CoLesIR at CLEF 2007: from English to French via Character N-Grams

Jesús Vilares
Dept. of Computer Science
University of A Coruña
Campus de Elviña
15071 – A Coruña (Spain)
jvilares@udc.es

Michael P. Oakes
School of Computing
University of Sunderland
St. Peter's Campus
Sunderland – SR6 0DD (UK)
Michael.Oakes@sunderland.ac.uk

Manuel Vilares
Dept. of Computer Science
University of Vigo
Campus As Lagoas s/n
32004 – Ourense (Spain)
vilares@uvigo.es

Abstract

This work is an extension of our proposal originally presented in CLEF 2006, which, unfortunately, could not be ready on time for the workshop. We describe here a knowledge-light approach for query translation in Cross-Language Information Retrieval systems. This proposal itself can be considered as an extension of the previous work of the Johns Hopkins University Applied Physics Lab, preserving its advantages but avoiding its main drawbacks. As in their original proposal, our work is based on the direct translation of character n-grams, avoiding in this way the need for word normalization during indexing or translation, and also dealing with out-of-vocabulary words. Moreover, since such a solution does not rely on language-specific processing, it can be used with languages of very different natures even when linguistic information and resources are scarce or unavailable. Nevertheless, in contrast with the original approach, our proposal is much faster and transparent, making extensive use of freely available resources. The system has been tested in the robust ad-hoc English-to-French bilingual task, obtaining encouraging results.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing—*Indexing methods*; H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval—*Query formulation*; I.2 [**Artificial Intelligence**]: I.2.7 Natural Language Processing—*Machine translation; Text analysis*

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

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

1 Introduction

This work is an extension of the proposal originally presented by our group in the previous CLEF edition, a new knowledge-light approach for query translation in Cross-Language Information

Retrieval (CLIR) systems based on the direct translation of character n-grams. Such a proposal itself can be considered as an extension of the previous work of the Johns Hopkins University Applied Physics Lab (JHU/APL) on the employment of overlapping character n-grams for indexing documents [7, 8].

The interest in using overlapping character n-grams comes from the fact that it provides a surrogate means to normalize word forms and to allow to manage languages of very different natures without further processing. Such a 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 straightforward, since both queries and documents are just tokenized into overlapping n-grams instead of words: the word `tomato`, for example, is split into `-tom-`, `-oma-`, `-mat-` and `-ato-`. The resulting n-grams are then processed by the retrieval engine either for indexing or querying. Nevertheless, when extending its use to the case of CLIR, an extra translation phase is needed during querying.

Aiming to avoid some of the limitations of classic dictionary-based translation methods, such as the need for word normalization or the inability to handle out-of-vocabulary words, JHU/APL researchers developed a direct n-gram translation algorithm which allows translation not at the word level but at the n-gram level [8]. This n-gram translation algorithm takes as input a parallel corpus, aligned at the paragraph (or document) level and extracts candidate translations as follows. 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 an ad-hoc statistical measure, a translation score is calculated for each of the terms occurring in the target language texts. Finally, the target n-gram with the highest translation score is selected as the potential translation of the source n-gram. Nevertheless, the whole process was found to be very slow, making the testing of new developments difficult: it could take several days in the case of working with 5-grams, for example.

This paper describes a new direct n-gram alignment proposal we have developed both to speed up the process and to make the system more transparent. The article is structured as follows. Firstly, Sect. 2 describes our approach. Next, in Sect. 3, our proposal is evaluated. Finally, in Sect. 4, we present our conclusions and future work.

2 Description of the system

Taking as our model the system designed by JHU/APL, we developed our own n-gram based retrieval system, trying to preserve the advantages of the original proposal but avoiding its main drawbacks. Moreover, instead of the ad-hoc resources developed for the original system [7, 8], our system has been built using freely available resources when possible in order to make it more transparent and to minimize effort. This way, we use the open-source retrieval platform Terrier [1] instead of the ad-hoc retrieval system employed by the original design, and the well-known Europarl parallel corpus[1] [4] is used as training data instead of the ad-hoc corpus employed by JHU/APL.

