CoLesIR at CLEF 2006: Rapid Prototyping of an N-gram-Based CLIR System

Jesús Vilares

Departamento de Computación
Universidade da Coruña
Campus de Elviña
15071 - La Coruña (Spain)
jvilares@udc.es

Michael P. Oakes
John I. Tait
School of Computing and Technology
University of Sunderland
St. Peter's Campus, St. Peter's Way
Sunderland - SR6 0DD (United Kingdom)
{Michael.Oakes, John.Tait}@sunderland.ac.uk

Abstract

In this our first joint participation as the CoLesIR group, our team has participated in the Portuguese monolingual ad-hoc task and in all robust ad-hoc tasks —all monolingual tasks, the English-to-German bilingual task, and the multilingual task.

We have developed an n-gram model inspired by the previous work of the Johns Hopkins University Applied Physics Lab. Our approach makes generalized use of freely available resources —such as the Europarl parallel corpus, the GIZA++ word-alignment toolkit, and the Terrier retrieval platform—, and employs a new n-gram direct translation technique. This new technique takes as input previously existing aligned word lists and obtains as output aligned n-gram lists. It can also handle word translation probabilities, as in the case of statistical word alignments.

This new n-gram-based approach shares the main advantages of the original proposal. This solution avoids the need for word normalization during indexing or translation, and it can also deal with out-of-vocabulary words. Since it does not rely on language-specific processing, it can be applied to very different languages, even when linguistic information and resources are scarce or unavailable. Our proposal adds to these characteristics a higher speed during the n-gram alignment process.

Unfortunately, lack of time did not allow us to get our n-gram direct translation system ready on time. This way, we could submit only those initial results to be used as baselines in the future evaluation of our approach.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods, Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation*; J.5 [**Arts and Humanities**]: Language translation

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Cross-Language Information Retrieval, character n-grams, translation algorithms, alignment algorithms, association measures

1 Introduction

CoLesIR group is an interuniversity research group created for joint participation in the CLEF competition. It is composed of members of the Compilers and Languages Research Group (CoLe)[1] of the Universities of A Coruña and Vigo (Spain), and members of the Information Retrieval Group[2] of the University of Sunderland (United Kingdom). Although we have participated separately in CLEF [17, 16, 13, 12], this is our first joint participation.

The Spanish CoLe group has been working for several years on the application of Natural Language Processing (NLP) techniques to Information Retrieval (IR) [18, 17, 16], and recently has entered into the field of Machine Translation (MT) [3]. The possibility of applying its experience in these fields to the field of Cross-Language Information Retrieval (CLIR), and the support given for this purpose by the Sunderland University IR Group, led to the birth of this joint group

So, this paper describes our first research experiences in the field of Cross-Language Information Retrieval. A CLIR system based in the employment of n-grams not only as indexing units, but also as translating units, is presented.

The article is outlined as follows. Section 2 presents previous work on the application of n-grams to CLIR systems. Next, section 3 describes our n-gram-based CLIR system. Section 4 shows the results obtained in our participation in both the Portuguese monolingual and robust tasks of the CLEF 2006 ad-hoc track. Unfortunately, lack of time did not allow us to get our n-gram direct translation system ready on time. This way, we could only submit those initial results to be used as baselines for the future evaluation of our approach. Finally, Section 5 presents our conclusions and future work.

2 Previous Approaches on N-gram-Based Translation

Our proposal has been inspired by the previous work of the Johns Hopkins University Applied Physics Lab (JHU/APL) about the employment of overlapping character n-grams for indexing documents [9, 8, 10, 11]. Their interest came from the possibilities that overlapping character n-grams may offer particularly in the case of non-English languages [8]: to provide a surrogate means to normalize word forms and to allow one to manage languages of very different natures without further processing, such as agglutinative languages as in the case of Turkish, or languages lacking word separator characters such as Japanese. Moreover, this knowledge-light approach does not rely on language-specific processing, and it can be used even when linguistic information and resources are scarce or unavailable.

In the case of monolingual retrieval, the employment of n-grams is quite simple, since the documents to be indexed are just tokenized into overlapping n-grams instead of the usual words. This way, the word `potato`, for example, is split into its different overlapping compounding n-grams: `-pot-`, `-ota-`, `-tat-` and `-ato-`. These resulting n-grams are then indexed by the retrieval engine. The same tokenizing process will be made with queries, allowing matching between documents and queries.

