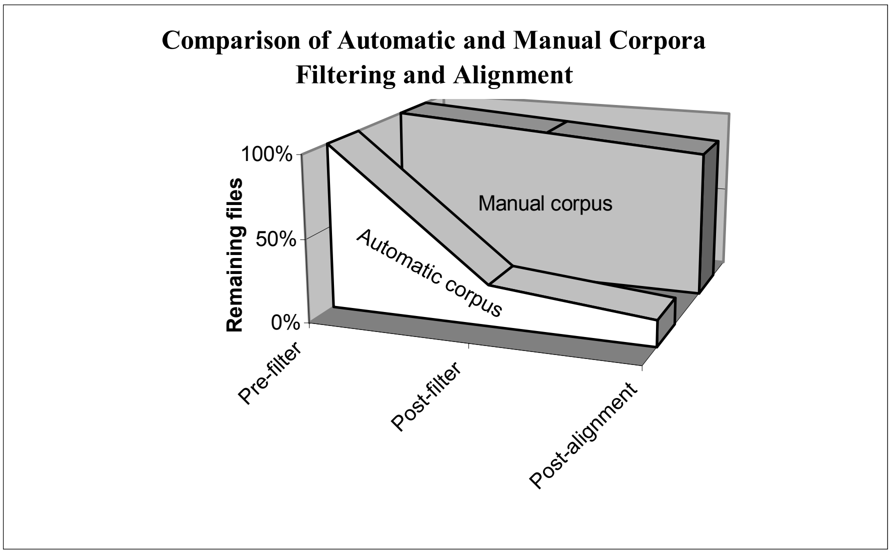


Fig. 6. A graphical comparison of the number of files passing all of the filtering

The high percentage of manual corpus files (88%) which pass filtering and alignment criteria compared with the low percentage (17%) of automatic corpus files which pass all of the filtering and alignment criteria is interpreted to reflect the quality of the translations of the corpora. The manual corpus is a collection of high quality parallel documents that have good translations and very similar HTML code around the text.

These documents were collected from university and governmental websites. The automatic corpus is a collection of web pages from a wide variety of sources. The quality of the translations of the parallel web pages varies from good to poor, and the HTML code around the text is often very different between parallel pages. This results in a low number of files passing all of the filtering and alignment criteria.

The threshold levels for file and sentence length as well as the alignment threshold may be adjusted to allow a greater or lesser number of files pass. Testing the system with different threshold levels for these variables combined with an evaluation of the final bilingual dictionary would be the best way to improve the overall system performance.

6 Building a Dictionary

Once the documents and sentences have been filtered and aligned the system can translate the terms in the sentences. The principal behind the automatic translation of terms is simple – if an English term and a Spanish term both occur in many translated parallel sentences, then the probability that they are translations of one another is higher than an English term and a Spanish term which do not co-occur in many sentences. Automatic construction of thesauri using statistical techniques is a widely used Information Retrieval technique [3][4][16].

The dictionary building stage is divided into three steps; building a matrix of words, section 6.1, normalising the raw co-occurrence scores in the matrix, section 6.2, lastly making a dictionary listing by extracting the Spanish terms with the highest co-occurrence probability for each English term, section 6.3.

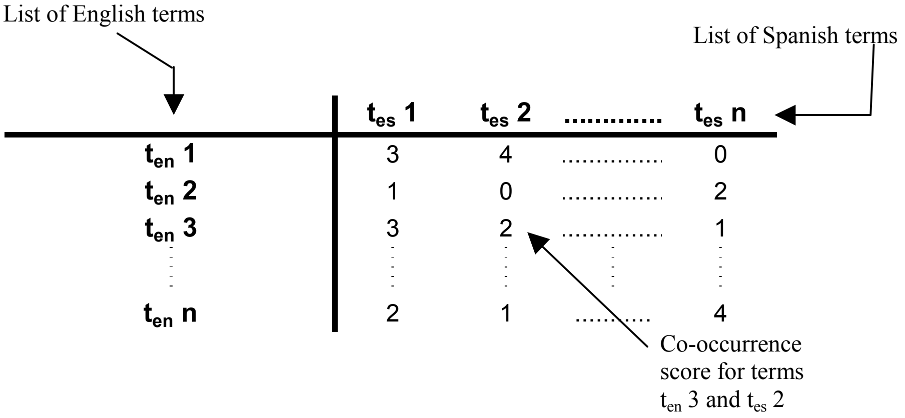
6.1 Building a matrix of English and Spanish words

