A First Approach to CLIR Using Character *N*-Grams Alignment

Jesús Vilares[1], Michael P. Oakes[2], and John I. Tait[2]

[1] Departamento de Computación, Universidade da 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, John.Tait}@sunderland.ac.uk

Abstract. This paper describes the technique for translation of character n-grams we developed for our participation in CLEF 2006. 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 makes considerable use of freely available resources and also tries to achieve a higher speed during the n-gram alignment process with respect to other similar approaches.

Key words: Cross-Language Information Retrieval, character n-grams, translation algorithms, alignment algorithms, association measures.

1 Introduction

This work has been inspired by the previous approach of the Johns Hopkins University Applied Physics Lab (JHU/APL) on the employment of overlapping character n-grams for indexing documents [7, 8]. Their interest came from the possibilities that overlapping character n-grams may offer particularly in the case of non-English languages: to provide a surrogate means to normalize word forms and to allow one to manage languages of very different natures without further processing. 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 both queries and documents are just tokenized into overlapping n-grams instead of words. In the case of cross-language retrieval, two phases are required during query processing: translation and n-gram splitting. In their later experiments, JHU/APL developed a new *direct n-gram translation* technique that used n-grams instead of words as translation units. Their goal was to avoid some of the limitations of classical dictionary-based translation, such as the need for word normalization or the inability to handle out-of-vocabulary words.

This n-gram translation algorithm takes as input a parallel corpus, aligned at the paragraph (or document) level and extracts candidate translations as follows [8]. 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 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. 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.

This paper describes our first experiments in the field of Cross-Language Information Retrieval (CLIR) employing our own direct n-gram translation technique. This new approach also tries to achieve a higher speed during the n-gram alignment process in order to make easier the testing of new statistical measures. The article is structured as follows. Firstly, Sect. 2 describes our n-gram-based CLIR system. Next, Sect. 3 shows the results of our first tests, still in development. Finally, Sect. 4 presents our preliminary conclusions and future work.

2 The Character *N*-Gram Alignment Algorithm

Taking as our model the system designed by JHU/APL, we developed our own n-gram based retrieval system. 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 have opted for using the open-source retrieval platform Terrier [1]. This decision was supported by the satisfactory results obtained with n-grams using different indexing engines [10].

A second difference comes from the translation resources to be used, in our case the well-known Europarl parallel 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: Romance (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish.

Finally, with respect to the n-gram translation algorithm itself, it now consists of two phases. In the first phase, the slowest one, the input parallel corpus is aligned at word-level using the well-known statistical aligner GIZA++ [9], obtaining the translation probabilities between the different source and target language words. 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, the least probable candidate translations, or combine the scores of bidirectional alignments [5] —source-target language and target-source language— instead of using just the direct one —source-target language. Finally, in the second phase, n-gram translation scores are computed employing statistical association measures [6], taking as input the translation probabilities calculated by GIZA++.

This approach increases the speed of the process by 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.

2.1 Word-Level Alignment Using Association Measures

In order to better illustrate the process involved in this second phase, we will take as basis how association measures could be used for creating bilingual dictionaries taking as input parallel collections aligned at paragraph level. In this 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}	$= R_2$
	$= C_1$	$= C_2$	$= N$

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 are called the *observed frequencies*: O_{11}, for example, stands for the number of aligned paragraphs where the source language paragraph contains $word_u$ and the target language paragraph contains $word_v$. The sum of the observed frequencies —or *sample size* N— is the total number of word pairs considered. 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, will take part of the bilingual dictionary. Our system employes two classical measures: *mutual information* and the *Dice coefficient*, defined by equations 1 and 2, respectively:

$$MI(word_u, word_v) = log\frac{NO_{11}}{R_1C_1} \quad (1) \qquad Dice(word_u, word_v) = \frac{2O_{11}}{R_1 + C_1} \quad (2)$$

2.2 Adaptations for *N*-Gram-Level Alignment

We have described how to compute and employ association measures for generating bilingual dictionaries from parallel corpora aligned at the paragraph level. However, in our proposal, we do not have aligned paragraphs but aligned words —a source word and its candidate translation—, both composed of n-grams. Our first idea could be just to adapt the contingency table to this context, by considering that we are now dealing with n-gram pairs $(n\text{-}gram_u, n\text{-}gram_v)$ cooccurring in aligned words instead of word pairs $(word_u, word_v)$ cooccurring in 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 word pairs where the source language word contains $n\text{-}gram_u$ and the target language word contains $n\text{-}gram_v$.