In the case of translingual retrieval, the document indexing process remains the same, but two phases are now required during query processing: one for translation and another one for n-gram splitting. In their initial cross-language experiments, JHU/APL firstly translated the source language query into the target language using Machine Translation (MT) techniques, parallel collections or bilingual dictionaries [9]. The resulting translated query was then split into n-grams, which were submitted to the retrieval engine.

Further experiments were made using a new n-gram-based translation approach. This so-called *direct n-gram translation* technique used n-grams instead of words as translation units. The objective pursued was to avoid some of the limitations of classical dictionary-based translation, such as the need for word normalization, the problems of translating multiple word expressions and the inability to handle out-of-vocabulary words [11]. This n-gram translation algorithm takes as input

[1] `http://www.grupocole.org`

[2] `http://www.cet.sunderland.ac.uk/IR/ir.html`

a parallel corpus, aligned at the paragraph (or document) level and extracts candidate translations as follows [10]. Firstly, for each candidate n-gram term to be translated, paragraphs containing this term in the source language are identified. Next, their corresponding paragraphs in the target language are also identified and, using a statistical measure similar to mutual information, a translation score is calculated for each of the terms occurring in one of such target language texts. Finally, the target n-gram with the higher translation score is selected as the potential translation of the source n-gram. The whole process is quite slow: it is said that the process takes several days in the case of working with 5-grams, for example [10].

3 Our approach

Taking as our model the system developed by JHU/APL, we have developed our own n-gram based retrieval system for testing our ideas. This system has been built using freely available resources when possible in order to make it more transparent and to minimize effort.

This way, instead of the ad-hoc retrieval system employed by the original design [9], we have opted for using the open-source TERRIER platform [1]. This decision was supported by the satisfactory results obtained with n-grams using other indexing engines [15].

The second point of difference with respect to the original approach comes from the translation resources to be used. JHU/APL employed bilingual word-lists extracted from a huge parallel corpus of their own [10] created by mining the web of the Official Journal of the European Union[3]. However, since our group has no access to such a large parallel corpus, we had to employ a smaller one, the well-known Europarl corpus [4]. This corpus was extracted from the proceedings of the European Parliament covering April 1996 to September 2003, containing up to 28 million words per language. It includes versions in 11 European languages: Romanic (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish.

Finally, with respect to the n-gram translation algorithm itself, since the alignment algorithm of the original approach was too slow for our purposes, we opted for a slightly different approach consisting of two phases. In the first phase, the slowest one, word-level alignment of the text was made employing a statistical alignment model. For this purpose, the parallel corpus was processed through the well-known GIZA++ toolkit [14], obtaining the translation probabilities between the different source and target language words which have been aligned by the software tool. Next, prior to the second phase, several heuristics can be applied —if desired— for refining or modifying such word-to-word translation scores. We can remove, for example, those candidate translations with a translation probability less than a previously established threshold, or we can combine the scores of bidirectional alignments [5] —source-target language and target-source language— instead of just the direct one —source-target language. Finally, in the second phase, n-gram translation scores are computed employing associative measures [7], taking as input the translation probabilities calculated by GIZA++.

At this point, and in order to illustrate accurately the process involved during this second phase, we will take as basis how associative measures are calculated and how they could be used for generating bilingual dictionaries automatically taking as input parallel collections aligned at paragraph level. In this illustrating context, given a word pair $(word_u, word_v)$ —$word_u$ standing for the source language word, and $word_v$ for its candidate target language translation—, their cooccurrence frequency can be organized in a *contingency table* resulting from a cross-classification of their cooccurrences in the aligned corpus:

	$V = word_v$	$V \neq word_v$	
$U = word_u$	O_{11}	O_{12}	$= R_1$
$U \neq word_u$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

[3] `http://europa.eu`

In this table, instances whose first component belongs to type $word_u$ —i.e., the number of aligned paragraphs where the source language paragraph contains $word_u$— are assigned to the first row of the table, and tokens whose second component belongs to type $word_v$ —i.e., the number of aligned paragraphs where the target language paragraph contains $word_v$— are assigned to the first column. The cell counts of this contingency table are called the *observed frequencies*:

O_{11}**:** Number of aligned paragraphs where the source language paragraph contains $word_u$ and the target language paragraph contains $word_v$.