The assumption was made in the filtering and alignment stages that a single sentence in English will be translated to a single Spanish sentence. To build the matrix of English and Spanish terms, it is further assumed that a single English term will translate to a single Spanish term. This is clearly not the case for many English and Spanish words, but it is a simplifying assumption that allows us to create a first implementation of our techniques.

Our approach to translating English to Spanish terms is based on statistical co-occurrence techniques. These, in our implementation, depend on the creation of a co-occurrence matrix which shall be described in the remainder of this section.

The word matrix can be imagined as a huge spreadsheet (Fig. 7).

The matrix itself is constructed as follows. For each word in an English sentence, it is assumed that the translation of the word is one of the Spanish terms in the parallel Spanish sentence. Therefore for each English term in the sentence, the co-occurrence score with every term in the parallel Spanish sentence is incremented by one.



List of English terms

List of Spanish terms

	t_{es} 1	t_{es} 2	....	t_{es} n
t_{en} 1	3	4	....	0
t_{en} 2	1	0	....	2
t_{en} 3	3	2	....	1
⋮	⋮	⋮	⋮	⋮
t_{en} n	2	1	....	4

Co-occurrence score for terms t_{en} 3 and t_{es} 2

Fig. 7. A schematic view of a word matrix. Each cell in the matrix contains the number of times an English term co-occurs with a Spanish term. t_{es}i is the *i*th Spanish word, $t_{en}i$ is the *i*th English word.

We shall illustrate this process below (Fig. 8a-e), using two English sentences 'The dog runs.' and 'The happy dog jumps.' and their Spanish translations '*El perro corre.*' and '*El perro feliz salta.*'.

The stopwords are removed from the sentences leaving 'dog runs' and 'happy dog jumps' and the Spanish versions '*perro corre*' and '*perro feliz salta*'.

	perro	corre
dog	1	1

Fig. 8a. Constructing a word matrix Step 1. After removing English and Spanish stopwords, the first English term 'dog' is added to the matrix with all the remaining Spanish terms in the parallel sentence '*perro corre*' and the co-occurrence score is incremented for each word pair.

	perro	**corre**
dog	1	1
runs	**1**	**1**

Fig. 8b. Step 2. The remaining term in the English sentence is added. The Spanish terms in the parallel sentence are already in the matrix, so only the co-occurrence scores for the new word pairs are incremented

	perro	corre	**feliz**	**salta**
dog	1	1	0	0
runs	1	1	0	0
happy	**1**	0	**1**	**1**

Fig. 8c. Step 3. The first term of the second English sentence 'happy dog jumps' is added to the matrix with the Spanish terms from the parallel sentence '*perro feliz salta*'. Since '*perro*' is also already in the matrix, the other new terms '*feliz*' and '*salta*' are added to the matrix and then all of the co-occurrence scores are incremented to 1.

	perro	corre	**feliz**	**salta**
dog	**2**	1	**1**	**1**
runs	1	1	0	0
happy	1	0	1	1

Fig. 8d. Step 4. The next English term in the second sentence, 'dog' is added to the matrix with the Spanish terms from the parallel sentence '*perro feliz salta*'. Since all of the English and Spanish terms are already in the matrix, the co-occurrence scores for the English term and all the Spanish terms are incremented.

	perro	corre	**feliz**	**salta**
dog	2	1	1	1
runs	1	1	0	0
happy	1	0	1	1
jumps	**1**	0	**1**	**1**

Fig. 8e. Step 5. The final term in the second English sentence 'jumps' is added to the matrix and the co-occurrence scores with the terms in the parallel Spanish sentence are incremented.

