Translating Pieces of Words

Paul McNamee
University of Maryland Baltimore County
1000 Hilltop Road
Baltimore, MD 21250 USA
mcnamee@cs.umbc.edu

James Mayfield
Johns Hopkins University Applied Physics Lab
11100 Johns Hopkins Road
Laurel, MD 20723-6099 USA
james.mayfield@jhuapl.edu

ABSTRACT
Translation for cross-language information retrieval need not be word-based. We show that character n-grams in one language can be 'translated' into character n-grams of another language. We demonstrate that such translations produce retrieval results on par with, and often exceeding, those of word-based and stem-based translation.

Categories and Subject Descriptors
H.3.1 [**Information Systems**]: Content Analysis and Indexing – *linguistic processing, indexing;* H.3.3 [**Information Systems**]: Information Search and Retrieval – *query formulation.*

General Terms
Experimentation.

Keywords
Cross-language information retrieval, character n-gram tokenization, parallel corpora, translation.

1. INTRODUCTION
Most cross-language information retrieval (CLIR) systems work by translating words from the source (*i.e.,* query) language to the target (*i.e.,* document) language. We propose that translating pieces of words (sequences of *n* characters in a row, called *character n-grams*) can be as effective as translating words while conveying additional benefits for CLIR. Translating pieces of words seems odd. To translate a word means to select a word in another language that, to a person, carries the same meaning. But word fragments do not (necessarily) carry meaning for a person, so how can we claim to translate them?

We avoid this difficulty by adopting a functional view of meaning. We define the meaning of an indexing term broadly as the range of documents the term allows us to access. Given this definition, a good translation of an indexing term is a term in the target language that means the same thing, *i.e.,* one that provides access to target language documents that are similar to those accessible through the source language term. Similarity can be defined in a variety of ways, such as 'describing the same concepts,' or 'equally relevant,' or even 'direct translations' if a parallel corpus is in use for evaluation.

This view of meaning does not commit to words as indexing terms; it admits the use of stems, phrases, or any other type of indexing term. In this study we use character n-grams. The translation of an n-gram is a target language term that provides access to target language documents that are similar to the source language documents accessible through the source n-gram.

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

2. PRODUCING TRANSLATIONS
A parallel collection allows us to assess the quality of a translation given this definition of translation. Given a retrieval approach, we can empirically determine the translation quality of a pair of terms by assessing the degree to which they provide access to the "same" documents. Alternatively, we can use a parallel collection to generate translations of query words, and use a cross-language retrieval task to evaluate the quality of those translations. We adopt the latter approach.

To build a parallel collection that supports translation of European languages, we mined the Official Journal of the European Union over the past six years. The Journal is available in the official EU languages, and is published electronically in Adobe PDF format. We downloaded content nightly, converted the PDF to plain text using a publicly available tool, and aligned the documents using the char_align software developed by Church [3]. In this way we obtained about 500MB of text in each language, which can be aligned with any of the other languages.

Once the aligned collection has been indexed, a statistical translation lexicon can be extracted, mapping terms in one language to a set of alternative translations, possibly with translation probabilities for each. Alternatively, the most probable match, or best *k* matches can be extracted [1].

Our method for producing translations of a source language term *s* starts with finding the set *S* of all source language documents that contain term *s*. We then select the set *T* of target language documents that are translations of documents in *S*. Next, we identify terms that occur much more frequently in *T* than they do in the target language collection as a whole; the best candidate translation for *s, t*, is the term exhibiting the largest such difference. This measure is related to mutual information; however, we believe our technique is more general as it permits the set of documents to be identified through any means. Because retrieval is an integral part of the process, we call this style of translation *retrieval-based translation*. Table 1 contains sample n-gram mappings from French to Dutch that were produced in this way. The French n-grams were taken from the words lait (*milk*), Olympique (*Olympic*), and Pays Bas (*Netherlands*).

3. EVALUATION
We conducted an evaluation of the effectiveness of n-gram translation using six language pairs in CLEF-2004. We used the HAIRCUT retrieval engine [4] with a language model similarity metric. The effectiveness of corpus-based, one-best translation of words, stems (as produced by the Snowball stemmers [7]), and character 5-grams, measured by mean average precision, is

compared in Figure 1. The figure shows the remarkable effectiveness of n-gram translation. In all cases save Spanish to Portuguese, n-gram translation outperforms both word translation and stem translation. The more disparate the language pairs, the bigger the advantage held by n-gram translation.