O_{12}**:** Number of aligned paragraphs where the source language paragraph contains $word_u$ but the target language paragraph does not contain $word_v$.

O_{21}**:** Number of aligned paragraphs where the source language paragraph does not contain $word_u$ but the target language paragraph contains $word_v$.

O_{22}**:** Number of aligned paragraphs where the source language paragraph does not contain $word_u$ and the target language paragraph does not contain $word_v$ either.

The sum of all these four observed frequencies —or *sample size* N— is equal to the total number of pairs of words considered. R_1 and R_2 are the row totals of the observed contingency table, while C_1 and C_2 are the corresponding column totals. Such row and column totals are also called *marginal frequencies*, and O_{11} is called the *joint frequency*. Equations for all association measures are given in terms of the observed frequencies, marginal frequencies, and the expected frequencies E_{11}, ..., E_{22} (under the null hypothesis that $word_u$ and $word_v$ are statistically independent). The expected frequencies can easily be computed from the row and column totals:

	$V = word_v$	$V \neq word_v$
$U = word_u$	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$U \neq word_u$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

Once the contingency table has been built, different association measures can be easily calculated for each pair of words. After this, the most promising pairs can be inserted into the automatically generated bilingual dictionary by selecting them from those with the highest association measures. In our case, two classical measures will be applied for this purpose: mutual information and Dice coefficient, defined by equations 1 and 2, respectively:

$$MI(word_u, word_v) = log \frac{O_{11}}{E_{11}} \qquad (1) \qquad\qquad Dice(word_u, word_v) = \frac{2O_{11}}{R_1 + C_1} \qquad (2)$$

At this point, we have described how to compute and employ association measures for the automatic generation of bilingual dictionaries from parallel corpora aligned at paragraph level. However, in our proposal, we do not have aligned paragraphs but aligned words —a source word and its candidate translation—, both composed by n-grams. Our first idea could be just to adapt the contingency table to that context. Consequently, we can consider that we are now dealing with n-gram pairs ($n\text{-}gram_u, n\text{-}gram_v$) cooccurring at aligned words instead of word pairs ($word_u, word_v$) cooccurring at aligned paragraphs. So, contingency tables should be redefined according to this new situation: O_{11}, for example, should be re-formulated as the number of aligned words where the source language word contains $n\text{-}gram_u$ and the target language word contains $n\text{-}gram_v$.

This first solution seems logical, and it is intuitive and easy to understand. Nevertheless, we find a problem. In the case of aligned paragraphs formed of words, we had *real* instances of word cooccurrences at the paragraphs aligned. However, in our proposal we do not have *real* instances of n-gram cooccurrences at aligned words —as it may be expected—, but just *probable* ones, since GIZA++ —the tool employed for the initial word-level alignment— is based on a statiscal alignment model which computes a translation probability for each cooccurring pair of words. So,

the same word may appear as being aligned with several translation candidates, each one with its corresponding probability. For example, taking the English words `milk` and `milky`, and the Spanish words `leche` *(`milk`)*, `lechoso` *(`milky`)* and `tomate` *(`tomato`)*, a possible output alignment would be:

source word	**candidate translation**	**probability**
`milk`	`leche`	0.98
`milky`	`lechoso`	0.89
`milk`	`tomate`	0.15

This way, it may be considered that the source 4-gram `-milk-` does not really cooccur with the target 4-gram `-lech-`, since the alignment between its containing words `milk` and `leche`, and `milky` and `lechoso` is not certain. Nevertheless, it seems much more probable that the *translation* of the 4-gram `-milk-` was `-lech-` rather than `-toma-`, for example, since the probability of the alignment of their containing words —`milk` and `tomato`— is much smaller than that of the words containing `-milk-` and `-lech-` —the pairs `milk` and `leche` and `milky` and `lechoso`. Taking this idea as basis, our proposal consists of weighting the likelihood of a cooccurrence according to the probability of its corresponding alignment.

