On the Feasibility of Character n-Grams Pseudo-Translation for Cross-Language Information Retrieval Tasks

Jesús Vilares[a,*], Manuel Vilares[b], Miguel A. Alonso[a], Michael P. Oakes[c]

[a] *Grupo LYS, Departamento de Computación, Facultade de Informática, Universidade da Coruña, Campus de A Coruña, 15071 – A Coruña (Spain)*
[b] *Grupo COLE, Departamento de Informática, Escola Superior de Enxeñaría Informática, Universidade de Vigo, Campus de As Lagoas, 32004 – Ourense (Spain)*
[c] *Research Institute of Information and Language Processing, University of Wolverhampton, Stafford St., Wolverhampton – WV1 1NA (United Kingdom)*

Abstract

The field of Cross-Language Information Retrieval relates techniques close to both the Machine Translation and Information Retrieval fields, although in a context involving characteristics of its own. The present study looks to widen our knowledge about the effectiveness and applicability to that field of non-classical translation mechanisms that work at character n-gram level. For the purpose of this study, an n-gram based system of this type has been developed. This system requires only a bilingual machine-readable dictionary of n-grams, automatically generated from parallel corpora, which serves to translate queries previously n-grammed in the source language. n-Gramming is then used as an approximate string matching technique to perform monolingual text retrieval on the set of n-grammed documents in the target language.

The tests for this work have been performed on CLEF collections for seven European languages, taking English as the target language. The performance attained, close to the upper baseline, confirms the validity of character n-gram based approaches for Cross Language Information Retrieval tasks, both for indexing–retrieval and translation purposes, these not being tied to a given implementation.

NOTICE: this is the authors version of a work that was accepted for publication in Computer Speech and Language. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in *Computer Speech and Language, 36:136-164, 2016. DOI 10.1016/j.csl.2015.09.004*

[*] Corresponding author: tel. +34 981 167 000 ext. 1364, fax +34 981 167 160.
Email addresses: jesus.vilares@udc.es (Jesús Vilares), vilares@uvigo.es (Manuel Vilares), miguel.alonso@udc.es (Miguel A. Alonso), michael.oakes@wlv.ac.uk (Michael P. Oakes)

Keywords:
Cross-Language Information Retrieval, character n-grams, alignment algorithms for Machine Translation

1. Introduction

Nowadays, not only has the amount and diversity of information available online risen dramatically, but users worldwide can also easily and instantly access and publish data. An immediate consequence is that data exists in many different languages, a fact that will remain over time and which justifies the increasing interest in finding ways of retrieving information across language boundaries. In response to this need, the aim of *Cross-Language Information Retrieval* (CLIR) is to provide techniques to return relevant documents written in a language (named the *target language*) different from the language in which the query was written (named the *source language*). Most current approaches manage CLIR by reducing it to well-known monolingual *Information Retrieval* (IR) counterparts (Nie, 2010; Grefenstette, 1998). This implies that we must answer three enchained questions (Kwok et al., 2005):

1. How a term expressed in one language might be expressed in another?
2. Which of the possible translations should be retained for the subsequent IR task?
3. How to properly weight the importance of translation candidates (in the event that more than one is retained)?

Depending on whether it is the queries, the documents, or both that are translated, we talk about query translation, document translation or interlingual-based CLIR (Wu et al., 2008).

In practice, study in this domain has focused mainly on query translation because it is computationally expensive to translate large-scale text collections (Nie, 2010; Gao et al., 2010b; McCarley, 1999; Hull and Grefenstette, 1996). In spite of this drawback, document translation has also deserved the attention of researchers. This is because a translation system can better exploit linguistic context to choose correct translations in documents than in queries. In particular, this kind of technique has proved from the beginning to be capable of generating competitive search results to monolingual searches (Nie, 2010; McCarley, 1999; Oard, 1998) when it works in combination with *Machine Translation* (MT) techniques.

The interlingual-based CLIR approach is the least popular of the three, although from a theoretical point of view (Dorr et al., 2004) it has many advantages. It is commonly associated with the generation of a language-independent representation for both query and documents. The assumption in this case is that one is able to represent sentences in every language using a standard common descriptive formalism. This should provide us with a robust starting point not only to bilingual CLIR, but also to multilingual CLIR. Unfortunately, the creation of such a language-independent representation turns out to be an

unattainable goal for the moment, which limits in practice the interest of these techniques.

Whatever the approach used, CLIR systems require the use of language resources to achieve their goal, namely machine-readable bilingual dictionaries, *corpus*-based resources and MT systems.

1.1. Character n-gram translation

An *n-gram* is a sub-sequence of n characters from a given word (Robertson and Willett, 1998). For example, `removal` can be split into four overlapping character 4-grams: `-remo-`, `-emov-`, `-mova-` and `-oval-`. In the context of textual information systems, n-gram level processing provides an intermediate level of representation that has advantages in terms of efficiency and effectiveness over the conventional character-based or word-based approaches to text processing (Robertson and Willett, 1998). Today n-grams are used as index terms for IR applications because of these advantages (Vilares et al., 2011; McNamee and Mayfield, 2004a; Robertson and Willett, 1998; Cavnar, 1994).

In this context, McNamee and Mayfield (2004b) were pioneers in the use of character n-grams as translation units for CLIR purposes. Their objective was to avoid some of the limitations of classical dictionary-based translation, such as the need for word normalization, translating multiple word expressions and handling *out-of-vocabulary* (OOV) words (McNamee and Mayfield, 2005). At this point we should clarify that, from a linguistic point of view, they were not *translating* the query, properly speaking, since they were obtaining neither words nor phrases at the output, but character n-grams, i.e. mere pieces of words with no proper meaning. However, from a retrieval perspective, such an approach does work as an actual translation since the query obtained at the output of the direct n-gram translation system, when submitted to the retrieval engine, allows us to obtain the documents we are searching for. This is why although we will abuse the term *translation* throughout this paper, it would in fact be more accurate to talk about *pseudo-translation* instead.

In principle, the use of direct translation of character n-grams provides CLIR systems with a number of significant advantages:

1. The overlapping of n-grams corresponding to a given word provides a way to normalize word forms, avoiding the need for explicit normalization during indexing or translation.
2. It supports the handling of OOV words and the management of languages of very different natures without further processing.
3. It does not rely on language-specific processing and, since only raw text is needed, it can be used even when linguistic information and annotated language resources are scarce or unavailable. Unfortunately, surprising though it may seem, this is by no means an uncommon situation. The Multilingual Europe Technology Alliance Network of Excellence (META-NET) research network,[1] founded by the European Commision and dedi-

[1] `http://www.meta-net.eu`.

cated to fostering the technological foundations of a multilingual European information society, has recently published a study about it (Rehm and Uszkoreit, 2011). This report shows that the state of language technology for European languages, even official ones, is still far from being accurate for most of them, especially in the case of Machine Translation, where fragmentary, weak or even no support at all is the common rule. This situation is even worse for most of the rest of world languages (Nakov and Ng, 2012). Luckily, parallel raw text can be still obtained from either the Web (Resnik and Smith, 2003), legal or administrative texts (Koehn, 2005) or other varied sources (Chew et al., 2006).

Taking the model of McNamee and Mayfield (2004b) as source of inspiration, we have implemented for this study a CLIR system based on a knowledge-light query translation module which uses character n-grams as processing units, not only for indexing purposes, but also during the query translation process. In this system, the n-grammed source language query is translated at n-gram level too before being submitted to the retrieval engine in order to search the target collection, which has also been indexed using character n-grams. This implementation maintains a fundamental difference with regard to the original system developed by McNamee and Mayfield (2004b), which concerns the type of the n-gram alignment applied, this being in fact the kernel of the system. Although both implementations take as input a parallel corpus for training the n-gram based translator, in the case of our implementation the corpus is aligned in two phases:

- Firstly, the parallel corpus is aligned at the word-level using statistical techniques (Och and Ney, 2003), allowing us to obtain the lexical translation probabilities between the different source and target language words. Such basic alignment can be refined by applying threshold-based filtering and bidirectional word-level alignment (Koehn et al., 2003), providing increased robustness and reliability. Subsequent processing then focuses only on those words whose translation is less ambiguous, considerably reducing the number of input word pairs to be processed and, consequently, the noise introduced into the system.

- In a second phase we focus on the n-gram translation level, for which scores are computed using statistical association measures (Manning and Schütze, 1999) taking as input the translation probabilities previously calculated at word level, and weighting the likelihood of a co-occurrence according to the probability of its containing word-level alignments.

All these processes and their corresponding configurations will be later explained in detail in Section 3.1.

A first try to make a study like the one we present in this work, in that case for English-to-Spanish text retrieval, was presented by the authors in Vilares et al. (2007a). These initial experiments were limited since only one association measure, the Dice coefficient, and only one word alignment configuration,

bidirectional alignment, were tested. These experiments were later extended to new association measures in Vilares et al. (2007b) and Vilares et al. (2009), but they were not as complete as desired since the use of unidirectional word alignment was only partially tested, and no tuning experiments were analyzed. Moreover, the experiments on the use of pointwise mutual information as association measure should be dismissed since, as we have recently discovered during the development of the present article, the range of values employed for those tests was too narrow, thus attaining a much lower performance than it should have been. Finally, the authors also showed in Vilares et al. (2008) some preliminary experiments for English-to-French CLIR using a different test set with a few configurations. A critical drawback of all these preliminary works is that since they were not made according to the needs of a proper testing of our proposed approach but according to the specific requirements of the conference or publication in question, there does not exist a common framework that allowed for an accurate comparison and analysis of the results obtained. Thus the generalization and validity of these previous conclusions of Vilares et al. are arguable. Therefore, there is a need for a common framework specifically designed for allowing us to make an extensive and homogeneous comparative study of the feasibility of n-gram based CLIR for a wide range of languages and a wide range of running configurations. The present work gives a response to this need and allows us to perform a wide range of experiments also involving languages from different language-families in order to analyze CLIR approaches based on approximate word matching.

1.2. Research Objective

The main goal of this work is to make an extensive study of the applicability of character n-gram based translation in the context of Cross-Language Information Retrieval. The questions we are looking to answer are:

1. Is the behavior of n-gram based translation consistent across different languages?
2. Which is the most effective way of applying it?

Moreover, in order to avoid any distortion in the results, no improvement techniques such as query expansion or relevance feedback have been introduced during our experiments, thus allowing us to study the performance of this approach on its own.

1.3. Outline

The structure of the rest of this article is as follows. Firstly, Section 2 introduces the reader to the use of character n-grams in text processing tasks. Next, Section 3 presents a framework for CLIR based on character n-gram translation. Section 4 introduces the methodology followed in our tests, while the following sections deal with our experiments and their discussion: Sections 5 and 6 present and discuss in detail, respectively, the results obtained for our first set of experiments, corresponding to the tuning of the system in a Spanish-to-English CLIR context. The experiments corresponding to the remaining

languages of our study, which can be seen as our test experiments, are presented and analyzed in a more concise way in Section 7. After these language-specific studies, a general discussion on the results obtained as a whole is presented in Section 8. Finally, Section 9 presents our conclusions and proposals for future work.

2. Background and Related Work

Character n-grams have been successfully used for a long time in a wide variety of text processing problems and domains, including the following: approximate word matching (Zobel and Dart, 1995), language identification (Lui et al., 2014) spelling-error detection (Salton, 1989), author attribution and profiling (Stamatatos, 2009; Escalante et al., 2011; Sapkota et al., 2013), and bioinformatics (Tomović et al., 2006). More recently, character n-grams have been drawing increasing attention in the field of automatic processing of SMS and microblog (e.g. Twitter) texts —which tend to be noisy by nature—, including tasks such as text normalization (Pennell and Liu, 2014), sentiment analysis (Aisopos et al., 2012) or language identification (Lui and Baldwin, 2014).

In this way, n-gram based processing has become a standard state-of-the-art text processing approach, whose success comes from its positive features (Tomović et al., 2006):

- Simplicity: no linguistic knowledge or resources are required.
- Robustness: relatively insensitive to spelling variations and errors.
- Domain independence: language and topic independent.
- Efficiency: one pass processing.

This fact has not been ignored by the IR community either (Büttcher et al., 2010, Ch. 3). In the following, we explain in some detail these advantages for the particular case of IR.