Nevertheless, the main difference is the n-gram alignment algorithm itself, which now consists of two phases. In the first phase, the slowest one, the input parallel corpus is aligned at the word-level using the well-known statistical tool GIZA++ [9], obtaining as output the translation probabilities between the different source and target language words. In our case, after some initial experiments [11], we have opted for a bidirectional alignment [5] which considers a (w_{EN}, w_{SP}) English-to-Spanish word alignment only if there also exists a corresponding (w_{SP}, w_{EN}) Spanish-to-English alignment. This way the subsequent processing will be focused only on those words whose translation seems less ambiguous, reducing both the number of input word pairs to be

[1] This corpus was extracted from the proceedings of the European Parliament, containing up to 28 million words per language. It includes versions in 11 European languages: Romance (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish.

processed and output n-gram pairs to be obtained by more than 60%. This reduction allows us to reduce greatly both computing and storage resources —including processing time.

Next, prior to the second phase, heuristics can be applied —if desired— for refining or modifying the word-to-word translation scores calculated by GIZA++. In our case, we have removed those least-probable word alignments from the input (those with a word translation probability less than a threshold W, with W=0.15) [11]. Such pruning leads to a considerable extra reduction of processing time and storage space: a reduction of over 90% in the number of both input word pairs processed and output n-gram pairs aligned.

Finally, in the second phase, n-gram translation scores are computed employing statistical association measures [6], taking as input the translation probabilities previously calculated by GIZA++.

As we can see, this first step acts as an 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. This approach increases the speed of the process by concentrating most of the complexity in the word-level alignment phase, allowing n-gram alignment techniques to be easily tested.

2.1 Word-level alignment using association measures

Our n-gram alignment algorithm is an extension of the way association measures can be used for creating bilingual word dictionaries taking as input parallel collections aligned at the paragraph level [10]. In this context, given a word pair (w_s, w_t) —w_s standing for the source language word, and w_t 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 input aligned corpus:

	$T = w_t$	$T \neq w_t$	
$S = w_s$	O_{11}	O_{12}	$= R_1$
$S \neq w_s$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

As shown, the first row accounts for those instances where the source language paragraph contains w_s, while the first column accounts for those instances where the target language paragraph contains w_t. The cell counts are called the *observed frequencies*: O_{11}, for example, stands for the number of aligned paragraphs where the source language paragraph contains w_s and the target language paragraph contains w_t; O_{12} stands for the number of aligned paragraphs where the source language paragraph contains w_s but the target language paragraph does not contain w_t; and so on. The total number of word pairs considered —or *sample size* N— is the sum of the observed frequencies. The row totals, R_1 and R_2, and the column totals, C_1 and C_2, are also called *marginal frequencies* and O_{11} is called the *joint frequency*.

Once the contingency table has been built, different association measures can be easily calculated for each word pair. The most promising pairs, those with the highest association measures, are stored in the bilingual dictionary.

2.2 Adaptations for n-gram-level alignment

We have described how to compute and use association measures for generating bilingual word dictionaries from parallel corpora. However, in our case we do not start with aligned paragraphs composed of words, but aligned words —previously aligned through GIZA++— composed of character n-grams. A first choice for adapting the previous word-level alignment algorithm to the case of n-grams could be just to adapt the contingency table to the new context, by considering that we are managing n-gram pairs (g_s, g_t) cooccurring in aligned words instead of word pairs (w_s, w_t) cooccurring in aligned paragraphs. So, contingency tables should be adapted accordingly:

O_{11}, for example, should be re-formulated as the number of aligned word pairs where the source language word contains n-gram g_s and the target language word contains n-gram g_t.