Taking again the previous `milk-milky` example, we can consider the overlapping 4-grams that compose each word. Thus, we would obtain an alignment like this:

source word	**candidate translation**	**probability**
`-milk-`	`-lech- -eche-`	0.98
`-milk- -ilky`	`-lech- -echo- -chos- -hoso-`	0.92
`-milk-`	`-toma- -omat- -mate-`	0.15

So, the contingency tables corresponding to the n-gram pairs *(`-milk-`, `-lech-`)* and *(`-milk-`, `-toma-`)* are as follows:

	$V =$ `-lech-`	$V \neq$ `-lech-`	
$U =$ `-milk-`	$O_{11} = 0.98 + 0.92 =$**1.90**	$O_{12} = 0.98 + 3 * 0.92 + 3 * 0.15 =$**4.19**	$R_1 =$**6.09**
$U \neq$ `-milk-`	$O_{21} =$**0.92**	$O_{22} = 3 * 0.92 =$**2.76**	$R_2 =$**3.68**
	$C_1 =$**2.82**	$C_2 =$**6.95**	$N =$**9.77**

	$V =$ `-toma-`	$V \neq$ `-toma-`	
$U =$ `-milk-`	$O_{11} =$**0.15**	$O_{12} = 2 * 0.98 + 4 * 0.92 + 2 * 0.15 =$**5.94**	$R_1 =$**6.09**
$U \neq$ `-milk-`	$O_{21} =$**0**	$O_{22} = 4 * 0.92 =$**3.68**	$R_2 =$**3.68**
	$C_1 =$**0.15**	$C_2 =$**9.62**	$N =$**9.77**

It can be seen in the example that the O_{11} observed frequency corresponding to the n-gram pair *(`-milk-`, `-lech-`)* is not 2 as it could be expected, but 1.90. This is because it appears in 2 alignments, `milk` with `leche` and `milky` with `lechoso`, but each cooccurrence in a alignment must also be weighted according to its translation probability like this: 0.98 (probability of the alignment of `milk` with `leche`) + 0.92 (probability of the alignment of `milky` with `lechoso`) = 1.90.

Once the contingency tables have been obtained, the Dice coefficients corresponding to each n-gram pair can be computed. As expected, the association measure of the pair *(`-milk-`, `-lech-`)* —the correct one— is much higher than that of the pair *(`-milk-`, `-toma-`)* —the wrong one:

$$Dice(\texttt{-milk-}, \texttt{-lech-}) = \frac{2*1.90}{6.09+2.82} = \mathbf{0.43} \qquad Dice(\texttt{-milk-}, \texttt{-toma-}) = \frac{2*0.15}{6.09+0.15} = \mathbf{0.05}$$

If we consider that a real existing cooccurrence instance —such as those of the word-based algorithm used for illustrating— corresponds to a 100% probability, we can think about the original

Task	Monolingual	Robust					
Language	***PT***	***DE***	***EN***	***ES***	***FR***	***IT***	***NL***
Size (in MB)	564	668	579	1,086	331	363	540
# of docs.	210,734	294,809	169,477	454,045	129,806	157,558	190,604

Table 1: Statistics of test collections (by task and language)

word-based algorithm for building the contingency table and calculating word-level associative measures as a particular case of the generalized algorithm we have proposed.

This new approach we have proposed for n-gram direct translation increases the speed of the process, concentrating most of the complexity in the word-level alignment phase. This first step acts as a initial filter, since only those n-gram pairs corresponding to aligned words will be considered, whereas in the original JHU/APL approach all n-gram pairs corresponding to aligned paragraphs were considered. On the other hand, since the n-gram alignment phase is much faster, different n-gram alignment techniques can be easily tested. Another advantage of this approach is that the n-gram alignment process can take as input previously existing lists of aligned words or even bilingual dictionaries, theoretically improving the results.

4 Experiments

In this our first joint participation as CoLesIR group, we have taken part in two tasks of the ad-hoc track: the Portuguese monolingual task, and the robust task. Unfortunately, the lack of time did not allow us to tune accurately our retrieval system, either to complete our n-gram direct translation tool. So, we can show here only the results intended to be used as baselines for future tests. This way, no tuning has been made with respect to the possibility of removing high or low-frequency n-grams, the employment of relevance feedback, or the use of pre or post-translation expansion techniques in the case of translingual runs [10].