A first major advantage of character n-grams when applied to IR is their inherent simplicity and ease of application (Foo and Li, 2004). IR systems typically utilize language-specific linguistic tools and resources to facilitate retrieval: stopword lists, phrase lists, stemmers, decompounders, lexicons, thesauri, part-of-speech taggers, etc. Obtaining and integrating these resources into the system may be costly (McNamee and Mayfield, 2004a). In contrast, character n-gram tokenization is a knowledge-light approach which does not rely on language-specific processing (Damashek, 1995; Cavnar, 1994), thus requiring no prior information about document contents or language. Basically, both queries and documents are simply tokenized into overlapping n-grams instead of words, and the resulting terms are then processed as usual by the retrieval engine. So, this n-gram based approach can be easily incorporated into traditional IR systems.

A second major factor for the usefulness of n-grams in IR is their robustness, which comes from the redundancy derived from the tokenization process itself.

Since every string is decomposed into overlapping small parts, any spelling errors that are present tend to affect only a limited number of those parts, leaving the remainder intact, thus still making matching possible. Therefore, these systems are able to cope not only with spelling errors, but also with OOV words and variants (Vilares et al., 2011; Lee and Ahn, 1996; Mustafa and Al-Radaideh, 2004), in contrast to classical conflation techniques based on stemming, lemmatization or morphological analysis, which are negatively affected by these phenomena.

A third major positive factor to be taken into account with regard to n-grams is their inherent language-independent nature, since no linguistic knowledge is taken into account (Robertson and Willett, 1998; Damashek, 1995). No prior information about stopwords, grammars for stemming, lemmatization, morphological analysis or even tokenization is required for their application. This is because n-gram based matching itself provides a surrogate means of normalizing word forms, thus allowing languages of very different natures to be managed without further processing (McNamee and Mayfield, 2004a). This is a very important factor, particularly in the case of multilingual environments or when linguistic resources are scarce or unavailable which, as we have explained in Section 1.1, is not unusual.

However, the use of n-gram based indexing, as with any other technique, is not free of drawbacks, the main one being the need for higher response times and storage space requirements due to the larger indexing representations they generate (Miller et al., 2000; McNamee and Mayfield, 2004a). The logical choice for minimizing this problem would be to reduce the index by using some kind of pruning (Carmel et al., 2001) or term selection (Zeman, 2009) technique.

Monolingual n-gram based retrieval has been successfully applied to a wide range of languages of very different natures and widely differing morphological complexity. It has been successfully applied, for example, with most European languages (McNamee, 2008; McNamee and Mayfield, 2004a; Savoy, 2003; Hollink et al., 2004) —being particularly accurate for compounding and highly inflectional languages—, but also with many others such as Turkish (Ekmekcioglu et al., 1996), Arabic (Khreisat, 2009; Mustafa and Al-Radaideh, 2004) and several Indian languages (Dolamic and Savoy, 2008), being particularly popular and effective in Asian IR (Foo and Li, 2004; Ogawa and Matsuda, 1999; Lee and Ahn, 1996) because of their unsegmented and agglutinative nature.

A related approach which makes use of the ability of n-grams to manage variants is their application to CLIR over closely-related languages using no translation, but only cognate matching.[2] Such an approach has been applied not only to classical CLIR tasks (McNamee and Mayfield, 2004a), but also in cross-language plagiarism detection (Potthast et al., 2011), for example.

Other IR-related, but more complex, application of n-grams are the use of *skipgrams* (McNamee, 2008), also referred to as *gap-n-grams* (Mustafa, 2005) or *s-grams* (Järvelin et al., 2008) by other authors. This is a generalization of

[2] *Cognates* are words with a common etymological origin. For example: *"traducción"* (‘‘`translation`’’) in Spanish vs. *"tradución"* in Galician vs. *"tradução"* in Portuguese.

the concept of n-gram by allowing *skips* during the matching process. However, McNamee (2008) shows that skipgrams are dramatically more costly than traditional n-grams without being demonstrably more effective. Moreover, their application is much more complex than for regular n-grams, since they require considerable modifications in the IR system. For these reasons their use here has been discarded.

3. A System for CLIR Based on Character n-Gram Translation

The so-called *direct n-gram translation* algorithm proposed by McNamee and Mayfield (2004b) takes as input a parallel corpus, aligned at the sentence (or document) level, and extracts candidate translations as follows. Firstly, for each candidate n-gram term to be translated, sentences containing this term in the source language are identified. Next, their corresponding sentences 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 these target language texts. Finally, the target n-gram with the highest translation score is selected as the potential translation of the source n-gram.

However, this first proposal proved to be improvable. Firstly, it lacked of flexibility, at least from an experimental perspective, since a single statistical measure is available for calculations and, for a given source n-gram, only the top-scored translation candidate is returned. So, what if we want to use other association measures or to try to expand the query with more translation candidates? Secondly, the way their n-gram alignment algorithm works is not very efficient. As explained above, their system takes as input a parallel corpus and every source language n-gram g_s of the input is cross-checked with every n-gram g_t of every target language sentence (or, even worse, document) aligned with a source language sentence (or document) of the corpus containing g_s. The number of the resulting combinations to be checked will rise dramatically, thus reducing the efficiency and increasing the consumption of computational resources. All of this constitutes a problem when trying new solutions or modifications. Moreover, it also integrated numerous closed-source resources and non-standard solutions, thus hampering its applicability and the reproducibility of the experiments. For this reason we decided to use for this study an n-gram based CLIR system of our own, looking for a more flexible experimentation platform for future developments, preserving the advantages of the original solution but at the same time avoiding its main drawbacks. Our immediate goals were to speed up the training process, to retrieve multiple translation candidates when available (the original system used a one-to-one translation policy) and to make use of freely available resources when possible. This allows us to minimize effort, to make it more transparent and to facilitate replication of the experiments by the research community.

3.1. Overview of the system

We have opted for a query translation based approach that uses as linguistic resource a parallel corpus, and character n-grams as terms. Essentially, the source language n-grammed query is translated into the target language to later perform the IR task on the collection of target documents, which is also indexed using character n-grams. This method maintains a fundamental difference from the original model proposed by McNamee and Mayfield (2004b) due to the type of the n-gram alignment to be applied, the kernel of the system, which now consists of two phases:

1. In the first phase, the input parallel corpus is aligned at the word level using a statistical aligner, the well-known statistical tool GIZA++ (Och and Ney, 2003), obtaining as output the lexical translation probabilities between the different source and target language words.[3] This first step acts as an initial filter, since only those n-gram pairs corresponding to aligned words will be considered in the subsequent process, thus focusing only on those words whose translation is less ambiguous. In this way, we will be reducing considerably the number of input word pairs to be processed and, consequently, both the noise introduced in the system and the number of entries to be processed, thus improving efficiency too. This contrasts with the original system developed by McNamee and Mayfield (2004b), which has a coarse-grained granularity since it works directly at sentence (or even document) level alignment, thus processing all the n-gram pairs contained in two aligned sentences (or documents).
2. In a second phase we focus on the n-gram translation level. Taking as input the resulting word-level alignments obtained in the previous phase and their probabilities, we compute the n-gram alignment scores by employing statistical association measures (Manning and Schütze, 1999).[4]

This two-step solution allows us to speed up the training process, since it concentrates most of the complexity in the initial word-level alignment phase, thus making the testing of new association measures or new procedures for n-gram alignment easier.

There are other important differences with regard to the implementation of McNamee and Mayfield (2004b). Freely available resources are used this time, which allows us to minimize effort and increase transparency. This way, as explained before, the initial word-level alignment is performed through the widely-used statistical translation tool GIZA++ (Och and Ney, 2003). Moreover, instead of the closed-source retrieval system employed by the original system, the TERRIER open-source retrieval platform (Ounis et al., 2007) is used here. Regarding the translation resources to be used, while McNamee and Mayfield employed a parallel corpus of their own, the well-known EUROPARL (release

[3] From this point forward, when referring to this type of alignment, we will talk about *word-level alignments* or, simply, *word alignments*.

[4] In this case we will talk about *n-gram level alignments* or *n-gram alignments*.

v6) parallel corpus (Koehn, 2005)[5] has been used in this work.[6] More detailed information about the resources integrated in our implementation is given in Section 4.1. Finally, as will be described below, the system uses three different standard association measures (Manning and Schütze, 1999) in its calculations, making our implementation more transparent and flexible.

In the following subsections, we will describe our implementation in more detail.

3.2. Processing parallel corpora using association measures: a generic example

In order to better illustrate the n-gram level alignment algorithm used in our implementation, we introduce a generic and simpler case first, where we take as input a parallel corpus of aligned *sequences* of *items*, and we obtain as output a list of pairs of aligned *items*.

In this initial context, given an item pair (x_s, x_t) —x_s standing for the source language item, and x_t for its candidate target language translation—, their co-occurrence frequency can be organized in a *contingency table* like this resulting from a cross-classification of their co-occurrences in the input aligned corpus:

	$x_t \in T$	$x_t \notin T$	
$x_s \in S$	O_{11}	O_{12}	$= R_1$
$x_s \notin S$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

The first row accounts for those instances where the source language sequence S contains item x_s, while the second row accounts for those instances where the source language sequence S does not contain x_s; in the same way, the first column accounts for those instances where the target language sequence T contains item x_t, while the second column accounts for those instances where the target language sequence T does not contain x_t. The cell counts are called the *observed frequencies*: O_{11}, for example, stands for the number of aligned sequences where the source language sequence contains x_s and the target language sequence contains x_t; O_{12} stands for the number of aligned sequences where the source language sequence contains x_s but the target language sequence does not contain x_t; and so on. *Sample size* N, the total number of item pairs considered, 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 (Manning and Schütze, 1999) can easily be calculated for each item pair (x_s, x_t).

[5] NOTE FOR THE REFEREES: Although the reference is much older than release v6, it is the one required to be used by Prof. Koehn, EUROPARL author.

[6] It should be noted that McNamee et al. (2009) did use EUROPARL v3 corpus for their experiments, but without using word-level alignment.

The most promising correspondences, those pairs with the highest association measures, would be selected for generating a bilingual dictionary. Thus, we would have obtained aligned items from aligned sequences.

3.3. Using association measures for n-gram level alignment

In the previous subsection, we described how to compute and use association measures for automatically generating bilingual item dictionaries taking as input parallel corpora of aligned sequences of items. Now, we will explain how to adapt this technique to our particular case: how to generate aligned character n-grams taking as input previously aligned words. This is the way the second phase of the n-gram level alignment algorithm employed in the system works: the word pairs previously aligned by GIZA++ in the first phase are processed in order to obtain the final output n-gram level alignments.

An easy choice could be simply to directly adapt the contingency table and the corresponding calculations to our context. We could consider that we are managing n-gram pairs (g_s, g_t) co-occurring in aligned words instead of item pairs (x_s, x_t) co-occurring in aligned sequences, as in the previous section. So, contingency tables should be adapted accordingly: O_{11}, for example, should be re-formulated as the number of aligned word pairs obtained through GIZA++ where the source language word w_s contains n-gram g_s and the target language word w_t contains n-gram g_t.

Unfortunately, although this simple solution seems logical, it is not correct. It must be noted that in this second phase we are taking as input the pairs of words previously aligned with GIZA++ and, since this tool uses a statistical alignment model which computes a lexical translation probability for each co-occurring word pair (Och and Ney, 2003), we will much probably find ourselves that, at its output, the same word may be aligned with several translation candidates, each one with a given probability and with only part of them being right. So, in case we had merely applied the adaptation explained above, the resulting noise introduced into the system would have been excessive since, for example, we would be giving the same credit, as an input evidence, to a word-level translation with only a 5% probability of being right as to another translation with a 95% probability, which does not seem very logical.

In order to explain how to proceed in this context, let us take as a toy example the case of the Spanish words `lluvia` *(`rain`)* and `lluvioso` *(`rainy`)*, and the English words `rain`, `rainy` and `snowy`. A possible input word-level alignment, with its corresponding probabilities and compounding 4-grams, would be:

source term	candidate translation	prob.
`lluvia`= {`-lluv-`,`-luvi-`,`-uvia-`}	`rain`= {`-rain-`}	0.87
`lluvioso`= {`-lluv-`,`-luvi-`,`-uvio-`,`-vios-`,`-ioso-`}	`rainy`= {`-rain-`,`-ainy-`}	0.80
`lluvioso`= {`-lluv-`,`-luvi-`,`-uvio-`,`-vios-`,`-ioso-`}	`snowy`= {`-snow-`,`-nowy-`}	0.22

Notice that these n-grams, those that will be used to calculate the n-gram alignments to be employed in later n-gram level translations, have been obtained by

tokenizing isolated words; as a result, no word-spanning n-gram level alignments may exist. This is the reason why that kind of character n-grams are ignored in our approach. In any case, as shown in (McNamee and Mayfield, 2004a), that will not harm the performance of the system.