From the illustrations above (Figs 8a – e) it is clear that the English term ‘dog’ and the Spanish term ‘*perro*’ have a higher co-occurrence score than the other word pairs in the matrix. It is therefore more likely that the English term ‘dog’ is translated to ‘*perro*’ than ‘*corre*’, ‘*feliz'* or ‘*salta*’.

When terms from many sentences are added to a matrix, the co-occurrence scores for all of the word pairs in the matrix increment and the contrast between different terms increases.

This trivial example highlights a major drawback with the approach. That is that nouns are likely to be associated with adjectives – words like ’happy’ and with verbs – words like ‘runs’. In order to distinguish between closely related words, the co-occurrence scores need to be normalised. We shall discuss this in the next section.

6.2 Normalising the co-occurrence scores

Normalising the co-occurrence scores is necessary to be able to distinguish between closely related terms in the lists of English and Spanish words. We used the Expected Mutual Information Measure (EMIM) [16] to calculate the degree of association between an English term and a Spanish term in a word pair in the matrix.

The EMIM measure was specifically suggested [16] as a means of calculating term dependencies within a document collection. In our system we re-interpret it for use in calculating how likely a term in one language is to be a translation of a term in another language.

An EMIM score is calculated for each word pair in the matrix, e.g. the terms ‘*perro'* and ‘*dog'* (see Fig. 8). The EMIM score is based on values contained within the contingency table shown in (Fig. 9). This contains four main pieces of information regarding the two terms:

i. how often both terms co-occur, i.e. how often two aligned sentence contain the terms, value (1) in Figure 9
ii. how often one term occurs in a sentence and the other term does *not* occur in the aligned sentences, values (2) and (3) in Figure 9
iii. how often *neither* term occurs in the set of aligned sentences being investigated. This count measures how rare the combination of terms are within the set of aligned sentences, value (4) in Figure 9.

	Spanish term t_{es} j present	Spanish term t_{es} j not present	
English term t_{en} i present	(1)	(2)	(7)
English term t_{en} i not present	(3)	(4)	(8)
	(5)	(6)	(9)

Fig. 9. Contingency table to calculate EMIM values.

The values required to calculate the EMIM scores are obtained from the matrix in the following way:

(1) – matrix score t_{en} i, t_{es} j
(2) – the difference between the maximum score and the matrix score for t_{en} i ((7)-(1))
(3) – the difference between the maximum score and the matrix score for t_{es} j ((5)-(1))
(4) – the part of the total score which is not from either t_{en} i or t_{es} j ((6)-(2) or (8)-(3))
(5) – maximum co-occurrence score for term t_{es} j
(6) – difference between twice the matrix maximum and the t_{es} maximum ((9)-(5))
(7) – maximum co-occurrence score for term t_{en} i
(8) – difference between twice the matrix maximum and the t_{en} maximum ((9)-(7))
(9) – twice the highest co-occurrence score in the matrix.

The EMIM score itself for each word pair is calculated using the following equation:

$$\text{EMIM} = (1)\log\frac{(1)}{(5)(7)} + (2)\log\frac{(2)}{(6)(7)} + (3)\log\frac{(3)}{(5)(8)} + (4)\log\frac{(4)}{(6)(8)} \qquad \textbf{(1)}$$

In this way a number can be assigned to each word pair which is an estimate of the strength of the association between the two terms t_{en} i and t_{es} j. The absolute value of the number is not important, it simply quantifies the association of the two terms t_{en} i and t_{es}j relative to all the other word pairs in the matrix.