However, we do not have *real* instances of n-gram cooccurrences at aligned words, but just *probable* ones, since GIZA++ uses a statistical alignment model which computes a translation probability for each cooccurring word pair [9]. So, the same word may be aligned with several translation candidates, each one with a given probability. Taking as example the case of the English words milk and milky, and the Spanish words leche *(milk)*, lechoso *(milky)* and tomate *(tomato)*, a possible output word-level alignment —with its corresponding probabilities— would be:

source word	**candidate translation**	**prob.**
milk	leche	0.98
milky	lechoso	0.92
milk	tomate	0.15

Our proposal consists of weighting the likelihood of a cooccurrence according to the probability of its containing word alignments. So, the resulting contingency tables corresponding to the n-gram pairs *(-milk-, -lech-)* and *(-milk-, -toma-)* are as follows:

	$T =$ -lech-	$T \neq$ -lech-	
$S =$ -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**
$S \neq$ -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**

	$T =$ -toma-	$T \neq$ -toma-	
$S =$ -milk-	$O_{11} =$**0.15**	$O_{12} = 2 * 0.98 + 4 * 0.92 + 2 * 0.15 =$**5.94**	$R_1 =$**6.09**
$S \neq$ -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**

Notice that, for example, the O_{11} frequency corresponding to *(-milk-, -lech-)* is not 2 as might be expected, but 1.90. This is because the pair appears in two word alignments —milk–leche and milky–lechoso—, but each cooccurrence in an alignment has been weighted according to its translation probability:

$$O_{11} = 0.98 \text{ (for milk–leche)} + 0.92 \text{ (for milky–lechoso)} = \mathbf{1.90} \ .$$

Once the contingency tables have been generated, the association measures corresponding to each n-gram pair can be computed. In contrast with the original JHU/APL approach [7, 8], which used an ad-hoc measure, ours uses three of the most extensively used standard measures: the *Dice coefficient* (*Dice*), *mutual information* (*MI*), and *log-likelihood* (*logl*), which are defined by the following equations [6]:

$$Dice(g_s, g_t) = \frac{2O_{11}}{R_1 + C_1} \ . \quad (1) \qquad MI(g_s, g_t) = log \frac{NO_{11}}{R_1 C_1} \ . \quad (2) \qquad logl(g_s, g_t) = 2 \sum_{i,j} O_{ij} \, log \frac{NO_{ij}}{R_i C_j} \ . \quad (3)$$

If using the Dice coefficient, for example, we find that the association measure of the pair *(-milk-, -lech-)* —the correct one— is much higher than that of the pair *(-milk-, -toma-)* —the wrong one:

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

Notice that if we consider that a real existing cooccurrence instance corresponds to a 100% probability, we can think about the original word-based algorithm described in Sect. 2.1 as a particular case of the generalized n-gram-based algorithm we have proposed here with n=∞.

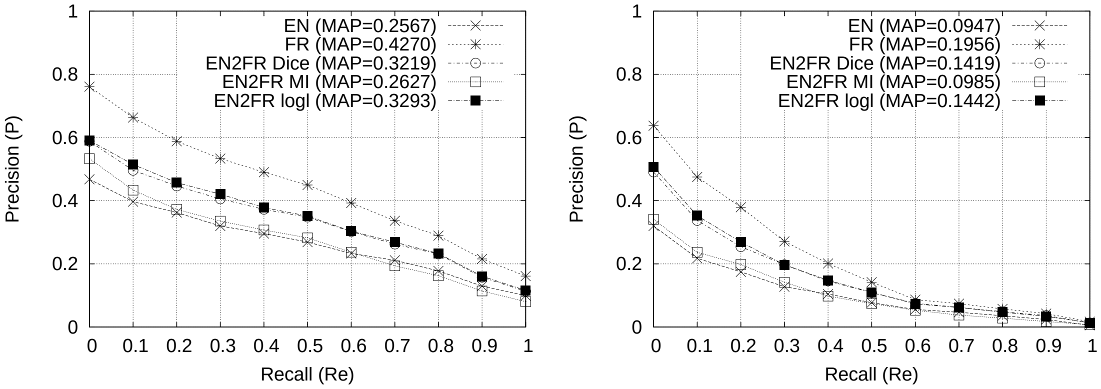

Figure 1: Precision vs. Recall graphs obtained for the *training* (left) and *test* topic sets (right).