Going back to our toy example, the source 4-gram `-lluv-` co-occurs with the target 4-gram `-rain-`, but the alignment between its containing words, `lluvia` and `rain` and `lluvioso` and `rainy`, is not certain (i.e. their translation probabilities are not 100%) and, besides, in the case of the word `lluvioso` there is also a second translation candidate: `snowy`. Nevertheless, it seems much more probable that the *translation* of `-lluv-` is `-rain-` rather than `-snow-`, since the probability of the alignment of their containing words —`lluvioso` and `snowy`— is much lower than that of the words containing `-lluv-` and `-rain-` —the pairs `lluvia` and `rain` and `lluvioso` and `rainy`. Taking this idea as a basis, the new algorithm we designed reflects this by weighting the likelihood of a co-occurrence according to the translation probability of its containing word alignments.

So, the resulting contingency tables which would correspond to the n-gram pairs *(-lluv-, -rain-)* and *(-lluv-, -snow-)* are as follows:

	-rain-$\in T$	-rain-$\notin T$	
-lluv-$\in S$	O_{11}=**1.67**	O_{12}=**1.24**	R_1=**2.91**
-lluv-$\notin S$	O_{21}=**4.94**	O_{22}=**4.96**	R_2=**9.90**
	C_1=**6.61**	C_2=**6.20**	N=**12.81**

	-snow-$\in T$	-snow-$\notin T$	
-lluv-$\in S$	O_{11}=**0.22**	O_{12}=**2.69**	R_1=**2.91**
-lluv-$\notin S$	O_{21}=**0.88**	O_{22}=**9.02**	R_2=**9.90**
	C_1=**1.10**	C_2=**11.71**	N=**12.81**

Notice that, for example, the O_{11} frequency corresponding to *(-lluv-, -rain-)* is not 2 as might be expected, but 1.67. This is because this n-gram pair appears in two word alignments, *(lluvia, rain)* and *(lluvioso, rainy)*, but each n-gram co-occurrence in these word alignments has been weighted according to its corresponding word translation probability:

O_{11}(`-lluv-`, `-rain-`) = 0.87 for *(lluvia, rain)* + 0.80 for *(lluvioso, rainy)* = **1.67** .

In the case of the O_{12} frequency, it corresponds to n-gram pairs *(-lluv-, g_t)*, with g_t different from `-rain-`. In our example we find: a single pair *(-lluv-, -ainy-)* in the word alignment *(lluvioso, rainy)*; and two pairs *(-lluv-, -snow-)* and *(-lluv-, -nowy-)* in the word alignment *(lluvioso, snowy)*. By weighting each occurrence according to the translation probability of its containing word alignment, we obtain:

O_{12}(`-lluv-`, $g_t \neq$ `-rain-`) = 0.80 for *(lluvioso,rainy)* + 2*0.22 for *(lluvioso,snowy)* = **1.24**.

The rest of the values can be calculated in a similar way.

Once the contingency tables have been generated, the association measures corresponding to each n-gram pair can be computed. In contrast with the implementation of McNamee and Mayfield (2004b), which used an ad-hoc measure, the current system uses three of the most extensively used standard association measures: the *Dice coefficient* (*Dice*), *pointwise mutual information* (*PMI*), and *log-likelihood* (*LogL*), which are defined by the following equations (Manning and Schütze, 1999):

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

Continuing with the previous example, notice that whatever the association measure to be used, we find that the output value obtained for the pair *(-lluv-, -rain-)* —the correct one— is much higher than that of the pair *(-lluv-, -snow-)* —the wrong one:

$$Dice(\texttt{-lluv-},\texttt{-rain-}) = \tfrac{2*1.67}{2.91+6.61} = \mathbf{0.35} \ > \ Dice(\texttt{-lluv-},\texttt{-snow-}) = \tfrac{2*0.22}{2.91+1.10} = \mathbf{0.11}\ ;$$

$$\text{PMI}(\texttt{-lluv-},\texttt{-rain-}) = log\tfrac{12.81*1.67}{2.91*6.61} = \mathbf{0.11} > \text{PMI}(\texttt{-lluv-},\texttt{-snow-}) = log\tfrac{12.81*0.22}{2.91*1.10} = \mathbf{-0.13}\ ;$$

$$LogL(\texttt{-lluv-},\texttt{-rain-}) =$$
$$2 * \left(1.67 * log\tfrac{12.81*1.67}{2.91*6.61} + 1.24 * log\tfrac{12.81*1.24}{2.91*6.20} + 4.94 * log\tfrac{12.81*4.94}{9.90*6.61} + 4.96 * log\tfrac{12.81*4.96}{9.90*6.20}\right) = \mathbf{0.05}$$
$$>$$
$$LogL(\texttt{-lluv-},\texttt{-snow-}) =$$
$$2 * \left(0.22 * log\tfrac{12.81*0.22}{2.91*1.10} + 2.69 * log\tfrac{12.81*2.69}{2.91*11.71} + 0.88 * log\tfrac{12.81*0.88}{9.90*1.10} + 9.02 * log\tfrac{12.81*9.02}{9.90*11.71}\right) = \mathbf{0.003}\ .$$

3.4. Word-level alignment filters

In addition to the two main phases of the alignment, word-level alignment and n-gram level alignment, an optional intermediate phase of filtering can be applied. The purpose of this extra phase is to reduce the noise introduced in the system by word-level translation ambiguities (e.g., if the same source language word has several candidate translations).

This way, two word-level filtering techniques will be tested. Firstly, we will try a simple threshold-based filtering by removing from the input the least probable word alignments, i.e., those with a word translation probability less than a given threshold we will note as W; in other words, a word-level pruning.

Secondly, we will try a more advanced bidirectional word-level alignment solution (Koehn et al., 2003), which considers a *(w_s, w_t) sourceLanguage-to-targetLanguage* word alignment only if there also exists a corresponding *(w_t, w_s) target-Language-to-sourceLanguage* word alignment.[7]

By applying these filters, subsequent processing will focus only on those words whose translation is less ambiguous, reducing both the noise introduced in the system and the number of input word pairs to be processed, thereby also increasing efficiency by reducing both computing and storage resources.

[7]It should be noted that according to the aligning algorithm employed by GIZA++ (Och and Ney, 2003), the obtaining of a word-level alignment (at some probability) from w_s to w_t when aligning the parallel corpora in the *sourceLanguage-to-targetLanguage* direction does not necessarily imply the existence of the corresponding w_t to w_s word-level alignment when processing the corpora in the reverse direction.

language	*#cognates*	*%cognates*	*difficulty*
Spanish	38/207	18.4%	2.25
German	121/207	58.5%	2.25
French	38/207	18.4%	2.50
Italian	38/207	18.4%	2.50
Dutch	130/207	62.8%	2.75
Finnish	3/207	1.5%	2.00
Swedish	109/207	52.7%	3.00

Table 1: Similarity measures of English with the rest of languages considered: percentages of cognates in the Swadesh lists (*left*); difficulty of learning a language for American English speakers where three means most similar/easiest to learn and one means least similar/most difficult to learn (*right*).

4. Experimental Set-up

We now describe the set-up used for the experiments made for this study and the decisions we have taken during their design.[8]

4.1. The evaluation framework

Following previous work in a multilingual context (Hollink et al., 2004; Savoy, 2003; McNamee and Mayfield, 2004a) and the restrictions due to our own availability of resources, we opted for testing our approach with a wide range of European languages for which parallel corpora are available in the EUROPARL (release v6) parallel corpus (Koehn, 2005)[9] and for which we also have available test collections from our past participation in several Cross-Language Evaluation Forum events (CLEF, 2014). The languages we have finally considered are the following, whose varied nature creates a representative test pool for our study: English (EN), German (DE), Dutch (NL) and Swedish (SW), all of them Germanic languages; Spanish (ES), French (FR) and Italian (IT), all of them Romance languages; and Finnish (FI), an Uralic Finnic language.

The inclusion of English as our common *target language* was convenient because, firstly, it is the dominant language on the Web[10] and, secondly, we can obtain directly comparable results since the same target collection is queried for the different query languages, which use the same (translated) query set. Moreover, many users, even if they understand English, prefer to use their mother tongue as *source language*.

[8] If more details were needed about the resources or configuration used by the system, we invite the reader to contact with the corresponding author.

[9] NOTE FOR THE REFEREES: As explained before, although the reference is much older than release v6, it is the one required to be used by Prof. Koehn, EUROPARL author.

[10] `http://www.internetworldstats.com/stats7.htm`.

```
<top>
<num> 154 </num>
<ES-title> Libertad de Expresión en Internet </ES-title>
<ES-desc> Encontrar documentos en los que se hable sobre la censura y la
libertad de expresión en Internet. </ES-desc>
<ES-narr> Los documentos en los que se discutan asuntos como la pornografía
o el racismo en Internet, sin mencionar el tema de la censura o libertad de
expresión, no se considerarían relevantes. </ES-narr>
</top>

<top>
<num> 154 </num>
<EN-title> Free Speech on the Internet </EN-title>
<EN-desc> Find documents which discuss censorship and freedom of speech on the
Internet. </EN-desc>
<EN-narr> Documents that discuss subjects such as pornography or racism on the
Internet without mentioning issues concerning censorship or freedom of speech
will not be considered relevant. </EN-narr>
</top>
```

Figure 1: Sample Spanish test topic and its English translation.

At this point, it may be useful for the interpretation and discussion of the results to calculate some kind of similarity measure between English, our common target language, and the different query languages to be used. Firstly, following the procedure described by Lehmann (1992), we estimated the percentages of cognates in the *Swadesh lists*[11] in order to calculate the degree of similarity between the different languages used. The resulting figures are shown in the left-hand side of Table 1. However, the Swadesh lists contain basic concepts, which are the words for which English most closely resembles the Germanic languages, so it is not an altogether fair test. So, as an alternative point of view, we also include in the right-hand side of Table 1 data published by Miller and Chiswick (2004), which are based on the difficulty Americans have in learning foreign languages.

With regard to the document collection employed in the evaluation process, as explained above, we have used an English collection, the English corpus of the so-called *robust task* celebrated within the CLEF 2006 *ad-hoc track*, which re-used test corpora (both collections and topics) from previous 2001, 2002 and 2003 CLEF editions (Nunzio et al., 2006). The English collection in question is formed by two subcollections: LA Times 94 (56,472 documents, 154 MB) and Glasgow Herald 95 (113,005 documents, 425 MB), totalling 169,477 documents with a size of 579 MB. Regarding the topics, we have used the 60 topics numbered C141 to C200 established for the robust task.[12] As shown in Fig-

[11] Available at `http://en.wiktionary.org/wiki/Appendix:Swadesh_lists`.

[12] Although the complete topic set for the robust task included topics C041 to C200, topics

languages (LX-EN)	*#sentences*	*#LX words*	*#EN words*
Spanish-English	1,786,594	51,551,485	49,411,045
German-English	1,739,154	45,607,269	47,978,832
French-English	1,825,077	54,568,499	50,551,047
Italian-English	1,737,081	49,065,283	49,981,015
Dutch-English	1,822,036	50,315,412	49,938,127
Finnish-English	1,742,553	34,123,013	47,601,416
Swedish-English	1,678,333	41,031,740	45,628,613

Table 2: Statistics for the parallel corpora used in this work: EUROPARL, release v6.

ure 1, topics are formed by three fields: a brief *title* statement, a one-sentence *description*, and a more complex *narrative* specifying the relevance assessment criteria. All topic sets, whatever the language, contain the same topics, which were translated manually by CLEF organization experts. Following CLEF standard policy, only *title* and *description* fields were used in the submitted queries.

For its implementation, the testing information retrieval system employed the open-source TERRIER platform (Ounis et al., 2007) as its core retrieval engine.

With respect to the subword level translation process introduced above, the n-gram based alignment system takes as input the release v6 of the EUROPARL parallel corpus. Table 2 shows the statistics for this parallel corpus.