It should be noted that the EMIM scores are all **negative** numbers because the denominator of the log term is always greater than the numerator. If the numerator of the log term is 0, then the log term is assigned 0 as its value *e.g.* for the term

$$(x)\log\frac{(x)}{(y)(z)} \ \text{if}\,(x) = 0,\ (x)\log\frac{(x)}{(y)(z)} = 0.$$

None of the denominator terms will be 0 as long as there is at least one word pair in the matrix. Therefore the smaller (more negative) the magnitude of the EMIM value, the greater the degree of normalised co-occurrence between the two terms and the more likely the terms can be regarded as translations of each other.

When the EMIM score has been calculated for each word pair, the original co-occurrence score in the matrix is replaced with the EMIM score.

6.3 Making a dictionary listing

A dictionary listing is made by taking each English term and finding each of the co-occurring Spanish terms that have the minimum and second lowest EMIM scores. A dictionary could also have been made by taking each Spanish term and finding an English term or terms with the minimum EMIM score. The system can be easily adapted to generate either or both types of dictionary listing.

The dictionary list of 1687 English terms was generated from the 35 aligned files of the manual corpus. A list of 1047 English terms was generated from the 72 aligned files of the automatic corpus. In the next section we shall evaluate the quality of the translations and the comparative quality of the translations from the two corpora.

7 Evaluation

We chose to evaluate the dictionaries which were created by counting the number of correctly translated words they contain rather than comparing the process of automatic dictionary construction with the corresponding manual process. If an acceptable dictionary can be constructed using our system, then there is no need to consider the construction process used. In this section we shall first describe *how* we evaluate the created dictionaries, section 7.1, and then present the results of the evaluation, section 7.2.

7. 1 Evaluating the dictionary lists

The initial hypothesis was that the manual corpus would produce a higher quality dictionary than the automatic corpus because at each stage of the collection, filtering and alignment, and translation the manual corpus documents were higher quality than the automatic corpus (Fig. 10).

Specifically, the manual corpus has a higher ratio of Spanish:English files collected, a higher ratio of files passing all of the filtering and alignment criteria and a higher ratio of words in the dictionary list per document in the corpus. All of these indices are taken to indicate that the manual corpus is of a higher quality than the automatic corpus.

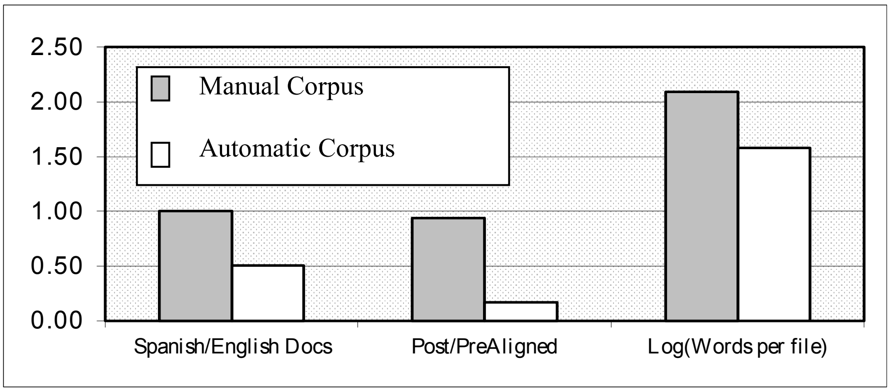


Fig. 10. Comparative statistics for the Manual and Automatic corpora. The histogram indicates that the manual corpus has more Spanish documents per English document than the Automatic corpus, that a far larger proportion of Manual corpus documents passed all of the filtering and alignment stages than was the case for the Automatic corpus, and that on average a document from the Manual corpus provided more words to the dictionary than a file from the Automatic corpus.

