Machine Translation-Supported Cross-Language Information Retrieval for a Consumer Health Resource

Graciela Rosemblat, Darren Gemoets, Allen C. Browne, Tony Tse

National Library of Medicine, Bethesda, Maryland

ABSTRACT

The U.S. National Institutes of Health, through its National Library of Medicine, developed ClinicalTrials.gov *to provide the public with easy access to information on clinical trials on a wide range of conditions or diseases. Only English language information retrieval is currently supported. Given the growing number of Spanish speakers in the U.S. and their increasing use of the Web, we anticipate a significant increase in Spanish-speaking users. This study compares the effectiveness of two common cross-language information retrieval methods using machine translation, query translation versus document translation, using a subset of genuine user queries from* ClinicalTrials.gov. *Preliminary results conducted with the* ClinicalTrials.gov *search engine show that in our environment, query translation is statistically significantly better than document translation. We discuss possible reasons for this result and we conclude with suggestions for future work.*

INTRODUCTION

ClinicalTrials.gov [1] is a Website[1] developed by the National Library of Medicine (NLM)[2] that provides the public with easy access to information on clinical trials for a wide variety of conditions and interventions. It contains nearly 8,000 records. In *ClinicalTrials.gov* only English-language information retrieval (IR) is supported at present. With the U.S. Spanish-speaking population ranking as the largest minority[3] and as the fastest growing segment of on-line users[4], NLM is committed to supporting this often underserved population[5]. The recent introduction of *MEDLINEplus en Español* [2] will accelerate the need for Spanish language support in *ClinicalTrials.gov*, an assumption supported by the results of a recent MEDLINEplus survey[6] and focus group members explicitly requesting that *ClinicalTrials.gov* be available in Spanish. Furthermore, providing information about trials to the Spanish-speaking community will contribute towards improving the diversity of clinical trials participants by facilitating the inclusion of people of Hispanic descent. Ultimately, the best way to serve the Spanish-speaking community will be to present retrieved documents in Spanish, the next phase of this project (see Directions for Future Work). We focus here on the IR aspects of the project.

The purpose of this study is to compare two well-known approaches for cross-language information retrieval (CLIR)— query translation and document translation— to determine which is more effective in assisting Spanish-speaking *ClinicalTrials.gov* users.

When the language of the queries does not match the language of the documents, one common approach is to use a machine translation (MT) system to reduce IR to a pseudo-monolingual level [3], that is, IR in a common language. For Spanish language queries and English language documents, this means either translating the queries into English to match the language of the documents, or translating the documents into Spanish to match the language of the queries. Document translation generally outperforms query translation [4] because translated documents provide greater linguistic context, which in turn facilitates part-of-speech disambiguation and sense selection. However, due to practical considerations such as speed and size of the collection, query translation is the prevailing CLIR method at present.

To evaluate these two CLIR methods in our own environment, we compared both to a monolingual retrieval system using our in-house search engine (SE). A small pilot project (described below) using human relevancy judgments on a subset of the *ClinicalTrials.gov* corpus allowed us to optimize the SE.

[1] http://www.ClinicalTrials.gov/

[2] NLM, National Institutes of Health (NIH), U.S. Department of Health and Human Services (DHHS).

[3] 38.8 million in 2000 (http://www.census.gov/).

[4] 11% of U.S. online users, comScore Media PR, 03/27/2003, http://www.comscore.com/press/pr.asp

[5] It is comforting to people to express health concerns in their native language.

[6] Internal NLM survey by Fulcrum Analytics, 2003.

METHODOLOGY

Pilot Study

The goal of the pilot study was to develop an automated mechanism for approximating human judgment of relevance to be used to create a Gold Standard. Limiting the number of documents allowed for the manual review of a representative set of documents for relevance. Human relevancy determination enabled us to create an automated IR method to act as ground truth, used to train a "surrogate" mechanism for determining relevance. Based on the results of the pilot study, the automated mechanism with optimized search parameters (criteria), was used in the full study of 7,170 records.

The pilot study was carried out with 25 randomly selected AIDS-related documents from a bilingual subset of *ClinicalTrials.gov*, and 100 natural language queries (NLQs) generated from generic clinical question templates [5]. An example of a generic question is: '*How effective is <Treatment X> for <Disease Y>?'*, where *<Treatment X>* and *<Disease Y>* represent terms corresponding to specific semantic types. We used NLQs in the pilot study to motivate human relevancy evaluation. To obtain the expressions to populate the templates, medical terms were extracted from the titles of the 25 documents using MetaMap [6], constrained by semantic types from the Unified Medical Language System® (UMLS®) Semantic Network®. MetaMap parses and maps text to concepts in the UMLS Metathesaurus® [7]. These terms also formed the basis for a parallel set of keyword queries. Human relevancy for each NLQ was determined by consensus among the first three authors and an AIDS professional. We wrote a Java program to determine the optimal SE parameters for the surrogate mechanism to best approximate human relevancy ranking. These parameters included consideration of:

- number of suggestions[7] for zero-hit queries;
- number of terms within a given suggestion; and
- minimum weighting score returned by the SE for the terms in each suggestion.