For the first phase of the alignment process, as explained before, a word-level alignment, the GIZA++ (Och and Ney, 2003) statistical aligner was used. During the iterative training of the alignment models, we used a pipeline configuration commonly used in diverse MT experiments (Huet and Lefévre, 2011; Ma and Way, 2010; Gao et al., 2010a): five iterations of IBM Model 1, five iterations of HMM, five iterations of IBM Model 3 and three iterations of IBM Model 4. Regarding the optional filtering phase, and threshold-based filtering in particular (previously described in Section 3.4), after studying the distribution of the input aligned word pairs, a minimal word translation probability threshold value of W=0.15 was chosen.

4.2. Indexing-retrieval processes

The indexing process is simple: documents are lowercased and punctuation marks, but not diacritics, are removed. The resulting text is then split into character n-grams and indexed using an InL2 ranking model[13] (Amati and van Rijsbergen, 2002) with the term frequency normalisation parameter value `c`

C041 to C140 could not be used in our experiments because no relevant assessments are available for them in the case of the GLASGOW HERALD subcollection.

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

set to its default value: c=1. According to the results of previous related work (McNamee and Mayfield, 2004a,b; Hollink et al., 2004; Vilares et al., 2011), 4-grams (n-grams of four characters) showed promising, so we decided to use n=4 as n-gram size. No stopword removal was applied at this point. The same running parameters have been used for all the experiments performed.

In the case of retrieval, the source language topic is firstly conflated by lowercasing and removing punctuation marks, and then split into 4-grams in the same way as documents. Next, the resulting 4-grams are replaced by their corresponding candidate translations (i.e., their target language n-gram level alignments) according to a selection algorithm we detail below. The resulting translated topics are then submitted to the retrieval system using neither query expansion nor relevance feedback in order to study the performance of n-gram level processing on its own, without introducing distortions in the results by integrating other techniques. Two selection algorithms are currently available:

1. **Top-rank-based:** which takes the H highest ranked n-gram alignments per source n-gram, according to their association measure. The range of values we have tested is:

$$H \in \{1,\ 2,\ 3,\ 5,\ 10,\ 20,\ 30,\ 40,\ 50,\ 75,\ 100\}.$$

2. **Threshold-based:** which takes those n-gram alignments whose association measure is greater than or equal to a given minimal threshold T. The way such a threshold is calculated depends on the association measure to be used. In the case of the Dice coefficient, since it takes values within the range [0..1], the thresholds can be fixed in a simple way, the following values being used in this case:

$$T \in \{0;\ 0.1;\ 0.2;\ \ldots\ 0.7;\ 0.8;\ 0.85;\ 0.9;\ 0.95;\ 0.975;\ 1\}.$$

   However, pointwise mutual information and log-likelihood measures can take any value within the range $(-\infty..\infty)$. Thus, in order to homogenize the tests as much as possible, in the case of such association measures the thresholds will be calculated according to the *mean* and *standard deviation* of their distributions:

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

   where T_i represents the i-th threshold, with $i \in \mathbb{Z}$; Δ_i represents the step to be used, whose granularity may vary according to i and the association measure used (the values of log-likelihood are much more dispersed than for pointwise mutual information); μ represents the *mean* of the association measure values of the n-gram pairs obtained for the present configuration; and σ represents their *standard deviation*.

Finally, the n-gram level translated query is submitted to the retrieval system.

4.3. Lower and upper baselines

Two baselines have been established for comparing and analyzing the results from different points of view, both using character n-grams as the processing unit:

- **EN 4-grams**: a target language (English) monolingual run using 4-grams as terms. For this purpose the original English topics were used. This is the *upper baseline*, the best result we could ever obtain by using n-gram based translation.

- **LX 4-grams** (where LX stands for the source language): the target (English) document collection is queried using the original n-grammed source language query (i.e. no translation is made). This kind of cognate matching allows us to measure the impact of casual matches and constitutes the *lower baseline*.

Apart from these baselines, which will be used for all languages, in the case of our Spanish-to-English (ES-to-EN) tuning experiments, which we will introduce in the next section, we have considered the convenience of using a larger set of baselines for comparative purposes, thus including the following extra runs:

- **EN stm**: another target language (English) monolingual run using the original English topics provided by CLEF, this time using a classical stemming-based approach. The Snowball stemmer,[14] based on Porter's algorithm (Porter, 1980), was used, while the stopword list was the one provided by the University of Neuchâtel.[15] Both resources are in common use among the IR research community.

- **Google stm**: a more classical cross-language run that uses Google Translate service[16] for translating the source language query into the target language (English) before conflating it using stemming and stopwords as before.

- **Google 4-grams**: our final baseline, it uses, as before, Google Translate for translating the source language query into English. However, once translated, instead of using a classical stemming-based approach, we use 4-grams as index terms. In other words, *Google 4-grams* is to *Google stm* as *EN 4-grams* is to *EN stm*.

5. System Tuning Using Spanish-to-English (ES-to-EN) CLIR

In order to get our system to work in a proper way, we need to tune a large number of different parameters. Moreover, we intend to demonstrate the generality of our n-gram based approach. Thus, we will use a Spanish-to-English (ES-to-EN) set-up to find the most promising configurations in terms of performance and efficiency. Such parameters will be used later for the remaining languages. This way, this initial set of ES-to-EN runs should be seen as the

[14] `http://snowball.tartarus.org`
[15] `http://www.unine.ch/info/clef/`
[16] `http://translate.google.es`

	unidirectional word alignment				*bidirectional word alignment*			
	W=0.00		***W=0.15***		***W=0.00***		***W=0.15***	
prob.	***#pairs***	***%pairs***	***#pairs***	***%pairs***	***#pairs***	***%pairs***	***#pairs***	***%pairs***
[0 .. 0.001)	1,127,088	38.11	0	0.00	331,916	30.74	0	0.00
[0.001 .. 0.01)	635,947	21.51	0	0.00	293,586	27.19	0	0.00
[0.01 .. 0.05)	575,542	19.46	0	0.00	219,016	20.28	0	0.00
[0.05 .. 0.10)	225,619	7.63	0	0.00	76,792	7.11	0	0.00
[0.10 .. 0.20)	246,764	8.34	82,540	36.08	70,311	6.51	26,921	23.37
[0.20 .. 0.30)	59,101	2.00	59,101	25.83	25,817	2.39	25,817	22.41
[0.30 .. 0.40)	32,744	1.11	32,744	14.31	17,642	1.63	17,642	15.31
[0.40 .. 0.50)	19,137	0.65	19,137	8.37	12,522	1.16	12,522	10.87
[0.50 .. 0.60)	6,861	0.23	6,861	3.00	6,465	0.60	6,465	5.61
[0.60 .. 0.70)	5,257	0.18	5,257	2.30	5,082	0.47	5,082	4.41
[0.70 .. 0.80)	4,277	0.14	4,277	1.87	4,195	0.39	4,195	3.64
[0.80 .. 0.90)	3,657	0.12	3,657	1.60	3,530	0.33	3,530	3.06
[0.90 .. 0.95)	1,567	0.05	1,567	0.68	1,460	0.14	1,460	1.27
[0.95 .. 0.975)	919	0.03	919	0.40	802	0.07	802	0.70
[0.975 .. 0.99)	12,708	0.43	12,708	5.55	10,764	1.00	10,764	9.34
[0.99 .. 1]	0	0.00	0	0.00	0	0.00	0	0.00
TOTAL:	2,957,188	100.00	228,768	100.00	1,079,900	100.00	115,200	100.00
avg. prob.:	0.04		0.34		0.06		0.42	

Table 3: Distribution of input aligned ES-to-EN word pairs across their word-to-word translation probabilities.

tuning phase of our system, while the runs for the remaining languages (see Section 7) should be seen as its *testing phase*.

At this point we note that because of problems of space and readability, it will not always be possible to show the results obtained for all the configurations tested, particularly in the case of threshold-base selection. So, we will restrict ourselves, when necessary, to those values which are most relevant to the analysis.

5.1. ES-to-EN alignment statistics

As explained above, our first test set corresponds to Spanish-to-English (ES-to-EN) cross-language runs. We will start our study by showing some statistics which do not depend on the particular association measure to be used.

Firstly, we will focus on the input word-level alignment, obtained by aligning the ES-EN Europarl parallel corpus (see Section 4.1) using Giza++. Table 3 shows the distribution of the input aligned ES-to-EN word pairs across their word-to-word translation probabilities, which exhibits a clear bimodal behavior with peaks at both ends, with the highest peak corresponding to low-probability translations. As previously described in Section 3.4, we have considered the use of both regular unidirectional alignment and bidirectional word-level alignment, and the application or not of a threshold-based filtering, with a W=0.00 threshold value meaning that no filtering has been done, and a W=0.15 value meaning that those word-level alignments whose word translation probability is less than 0.15 have been removed. Next, Table 4 shows, for those same aligned word pairs,

	unidirectional word alignment				*bidirectional word alignment*			
	W=0.00		***W=0.15***		***W=0.00***		***W=0.15***	
#transl.	*#words*	*%words*	*#words*	*%words*	*#words*	*%words*	*#words*	*%words*
[1 .. 1]	9,504	6.92	57,115	50.74	56,188	44.32	79,013	82.36
[2 .. 2]	9,277	6.75	25,855	22.97	17,365	13.70	14,850	15.48
[3 .. 4]	17,747	12.92	19,755	17.55	16,162	12.75	2,046	2.13
[5 .. 9]	40,637	29.59	9,837	8.74	15,087	11.90	30	0.03
[10 .. 19]	24,590	17.90	0	0.00	10,040	7.92	0	0.00
[20 .. 29]	10,642	7.75	0	0.00	4,095	3.23	0	0.00
[30 .. 39]	6,068	4.42	0	0.00	2,215	1.75	0	0.00
[40 .. 49]	4,173	3.04	0	0.00	1,362	1.07	0	0.00
[50 .. 74]	6,180	4.50	0	0.00	1,753	1.38	0	0.00
[75 .. 99]	3,095	2.25	0	0.00	935	0.74	0	0.00
[100 .. ∞)	5,443	3.96	0	0.00	1,584	1.25	0	0.00
TOTAL:	137,356	100.00	112,562	100.00	126,786	100.00	95,939	100.00
avg. #transl.:	21.53		2.03		8.52		1.20	

Table 4: Distribution of source words in the input aligned ES-to-EN word pairs across their number of possible translations.

the distribution of the source (Spanish) words across their number of possible (English) translations.

Finally, we will pay attention to the corresponding output n-gram level alignment obtained by the algorithm used in our implementation. Table 5 shows the distribution of source n-grams across their number of possible n-gram level alignments, i.e. their number of n-gram level translations.

Next, we will present the performance results obtained for the different configurations tested.