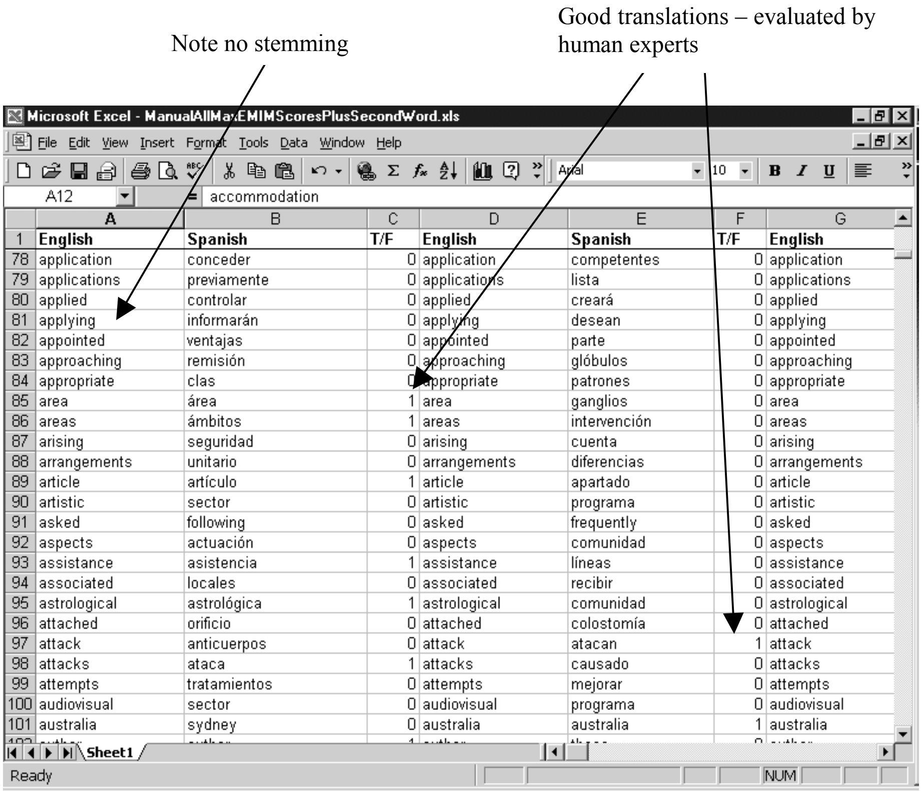
The evaluation experiment consisted of two fluent Spanish speakers reviewing the dictionary listings from both the manual and automatic corpora. These reviewers examined how many correct translations were found in the dictionaries.

For each English term in the listing, if any of the Spanish terms with the minimum or second lowest EMIM score was a good translation of that term, then the count of correct translations was incremented (Fig. 11).

If there was disagreement between the evaluators, a dictionary [1] was used to check the word in dispute.

7.2 Results of the evaluation

Our system was developed incrementally. The initial version included stopwords and did not remove numbers or words of <4 characters from the dictionary list. Only one term with the minimum EMIM score together with one term with the second lowest EMIM score were incorporated in the dictionary listing. Version 2 removed stopwords, but kept short terms (<4 characters) and again, used only single term with the minimum EMIM and second lowest EMIM scores. Version 3 removed stopwords and words with <4 characters, but only included single terms with the minimum EMIM and second lowest EMIM scores in the dictionary. The final version removed stopwords, only allowed words of >4 characters and included all of the terms with the minimum and second lowest EMIM scores in the dictionary listing.