These optimized parameters for monolingual retrieval provided a mechanism to derive a Gold Standard, representing "optimal" sets of retrieved documents for comparing alternative CLIR mechanisms. The criteria were optimized for obtaining the best F factor (a measure of performance combining both precision (P) and recall (R) values), prioritizing a high P in those cases where the same F was obtained.

Approach	*Language of Query*	*Corpus*
Gold Standard	**English queries** constructed with MetaMap keywords	original **English corpus**
Query translation	**English queries**, via **MT** of Spanish human translated (HT) queries	original **English corpus**
Document translation	**Spanish queries**, via **HT** of original English queries	**Spanish corpus** via MT

Table 1. Query and corpus language combinations

As summarized in Table 1, the query translation and document translation runs were compared against the Gold Standard document set. This established how closely the crosslingual searches approximated monolingual retrieval, traditionally taken as the best measure of effectiveness [4] in the CLIR literature.

To translate queries and documents, we used the machine translation (MT) software of the Pan American Health Organization (PAHOMTS) [8]. The sentence is the basic translation unit, but smaller units are parsed and translated as well. Selection of alternate translations is determined by context-specific rules. In addition to default and alternate translations (context-specific), the system provides alternate, topic-specific translations or glosses for expressions that already have a more general translation. These specialized subdictionaries for particular fields of knowledge, or microglossaries, can be selected at run-time to override an expression's more general translation [8]. For our pilot study, PAHOMTS was used with default dictionaries and Patient Education (consumer-oriented terminology) and SuperMedical (specialized medical translations) microglossaries. The language directions used were English-Spanish and Spanish-English.

Natural Language Queries	F	P	R
Query Translation	**0.734**	0.710	0.759
Document Translation	**0.647**	0.680	0.617
Keyword Queries			
Query Translation	**0.694**	0.761	0.638
Document Translation	**0.630**	0.597	0.668

Table 2. Initial scores, 100 queries, 25 documents

A detailed analysis of the data (Table 2) and the results obtained led to the following changes:

[7] Each suggestion presents a ranked list of candidate documents for retrieval.

- The SE retrieval procedure was adjusted at the basic monolingual level resulting in the generation of more suggestions.
- We filtered out some *semantic* types (UMLS qualitative/quantitative concepts) and certain *lexical* ones (*administration*, *recommending*, *dosage*, etc.) to reduce noise, false negatives / false positives.
- Spanish stopwords were added to the existing English-only list.
- The Gold Standard criteria for keyword queries was determined by running the queries against the Gold Standard document set for NLQs (instead of against the documents judged relevant by humans).

Re-executing the criteria-optimizing Java program resulted in a new Gold Standard document set, along with a new optimal set of Gold Standard search parameters (Table 3).

Natural Language Queries	**F**	**P**	**R**
Query Translation	**0.805**	0.835	0.777
Document Translation	**0.722**	0.743	0.701
Keyword Queries			
Query Translation	**0.876**	0.898	0.855
Document Translation	**0.765**	0.797	0.736

Table 3. Refined criteria, 100 queries, 25 documents

As a result of the changes specified above, the Gold Standard of keyword queries more closely matched human relevancy[8]. The keyword queries showed improved results across-the-board, scoring higher than the NLQs in both runs.

The results of the pilot study anticipate the full study results in that query translation scored significantly higher than document translation.

Full Study

The corpus for the full study consisted of 7,170 records, the entire collection of *ClinicalTrials.gov* records on January 15, 2003. We randomly selected 225 *ClinicalTrials.gov* queries logged between January 1-7, 2003, excluding a few malformed queries, duplicates, and those with spelling errors. Although there were some NLQs, the vast majority were keyword queries, typically a single noun or nominal expression indicating a condition or a drug.

[8] Keyword queries refined Gold Standard:
Initial score: F= 0.620. Final score: F= 0.799.

The Gold Standard parameters for keyword queries were used for all further comparisons. The manual filters from the pilot study were not applied because these queries did not contain distractors such as *recommending* and *dosage*.

The 225 English queries were run against the *ClinicalTrials.gov* corpus to produce the Gold Standard document set.