5.2. ES-to-EN results using the Dice coefficient

The first round of our ES-to-EN experiments was performed using the Dice coefficient (Eq. 1). Table 6 presents the performance results, measured in terms of *mean average precision* (MAP), obtained when applying the subword-level translation approach with Dice. The left-hand (sub)table corresponds to the results obtained using the top-rank-based selection algorithm (for the H values previously described in Section 4.2), while the right-hand (sub)table employed the threshold-based selection algorithm (for the threshold values T introduced in Section 4.2). For each (sub)table, the right-hand two-column group shows those results obtained when using a classical unidirectional ES-to-EN word-level alignment, while the left-hand two-column group shows the results obtained when applying one of the proposed refinements, the use of a bidirectional ES-to-EN word-level alignment (introduced in Section 3.4). Finally, for each of these two-column groups, the first column stands for the results obtained when no minimal word alignment probability W is required (i.e., W=0.00), while for the second column a word translation probability threshold W=0.15, the other of the proposed refinements (described in Section 3.4), has been applied.

	unidirectional word alignment				*bidirectional word alignment*			
	W=0.00		***W=0.15***		***W=0.00***		***W=0.15***	
#transl.	*#4-gr*	*%4-gr*	*#4-gr*	*%4-gr*	*#4-gr*	*%4-gr*	*#4-gr*	*%4-gr*
[1 .. 1]	811	1.47	3,006	5.73	2,856	5.46	3,530	7.19
[2 .. 2]	778	1.41	2,152	4.10	2,511	4.80	2,786	5.68
[3 .. 4]	2,110	3.84	4,889	9.32	6,494	12.42	7,176	14.62
[5 .. 9]	5,956	10.83	10,832	20.65	12,519	23.94	13,464	27.44
[10 .. 19]	8,741	15.89	10,881	20.74	7,944	15.19	8,716	17.76
[20 .. 29]	5,484	9.97	5,088	9.70	3,213	6.14	3,443	7.02
[30 .. 39]	3,494	6.35	2,794	5.33	1,822	3.48	1,894	3.86
[40 .. 49]	2,451	4.45	1,842	3.51	1,242	2.37	1,383	2.82
[50 .. 74]	3,773	6.86	2,828	5.39	1,977	3.78	2,142	4.36
[75 .. 99]	2,438	4.43	1,658	3.16	1,201	2.30	1,177	2.40
[100 .. ∞)	18,983	34.50	6491	12.37	10,524	20.12	3,363	6.85
TOTAL:	55,019	100.00	52,461	100.00	52,303	100.00	49,074	100.00
avg. #transl.:	360.12		56.29		172.37		32.83	

Table 5: Distribution of source 4-grams in the output aligned ES-to-EN 4-gram pairs across their number of possible translations.

top-rank-based				
	unidirectional		*bidirectional*	
H	***W=0.00***	***W=0.15***	***W=0.00***	***W=0.15***
1	**0.2561**	0.**2515**	**0.2475**	0.2432
2	0.2337	0.2377	0.2450	**0.2455**
5	0.2084	0.2002	0.2001	0.2065
10	0.1524	0.1554	0.1536	0.1593
20	0.1280	0.1238	0.1304	0.1285
30	0.0874	0.1037	0.0959	0.1110
40	0.0582	0.0782	0.0637	0.0686

threshold based				
	unidirectional		*bidirectional*	
T	***W=0.00***	***W=0.15***	***W=0.00***	***W=0.15***
0.00	0.0023	0.0015	0.0013	0.0026
0.10	**0.1660**	0.1635	**0.1769**	0.1582
0.20	0.1525	**0.1789**	0.1737	0.1628
0.30	0.1620	0.1618	0.1667	**0.1930**
0.40	0.1453	0.1633	0.1639	0.1616
0.50	0.1422	0.1377	0.1463	0.1542
0.60	0.1362	0.1332	0.1274	0.1327

Table 6: MAP results obtained for the ES-to-EN CLIR runs using the Dice coefficient with the top-rank-based selection algorithm (left table), and for the threshold-based selection algorithm (right table).

top-rank-based				
	unidirectional		*bidirectional*	
H	***W=0.00***	***W=0.15***	***W=0.00***	***W=0.15***
1	0.0842	0.1106	0.0824	0.1136
2	0.0961	0.1319	0.1066	0.1491
5	0.1386	0.1645	0.1523	0.1689
10	0.1265	0.1583	0.1571	**0.2021**
20	**0.1735**	**0.1758**	**0.1804**	0.1876
30	0.1640	0.1646	0.1646	0.1682
40	0.1389	0.1389	0.1449	0.1446

threshold based				
	unidirectional		*bidirectional*	
T	***W=0.00***	***W=0.15***	***W=0.00***	***W=0.15***
μ	0.0011	0.0440	0.0012	0.1423
μ**+0.5**σ	0.0019	0.1743	0.0046	0.1979
$\mu+\sigma$	0.0470	**0.1966**	0.0783	0.1923
μ**+1.5**σ	**0.2048**	0.1797	**0.2043**	**0.1997**
μ**+2**σ	0.1479	0.1763	0.1685	0.1428
μ**+2.5**σ	0.1443	0.1553	0.1447	0.1321
μ**+3**σ	0.1414	0.1307	0.1330	0.1330
μ**+3.5**σ	0.1330	0.1330	0.1330	–

Table 7: MAP results obtained for the ES-to-EN CLIR runs using pointwise mutual information with the top-rank-based selection algorithm (left table), and for the threshold-based selection algorithm (right table).

This way all possible configurations are covered. The best results for each <selection algorithm/word-level alignment/word-level probability threshold> configuration are shown in boldface.

During their analysis, statistical significance tests have been used for comparing, in terms of MAP, the performance of each of these possible running configurations; in particular, two-tailed T-tests over MAP values with α=0.05 have been applied throughout this work. At this point, those tests showed that:

(a) Results obtained using the top-rank-based selection algorithm are significantly better than those for threshold-based selection.

(b) The results obtained using unidirectional or bidirectional word-level alignments showed no significant difference.

(c) There is no significant difference between the optimal result (obtained using unidirectional word-level alignment, no word-level probability threshold and the top-rank-based selection algorithm) and the results for the remaining top-rank-based sub-optimal runs shown in boldface in the table (i.e. the best results obtained with the other configurations using top-rank-based selection).

5.3. ES-to-EN results using pointwise mutual information

Our second round of ES-to-EN runs tested the behavior of the system when using pointwise mutual information (Eq. 2) as the association measure. The detailed results can be seen all together in Table 7, with the same distribution as before. Again, the best results obtained are shown in boldface. The corresponding statistical significance tests (again two-tailed T-tests over MAP values with α=0.05) have shown that:

(a) This time we have not found significant differences between the results obtained with the top-rank-based and threshold-based selection algorithms.

top-rank-based				
	unidirectional		*bidirectional*	
H	***W=0.00***	***W=0.15***	***W=0.00***	***W=0.15***
1	**0.2785**	**0.2771**	**0.2703**	**0.2732**
2	0.2557	0.2590	0.2509	0.2590
5	0.1997	0.2068	0.1961	0.2023
10	0.1589	0.1636	0.1464	0.1640
20	0.1177	0.1202	0.1144	0.1227
30	0.0842	0.0948	0.0910	0.0957
40	0.0612	0.0813	0.0639	0.0685

threshold based				
	unidirectional		*bidirectional*	
T	***W=0.00***	***W=0.15***	***W=0.00***	***W=0.15***
μ	0.0045	0.0305	0.0061	0.0444
μ**+10**σ	0.0444	0.1156	0.0633	0.1288
μ**+20**σ	0.0798	0.1392	0.1189	**0.1495**
μ**+30**σ	0.1229	**0.1487**	0.1396	0.1472
μ**+40**σ	0.1354	0.1463	0.1443	0.1415
μ**+50**σ	0.1362	0.1446	0.1491	0.1398
μ**+60**σ	0.1410	0.1424	**0.1494**	0.1393
μ**+70**σ	**0.1478**	0.1396	0.1451	0.1371
μ**+120**σ	0.1446	0.1360	0.1394	0.1330
μ**+150**σ	0.1401	0.1330	0.1391	–

Table 8: MAP results obtained for the ES-to-EN CLIR runs using log-likelihood with the top-rank-based selection algorithm (left table), and for the threshold-based selection algorithm (right table).

(b) As before, the results obtained using unidirectional or bidirectional word-level alignments do not significantly differ between them.

(c) In general, there is no significant difference between the optimal runs obtained either for the top-rank-based and the threshold-based selection algorithms, and the remaining sub-optimal runs.

5.4. ES-to-EN results using log-likelihood

The last round for this first ES-to-EN test series uses log-likelihood (Eq. 3). The results obtained are presented in Table 8 with the usual distribution and the best results for each configuration shown in boldface.[17] For the log-likelihood experiments, significance tests showed similar behavior to those with the Dice Coefficient:

(a) Results obtained using the top-rank-based selection algorithm are significantly better than those for threshold-based selection.

(b) The results obtained using unidirectional or bidirectional word-level alignments do not significantly differ between them.

(c) There is no significant difference between the optimal result (obtained using unidirectional word-level alignment, no word-level probability threshold and the top-rank-based selection algorithm) and those for the remaining top-rank-based sub-optimal runs.

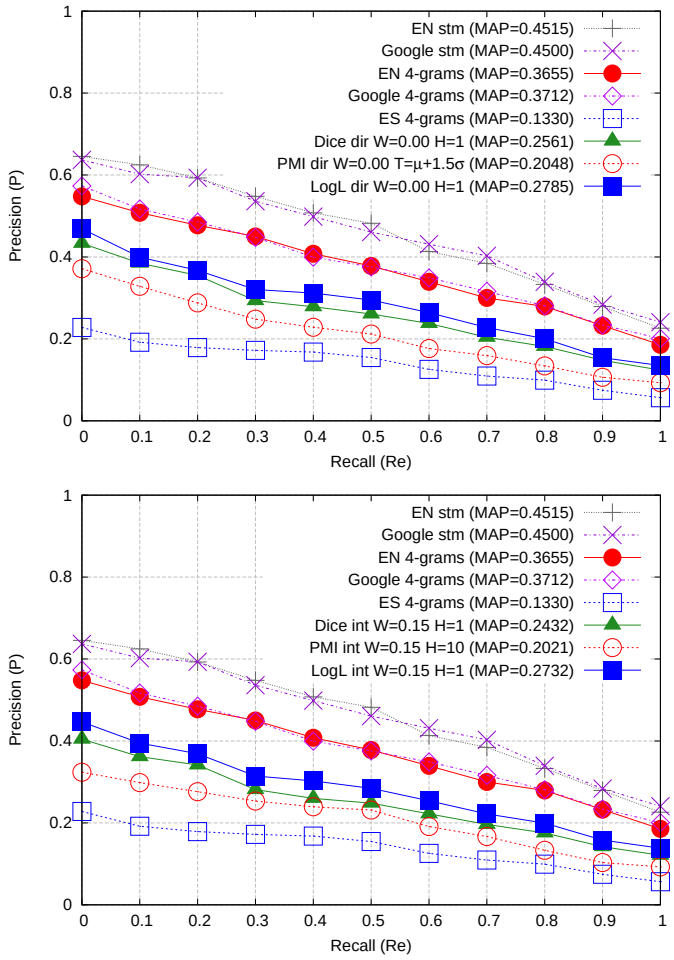

Figure 2: ES-to-EN Summary MAP results and Precision vs. Recall graphs: optimal runs (top) and most efficient sub-optimal runs (bottom).

5.5. ES-to-EN summary results

Finally, the left-hand graph of Figure 2 presents the results for the best configurations found compared with the baselines proposed in Section 4.3.[18] In this case, Precision vs. Recall graphs are also shown in addition to MAP values

[17]Notice that in the case of the threshold-based selection algorithm, since the standard deviation of the log-likelihood distribution values has been found to be much greater than for pointwise mutual information, the steps we have used are longer —see Δ_i parameter in Eq. 4.

[18]At this point we make notice that the right-hand graph of the figure shows sub-optimal configurations with no statistically significant difference with respect to the previous ones but

in order to make their analysis easier.

Regarding these figures, they show that:

(a) The performance of n-gram based approaches is satisfactory although it is still below that one of more complex classical word-based techniques which make use of their language knowledge.

(b) The performance of phrase-based MT runs (*Google stm* and *Google 4-grams*) is similar to that one of target language monolingual runs (*EN stm* and *EN 4-grams*, respectively).

(c) Our upper baseline, the n-gram based monolingual run (*EN 4-grams*), performs significantly better than n-gram based CLIR runs.

(d) The log-likelihood run performs similarly to the Dice Coefficient run — slightly outperforming it—, but improves on pointwise mutual information results significantly.

(e) Both log-likelihood and Dice Coefficient runs outperform significantly the lower baseline, *ES 4-grams*, which accounts for casual matching.

(f) Mutual information shows no significant difference to the lower baseline.

6. Discussion of Results for ES-to-EN CLIR

Now the results obtained for our ES-to-EN experiments have been presented, it is time to analyze them carefully.

6.1. Upper baselines

As stated before, there is still margin for improvement when comparing monolingual n-gram based IR (*EN 4-grams*) with classical monolingual word-based IR (*EN stm*). The results obtained are positive, but they can be improved. However, this is a question beyond the scope of this paper since we are focusing on the translation process. Moreover, the small difference attained when using phrase-based MT for query translation (*Google stm* and *Google 4-grams*) with respect to monolingual results (*EN stm* and *EN 4-grams*, respectively), shows us that this should be our role model for the future.

6.2. Using Dice coefficient

Following the same order as when presenting the experiments, we will study first those results obtained using the Dice coefficient in our Spanish-to-English experiments of Section 5.2.