So, documents were just split into n-grams and indexed, as were the queries. Before that, the text had been converted into lowercase and punctuation marks were removed [10]. Diacritics, however, have been kept in this first first set of experiments.

The open-source Terrier platform [1] has been employed as retrieval engine using a InL2[4] ranking model [2]. No stopword removal or query expansion have been applied at this point. The same running parameters have been used for all the experiments performed. With respect to the n-gram length, we decided to use 4-grams as a compromise size after studying the results previously obtained by the JHU/APL group [9, 8, 10, 11] using different n-gram lengths.

4.1 Portuguese Monolingual Task

The possibility of working with Portuguese caught our attention because of its proximity to Galician language. Our Spanish part, the CoLe group, has been working for many years on NLP and IR in Galician, a Romance language spoken in Galicia, in the North-West of Spain, where it is co-oficial language. Nevertheless, the lack of freely available resources for this language has limited such work, particularly in the case of IR. This way, the existence of a Portuguese corpus for IR evaluation is very interesting because of the proximity of both languages, Galician and Portuguese, since they were a single language in their origin, and their linguistic phenomena are still very similar today.

The document collection used for this task comprises news published during 1994 and 1995 by the newspapers Público —Portuguese— and Folha de São Paulo —Brazilian. See column PT of Table 1 for more details. The test set includes 50 topics (C301–C350). Only *title* and *description* fields were used in the submitted queries.

[4]Inverse Document Frequency model with Laplace after-effect and normalization 2.

	Portuguese		Robust monolingual											
			Training topics						*Test topics*					
	PT$_I$	***PT$_D$***	***DE***	***EN***	***ES***	***FR***	***IT***	***NL***	***DE***	***EN***	***ES***	***FR***	***IT***	***NL***
# Retr.	50k	50k	60k	60k	60k	60k	60k	60k	100k	100k	100k	100k	100k	100k
# Rel. exp.	2,677	2,677	2,252	1,150	2,908	1,351	1,197	1,946	3,641	1,533	5,008	2,190	1,930	2,717
# Rel. retr.	2,152	2,173	2,080	960	2,509	1,257	1,099	1,766	3,227	1,379	4,105	2,025	1,758	2,374
Non-int. Pr.	35.18	32.69	36.11	27.61	30.75	33.64	29.03	37.30	37.21	37.64	40.17	39.51	32.23	41.60
R-Pr.	35.73	32.23	35.99	27.11	30.74	31.62	27.63	37.84	36.49	36.21	40.03	37.56	32.06	39.85
Binary Pref.	35.26	33.22	38.32	32.23	39.44	34.39	29.33	36.24	40.78	39.58	50.27	42.19	36.38	40.17
Geo. Pr.	20.95	18.66	24.25	4.89	15.13	19.73	10.40	23.25	14.80	8.41	18.85	11.91	8.23	16.40
0% Re.	70.32	67.77	72,54	52,51	68,31	63,36	53,66	76,98	65,97	63,96	71,62	68,41	61,40	75,33
10% Re.	57.77	54.83	58,49	45,38	55,57	53,95	48,22	65,32	57,08	58,49	61,36	60,81	52,75	64,76
20% Re.	50.75	47.39	50,26	39,33	47,22	48,07	44,34	55,75	51,59	52,10	55,85	53,03	45,71	58,71
30% Re.	45.81	41.92	43,36	36,30	40,87	41,76	38,07	49,08	45,07	47,75	51,15	47,19	40,23	51,64
40% Re.	40.39	36.97	39,17	31,01	33,65	37,58	32,20	43,41	41,80	44,23	46,85	43,86	35,97	46,38
50% Re.	35.95	34.03	36,62	28,21	28,00	35,37	30,24	39,25	38,62	39,58	43,08	40,99	33,01	42,53
60% Re.	31.86	29.59	33,31	22,18	24,66	30,61	26,55	30,43	34,91	32,79	38,32	35,97	28,36	35,79
70% Re.	27.81	25.73	29,50	19,66	20,43	24,10	22,17	25,21	31,28	29,44	33,88	32,68	24,94	32,01
80% Re.	21.54	20.35	23,25	16,32	17,22	21,25	18,72	21,36	26,85	25,30	28,16	28,77	21,38	28,64
90% Re.	15.20	13.50	18,40	14,36	13,14	17,62	13,82	16,47	19,40	20,28	18,03	23,71	17,71	22,59
100% Re.	6.53	6.53	10,01	9,40	6,95	13,86	10,23	8,99	13,14	17,30	10,51	18,86	13,03	16,29
5 docs.	51.20	45.20	49.67	32.67	44.00	34.00	32.00	50.67	44.20	37.00	48.00	40.80	36.40	53.60
10 docs.	47.20	44.40	43.17	27.33	38.33	28.67	28.17	42.67	41.20	28.60	43.10	35.20	31.60	43.50
15 docs.	44.40	44.00	40.11	23.22	36.67	25.67	26.22	37.56	38.53	24.93	39.40	30.47	27.93	38.80
20 docs.	42.50	41.70	36.83	21.42	34.67	24.50	23.75	34.00	36.30	21.85	35.95	27.80	25.50	34.90
30 docs.	39.60	37.87	32.72	18.28	30.22	22.17	21.61	29.00	31.87	18.20	32.33	23.87	21.83	30.13
100 docs.	24.32	23.00	19.40	10.58	19.90	12.77	12.07	16.83	19.22	9.26	20.69	12.95	11.14	15.41
200 docs.	15.54	14.92	13.30	6.58	13.53	8.43	7.40	11.05	12.27	5.61	13.82	8.00	6.91	9.39
500 docs.	7.61	7.59	6.60	3.03	7.33	4.00	3.46	5.46	5.92	2.61	7.32	3.81	3.25	4.42
1,000 docs.	4.30	4.35	3.47	1.60	4.18	2.09	1.83	2.94	3.23	1.38	4.11	2.03	1.76	2.37