Note no stemming

Good translations – evaluated by human experts

Microsoft Excel - ManualAllMaxEMIMScoresPlusSecondWord.xls

A12 = accommodation

	A	B	C	D	E	F	G
1	**English**	**Spanish**	T/F	**English**	**Spanish**	T/F	**English**
78	application	conceder	0	application	competentes	0	application
79	applications	previamente	0	applications	lista	0	applications
80	applied	controlar	0	applied	creará	0	applied
81	applying	informarán	0	applying	desean	0	applying
82	appointed	ventajas	0	appointed	parte	0	appointed
83	approaching	remisión	0	approaching	glóbulos	0	approaching
84	appropriate	clas	0	appropriate	patrones	0	appropriate
85	area	área	1	area	ganglios	0	area
86	areas	ámbitos	1	areas	intervención	0	areas
87	arising	seguridad	0	arising	cuenta	0	arising
88	arrangements	unitario	0	arrangements	diferencias	0	arrangements
89	article	artículo	1	article	apartado	0	article
90	artistic	sector	0	artistic	programa	0	artistic
91	asked	following	0	asked	frequently	0	asked
92	aspects	actuación	0	aspects	comunidad	0	aspects
93	assistance	asistencia	1	assistance	líneas	0	assistance
94	associated	locales	0	associated	recibir	0	associated
95	astrological	astrológica	1	astrological	comunidad	0	astrological
96	attached	orificio	0	attached	colostomía	0	attached
97	attack	anticuerpos	0	attack	atacan	1	attack
98	attacks	ataca	1	attacks	causado	0	attacks
99	attempts	tratamientos	0	attempts	mejorar	0	attempts
100	audiovisual	sector	0	audiovisual	programa	0	audiovisual
101	australia	sydney	0	australia	australia	1	australia

Sheet1

Ready NUM

Fig. 11. An example of part of the Manual corpus dictionary listing in an Excel spreadsheet. Note that if any of the Spanish is a good translation of the English term, then the count of good translations increments.

The removal of stopwords and short words improved the percentage of correct translations slightly (Figure 12). A larger increase in the percentage of correct translations is seen when all of the terms with the minimum and second lowest EMIM scores are collected in the dictionary listing. Collecting all of these translation terms results in a large increase in the number of translation terms as well as the number of correct translations. For example in the first version a total of 1697 English terms were collected from the manual corpus. For each of these terms, 2 Spanish terms were collected resulting in a total of 3394 Spanish words. A total of 612 English terms had a correct translation in the list of Spanish terms (36.1%).

The precision (defined as $\dfrac{number\_of\_correct\_translations}{number\_of\_Spanish\_terms}$) is 18.0%.

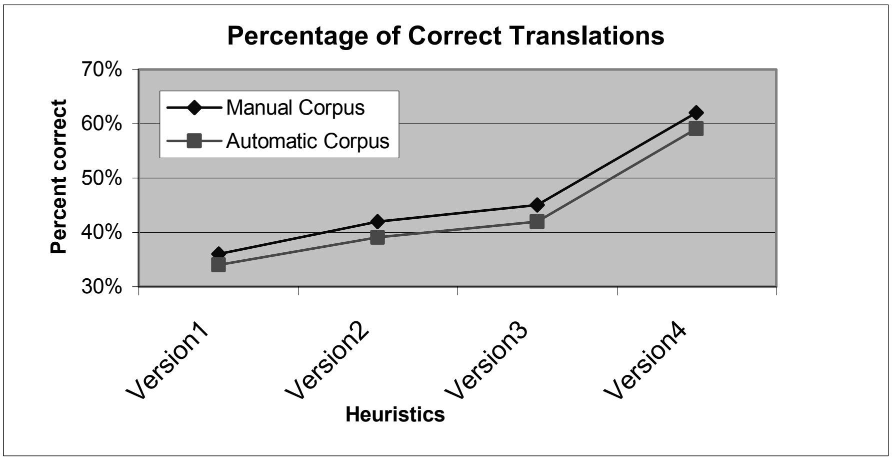


Fig. 12. Graph of the improvement in the percentage of correctly translated English terms with different versions of the system.

In the fourth version, 1688 English terms were collected from the manual corpus. Collecting all of the Spanish terms with either the minimum or the second lowest EMIM score results in the collection of 9136 Spanish terms – a much higher recall than the earlier version. A total of 1048 English terms have a correct translation in the list of Spanish terms (62.1%), but the precision is lower than the earlier versions (11.5%) because of the increase in the number of Spanish terms collected.

There appears then to be some kind of trade-off between number of correct translations and the precision of the translated terms. This balance is similar to the balance between recall and precision that occurs in IR systems.