We will consider the case of applying no refinements during word-level alignment, that is when using a unidirectional word-level alignment with no word-level alignment filtering (i.e. W=0.00) as our *basic configuration*. In this case,

with a more efficient performance. They will be discussed later in Section 7.

the best results for the top-rank-based selection algorithm are obtained for H=1, that is when minimizing the number of candidate translations. On the other hand, when employing threshold-based selection, the performance values obtained for the different thresholds show much less variation than for the top-rank-based algorithm, although they are significantly outperformed by it. This is because of the noise introduced by the extra n-grams added by the method based on thresholds.

Next, trying to reduce the noise introduced in the system by word-level translation ambiguities, we removed those least-probable word alignments from the input by applying the first of the proposed refinements: threshold-based word-level alignment filtering. After studying the distribution of the output word-level alignments obtained with GIZA++ —see Table 3—, we decided to dismiss those pairs with a word translation probability less than a threshold W=0.15. In this way we drastically reduced by more than 90% both the number of input word pairs processed —see Table 3— and the mean number of possible translations per source word —see Table 4.

Such a reduction in the number of input word pairs had, consequently, an inmediate effect on the output n-gram level alignments, reducing the mean number of possible translations per source n-gram by nearly 85% — see Table 5.

This reduction, both at word and n-gram level, resulted in a considerable reduction of both processing and storage resources.

As previously stated in Section 5.2, the results obtained by introducing this refinement, are, in general, not significantly different in terms of performance from those obtained for the basic configuration, whatever the selection algorithm used. So, it can be concluded that although word-level pruning does not really improve the results, it does considerably reduce those computing and storage resources required by the system, justifying its application. On the other hand, these results prove that this n-gram based solution has a robust behavior against the noise introduced by the very high percentage of low-probability word-level alignments of the input in the case of the basic configuration.

Next, we tested the second of the proposed refinements, the use of bidirectional word-level alignment. As explained in Section 3.4, its aim was to improve the accuracy of the n-gram alignment process by focusing the processing on those words whose translation is less ambiguous. We will take again our *basic configuration* as a reference.

At word level, when examining Tables 3 and 4 we can see that bidirectional word alignment attains a reduction of approximately 60% in both the number of input word pairs and the mean number of possible translations per input word.

Consequently, at the n-gram level, according to Table 5, the mean number of possible translations per source n-gram was reduced by more than 50% after applying this new refinement.

As before, this reduction at both input and output level allows us to reduce the computing and storage resources.

With respect to the results themselves, as stated in Section 5.2, they are not significantly different from those obtained with the original unidirectional word-level alignment. So, we can conclude that the use of bilingual alignment

neither improves nor degrades the performance of the system, but does allow us to reduce both computing and storage resources. Moreover, the system has once again demonstrated its robustness to inaccurate or ambiguous input word-level alignments.

Finally, because of their good behavior separately, we also studied the possibility of combining both refinements, word-level bilingual alignment and word-level pruning, looking for an additional reduction of both the level of ambiguity and the computing and storage resources consumed. We take, as usual, our initial *basic configuration* as the baseline.

At word level, Tables 3 and 4 show that, when combining both refinements, we obtain a increased reduction of approximately 95% in both the number of input word alignments and in the mean number of possible translations per input source word.

As a result, at the n-gram level, Table 5 show a reduction of more then 90% in the mean number of possible output n-gram translations per source n-gram.

The results obtained, as previously stated in Section 5.2, are still not significantly different from the initial ones, with the top-rank-based selection algorithm performing significantly better —although this time the best performance was obtained for H=2, the difference with the second best run, the one for H=1, is negligible. On the other hand, such results show no apparent deterioration in performance, allowing us to conclude that the combined use of both refinements minimizes the resources required by the system without harming its performance.

6.3. Using pointwise mutual information

Our second round of experiments, presented in Section 5.3, makes use of pointwise mutual information.

As before, our first test runs used the so-called *basic configuration*: single unidirectional word-level alignment with no word-level pruning (i.e., W=0.00). When examining the results obtained using the top-rank-based selection algorithm we found that, unlike before, results improved when progressively increasing the number of n-grams accepted up to a maximum at H=20. Nevertheless, these results are significantly worse than those obtained using the Dice coefficient. This is because PMI tends to overestimate low-frequency data, meaning that inaccurate but frequent n-gram alignments are assigned very high PMI values, even higher than more accurate alignments, thus introducing too much noise in the translated query and, therefore, visibly decreasing performance. Regarding threshold-based selection results, they tend to behave in a more homogeneous way between thresholds, with no significant differences with respect to top-ranked selection, thus also performing significantly worse than when using Dice.

When introducing the first refinement, word-level pruning according to a translation probability threshold W=0.15 —previously described in Section 3.4— the gains were exactly the same as in the case of the Dice coefficient, except for the mean n-gram association measure. This is because the gains at word-level,

both with respect to the reduction of input word pairs and the increase of the mean translation probability, depend only on the value of W, and are not affected by the association measure chosen. At the n-gram level, the reduction in the number of output n-gram pairs only depends on the input word pairs —and, consequently, also on the value of W. Nevertheless, the mean association measures vary, since we are now working with pointwise mutual information instead of the Dice coefficient.

As shown in Section 5.3, the behavior of the system and the results obtained for both selection algorithms do not significantly differ from those obtained for the basic configuration. As in the case of the Dice coefficient, the introduction of the word-level threshold W does not degrade the performance of the system, although does reduce the computing and storage resources required. On the other hand, the system continues to show its robustness against the distortion introduced by low-probability inputs.

Next, we tried the second proposed refinement: word-level bidirectional alignment. As shown in Section 5.3, the results obtained showed no significant differences from those for the regular unidirectional alignment, whether we apply word-level pruning or not —i.e. whether W=0.00 or W=0.15. As before, the gains obtained when using bidirectional word alignment, either in combination or not with the use of word-level pruning, were exactly the same as those with the Dice coefficient.

From this behavior we conclude that, as in the case of using the Dice coefficient, the introduction of a bidirectional word alignment not only has no effect on the performance of the system, but has the benefit of reducing the resources needed. On the other hand, the system again showed its robustness against inaccurate or ambiguous input word alignments.

6.4. Using log-likelihood

Our last round of ES-to-EN experiments tested the behavior of the system when employing log-likelihood for the different possible configurations, as described in Section 5.4.

As usual, our first test runs corresponded to our *basic configuration*. In the case of using the top-ranked selection algorithm, the behavior of the system is similar to that for the Dice coefficient, with the best results being obtained when limiting the number of candidate n-grams accepted, with H=1 as the best configuration, even outperforming Dice. However, in the case of the threshold-based selection algorithm, the results obtained were very poor, being the lowest performance obtained so far.

For the first refinement, word-level pruning, the gains were exactly the same as in the case of the Dice coefficient and PMI.

The behavior of the system and the results obtained, as stated in Section 5.4, did not significantly differ from those obtained with the basic configuration. As in the case of the rest of association measures, the introduction of the word-level threshold did not degrade performance, but did reduce both the computing and storage resources required. On the other hand, the system again demonstrated its robustness against the distortion introduced by low-probability inputs.

Our last bunch of test runs corresponded to those results obtained applying word-level bidirectional alignment. As shown in Section 5.4, the results obtained were not significantly different from those for the regular unidirectional alignment, both in the case of applying threshold-based pruning or not —i.e. for W=0.15 and W=0.00, respectively.

From this behavior we conclude that, as in the case of the other association measures, the introduction of a bidirectional word-level alignment not only has no effect on the performance of the system, but has the benefit of reducing the resources needed. On the other hand, the system continued to show its robustness against inaccurate or ambiguous input word-level alignments.

7. Testing Experiments with other Language Pairs

After a first *tuning* phase in a ES-to-EN context in order to find a proper running configuration for our system (see Sections 5 and 6), it is time to move to a second *testing* phase, properly speaking, in order to prove the generality and validity of our approach by trying with the remaining languages.[19]

For the purpose of selecting a common running configuration, we can benefit from the fact that, as stated during the previous discussion:

- The application of the proposed refinements —word-level pruning and bilingual word alignment— does not harm performance and, on the contrary, we gain in efficiency by reducing computing and storage resources.
- The top-rank-based selection algorithm outperforms the threshold-based one both in terms of performance and efficiency —except in the case of pointwise mutual information, where performance is similar.
- Log-likelihood and Dice perform similarly, being significantly superior to pointwise mutual information, whose performance was shown to be poor.

Thus, we finally decided to dismiss pointwise mutual information because of its poor performance and efficiency compared to Dice and log-likelihood because of the much higher number of candidate translation n-grams required. However, in the case of both the Dice coefficient and the log-likelihood, for the remainder of our experiments we will adopt the following running configuration as a compromise between performance and efficiency:

[19] NOTE FOR THE REFEREES: It must be noted that, in order to further check the generality of our approach, detailed experiments similar to those presented above corresponding to the tuning phase (i.e. ES-to-EN) were also performed for German-to-English (DE-to-EN), obtaining a similar behavior with similar results to those of the ES-to-EN runs, but because of the limitations of space in the article, they could not be shown here. However, in the case of the results we are going to present now, no special tuning was made for any individual language; all the results we are presenting now have been obtained using the same system configuration, which we will introduce ahead: Top-rank-based selection with H=1, bidirectional word-level alignment and word-level threshold pruning with W=0.15.

- Top-rank-based selection with H=1, bidirectional word-level alignment and word-level threshold pruning with W=0.15.

For the sake of completeness, the results obtained for ES-to-EN when using these sub-optimal but more efficient configurations are shown in the right-hand graphs of Figure 2.

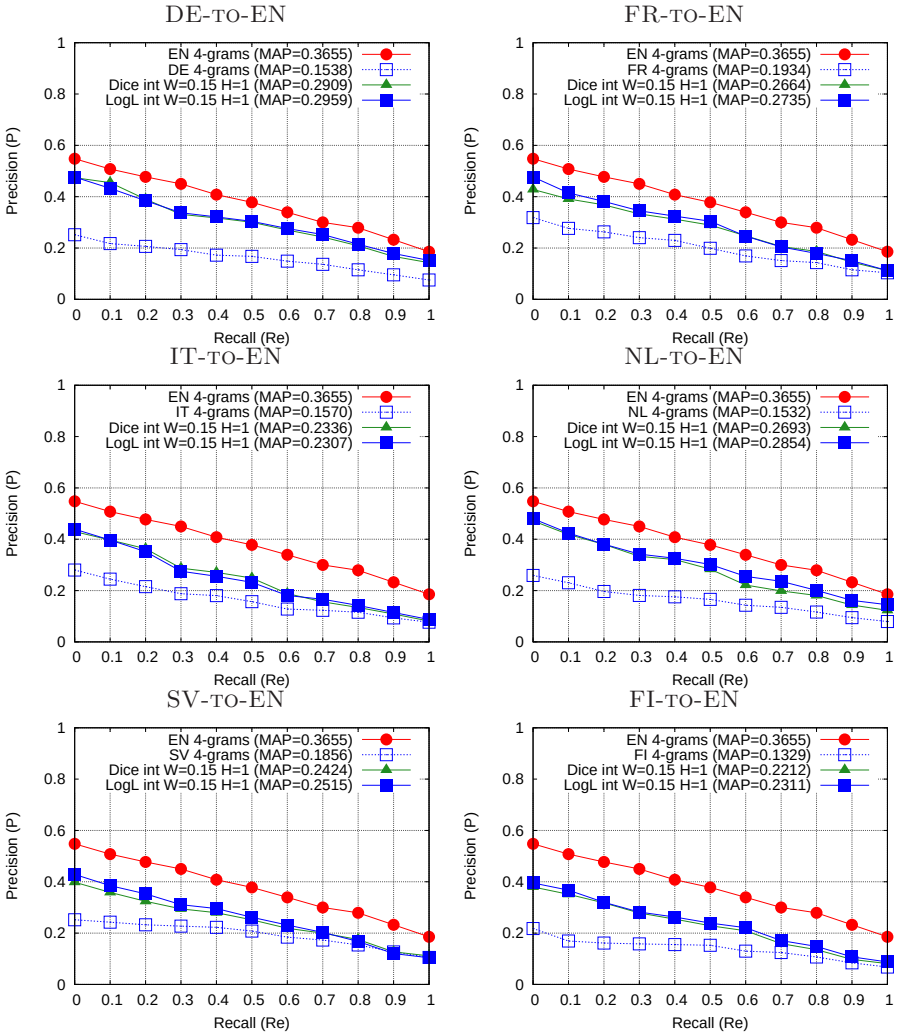

Figure 3: Summary MAP results and Precision vs. Recall graphs for the rest of source languages using the selected common configuration: German (DE-to-EN), French (FR-to-EN), Italian (IT-to-EN), Dutch (NL-to-EN), Swedish (SV-to-EN) and Finnish (FI-to-EN).