Table 2: Baseline results for monolingual runs: Portuguese and robust monolingual tasks

Unlike the rest of our experiments, two ranking models were used this time: the previously referred to InL2 model, and the hypergeometric model named DLH[5]. The results obtained are shown in Table 2, columns PT_I and PT_D, respectively. The performance of the system is measured using the parameters contained in each row: number of documents retrieved, number of relevant documents expected, number of relevant documents retrieved, average precision (non-interpolated) for all relevant documents (averaged over queries), R-precision, binary preference, geometric average precision, precision at 11 standard levels of recall, and precision at N documents retrieved.

4.2 Robust Task

Since our CLIR system is still in its first stages, we preferred to test it with the most commonly used languages in CLIR before trying more exotic or less-known languages. This is the main reason for participating in the robust task. The robust task is essentially an ad-hoc task which makes use of the topics and collections used from CLEF 2001 to CLEF 2003. The data collections, whose content is described in Table 1, are formed by newspapers and newswires written in six languages: German (*DE*), English (*EN*), Spanish (*ES*), French (*FR*), Italian (*IT*) and Dutch (*NL*). The test set is formed by 160 topics (C041–C200). This initial set has been divided into two subsets: a so-called *training topics* subset —formed by topics C050–C059, C070–C079, C100–C109, C120–C129, C150–159, C180–189—, to be used for tuning purposes, and a so-called *test topics* subset —formed by the rest of the topics—, for testing purposes. Again, only *title* and *description* fields were used in the submitted queries.

4.2.1 Monolingual experiments

We have participated in all the monolingual subtasks of the robust task: German (*DE*), English (*EN*), Spanish (*ES*), French (*FR*), Italian (*IT*) and Dutch (*NL*). Results are shown in Table 2.