The results of the evaluation of the final version showed that the manual corpus dictionary contained 1048 good translations out of 1687 English terms which is 62.1% of the total number of terms (Fig. 12). The automatic corpus contains 618 good translations out of 1047 English terms or 59.0% of the total number of terms. It can also be seen that in all of the versions, the percentage of good translations in the manual and automatically collected corpora are about the same (Fig. 12).

This was an unexpected result. As discussed above, we considered that the manual corpora would produce significantly higher quality dictionaries than the automatic corpora. This would be expressed as a higher number of good translations in the manual corpora dictionaries than in the automatic corpora dictionaries.

There are two possible explanations for this – either the alignment in this system is not sophisticated enough to discriminate between high and low quality parallel documents, or it shows that a dictionary can be made by collecting parallel documents from anywhere on the Internet without the need for sophisticated document collection software. A corpus gathered by a quick and simple collection generates a dictionary of similar quality to that of a high quality corpus of parallel documents.

8 Conclusions

The objective of this paper was to design and build a system that would allow the construction of a bilingual dictionary from parallel documents found on the World Wide Web. Any bilingual dictionary created can be put to a variety of uses including Cross Language Information Retrieval (CLIR).

English and Spanish were chosen as the languages for the bilingual dictionary to illustrate our approach. As well as building the dictionaries, an evaluation of the translations contained in the dictionaries was carried out by two bilingual people to assess the quality of the dictionaries produced.

Creating a dictionary requires three distinct and independent steps. Unlike other approaches, which use a combination of techniques, e.g. [3][4][14], our system was a unified system. Firstly a corpus of parallel English and Spanish documents is collected. In this system a query is sent to the AltaVista search engine that then searches for English language web documents containing a link to a ‘Spanish version’. To provide a corpus to compare the automatic collection system with, a corpus of 41 parallel documents was also collected manually.

The second step in the process is filtering the document pairs for length and language to ensure that they can be translations of one another, then the HTML tags of the documents are used to align the English and Spanish text. This process was carried out for both the automatic and manual corpora. Overall, a higher percentage of manual corpus documents (88%) passed the filtering and alignment process than documents from the automatic corpus (17%). This indicates that the manual corpus contains English and Spanish documents whose HTML structure is more alike and whose translations are of better quality.

The third step in the dictionary building process is to use statistical techniques to find translations of each of the English words in the corpora. A large matrix of English and Spanish word pairs is used to determine which English and Spanish words are most closely associated with each other in the corpora. The better the association score between the terms in a word pair, the more likely the words are to be translations of one another. The association scores have been normalised using an adaptation of the EMIM technique. A dictionary listing was produced by taking each English term and all of the Spanish terms with the two best association scores for each English word.

The latest version of our system returned a dictionary list from a manual corpus in which 62% of the English words were translated correctly. The automatic corpus dictionary contained 59% of correct translations.

Overall we have shown that it is possible to build a bilingual dictionary by mining parallel web pages. The percentage of good translations of words in the dictionary is relatively low using the current system parameters, but future work would focus on improving the heuristics used at each stage of the process.

The conclusion that an automatically collected corpus of relatively poor quality parallel documents can generate a dictionary that is as good as a dictionary generated by a high quality corpus is interesting. It raises the possibility that high quality dictionaries can be generated quickly and easily from the Internet without the need for sophisticated collection algorithms such as those used by some workers [3][4].

9 Future Work

The current system uses a simple query that retrieves a Spanish language page for up to 67% of the total number of English language pages collected. This percentage could be improved by collecting English pages with links to Spanish pages which themselves also have links back to the original English page. This would improve the likelihood that the pages are translations of one another.

The filtering and alignment stage could be improved by implementing more rigorous language checks. At the moment, the language filtering procedure leads to many English words being included as Spanish terms and vice versa. Removing some of the English words from the Spanish vocabulary and vice versa would improve the final dictionary. Other refinements to the filtering and alignment could include adjusting the length filters to reduce the chance of non-parallel documents passing this stage.