PAHOMTS software was used with NLM-updated dictionaries and the Patient Education and SuperMedical microglossaries. The updates to the dictionary consisted of several hundred vocabulary items and expressions, plus general and specific context-sensitive rules and sense selection for specific collocations. The updates were based on human review of the unedited MT output of approximately 70-100 randomly selected *ClinicalTrials.gov* records.

The Wilcoxon Signed Rank Test was used to compare the F-values obtained with the two methods, query translation and document translation. Its parametric alternative (paired t-test) was not applicable here because many of the differences between F-values were equal to zero.

Finally, to see whether our results held for genuine Spanish queries, an additional run was carried out with 119 randomly selected, non-cognate, well-formed queries from *MEDLINEplus en Español*, a similar consumer-oriented medical web-based system. These Spanish queries were logged in early 2003. The Gold Standard was established by running a human translation of these queries against the English *ClinicalTrials.gov* corpus, using the same SE parameters as in the main study. No manual filters were applied.

RESULTS

The full study results (Table 4) were largely consistent with those of the pilot study in that query translation outperformed document translation.

Keyword Queries	**F**	**P**	**R**
Query Translation	**0.859**	0.876	0.842
Document Translation	**0.762**	0.811	0.719

Table 4. Final scores, 225 queries, 7,170 documents

Query translation was statistically significantly better than document translation based on the non-parametric Wilcoxon Signed Rank Test ($p<0.0001$).

These results are directly related to the language in which the search was performed: the better capabilities built into the search engine for English resulted in query translation scoring significantly higher than document translation. These capabilities include English Lexical Variant Generation [10,11], synonymy, and conceptual (semantic) matching, among others.

The queries in the full study included a percentage of true cognates (*glaucoma*, *diabetes*, *endometriosis*) which resulted in perfect precision and recall values. There was some concern that including these query types would artificially inflate the results, and also dilute the difference between the two CLIR methods. An analysis of the queries in *ClinicalTrials.gov* logs showed that they broadly fall into these categories:

a) true cognates;
b) misspelled and/or malformed;
c) containing Boolean predicates;
d) well-formed, non-cognate queries;

The inclusion of queries a) and b) in the set, though it constitutes a fair sample representation, would fail to provide a true measure of MT system performance, since queries a) will always score 1.0 for P and R, and queries b) will not contribute to the results due to their undefined F value. The Boolean operators in c) no longer function if translated. Consequently, we limited the query set to those in d), which can be considered non-trivial queries.

Finally, to simulate typical user behavior, we truncated the retrieved document list at the top 10 ranked documents per query[9]. The process of considering only the non-trivial queries, combined with the 10 document cut-off, resulted in a new set of 119 queries (Table 5). Although these scores deviated greatly from those in Table 4, the margin of difference between the two retrieval methods was consistent with the previous runs.

Keyword Queries	**F**	**P**	**R**
Query Translation	**0.592**	0.792	0.473
Document Translation	**0.517**	0.729	0.401

Table 5. Final scores, 119 non-trivial queries, 7,170 documents

The preliminary findings from our pilot study also generalize to Spanish MEDLINEplus queries[10]. Interestingly, these queries were remarkably similar to the English ones (see Discussion).

DISCUSSION

CLIR literature generally reports that document translation outperforms query translation, since more linguistic context improves disambiguation [4]. However, our results show that the opposite is true in our environment. We believe this outcome is due to:

- The enhanced capabilities for English retrieval (lexical variant generation, synonymy, search engine design) play a very significant role[11].
- Most of our queries, both English and Spanish, are unambiguously ***nouns*** (treatments, conditions) which eliminates much of the need for part-of-speech and lexical disambiguation.
- Corpus and queries are domain-restricted. Selecting domain-specific glossaries at run-time prioritizes specialized translations over regular ones in cases of polysemy (multiple senses).
- The SE was optimized for keyword queries, which in our case far outnumber NLQs.

P, R, and F are likely to improve as more *ClinicalTrials.gov* vocabulary terms and language rules are added to the PAHO MT dictionaries.

