

English-to-French CLIR: A Knowledge-Light Approach through Character N -Grams Alignment

Jesús Vilares¹, Michael P. Oakes², and Manuel Vilares³

¹ Dept. of Computer Science, University of A Coruña
Campus de Elviña s/n, 15071 - A Coruña (Spain)
jvilares@udc.es

² School of Computing and Technology, University of Sunderland
St. Peter's Campus, St. Peter's Way, Sunderland - SR6 0DD (United Kingdom)
Michael.Oakes@sunderland.ac.uk

³ Dept. of Computer Science, University of Vigo
Campus As Lagoas s/n, 32004 - Ourense (Spain)
vilares@uvigo.es

Abstract. This paper describes an extension of our work presented in the robust English-to-French bilingual task of the CLEF 2007 workshop, a knowledge-light approach for query translation in Cross-Language Information Retrieval systems. Our work is based on the direct translation of character n -grams, avoiding 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 nature even when linguistic information and resources are scarce or unavailable. The results obtained have been very positive, and support the findings from our previous English-to-Spanish experiments.

1 Introduction

This work is an extension of our proposal originally developed for the robust English-to-French bilingual task of the CLEF 2007 workshop [1]. It consists of a knowledge-light approach for query translation in Cross-Language Information Retrieval (CLIR) systems based on the direct translation of character n -grams. This proposal itself can be considered as an extension of the previous work by [2].

The use of overlapping character n -grams both as indexing and translation units provides a means to normalize word forms. In addition, the approach supports the handling of out-of-vocabulary words and the management of languages of very different nature without further processing. Moreover, 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.

Since the architecture of our system has been described in depth in a previous CLEF publication [3], this paper focuses on the work performed after the workshop. The paper is structured as follows: firstly, Sect. 2 briefly introduces our

approach; next, Sect. 3 presents the new experiments; finally, Sect. 4 contains our conclusions and proposals for future work.

2 Description of the System

Taking as our model the system designed by JHU/APL [2], we have developed our own n -gram based retrieval system, trying to preserve the advantages of the original system but avoiding its main drawbacks.

The main difference with our proposal is the n -gram alignment algorithm, the basis of the system, which consists of two phases. In the first phase, the slowest one, the input parallel corpus is aligned at the word-level using the statistical tool GIZA++ [4], obtaining as output the translation probabilities between the different source and target language words. In our case, taking advantage of our previous experiments with English-to-Spanish [5,6], we have opted for a bidirectional alignment [7] which considers, for example, a (w_{EN}, w_{FR}) English-to-French word alignment only if there also exists a corresponding (w_{FR}, w_{EN}) French-to-English alignment. This way, subsequent processing is focused only on those words whose translation seems less ambiguous, considerably reducing the number of input word pairs to be processed —actually about 70%— and, consequently, the noise introduced in the system. This reduction allows us to greatly reduce both computing and storage resources.

Next, prior to the second phase, we have also removed those least-probable word alignments from the input (those with a word translation probability less than a threshold W , with $W=0.15$) [5,6]. Such pruning leads to a considerable reduction of processing time and storage space: a reduction of about 95% in the number of input word pairs processed.

Finally, in the second phase, n -gram translation scores are computed using statistical association measures [8], taking as input the translation probabilities previously calculated by GIZA++, and weighting the likelihood of a cooccurrence according to the probability of its containing word alignments [5,6].

For this purpose, our system employs 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 expressions [8]:

$$Dice(g_s, g_t) = \frac{2O_{11}}{R_1 + C_1}. \quad (1) \quad MI(g_s, g_t) = \log \frac{NO_{11}}{R_1 C_1}. \quad (2)$$

$$logl(g_s, g_t) = 2 \sum_{i,j} O_{ij} \log \frac{NO_{ij}}{R_i C_j}. \quad (3)$$

3 Evaluation

In the past CLEF 2007 workshop, our group took part in the robust English-to-French bilingual task. The *robust task* is essentially an ad-hoc task which re-uses the topics and collections from past CLEF editions [9].