Once the co-occurrence matrix is built, an iteration of the construction process would allow the terms with the highest co-occurrence scores to be selected over other terms in any given sentence. This could improve the mapping between terms compared with the initial co-occurrence matrix where there was no prior knowledge available. Additionally, the percentage of good translations in the dictionary may be improved if a much larger vocabulary is processed because the contrast between association scores for co-occurring terms would be improved if a larger number of sentences containing the co-occurring terms were processed.

All of these improvements are relatively straightforward to implement, and would allow a better test of the dictionary building system.

References

1. Appleton's New Cuyás English-Spanish and Spanish-English Dictionary 5th edition. 1972.
2. Brown, R.D.: Automatically extracted thesauri for cross-language IR: when better is worse, 1st Workshop on Computational Terminology (Computerm), p15-21, 1998.
3. Chen, J.: Parallel Text Mining for Cross-Language Information Retrieval using a Statistical Translation Model. M.Sc. thesis, University of Montreal, 2000. http://www.iro.umontreal.ca/~chen/thesis/node1.html.
4. Chen, J. and Nie, J-Y.: Parallel Web Text Mining for Cross-Language IR. In Proceedings of RIAO-2000: "Content-Based Multimedia Information Access", Paris, 12-14 April 2000.
5. Davies, M.W. and Ogden, W.C.: QUILT, Implementing a large-scale cross-language text retrieval system, 20th International Conference on Research and Development in Information Retrieval (ACM SIGIR'97), Philadelphia, p92-98, 1997.
6. Global Reach website. http://www.glreach.com/globstats/
7. Grefenstette, G. (ed.): Cross-Language Information Retrieval. Kluwer Academic Publisher, 1998.
8. Littman, M.L., and Dumais, S.T. and Landauer, T.K.: Automatic Cross-language Information Retrieval using Latent Semantic Indexing. In Grefenstette, G. (ed.): Cross-language Information Retrieval, Kluwer Academic Publishers, p51-62, 1998.
9. Nie, J-Y., Simard, M., Isabelle, P. and Durard, R.: Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (ACM SIGIR'99), Berkeley, p74-81. 1999.
10. Oakes, M.P.: Statistics for Corpus Linguistics. Edinburgh Textbooks in Empirical Linguistics. 1998.
11. Oard, D.: Language Distribution of the Web. Web site for Research Resources on Cross-Language Text Retrieval. http://www.clis2.umd.edu/dlrg/filter/papers/
12. Peters, C. and Sheridan, S.: Multilingual Information Access. In M. Agosti, F. Cresti, and G. Pasi (Eds.): Lectures on Information Retrieval/ESSIR 2000, LNCS 1980, pp. 51-80, 2000.
13. Picchi, E. and Peters, C.: Cross-Language Information Retrieval: A System for Comparable Corpus Querying. In Grefenstette, G. (ed.): Cross-language Information Retrieval, Kluwer Academic Publishers, p81-92, 1998.
14. Resnik, P.: Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text. In Proceedings of the AMTA-98 Conference, October, 1998.
15. Resnik, P.: Mining the Web for Bilingual Text. In Proceedings of the International Conference of the Association of Computational Linguistics (ACL-99), College Park, Maryland, 1999.
16. van Rijsbergen, C.J.: Information Retrieval. 2nd Edition. CD-ROM version, 1999. http://www.dcs.gla.ac.uk/Keith/Preface.html
17. Sheridan, P. and Ballerini, J.P.: Experiments in Multilingual Information Retrieval using the SPIDER system. In Proceedings of the 19th International Conference on Research and Development in Information Retrieval (ACM SIGIR'96), Zurich, p58-65. 1996.
18. Yang, Y. and Carbonell, J.G. and Brown, R.D. and Frederking, R.E.: Translingual information retrieval: learning from bilingual corpora, Artificial Intelligence, 103:323-345, 1998.