Limitations

The Wilcoxon Signed Rank Test has demonstrated that the superior performance of query translation compared to document translation is not by chance. However, the actual difference in F factor (approximately 0.1 and 0.75 in our respective samples) as well as the P, R, and F values obtained are data-driven. Sampling bias is a limitation: a different set of queries may result in different values. There is good evidence for generalizability: a repetition of the study with 119 different queries and translations by an external translator, unfamiliar with the PAHOMTS system, also resulted in query translation outscoring document translation.[12]

The document set returned by monolingual *ClinicalTrials.gov* IR served as the Gold Standard. This had the benefit of enabling us to compare

[9] With no cut-off, or with a 20-document cut-off, F increased in both runs. Query translation still scored higher by the same margin of difference (about 0.1).
[10] Query Transl. F=0.723, Document Transl. F=0.564.
[11] For example, English query *kidney* retrieved 434 documents in our study, but Spanish (HT) *riñón* retrieved 0 documents. Spanish prefers the adjectival form *renal* as a premodifier. Presently, we have no resource available to establish the semantic link between the noun (*riñón*) and the adjective (*renal*).
[12] Query Transl. F=0.588, Document Transl. F=0.468.

Spanish retrieval capabilities with the deployed *ClinicalTrials.gov* system, which contains its own limitations and reduces in turn the viability of the Gold Standard. In theory, a given query may produce *more* relevant results in Spanish than in English. In our study, such differences would appear as "incorrect" because they deviate from the Gold Standard.

Finally, human translations are a limitation. Queries lack context and are often confusing, causing a variance in interpretation. There are often several ways to translate a given query, especially given the limited context of many of the queries. At times a translator must "guess" the author's intent and translate accordingly. Other times a direct translation may not exist. The unavailability of genuine native *ClinicalTrials.gov* Spanish queries, which will not exist until the Spanish version is deployed, imposes another limitation. However, in our view, the Spanish MEDLINEplus queries provide an adequate surrogate to test the validity of our findings.

Directions for Future Work

We have just started the process of developing lexical variant generation in Spanish, and exploring sources of Spanish synonymy and spell-checking. Based on our findings to date, these tools should enable Spanish retrieval to match English retrieval more closely.

By selecting the document set as returned by the deployed *ClinicalTrials.gov* and with the knowledge that for present search capabilities, query translation outperforms document translation, we are in a position to study the performance of Spanish retrieval compared to the deployed English-only system. This will further enable us to know how well we are processing Spanish queries, and provide us with a basis for comparison as we develop Spanish versions of our English-only search tools.

MT is designed as a tool to aid human translators, and presenting unedited translations to the public is not recommended. Human post-editing is a labor- and time-intensive process, especially for a dynamic corpus like *ClinicalTrials.gov* that is updated nightly. Thus, we are now developing an abbreviated Spanish version of *ClinicalTrials.gov* records, with static Spanish translations for items that remain constant for all trial protocols (section titles, subtitles, names of treatments, etc). Non-static fields (trial description, eligibility criteria) will provide links to the English version of the trial. The main title will also be in Spanish, which will help Spanish-speakers decide whether a study applies to them and act accordingly[13]. In the longer term, efforts will also be directed toward optimizing document translation techniques.

ACKNOWLEDGEMENTS

The authors are indebted to John Frye for his contributions to relevance determination, Olivier Bodenreider for sharing his knowledge on statistical issues; and Russell Loane for changing the SE retrieval procedures to our specifications.

REFERENCES

[1] McCray AT, Ide NC. Design and implementation of a national clinical trials registry. *J Am Med Inform Assoc.* 2000 May-Jun; 7(3):313-23.

[2] MEDLINEplus GOES SPANISH. Press Release. *http://www.nlm.nih.gov/news/press_releases/medplusspanish.html*.

[3] Oard DW, Diekema AR. Cross-Language Information Retrieval. In: Williams HE, (Ed.) *Annual Review of Information Science and Technology (ARIST)*. Vol. 33. Medford, NJ: Information Today; 1998, 223-56.

[4] Oard DW. A comparative study of query and document translation or cross-language information retrieval. *Third Conference of the Association for Machine Translation in the Americas, AMTA, 1998: 472-83*.

[5] Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, et al. A taxonomy of generic clinical questions: classification study. 2000. *BMJ*. 321(7258):429-32.

[6] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp 2001:17-21*.

[7] *UMLS Knowledge Sources. Documentation*. U.S. National Library of Medicine. 14th Edition. 2003.

[8] PAHO Machine Translation System: *http://www.paho.org/English/AGS/MT/Machine_Trans.htm*

[9] León M. *A New Look for the PAHO MT System*. In: White JS, (Ed) AMTA 2000, pp. 219-22.

[10] SPECIALIST Lexicon. Fact Sheet. *http://www.nlm.nih.gov/pubs/factsheets/umlslex.html*

[11] Divita G, Browne AC, and Rindflesch T. Evaluating lexical variant generation to improve information retrieval. *Proc AMIA Symp*. 1998:775-9.

[13] Either by enlisting help with the English translation, or by reading it in English, as often reading ability well exceeds writing ability.