The results obtained with this configuration for German (DE), French (FR), Italian (IT), Dutch (NL), Swedish (SV) and Finnish (FI) source languages,

compared with their corresponding baselines, are presented in Figure 3. Note that, aiming to improve readibility, only the monolingual n-gram based baselines have been used, thus focusing on n-gram based retrieval performance. English stemming based results (*EN stm*), which would be common to all figures, are available in Figure 2. With regard to phrase-based MT baselines (i.e. *Google stm* and *Google 4-grams*), experiments were made for all the languages involved, although they have not been displayed here in order not to overload the figures. The results obtained for the different languages showed qualitatively similar results to those previously obtained for ES-to-EN.

Going back to Figure 3, as it can be seen, such results are very similar to those previously obtained for Spanish. Moreover, according to the significance tests performed —remember that two-tailed T-tests over MAP values with α=0.05 have been applied throughout this work—, we can state, in general, that:

(a) Our upper baseline, the n-gram based monolingual run (*EN 4-grams*), performs significantly better than the corresponding n-gram based CLIR runs.

(b) The log-likelihood runs perform similarly, with no significant differences, to the corresponding Dice Coefficient runs, usually outperforming them — except for Italian-to-English, where Dice slightly outperforms log-likelihood.

(c) Both the log-likelihood and Dice Coefficient runs significantly outperform their corresponding lower baselines, *LX 4-grams* (where LX stands for the source language), which accounts for casual matching.

7.1. Discussion of the Results with other Language Pairs

As can be seen when analyzing the results presented in the previous section for the different languages tested, they are no different, from a qualitative point of view, from those previously obtained for the Spanish-to-English CLIR runs. From a quantitative point of view, and focusing on MAP, the results obtained are also quite close to those obtained before, since they vary, in general, within the range of the values previously obtained for Spanish. The lowest MAP was obtained for the non-Romance and non-Germanic language, Finnish, but even in that case we are talking about 0.22–0.23 MAP values, which are close to the expected values according to our experiments on Spanish.

So, this n-gram based approach has been able to perform effective retrieval, thus proving that the validity of this technique is independent of the languages involved.

8. General Discussion

As we have stated, n-gram based translation avoids some of the limitations of classic dictionary-based translation methods, such as the need for word normalization or the inability to handle misspellings and out-of-vocabulary words.

In the case of normalization, the overlapping of n-grams corresponding to a given word provides in itself a surrogate means to normalize word forms during indexing and translation. This is because those parts shared by a word and its morphological variants, their roots and possibly other morphemes, will be translated into the same target n-gram and then matched, making retrieval possible. In a similar way, n-gram based translation approaches allow the translation and matching of both misspellings and out-of-vocabulary words, since those parts of the unknown word which are shared with other known words —either because they are shared roots or morphemes or because they have not been affected by the misspelling— can still be translated and matched at the n-gram level.

Moreover, since this is a knowledge-light approach which does not rely on language-specific processing, it can be used for a wide range of languages of very different natures, even in the face of the lack of linguistic information and language resources available; in contrast, other more classical CLIR approaches need language-specific resources for their application, such as stemmers, stopword lists, lexicons, tagged corpora and bilingual dictionaries, which are not always available, even for main European languages (Rehm and Uszkoreit, 2011).

The results obtained throughout our experiments, which are consistent across languages, have demonstrated the validity and applicability of this kind of n-gram based translation approaches for CLIR, although there is still a margin for improvement with respect to more complex classical word-based techniques. Regarding our particular implementation, these results indicate that both the log-likelihood and the Dice coefficient significantly outperform pointwise mutual information, the former performing slightly better. Our tests also showed the top-rank-based selection algorithm to be, in general, significantly better not only from a performance but also from an efficiency point of view, since the number of translation n-grams to be processed is fewer.

As a final summary, and according to the results we have obtained, it is time to answer the questions we had formulated at the beginning of this study (see Section 1.2):

Q: *Is n-gram based translation a valid approach for CLIR and other cross-language text processing applications?*

A: Yes, this approach has proven its validity throughout the tests performed, attaining an intermediate performance closer to the upper baseline than to the lower baseline.

Q: *Is the behavior of this approach consistent across different languages?*

A: Yes, the results obtained for the different languages used in our experiments have shown a consistent behavior across them.

Q: *Which is the most efficient way of applying it?*

A: We have found large differences depending on the running configuration used, but when taking as criteria a compromise between effectiveness and

efficiency, the most promising ones are the following: at word-level, the application of a non-standard bilingual alignment and the pruning of input word alignments according to a word translation probability threshold (W=0.15); and at character n-gram level, the use of the Dice coefficient or log-likelihood as an association measure and the employment of the top-rank-based selection algorithm restricting H to a maximum.

As a result, we can confirm the validity of character n-gram based approaches for CLIR tasks, both for indexing-retrieval and translation purposes, these not being tied to certain implementations such as that of McNamee and Mayfield (2004b) or the present one. For all these reasons we believe that this study constitutes an interesting contribution in the state of the art of this particular field of n-gram based processing.

9. Conclusions and Future Work

This article presents a study of the feasibility of Cross-Language Information Retrieval (CLIR) systems which use character n-grams not only as indexing units, but also as translation units, looking to extend the main advantages of n-grams (simplicity, independency and robustness) not only in the indexing-retrieval process, but also in the query translation process.

For this purpose, we have made use of a implementation of our own which integrates a novel algorithm for parallel text alignment at the subword (i.e. character n-gram) level. This algorithm consists of two phases. In the first phase, the most time-consuming, the input parallel corpus is aligned at the word level using a statistical aligner. In the second phase, association measures existing between the character n-grams compounding each aligned word pair are computed taking as input the translation probabilities calculated in the previous phase. This two-level proposal allows us to speed up the training process, concentrating most of the complexity in the word-level alignment phase and making the testing of new techniques and new association measures for n-gram alignment easier. Three of the most widely used association measures have been considered in this work: the Dice coefficient, pointwise mutual information and log-likelihood. The resulting n-gram level alignments were used for query translation at character n-gram level. For this purpose, two algorithms for the selection of candidate translations have been tested: a top-rank-based algorithm, which takes the H highest ranked n-gram alignments; and a threshold-based algorithm, which selects the n-gram level alignments according to a minimal threshold T.

Two techniques have been also considered for improving the system: the use of a bidirectional alignment during the input word-level alignment, and the introduction of a minimal word-level translation probability threshold for word-level pruning. Both techniques have allowed us to increase efficiency by drastically reducing the number of input word alignments to be processed and, consequently, the number of output n-gram alignments. This is done without

degrading the performance of the system. This way, computing and storage resources needed by the system can be considerably reduced.

The results obtained throughout our study confirm not only the feasibility of character n-gram based approaches for CLIR tasks, both for indexing–retrieval and translation purposes, but also that this validity not being tied to a given implementation. Moreover, our experiments show the remarkable robustness of these approaches against noisy or ambiguous input alignments. This factor, together with the inherent language-independent nature of n-grams, make this kind of solutions particularly interesting when dealing with multilingual environments where annotated language resources are scarce or unavailable.

With respect to future work, we plan to continue advancing on our study of the applicability of character n-gram based processing to IR and CLIR tasks. With respect to n-gram based IR in general, we intend to address some aspects that, at this point, still require attention: firstly, how to properly apply query expansion and relevance feedback in this context and, secondly, how to reduce the larger storage space required by n-gram based indexes and the resulting extra processing time, in order to both increase the performance of the system and reduce processing and storage resources. With regard to this later aspect, we propose to extend the concept of *stopword* to the case of n-grams. Savoy and Rasolofo (2002) made a similar proposal, the use of a *stop-n-gram* list for eliminating those most frequent and least discriminative n-grams. However, their list was not automatically generated, but obtained from n-grams created from a previously existing stopword list, which means that the system would become language-dependent, in their case from Arabic. Foo and Li (2004) used a similar manually created list for Chinese. We propose that such *stop-n-grams* should be generated automatically from the input texts (Blanco and Barreiro, 2007; Lo et al., 2005) in order to preserve the language-independent nature of n-gram based approaches. Regarding to CLIR in particular, we also intend to study the effects of the input parallel corpus on the alignment process with respect to: (a) the minimal input required, following the example of McNamee et al. (2009); and (b) in the particular case of the n-gram alignment algorithm presented in the this work, the quality of the first phase word-level alignment, that is, if this word alignment can be simplified in order to reduce its associated computational costs. Moreover, we want to take advantage of our experience in the study of the impact of misspellings in monolingual IR systems (Vilares et al., 2011) and to extend that work to the case of CLIR systems.

Finally, from a more practical point of view, we believe it would be interesting to use n-gram based translation for supporting the generation process of multilingual thesaurus for technical domains, as in the case of MorphoSaurus (Schulz et al., 2006) in Medicine, and its application to CLIR tasks (Markó et al., 2005). Twitter and other microblogging services will deserve special attention since it is a very noisy multilingual environment, for which specialized linguistic resources are still very scarce, particularly for non-English languages. This way, following the example of the research community, we intend to study the application of our n-gram based approach to our current research lines in microblog text processing for text normalization (Pennell and

Liu, 2014), sentiment analysis (Aisopos et al., 2012) and language identification tasks (Lui and Baldwin, 2014).

Biographies of the Authors

Jesús Vilares graduated in Computer Science Engineering from the University of A Coruña (Spain) in 2000. After a short period as a lecturer at the University of Vigo (Spain), he obtained a PhD Grant from the Spanish Ministry of Education (FPU Grant) at the University of A Coruña, where he obtained his PhD. in Computer Science in 2005. He is currently an Associate Professor at this university and he has been a member of the founding committee of the Spanish Society for Information Retrieval until this year. His research work focuses on Natural Language Processing —currently focused on microblog processing—, Text Mining and Information Retrieval.

Manuel Vilares has an MSc. in Applied Mathematics from the University of Santiago de Compostela (Spain, 1987), an MSc. in Software Engineering from CERICS (France, 1988), and a PhD. in Computer Science from the University of Nice–Sophia-Antipolis (France, 1992). He initially worked at INRIA (France) and later in Spain (1992), where he became Full Professor in Computer Science at the University of Vigo (2002). His research work focuses on Natural Language Processing, Logic Programming, Programming Language Design and Information Extraction.

Miguel A. Alonso graduated in Computer Science from the University of A Coruña (Spain) in 1993, where he started his research career. He spent a year at the Ramón Piñeiro Center for Research on Humanities (Santiago de Compostela, Spain) and then a year at the Rocquencourt research unit of INRIA (French National Institute for Research in Computer Science and Control). Since 1997 he has been a member of the Faculty of Computer Science of the University of A Coruña, where he obtained his Ph.D. degree in 2000, and since 2003 he has been an Associate Professor at that university. His research work focuses on Natural Language Processing and its application to Information Retrieval and Text Mining (particularly Opinion Mining).

Michael Oakes received a Ph.D. in Computer Science from the University of Liverpool in 1994. After his recent posts as Senior Lecturer in Computing at the University of Sunderland and Visiting Researcher at Uni Research, Bergen, he joined the Research Group of Computational Linguistics of the University of Wolverhampton as Reader in Computational Linguistics. His research work focuses on Information Retrieval, Computational Stylometry, and Statistics for Linguistics.

Acknowledgements

This research has been partially funded by the Spanish Ministry of Economy and Competitiveness and FEDER (through projects FFI2014-51978-C2-1-R and FFI2014-51978-C2-2-R) and by the Autonomous Government of Galicia (through grant R2014/034).

References

Aisopos, F., Papadakis, G., Tserpes, K., Varvarigou, T., 2012. Content vs. context for sentiment analysis: A comparative analysis over microblogs. In: Proceedings of the 23rd ACM Conference on Hypertext and Social Media. HT'12. ACM, pp. 187–196.