Table 1. Sample 4-gram translations

Source (French)	Target (Dutch)
lait	melk
olym	olym
ique	isch
pays	_lan
ys_b	_ned

4. DISCUSSION

Translation of n-grams for CLIR has a number of advantages over word translation:

- The traditional advantages of n-grams for monolingual retrieval, including language neutrality, automatic handling of morphological variation, robustness against typographical errors, and capture of phrasal information (when n-grams span word boundaries), hold for CLIR.

- Problems of data sparseness are mitigated. Rare and out-of-vocabulary terms are a problem for all term types. However, n-grams present more opportunities to find useful matches than do words or stems. Consider a query on *Gaddafi* or *Gadaffi* or *Qaddafi* or *Khadafi,* or any of the many other variants of the name. A word-based system is likely to fail to find a translation for most of these; an n-gram-based system is likely to find useful translations for some component n-grams for many such variants.

Pirkola *et al.* used small n-grams (n=2 or 3) in concert with statistically derived rules for mapping orthography to achieve translation of out-of-vocabulary words between closely related languages [6]. This present work can be distinguished from that study in several ways. First, we have extrinsically validated the efficacy of our translations using bilingual test sets. Second, our approach should be effective even in languages without a common alphabet. Third, Pirkola *et al.*'s method uses a bilingual dictionary, while our approach requires a parallel corpus. One thing that is not clear at present is how large of a corpus is needed to derive accurate n-gram mappings, though this can be addressed in future work.

We see several benefits from the use of retrieval-based translation. For one thing, term types need not be the same on each side of the translation. For example, one might translate a French word into a Chinese 2-gram, thereby obviating the need to perform Chinese word segmentation. Also, translations may be calculated one term at a time. This contrasts with IBM model translation [2] such as performed by Giza++ [5], which requires large amounts of memory to translate many terms simultaneously.

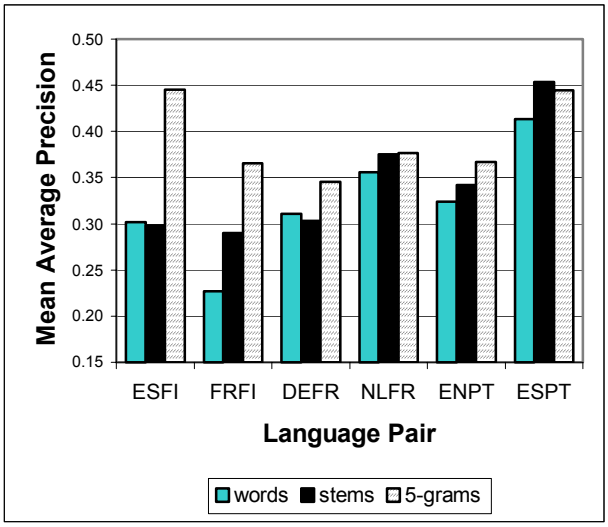

Figure 1. Comparison of word, stem, and n-gram translation, for six language pairs. Differences from n-gram translation are statistically significant at the 0.05 level (Wilcoxon test) for ESFI, FRFI, DEFR, and (words only) NLFR.

One of our goals for this experiment is to demonstrate that an end-to-end system for cross-language information retrieval can avoid language-specific processing, even in the translation component. N-gram translation clearly supports this contention.

ACKNOWLEDGMENTS

The authors thank Charles Nicholas (UMBC) and several anonymous reviewers for their constructive feedback.

REFERENCES

[1] Braschler, M. and Schäuble, P. 'Experiments with the Eurospider Retrieval System for CLEF 2000.' *Revised Papers from the Workshop of the Cross-Language Evaluation Forum* pp. 140-148. 2000.

[2] Brown, P., Della Pietra, V., Della Pietra, S. and Mercer, R. 'The mathematics of statistical machine translation: Parameter estimation.' *Computational Linguistics* **19**(2):263-311. 1993.

[3] Church, K. 'Char_align: A program for aligning parallel texts at the character level.' *Proceedings of the 31st conference of the Association for Computational Linguistics,* pp. 1-8. 1993.

[4] McNamee, P., and Mayfield J. Character N-gram Tokenization for European Language Retrieval. *Information Retrieval,* **7**(1-2):73-97. 2004.

[5] Och, F. *GIZA++: Training of statistical translation models.* <http://www.fjoch.com/GIZA++.html>.

[6] Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K., and Järvelin K. 'Fuzzy Translation of Cross-Lingual Spelling Variants.' *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 345-352. 2003.

[7] Porter, M. *Snowball.* <http://snowball.tartarus.org/>