Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 136–164

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

# 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*

Received 5 August 2014; received in revised form 11 May 2015; accepted 18 September 2015
Available online 1 October 2015

**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. After an initial tuning phase in order to analyze the most effective way for its application, the results obtained, close to the upper baseline, not only confirm the consistency across languages of this kind of character *n*-gram based approaches, but also constitute a further proof of their validity and applicability, these not being tied to a given implementation.


*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 considerably, 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*

☆ This paper has been recommended for acceptance by Roger K. Moore.

* Corresponding author. Tel.: +34 981 167 000x1364; fax: +34 981 167 160.
*E-mail addresses:* jesus.vilares@udc.es (J. Vilares), vilares@uvigo.es (M. Vilares), miguel.alonso@udc.es (M.A. Alonso), michael.oakes@wlv.ac.uk (M.P. Oakes).

http://dx.doi.org/10.1016/j.csl.2015.09.004
0885-2308/

*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 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, respectively (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 right 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 it has many advantages (Dorr et al., 2004). 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.