Unfortunately, our system could not be accurately tuned for the workshop. So, we had to use the parameters employed in our previous English-to-Spanish experiments [5,6]. Moreover, only one of the two selection algorithms available was used, the so-called *top-rank-based* algorithm. This section presents the work developed after the CLEF 2007 workshop for tuning the system for the new target language. This new experiments also include the *threshold-based* selection algorithm.

With respect to the indexing process, documents were simply split into n -grams and indexed. We used 4-grams as a compromise n -gram size [5,6]. Before that, the text was lowercased and punctuation marks were removed [2], but not diacritics. The open-source TERRIER platform [10] was used as retrieval engine with a InL2¹ ranking model [11]. No stopword removal or query expansion were applied at this point.

For querying, the source language topic² is firstly split into n -grams. Next, these n -grams are replaced by their candidate translations according to a selection algorithm, and the resulting translated topics are then submitted to the retrieval system. Two alternative selection algorithms were implemented: a *top-rank-based* algorithm, that takes the N highest ranked n -gram alignments according to their association measure, and a *threshold-based* algorithm, that takes those alignments whose association measure is greater than or equal to a threshold T .

The work presented in this paper was developed in two phases. Firstly, the *training topics* subset was used for tuning the system for the different association measures implemented: the Dice coefficient, mutual information and log-likelihood. Next, the performance was tested using the *test topics* subset.³

3.1 Tuning Runs Using the Dice Coefficient

The first tuning runs were made for the Dice coefficient and the top-rank-based selection algorithm, that is, by taking the target n -grams from the N top n -gram-level alignments with the highest association measures. Different values were tried, with $N \in \{1, 2, 3, 5, 10, 20, 30, 40, 50, 75, 100\}$. The results obtained are shown in the left hand graph of Fig. 1⁴ —notice that mean average precision (MAP) values are also given. The best results were obtained when using a limited number of translations, those obtained with $N=1$ being the best.

The next tuning runs were made for the threshold-based selection algorithm, that is, by fixing a minimal association measure threshold T . Since the Dice coefficient takes values in the range [0..1], we tried different values $T \in \{0.00, 0.001, 0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 1.00\}$. The results obtained are shown in the right hand graph of Fig. 1, with the best ones at $T=0.40$. Nevertheless, these results were significantly less good than those for the top-rank-based algorithm.⁵

¹ Inverse Document Frequency model with Laplace after-effect and normalization 2.

² Only *title* and *description* topic fields were used in the submitted queries.

³ All these experiments must be considered as *unofficial* experiments.

⁴ Only a subset of the results are shown in order not to crowd the figures.

⁵ Two-tailed T-tests over MAPs with $\alpha=0.05$ have been used along this work.

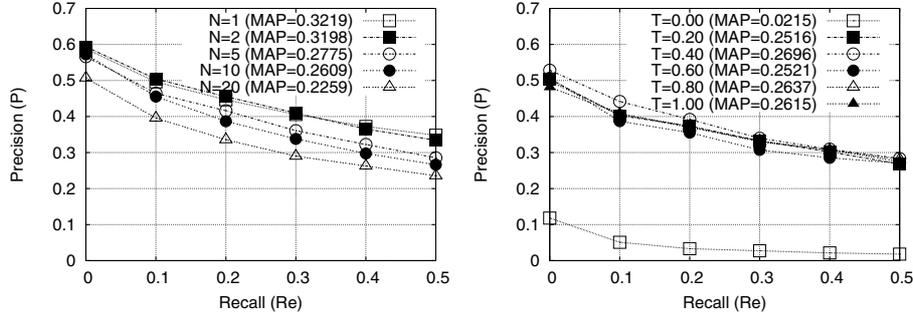


Fig. 1. Tuning precision vs. recall graphs for the Dice coefficient when using the top-rank-based (left) and threshold-based (right) selection algorithms