3 Evaluation

Since the lack of time did not allow us to have our n-gram direct translation tool ready on time for the past CLEF 2006 workshop [10], this year we have taken part again in the robust ad-hoc task, specifically in the English-to-French bilingual task, in order to present the current development of our work.

The *robust task* is essentially an ad-hoc task which re-uses the topics and collections from past CLEF editions [2]. In this case, the French document collection is formed by 87,191 news reports (243 MB) provided by Le Monde and SDA and corresponding to the year 1994. The English topics set consists of 200 topics divided into two subsets: a *training topics* subset to be used for tuning purposes, formed by 100 topics (C041–C140); and a *test topics* subset for testing purposes, formed by the remaining 100 topics (C251–C350). Moreover, only *title* and *description* topic fields were used in the submitted queries.

With respect to the indexing process, documents were simply split into n-grams and indexed, as were the queries. We have used 4-grams as a compromise n-gram size after studying the results previously obtained by the JHU/APL group [7, 8] using different lengths. Before that, the text had been lowercased and punctuation marks were removed [8], but not diacritics. The open-source Terrier platform [1] was used as retrieval engine with a InL2[2] ranking model [3]. No stopword removal or query expansion were applied at this point.

For querying, the source language topic is firstly split into n-grams. Next, these n-grams are replaced by their N most probable alignments.[3] The resulting translated topics are then submitted to the retrieval system.[4] Because of the lack of time, we could not tune the N value for this new set of English-to-French experiments, so we decided to take those values used in our previous English-to-Spanish experiments [11]:

Dice coefficient	N=1
Mutual Information	N=10
Log-likelihood	N=1

Finally, Fig. 1 and Fig. 2 show the results obtained for each association measure: the Dice coefficient (`EN2FR Dice`), Mutual Information (`EN2FR MI`), and log-likelihood (`EN2FR logl`). We also show the results for two baselines: by querying the French index with the initial English topics split into 4-grams (`EN`) —allowing us to measure the impact of casual matches—, and by querying the index using the French topics split into 4-grams (`FR`) —i.e. a French monolingual run and our ideal performance goal. Notice that mean average precision (MAP) values are also given.

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

[3] With $N \in \{1, 2, 3, 5, 10, 20, 30, 40, 50, 75, 100\}$.

[4] A second selection algorithm, consisting of taking those alignments with a probability greater or equal than a threshold T, was also used in previous experiments [11]. Nevertheless, this threshold-based approach has been dismissed because of the difficulty for fixing T and because if performed not as well as this top-rank-based approach.