This solution seems logical, but there is a problem. In the case of aligned paragraphs, we had *real* instances of word cooccurrences at the paragraphs aligned. However, now we do not have *real* instances of n-gram cooccurrences at aligned words —as may be expected—, but just *probable* ones, since GIZA++ uses a statistical alignment model which computes a translation probability for each cooccurring word pair. So, the same word may appear as being aligned with several translation candidates, each one with a given 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`	0.92
`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 `-milk-` was `-lech-` rather than `-toma-`, 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 a basis, our proposal consists of weighting the likelihood of a cooccurrence according to the probability of its corresponding alignment. 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 that the O_{11} frequency corresponding to the n-gram pair *(`-milk-`, `-lech-`)* is not 2 as might 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 (for `milk` with `leche`) + 0.92 (for `milky` with `lechoso`) = 1.90.

Once the contingency tables have been obtained, the Dice coefficients corresponding to each n-gram pair can be computed, for example. 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:

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

3 Evaluation

Our group took part in the CLEF 2006 ad-hoc track [11], but the lack of time did not allow us to complete our n-gram direct translation tool. So, we could submit only those results intended to be used as baselines for future tests. Since these results are publicly available in [2], we will discuss here only those new results obtained in later experiments.

These new experiments were made with our n-gram direct translation approach using the English topics and the Spanish document collections of the robust task —i.e., a English-to-Spanish run. The *robust task* is essentially an ad-hoc task which takes the topics and collections used from CLEF 2001 to CLEF 2003. In the case of the Spanish data collection, it is formed by 454,045 news reports (1.06 GB) provided by EFE, a Spanish news agency, corresponding to the years 1994 and 1995. The test set consists of 160 topics (C041–C200). This initial set is divided into two subsets: a *training topics* subset to be used for tuning purposes and formed by 60 topics (C050–C059, C070–C079, C100–C109, C120–C129, C150–159, C180–189), and a *test topics* subset for testing purposes. Since the goal of these first experiments is the tuning and better understanding of the behavior of the system, we will only use here the *training topics* subset. Moreover, only *title* and *description* 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 use 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[3] ranking model [3]. No stopword removal or query expansion have been applied at this point.

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

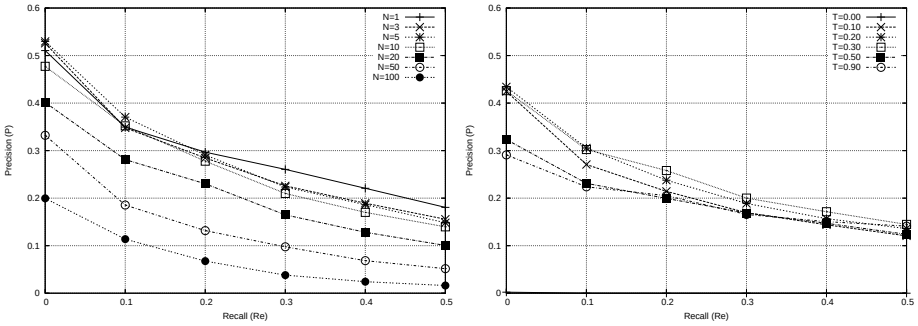

Fig. 1. Precision vs. Recall graphs when taking the N most probable n-gram translations (*left*) and when using a minimal probability threshold T (*right*).

For querying, the source language topic is firstly split into n-grams, 4-grams in our case. Next, these n-grams are replaced by their candidate translations according to a selection algorithm. Two algorithms are available: the first one takes the N most probable alignments, and the second one takes those alignments with a probability greater or equal than a threshold T. The resulting translated topics are then submitted to the retrieval system.

With respect to the n-gram alignment itself, we have refined the initial word-level alignment by using bidirectional alignment. That is, we will consider a ($word_{English}$, $word_{Spanish}$) English-to-Spanish alignment only if there is a corresponding ($word_{Spanish}$, $word_{English}$) Spanish-to-English alignment. Finally, we have used only one of the association measures available, the Dice coefficient.

The results for this first approach are shown in the two Precision vs. Recall graphs of Fig. 1[4]. The figure on the left shows the results obtained when taking the N most probable n-gram translations, with $N \in \{1, 3, 5, 10, 20, 50, 100\}$. The figure on the right shows the results obtained when using a minimal probability threshold T, with $T \in \{0.00, 0.10, 0.20, 0.30, 0.50, 0.90\}$. As it can be seen, the results taking the N most probable alignments are better, particularly when using few translations.

Next, trying to improve the accuracy of the n-gram alignment process, we removed those least-probable word alignments from the input (those with a word translation probability less than a threshold W, with W=0.15). The new results obtained, shown in Fig. 2, are very similar to those without pruning. Nevertheless, such pruning led to a considerable reduction of processing time and storage space: 95% reduction in the number of input word pairs processed and 91% reduction in the number of output n-gram pairs aligned. The results taking the N most probable alignments impove as N is reduced.