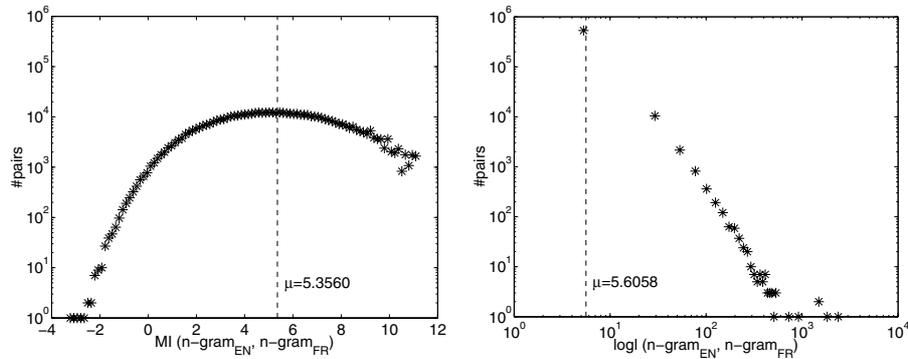


Fig. 2. Distribution of aligned n -gram pairs across their association measures: mutual information (MI , left) and log-likelihood ($logl$, right). Mean (μ) is also shown.

3.2 Tuning Runs Using Mutual Information

The main difference of mutual information (MI) with respect to the Dice coefficient is that the former can take any value within $(-\infty.. + \infty)$ —the distribution found with our data is shown in the left hand graph of Fig. 2—, while the latter takes values within the range $[0..1]$. This had to be taken into account in order to adapt our testing methodology.

In the case of the top-rank-based selection algorithm, we continued taking the N top-ranked n -gram alignments, even if their MI value was negative. The results obtained, shown in the left hand graph of Fig. 3, and with the best performance at $N=10$, were not as good as those obtained with the Dice coefficient.

In the case of the threshold-based algorithm, we had to take into account that the range of MI values may vary considerably for each run. So, in order to homogenize the experiments, the threshold values were not fixed according to concrete values as before, but according to the following formula:

$$T_i = \mu + 0.5 i \sigma . \tag{4}$$

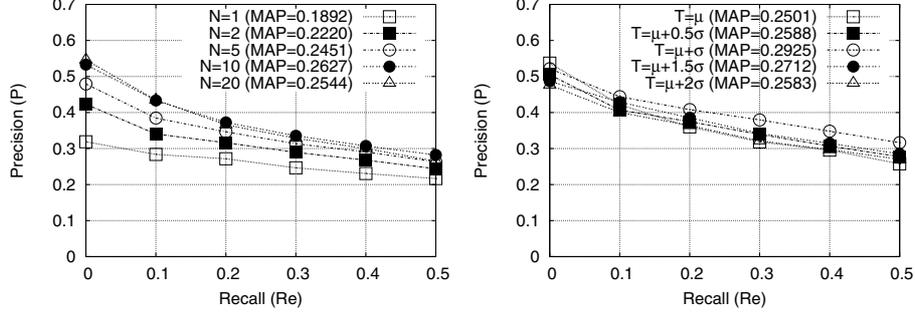


Fig. 3. Tuning precision vs. recall graphs for mutual information when using the top-rank-based (left) and threshold-based (right) selection algorithms

where T_i represents the i -th threshold —with $i \in \mathbb{N}$ —, μ represents the *mean* of the MI values of the aligned n -gram pairs, and σ represents their *standard deviation*. The resulting thresholds are as follows:

$$\mu, \mu + 0.5\sigma, \mu + \sigma, \mu + 1.5\sigma, \dots$$

The right hand graph of Fig. 3 shows the results obtained for this algorithm, which were very similar, although the best ones were obtained for $T = \mu + \sigma$. These results were better than those obtained for the top-rank-based algorithm, but this difference was not statistically significant. However, these results were still not as good as those obtained for the Dice coefficient.