[5] http://ir.dcs.gla.ac.uk/wiki/HypergeometricModel

	ENDE		*ENxx*	
	train	***test***	***train***	***test***
# Retr.	60k	100k	60k	100k
# Rel. exp.	2,252	3,641	10,804	17,019
# Rel. retr.	1,710	2,599	5,968	8,877
Non-int. Pr.	24.26	25.24	19.37	22.63
R-Pr.	26.50	26.40	26.57	29.02
Binary Pref.	31.24	31.57	27.86	30.82
Geo. Pr.	11.58	4.31	11.98	11.24
0% Re.	54.38	51.63	71.28	66.50
10% Re.	44.14	40.98	39.63	44.44
20% Re.	36.21	36.31	32.20	37.27
30% Re.	31.39	32.49	27.27	30.78
40% Re.	27.83	29.44	23.33	26.89
50% Re.	24.46	27.38	18.28	22.05
60% Re.	21.51	22.85	13.45	17.04
70% Re.	18.51	19.63	9.56	13.26
80% Re.	13.72	16.50	4.67	8.72
90% Re.	9.02	10.66	2.60	4.73
100% Re.	5.76	6.45	0.33	0.24
5 docs.	35.00	32.60	43.67	43.60
10 docs.	33.17	28.70	41.00	43.60
15 docs.	30.89	25.80	39.44	41.67
20 docs.	27.83	24.90	38.08	40.40
30 docs.	23.61	22.80	36.33	37.33
100 docs.	14.00	13.54	27.52	28.04
200 docs.	9.76	9.02	22.13	22.10
500 docs.	5.16	4.58	14.94	13.66
1,000 docs.	2.85	2.60	9.95	8.88

Table 3: Baseline results for translingual runs: English to German robust bilingual task (*ENDE*) and robust multilingual task with English as source language (*ENxx*)

4.2.2 Bilingual experiments

In this case, we have just participated in the English-to-German bilingual subtask. Since our direct n-gram translation tool was not ready on time, we opted for a similar approach to that used by JHU/APL group in their first translingual retrieval experiments [9]. This way, the source language query is first translated into the target language before splitting it into n-grams to be submitted to the retrieval engine. In our case we have used Altavista's Babel Fish[6] for translating the queries. Columns *ENDE* of Table 3 show the results obtained.

4.2.3 Multilingual experiments

In this final multilingual task, English has been used as the source language, whereas all six available collections are used for retrieving. As we explained, our direct n-gram translation tool could not be used because it was not ready on time. Our initial baseline runs were submitted instead.

As before, source language —English— queries were translated into each of the target languages using Altavista's Babel Fish. Once translated, they were split into n-grams for querying their corresponding target language collection. Next, the different rankings retrieved for each target language collection are normalized. The similarity value or *retrieval status value* (RSV) of the ith document retrieved is normalized by the maximum and minimum of the ranking [6] as follows:

$$RSV_i' = \frac{RSV_i - RSV_{min}}{RSV_{max} - RSV_{min}} \tag{3}$$

where RSV_i is the original similarity value, RSV_i' is the normalized one, and RSV_{min} and RSV_{max} are the minimal and maximal similarity values of that ranking, respectively. Once normalized, all individual rankings are merged into the final output ranking to be retrieved.

[6] `http://babelfish.altavista.com`

Columns *ENxx* of Table 3 show the results obtained.

5 Conclusions and future work

This paper describes our initial work in the field of Cross-Language Information Retrieval. Using our past experience in the application of Natural Language Processing techniques to Information Retrieval, and our recent work in Machine Translation (MT), we have developed an n-gram-based system which uses such subwords not only as indexing units, but also as translating units.

This work has been inspired in the previous work of the Johns Hopkins University Applied Physics Lab [9, 8, 10, 11]. However, its the training algorithm was too slow for our purposes. Thus we decided to develop our own n-gram based retrieval system for testing our ideas. Freely available resources have been used when possible in its design in order to make it more transparent and to minimize effort. For speeding up the training process, we have opted for a slightly different algorithm to the original one, now consisting of two phases. In the first phase, the slowest one, word-level alignment of the text is made through a statistical alignment tool. In the second phase, n-gram translation scores are computed employing association measures taking as input the translation probabilities calculated in the previous phase. This new approach increases the speed of the training of the process, concentrating most of the complexity in the word-level alignment phase. Another advantage is that the n-gram alignment process can take as input previously existing aligned word lists or even bilingual dictionaries, which should improve the results.

Unfortunately lack of time did not allow us to get our n-gram direct translation system ready on time. Thus we have only included those baseline results to be used in the future evaluation of our approach.

With respect to future work, we intend to complete and test both our n-gram direct translation system and our retrieval module as soon as possible. Once the base system is working, we intend to test the behavior of new association measures [19].