Finally, Fig. 3 shows our best results with respect to several baselines: by querying the Spanish index with the English topics split into 4-grams (*EN_4gr*) —for measuring the impact of casual matches—, by using stemmed Spanish

[4] Highest recall levels, lesser in importance, have been removed for improving reading.

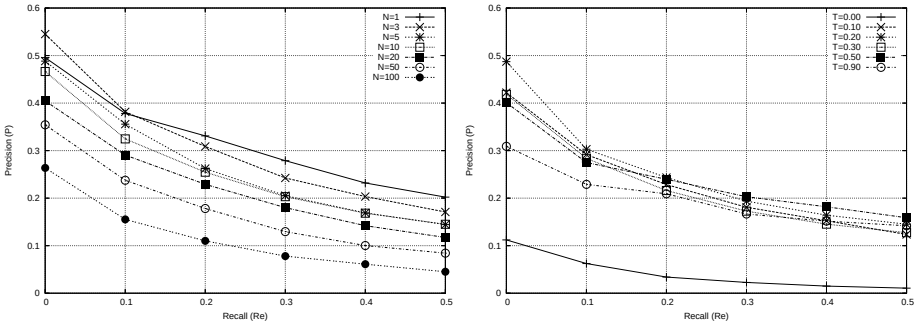

Fig. 2. Precision vs. Recall graphs when taking the N most probable n-gram translations (*left*) and when using a minimal probability threshold T (*right*). Input word alignments with a probability lesser than W=0.15 have been dismissed.

topics (*ES_stm*), and by using Spanish topics split into 4-grams (*ES_4gr*) —our performance goal. Although current performance is not as good as expected, these results are encouraging, since it must be taken into account that these are our very first experiments, so the margin for improvement is still great.

4 Conclusions and Future Work

This paper describes our initial work in the field of Cross-Language Information Retrieval in developing a 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 freely available resources were used when possible. Our n-gram alignment algorithm 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 future work, once tuned, the system will be tested with other bilingual and multilingual runs. The employment of relevance feedback, and the use of pre or post-translation expansion techniques, are also planned. Finally, we also intend to try new association measures [12] for n-gram alignment.

Acknowledgment

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).

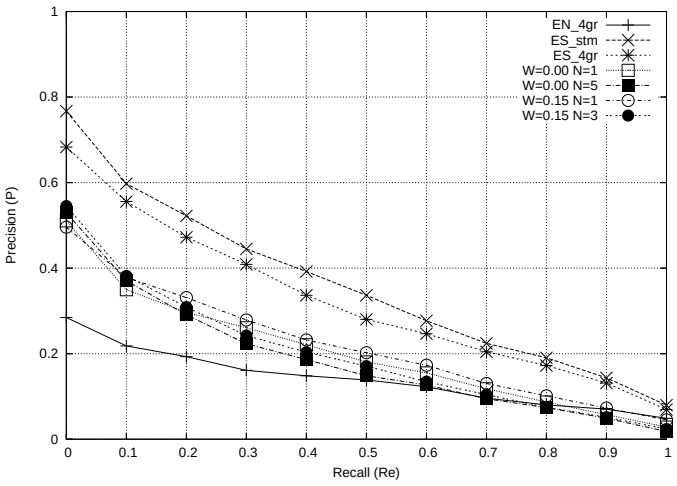

Fig. 3. Summary Precision vs. Recall graph.

References

1. `http://ir.dcs.gla.ac.uk/terrier/`.
2. `http://www.clef-campaign.org`.
3. Gianni Amati and Cornelis Joost van Rijsbergen. Probabilistic models of Information Retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, 2002.
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*, pp. 79–86, 2005. Corpus available in `http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl/`.
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*, pp. 48–54, Morristown, NJ, USA, 2003. ACL.
6. Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.
7. Paul McNamee and James Mayfield. Character N-gram Tokenization for European Language Text Retrieval. *Information Retrieval*, 7(1-2):73–97, 2004.
8. Paul McNamee and James Mayfield. JHU/APL experiments in tokenization and non-word translation. Vol. 3237 of *Lecture Notes in Computer Science*, pp. 85–97. Springer-Verlag, Berlin-Heidelberg-New York, 2004.
9. F. J. Och and H. Ney. A systematic comparison of various statistical alignment models, 2003. Toolkit available at `http://www.fjoch.com/GIZA++.html`.
10. Jacques Savoy. Cross-Language Information Retrieval: experiments based on CLEF 2000 corpora. *Information Processing and Management*, 39:75–115, 2003.
11. Jesús Vilares, Michael P. Oakes, and John I. Tait. CoLesIR at CLEF 2006: rapid prototyping of a N-gram-based CLIR system. In *Working Notes of the CLEF 2006 Workshop, 20-22 September, Alicante, Spain*, 2006. Available at [2].
12. Julie Weeds and David Weir. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475, 2005.