3.3 Tuning Runs Using Log-Likelihood

As before, the first runs used the top-rank-based algorithm. These results, shown in the left hand graph of Fig. 4, and with the best performance at $N=2$, were similar to those obtained for the Dice coefficient.

Regarding the threshold-based selection algorithm, log-likelihood, like MI, does not have a fixed range of possible values. So, as with MI, we established the thresholds according to the *mean* and *standard deviation* of the association measures. Nevertheless, after studying the distribution of the output aligned n -gram pairs across their log-likelihood values —see right hand graph of Fig. 2—, we realized that this distribution was clearly biased towards low values just slightly less than the mean. As a consequence, we worked with varying granularities and developed the following formula for calculating the threshold values:

$$T_i = \begin{cases} \mu + 0.05 i \sigma & -\infty < i \leq 2 \\ \mu + 0.50 (i - 2) \sigma & 2 < i < +\infty \end{cases} \quad (5)$$

where, as before, T_i represents the i -th threshold —this time with $i \in \mathbb{Z}$ —, μ represents the *mean* of the log-likelihood values of the aligned n -gram pairs, and σ represents their *standard deviation*. The resulting thresholds are as follows:

$$\dots \mu - 0.05\sigma, \mu, \mu + 0.05\sigma, \mu + 0.1\sigma, \mu + 0.5\sigma, \mu + \sigma \dots$$

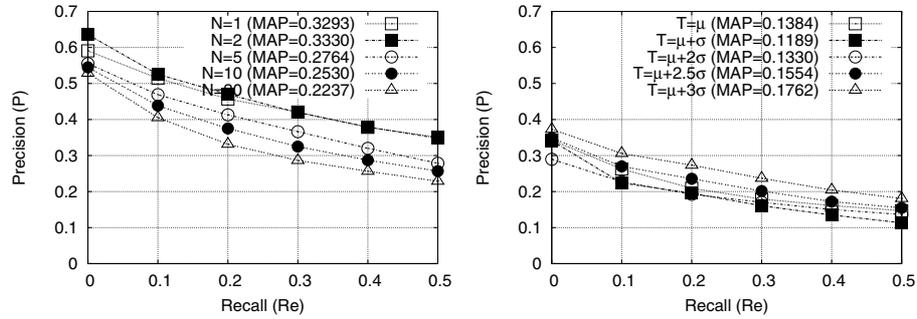


Fig. 4. Tuning precision vs. recall graphs for log-likelihood when using the top-rank-based (left) and threshold-based (right) selection algorithms

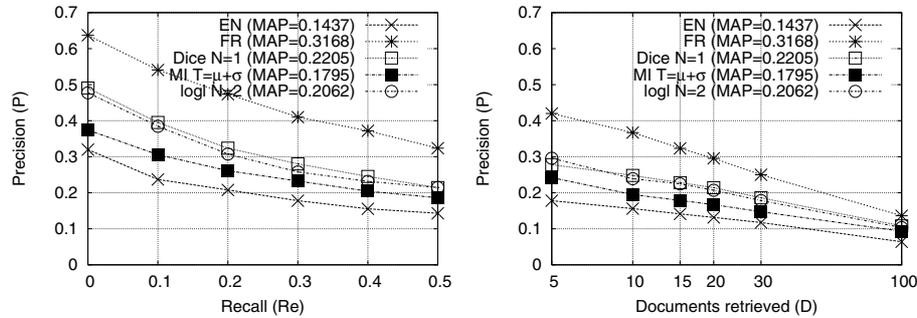


Fig. 5. Precision vs. recall graphs (left) and precision at top D documents graphs (right) for the *test topics* set

The results obtained, shown in the right hand graph of Fig. 4, and with the best performance for $T = \mu + 3\sigma$, were significantly worse than those for the top-rank-based algorithm.