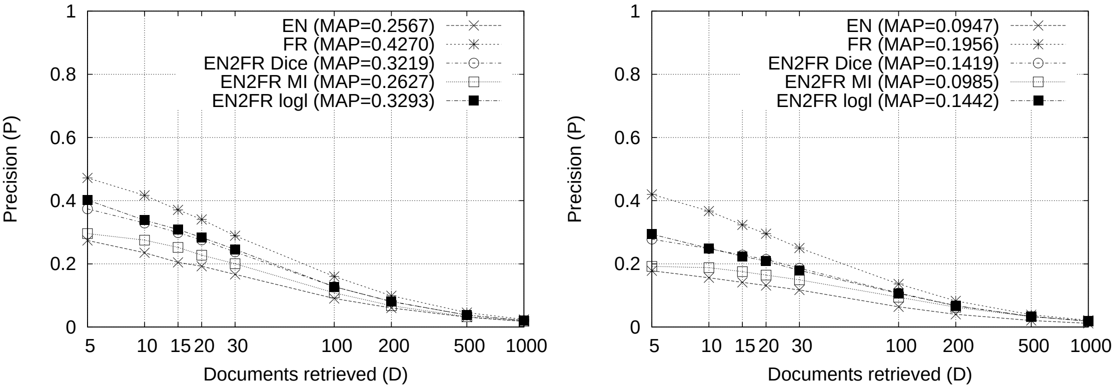

Figure 2: Precision at top D documents graphs obtained for the *training* (left) and *test* topic sets (right).

These results show that the log-likelihood measure obtains the best results for both topic sets, although no significant difference is found with respect to Dice.[5] On the other hand, both approaches perform significantly better than mutual information.

Although we still need to improve our results in order to reach our ideal performance goal, our current results are encouraging, since it must be taken into account that these are still our first experiments, so the margin for improvement is still great.

4 Conclusions and Future Work

This paper extends the proposal originally presented in CLEF 2006 for the development of a CLIR system which uses character n-grams not only as indexing units, but also as translation units. This system was inspired by the work of the Johns Hopkins University Applied Physics Lab [7, 8], but tries to preserve its advantages while avoiding its main drawbacks. As in their original proposal, our work is based on the direct translation of character n-grams, avoiding in this way the need for word normalization during indexing or translation, and also dealing with out-of-vocabulary words.

Moreover, since such a solution does not rely on language-specific processing, it can be used with languages of very different natures even when linguistic information and resources are scarce or unavailable. Nevertheless, in contrast with the original approach, our proposal is much faster and transparent, making extensive use of freely available resources.

So, the n-gram alignment algorithm described consists 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 statistical association measures, taking as input the translation probabilities calculated in the previous phase. This new approach speeds up the training process, concentrating most of the complexity in the word-level alignment phase, making the testing of new association measures for n-gram alignment easier.

With respect to our future work, new tests with other languages of different characteristics are being prepared in order to complete the tuning of the system, including 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 [8].

Acknowledgments

This research has been partially funded by the European Union (FP6-045389), Ministerio de Educación y Ciencia and FEDER (TIN2004-07246-C03 and HUM2007-66607-C04), and Xunta de Galicia

[5] Two-tailed T-tests over MAPs with α=0.05 have been used along this work.

(PGIDIT05PXIC30501PN, PGIDIT05SIN044E, and *Rede Galega de Procesamento da Linguaxe e Recuperación de Información*). The authors would also like to thank John I. Tait for his support.

References

[1] http://ir.dcs.gla.ac.uk/terrier/ (visited on August 2007).

[2] http://www.clef-campaign.org (visited on August 2007).

[3] 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.

[4] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proc. of the 10th Machine Translation Summit (MT Summit X)*, pp. 79–86, 2005. Corpus available in http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl/ (visited on August 2007).

[5] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *NAACL '03: Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 48–54, 2003.

[6] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing.* The MIT Press, 1999.

[7] P. McNamee and J. Mayfield. Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1-2):73–97, 2004.

[8] P. McNamee and J. Mayfield. JHU/APL experiments in tokenization and non-word translation. Vol. 3237 of *Lecture Notes in Computer Science*, pp. 85–97. Springer-Verlag, 2004.

[9] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003. Source code available at http://www.fjoch.com/GIZA++.html (visited on August 2007).

[10] J. Vilares, M. P. Oakes, and J. I. Tait. CoLesIR at CLEF 2006: rapid prototyping of a n-gram-based CLIR system. In *Working Notes of the CLEF 2006 Workshop*, 2006. Available at [2].

[11] J. Vilares, M. P. Oakes, and M. Vilares. Character n-grams translation in cross-language information tetrieval. Vol. 4592 of *Lecture Notes in Computer Science*, pp. 217–228. Springer-Verlag, 2007.