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

Blanco, R., Barreiro, A., 2007. Static pruning of terms in inverted files. In: Proceedings of the 29th European Conference on IR Research (ECIR 2007). Vol. 4425 of Lecture Notes in Computer Science. Springer-Verlag, pp. 64–75.

Büttcher, S., Clarke, C. L., Cormack, G. V., 2010. Information Retrieval: Implementing and Evaluating Search engines. MIT Press.

Carmel, D., Cohen, D., Fagin, R., Farchi, E., Herscovici, M., Maarek, Y. S., Soffer, A., 2001. Static index pruning for information retrieval systems. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01). ACM, pp. 43–50.

Cavnar, W. B., 1994. Using an n-gram-based document representation with a vector processing retrieval model. In: NIST Special Publication 500-225: Overview of the Third Text REtrieval Conference (TREC 3). pp. 269–278.

Chew, P. A., Verzi, S. J., Bauer, T. L., McClain, J. T., 2006. Evaluation of the Bible as a resource for Cross-Language Information Retrieval. In: Proceedings of the Workshop on Multilingual Language Resources and Interoperability. ACL, pp. 68–74.

CLEF, 2014. The CLEF Initiative. `http://www.clef-initiative.eu`

Damashek, M., 1995. Gauging similarity with n-grams: Language-independent categorization of text. Science 267 (5199), 843–848.

Dolamic, L., Savoy, J., 2008. UniNE at FIRE 2008: Hindi, Bengali, and Marathi IR. In: Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2008).

Dorr, B., Hovy, E., Levin, L., 2004. Machine Translation: Interlingual methods.

Ekmekcioglu, F. C., Lynch, M. F., Willett, P., 1996. Stemming and n-gram matching for term conflation in Turkish texts. Information Research 2 (2).

Escalante, H. J., Solorio, T., Montes-y Gómez, M., 2011. Local histograms of character n-grams for authorship attribution. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'11) – Volume 1. ACL, pp. 288–298.

EUROPARL, 2014. European Parliament Proceedings Parallel Corpus 1996–2011. http://www.statmt.org/europarl/

Foo, S., Li, H., 2004. Chinese word segmentation and its effect on Information Retrieval. Information Processing and Management 40 (1), 161–190.

Gao, Q., Bach, N., Vogel, S., 2010a. A semi-supervised word alignment algorithm with partial manual alignments. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR (WMT'10). ACL, pp. 1–10.

Gao, W., Niu, C., Nie, J.-Y., Zhou, M., Wong, K.-F., Hon, H.-W., 2010b. Exploiting query logs for cross-lingual query suggestions. ACM Transanctions on Information Systems 28, 1–33.

GIZA, 2014. giza-pp: GIZA++ statistical translation models toolkit. http://code.google.com/p/giza-pp/

Grefenstette, G. (Ed.), 1998. Cross-Language Information Retrieval. Vol. 2 of The Kluwer International Series on Information Retrieval. Kluwer Academic Publishers.

Hollink, V., Kamps, J., Monz, C., De Rijke, M., 2004. Monolingual document retrieval for European languages. Information Retrieval 7 (1-2), 33–52.

Huet, S., Lefévre, F., 2011. Unsupervised alignment for segmental-based language understanding. In: Proceedings of the First Workshop on Unsupervised Learning in NLP (EMNLP'11). ACL, pp. 97–104.

Hull, D., Grefenstette, G., 1996. Querying across languages: A dictionary-based approach to multilingual Information Retrieval. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96). ACM, pp. 49–57.

Järvelin, A., Talvensaari, T., Järvelin, A., 2008. Data driven methods for improving mono- and cross-lingual IR performance in noisy environments. In: Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data (AND'08). Vol. 303 of ACM International Conference Proceeding Series. ACM, pp. 75–82.

Khreisat, L., 2009. A machine learning approach for Arabic text classification using n-gram frequency statistics. Journal of Informetrics 3 (1), 72 – 77.

Koehn, P., 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In: Proceedings of the 10th Machine Translation Summit (MT Summit X), pp. 79–86. Corpus available at EUROPARL (2014).

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

Kwok, K., Choi, S., Dinstl, N., 2005. Rich results from poor resources: NTCIR-4 monolingual and cross-lingual retrieval of Korean texts using Chinese and English. ACM Transactions on Asian Language Information Processing 4, 136–162.

Lee, J. H., Ahn, J. S., 1996. Using n-grams for Korean text retrieval. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96). ACM, pp. 216–224.

Lehmann, W. P., 1992. Historical Linguistics. Taylor & Francis, London, Ch. 9.

Lo, R., He, B., Ounis, I., 2005. Automatically building a stopword list for an Information Retrieval system. In: Proceedings of the 5th Dutch-Belgian Information Retrieval Workshop (DIR'05).

Lui, M., Baldwin, T., 2014. Accurate language identification of Twitter messages. In: Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM 2014). ACL, pp. 17–25.

Lui, M., Lau, J. H., Baldwin, T., 2014. Automatic detection and language identification of multilingual documents. Transactions of the Association for Computational Linguistics 2, 27–40.

Ma, Y., Way, A., 2010. HMM word-to-phrase alignment with dependency constraints. In: Proceedings of the COLING 2010/SIGMT Fourth Workshop on Syntax and Structure in Statistical Translation (SSST-4), pp. 101–109.

Manning, C. D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. The MIT Press.

Markó, K., Schulz, S., Medelyan, O., Hahn, U., 2005. Bootstrapping dictionaries for cross-language information retrieval. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05). ACM, pp. 528–535.

McCarley, J., 1999. Should we translate the documents or the queries in cross-language information retrieval? In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99). ACL, pp. 208–214.

McNamee, P., 2008. Textual representations for corpus-based bilingual retrieval. Ph.D. thesis, University of Maryland at Baltimore County.

McNamee, P., Mayfield, J., 2004a. Character N-gram Tokenization for European Language Text Retrieval. Information Retrieval 7 (1-2), 73–97.

McNamee, P., Mayfield, J., 2004b. JHU/APL experiments in tokenization and non-word translation. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (Eds.), Comparative Evaluation of Multilingual Information Access Systems. Vol. 3237 of Lecture Notes in Computer Science. Springer-Verlag, pp. 85–97.

McNamee, P., Mayfield, J., 2005. Cross-Language Retrieval Using HAIRCUT at CLEF 2004. Vol. 3491 of Lecture Notes in Computer Science. Springer-Verlag, pp. 50–59.

McNamee, P., Mayfield, J., Nicholas, C., 2009. Translation corpus source and size in bilingual retrieval. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. NAACL-Short'09. ACL, pp. 25–28.

Miller, E., Shen, D., Liu, J., Nicholas, C., 2000. Performance and scalability of a large-scale n-gram based information retrieval system. Journal of Digital Information 1 (5), 1–25.

Miller, P. W., Chiswick, B. R., 2004. Linguistic distance: A quantitative measure of the distance between english and other languages. Discussion Paper 1246, Institute for the Study of Labor (IZA).

Mustafa, S. H., 2005. Character contiguity in n-gram-based word matching: The case for Arabic text searching. Information Processing and Management 41 (4), 819–827.

Mustafa, S. H., Al-Radaideh, Q. A., 2004. Using n-grams for Arabic text searching. Journal of the American Society for Information Science and Technology (JASIST) 55 (11), 1002–1007.

Nakov, P., Ng, H. T., 2012. Improving statistical machine translation for a resource-poor language using related resource-rich languages. Journal of Artificial Intelligence Research (JAIR) 44, 179–222.

Nie, J.-Y., 2010. Cross-Language Information Retrieval. Vol. 8 of Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Nunzio, G. M. D., Ferro, N., Mandl, T., Peters, C., 2006. CLEF 2006: Ad Hoc Track Overview. In: Working Notes of the CLEF 2006 Workshop, pp. 21–34. Available at CLEF (2014).

Oard, D., 1998. A comparative study of query and document translation for Cross-Language Information Retrieval. In: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup (AMTA'98). Springer-Verlag, pp. 472–483.

Och, F. J., Ney, H., 2003. A systematic comparison of various statistical alignment models. Computational Linguistics 29 (1), 19–51. Toolkit available at GIZA (2014).

Ogawa, Y., Matsuda, T., 1999. Overlapping statistical segmentation for effective indexing of Japanese text. Information Processing and Management 35 (4), 463–480.

Ounis, I., Lioma, C., Macdonald, C., Plachouras, V., 2007. Research directions in Terrier: A search engine for advanced retrieval on the Web. Novática/UPGRADE Special Issue on Web Information Access 8 (1), 49–56. Terrier toolkit available at TERRIER (2012).

Pennell, D. L., Liu, Y., 2014. Normalization of informal text. Computer Speech and Language 28 (1), 256–277.

Porter, M. F., 1980. An algorithm for suffix stripping. Program 14 (3), 130–137.

Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P., 2011. Cross-language plagiarism detection. Language Resources and Evaluation 45, 45–62, 10.1007/s10579-009-9114-z.

Rehm, G., Uszkoreit, H. (Eds.), 2011. META-NET White Paper Series. Springer. Available online at `http://www.meta-net.eu/whitepapers`.

Resnik, P., Smith, N. A., 2003. The Web as a parallel corpus. Computational Linguistics 29 (3), 349–380.

Robertson, A. M., Willett, P., 1998. Applications of n-grams in textual information systems. Journal of Documentation 54 (1), 48–69.

Salton, G., 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley.

Sapkota, U., Solorio, T., Montes-y Gómez, M., Ramírez-de-la Rosa, G., 2013. Author profiling for English and Spanish text. In: Notebook for PAN at CLEF 2013.

Savoy, J., 2003. Cross-Language Information Retrieval: Experiments based on CLEF 2000 corpora. Information Processing and Management 39, 75–115.

Savoy, J., Rasolofo, Y., 2002. Report on the TREC 11 experiment: Arabic, named page and topic distillation searches. In: NIST Special Publication 500-251: The Eleventh Text Retrieval Conference (TREC 11), pp. 765–774.

Schulz, S., Markó, K., Daumke, P., Hahn, U., Hanser, S., Nohama, P., de Andrade, R. L., Pacheco, E., Romacker, M., 2006. Semantic atomicity and multilinguality in the medical domain: Design considerations for the morphosaurus subword lexicon. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). European Language Resources Association (ELRA).

Stamatatos, E., 2009. A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60 (3), 538–556.

Terrier, 2012. Terrier IR Platform. `http://www.terrier.org`

Tomović, A., Janičić, P., Kešelj, V., 2006. n-Gram-based classification and unsupervised hierarchical clustering of genome sequences. Computer Methods and Programs in Biomedicine 81 (2), 137–153.

Vilares, J., Oakes, M. P., Tait, J. I., 2007a. A first approach to CLIR using character n-grams alignment. In: Evaluation of Multilingual and Multimodal Information Retrieval. Vol. 4730 of Lecture Notes in Computer Science. Springer-Verlag, pp. 111–118.

Vilares, J., Oakes, M. P., Vilares, M., 2007b. Character N-Grams Translation in Cross-Language Information Retrieval. In: Natural Language Processing and Information Systems. Vol. 4592 of Lecture Notes in Computer Science. Springer-Verlag, pp. 217–228.

Vilares, J., Oakes, M. P., Vilares, M., 2008. English-to-French CLIR: A knowledge-light approach through character n-grams alignment. Vol. 5152 of Lecture Notes in Computer Science. Springer-Verlag, pp. 148–155.

Vilares, J., Oakes, M. P., Vilares, M., 2009. Recent Advances in Natural Language Processing V. Vol. 309 of Current Issues in Linguistic Theory. John Benjamins Publishing Company, Ch. Character N-Grams as Text Alignment Unit: CLIR Applications.

Vilares, J., Vilares, M., Otero, J., 2011. Managing Misspelled Queries in IR Applications. Information Processing & Management 47 (2), 263–286.

Wu, D., He, D., Ji, H., Grishman, R., 2008. A study of using an out-of-box commercial MT system for query translation in CLIR. In: Proceedings of the 2nd ACM Workshop on Improving non English Web Searching (iNEWS'08). ACM, pp. 71–76.

Zeman, D., 2009. Using unsupervised paradigm acquisition for prefixes. In: Evaluating Systems for Multilingual and Multimodal Information Access. Vol. 5706 of Lecture Notes in Computer Science. Springer, pp. 983–990.

Zobel, J., Dart, P., 1995. Finding approximate matches in large lexicons. Software–Practice & Experience 25 (3), 331–345.