3.4 Test Runs

Once the system had been tuned for the new target language, the proper tests could be performed using the *test topics* set. The best configurations found for each association measure were used in these runs:

- Dice coefficient (EN2FR Dice): top-rank-based selection algorithm ($N = 1$)
- Mutual Information (EN2FR MI): threshold-based selection algorithm ($T = \mu + \sigma$)
- Log-likelihood (EN2FR logl): top-rank-based selection algorithm ($N = 2$)

Fig. 5 presents the results obtained for the test runs with respect to two baselines: the first 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 the other obtained by querying the French index using the French topics split into 4-grams (FR) —i.e. a French monolingual run and our *ideal performance* goal.

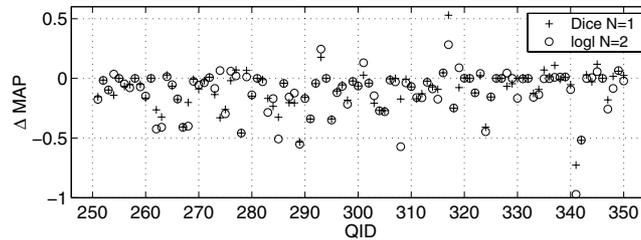


Fig. 6. MAP differences with the French monolingual run for the *test topics* set

These results showed that the Dice coefficient and the log-likelihood measure produced the best results —when using the top-rank-based algorithm. Both approaches performed significantly better than mutual information —the latter using a threshold-based algorithm. Regarding the baselines, all configurations performed significantly better than the English topics run, showing that our positive results were not coincidental. On the other hand, these results were still not as good as the French monolingual run, our *ideal performance* goal, but it must be taken into account that this approach is not still fully developed, so there is margin for improvement. Finally, Fig. 6 shows the MAP differences with the French monolingual run (FR) obtained for each topic in the case of our best configurations: the Dice coefficient (Dice) and the log-likelihood (logl) measure when using the top-rank-based algorithm.

4 Conclusions and Future Work

This work presents a knowledge-light approach for query translation in Cross-Language Information Retrieval systems based on the direct translation of character n -grams. The experiments shown in this paper are an extension of those performed in the robust English-to-French task of the CLEF 2007, and confirm the positive results previously obtained in our English-to-Spanish experiments [5,6], thus demonstrating the validity of our approach.

With respect to our future work, new tests with other languages of different characteristics are being prepared. We also intend to simplify the processing for reducing the computational costs even more. Finally, the employment of relevance feedback, or the use of pre or post-translation expansion techniques in the case of translanguing runs [2] are also being considered.

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 (PGIDIT07SIN005206PR, PGIDIT05PXIC30501PN, and *Rede Galega de Procesamento da Linguaxe e Recuperación de Información*).

References

1. <http://www.clef-campaign.org> (visited on November 2007)
2. McNamee, P., Mayfield, J.: JHU/APL experiments in tokenization and non-word translation. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 85–97. Springer, Heidelberg (2004)
3. Vilares, J., Oakes, M.P., Tait, J.I.: A first approach to CLIR using character n-grams alignment. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 111–118. Springer, Heidelberg (2007)
4. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1), 19–51 (2003), <http://www.fjoch.com/GIZA++.html> (visited on November 2007)
5. Vilares, J., Oakes, M.P., Vilares, M.: A knowledge-light approach to query translation in cross-language information retrieval. In: Proc. of International Conference on Recent Advances in Natural Language Processing (RANLP 2007), pp. 624–630 (2007)
6. Vilares, J., Oakes, M.P., Vilares, M.: Character n-grams translation in cross-language information retrieval. In: Kedad, Z., Lammari, N., Métais, E., Meziane, F., Rezgui, Y. (eds.) NLDB 2007. LNCS, vol. 4592, pp. 217–228. Springer, Heidelberg (2007)
7. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: NAACL 2003: Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 48–54 (2003)
8. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge (1999)
9. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007 ad hoc track overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2007)
10. <http://ir.dcs.gla.ac.uk/terrier/> (visited on November 2007)
11. Amati, G., van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems* 20(4), 357–389 (2002)