Acknowledgments

This research has been partially supported by Ministerio de Educación y Ciencia and FEDER (TIN2004-07246-C03-02), Xunta de Galicia (PGIDIT05PXIC30501PN, PGIDIT05SIN044E), and Dirección Xeral de Investigación, Desenvolvemento e Innovación (*Programa de Recursos Humanos* grants).

References

[1] `http://ir.dcs.gla.ac.uk/terrier/` (visited on August 2006).

[2] G. Amati and C.J. van Rijsbergen. Probabilistic models of Information Retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, 2002.

[3] Víctor M. Darriba, Gabriel P. Lopes, and Tiago Ildefonso. Measuring the impact of cognates in parallel text alignment. In *Proc. of 12th Portuguese Conference on Artificial Intelligence (EPIA 2005), Covilha, Portugal, December 5-8*, pages 334–343. IEEE Press, 2005.

[4] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of the 10th Machine Translation Summit (MT Summit X), September 12-16, 2005: Phuket, Thailand*, pages 79–86, 2005. Corpus available in `http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl/` (visited on August 2006).

[5] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *NAACL '03: Proc. of the 2003 Conference of the North American Chapter of the ACL*, pages 48–54, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[6] Joon Ho Lee. Analyses of multiple evidence combination. In *Proc. of SIGIR '97, July 27-31, Philadelphia, PA, USA*, pages 267–276. ACM Press, 1997.

[7] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge (Massachusetts) and London (England), 1999.

[8] Paul McNamee and James Mayfield. Scalable multilingual information access. In volume 2785 of *Lecture Notes in Computer Science*, pages 207–218. Springer-Verlag, Berlin-Heidelberg-New York, 2003.

[9] Paul McNamee and James Mayfield. Character N-gram Tokenization for European Language Text Retrieval. *Information Retrieval*, 7(1-2):73–97, 2004.

[10] Paul McNamee and James Mayfield. JHU/APL experiments in tokenization and non-word translation. In volume 3237 of *Lecture Notes in Computer Science*, pages 85–97. Springer-Verlag, Berlin-Heidelberg-New York, 2004.

[11] Paul McNamee and James Mayfield. Cross-Language Retrieval Using HAIRCUT at CLEF 2004. In volume 3491 of *Lecture Notes in Computer Science*, pages 50–59. Springer-Verlag, Berlin-Heidelberg-New York, 2005.

[12] Enrique Méndez, Jesús Vilares, and David Cabrero. COLE experiments at QA@CLEF 2004 Spanish monolingual track. In volume 3491 of *Lecture Notes in Computer Science*, pages 544–551. Springer-Verlag, Berlin-Heidelberg-New York, 2005.

[13] Michael P. Oakes and Souvik Banerjee. Regular sound changes for Cross-Language Information Retrieval. In volume 3237 of *Lecture Notes in Computer Science*, pages 263–270. Springer-Verlag, Berlin-Heidelberg-New York, 2004.

[14] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models, 2003. Source code available at `http://www.fjoch.com/GIZA++.html` (visited on August 2006).

[15] Jacques Savoy. Cross-Language Information Retrieval: experiments based on CLEF 2000 corpora. *Information Processing and Management*, 39:75–115, 2003.

[16] Jesús Vilares, Miguel A. Alonso, and Francisco J. Ribadas. COLE experiments at CLEF 2003 Spanish monolingual track. In volume 3237 of *Lecture Notes in Computer Science*, pages 345–357. Springer-Verlag, Berlin-Heidelberg-New York, 2004.

[17] Jesús Vilares, Miguel A. Alonso, Francisco J. Ribadas, and Manuel Vilares. COLE experiments at CLEF 2002 Spanish monolingual track. In volume 2785 of *Lecture Notes in Computer Science*, pages 265–278. Springer-Verlag, Berlin-Heidelberg-New York, 2003.

[18] Manuel Vilares, Francisco J. Ribadas, and Jorge Graña. On pattern-matching as query facility. In Alexander Gelbukh, editor, *Topics in Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science. Springer-Verlag, Berlin-Heidelberg-New York, 2006.

[19] Julie Weeds and David Weir. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475, 2005.