Graciela Rosemblat, Tony Tse, Darren Gemoets
National Library of Medicine, Bethesda, Maryland, USA

Adapting a Monolingual Consumer Health System for Spanish Cross-Language Information Retrieval

Abstract: This preliminary study applies a bilingual term list (BTL) approach to cross-language information retrieval (CLIR) in the consumer health domain and compares it to a machine translation (MT) approach. We compiled a Spanish-English BTL of 34,980 medical and general terms. We collected a training set of 466 general health queries from MedlinePlus en español and 488 domain-specific queries from *ClinicalTrials.gov* translated into Spanish. We submitted the training set queries in English against a test bed of 7,170 ClinicalTrials.gov English documents, and compared MT and BTL against this English monolingual standard. The BTL approach was less effective (F = 0.420) than the MT approach (F = 0.578). A failure analysis of the results led to substitution of BTL dictionary sources and the addition of rudimentary normalization of plural forms. These changes improved the CLIR effectiveness of the same training set queries (F = 0.474), and yielded comparable results for a test set of new 954 queries (F= 0.484). These results will shape our efforts to support Spanish-speakers' needs for consumer health information currently only available in English.

1 Introduction

Language is a barrier for non-English speakers seeking health information. Many high-quality online consumer health systems offer information in English only. There is a growing need to provide access to health information for non-English speaking consumers: in March 2002, over 13.3% of the population in the United States was Spanish-speaking (Ramirez & de la Cruz, 2002). While Spanish-language health sites exist, a survey by Berland et al. (2001) found some to be more difficult to read and less comprehensive than comparable English-language sites.

Cross-language information retrieval (CLIR) provides one way to leverage existing consumer health resources by matching queries in one language with documents written in another, either by translating the queries or the documents (Adriani & Croft, 1997). While translating the queries requires considerably fewer resources than translating full-length documents, this approach increases the likelihood of mistranslation due to word-sense and part-of-speech ambiguity (Oard, 1998). In an earlier CLIR study with a machine translation (MT) approach, we found that query translation was more effective than document translation in our environment (Rosemblat, Gemoets, Browne, & Tse, 2003). The present study describes preliminary query-translation CLIR strategies using bilingual term lists (BTL) with *ClinicalTrials.gov* (http://clinicaltrials.gov/), an existing consumer health site.

The *ClinicalTrials.gov* Web site provides the public with easy access to information about clinical research protocols for a variety of conditions and interventions. The system integrates several NLM products, including a flexible, custom-designed search engine (McCray, Ide, Loane, & Tse, 2004); a monolingual terminology server based on the Unified Medical Language System® (NLM, 2003) for synonym expansion; and normalization of inflectional variants using the Lexical Variant Generator (Divita, Browne, & Rindflesch, 1998).

While only English-language retrieval is supported at present, the results of a recent in-house survey of visitors to MedlinePlus, NLM's primary consumer health site, and focus groups of Spanish-speaking consumers indicate a strong need for information in Spanish about clinical research. Making such information accessible to Spanish speakers would empower members of this underserved population to make informed decisions about clinical research participation. It would also facilitate the inclusion of people of Hispanic descent in human studies.

2 Background

Our earlier study compared query- and document-translation through an MT approach (using the Pan American Health Organization MT system), with a test bed of 7,170 records from *ClinicalTrials.gov* as of 15 January 2003 (Rosemblat, Gemoets, Browne, & Tse, 2003). For that study, we randomly selected 119 queries from *ClinicalTrials.gov* log files, excluding malformed and non-retrieving queries, cognates, and misspellings. A professional medical translator translated the queries into Spanish. Based on the F-factor retrieval measure, which combines precision and recall, query translation (F = 0.592) was found to be more effective than document translation (F = 0.517). Detailed analysis of the results revealed that this outcome was due to:

- Nouns in Queries: Most of the test queries were unambiguously nouns, eliminating the need for part-of-speech and lexical disambiguation.
- Built-in English-Only Search Enhancements: Translating the queries into English allowed us to take advantage of our system's enhanced capabilities for English retrieval, namely, lexical variant generation, synonymy, and search engine design.

In our current study, a BTL approach is compared to an MT method. The primary reasons for desiring to move from a proprietary MT system to BTL are (1) customizability and transparency needed to evaluate the detailed effects of various linguistic and terminological parameters on CLIR and (2) system compatibility with existing software.

3 Methods

We constructed a BTL of single- and multi-word expressions that covered the general and medical domains (Table 1).

Name	Description	Terms	Source
Medical Subjects Heading (MeSH®)[1]	Diseases (C-tree), Psychiatry and Psychology (F-tree)	9,128	Bireme; NLM
ClinicalTrials.gov	Biomedical, Diseases, General Terms	1,132	Human translation; NLM
UMLS Specialist[2]	Biomedical, General Terms	10,337	Machine translation; NLM
Kspan (non-verbs)	General Language Terms	10,501	Bonnie Dorr, UMD
Ergane	General Language Terms	4,483	http://www.travlang.com/
IDP*	General Language Terms	6,475	Internet Dictionary Project
CPT5*	Technical Medical Terms	1,280	2003 Physicians' Current Procedural Terminology
Freelang**	General Language Terms	18,307	http://www.freelang.net/

Table 1. Major sources of term-pair entries in the BTL. *Sources in the initial BTL, but removed upon analysis. **Source added to the BTL after analysis.

For the present study, a training set of 954 anonymized queries was collected, excluding misspelled, cognate, and malformed queries: 488 queries were randomly extracted from *ClinicalTrials.gov* log files (21 February 2003) in English, to represent actual consumer needs for clinical research information. The remaining 466 queries were randomly extracted from MedlinePlus en español log files[3] (4 March 2003) in Spanish, to represent actual Spanish speakers' general health information needs. Professional translators created the corresponding training sets of 954 queries in Spanish and English, respectively, although we edited a few mistranslations. Each of the queries in the training set retrieved documents in the English monolingual test bed system.

Using search engine parameters from our previous study, the training set of 954 English queries was submitted to the *ClinicalTrials.gov* monolingual system against the same corpus of 7,170 *ClinicalTrials.gov* English-language records (15 January 2003). The resulting document set served as the standard against which we measured CLIR retrieval effectiveness.

The corresponding 954 Spanish queries were used in both CLIR approaches, BTL and MT, to compare their effectiveness against the English monolingual standard. Prior to BTL look-up and matching, Spanish queries underwent removal of diacritics and conversion to all lower case. Unmatched multi-word expressions were broken into smaller consecutive components, down to individual words. If still unmatched, they were passed through untranslated. Stopwords were left untranslated as well.

To simulate typical users' behavior, we limited CLIR retrieval sets to the 10 top-ranked documents per query. Those queries that failed to retrieve documents in both CLIR approaches were removed to reduce noise. We compared the retrieved document sets from both CLIR methods to the results from the English monolingual standard. We then analyzed the significant "failures" —queries where BTL was much less effective than MT. The analysis led to changes to the BTL method (see Results), which resulted in improved effectiveness when we again ran the same training set queries. A test set of 954 different queries run with the modified BTL produced comparable results. The test set queries were extracted from the same sources with the same distribution as the training set queries (*ClinicalTrials.gov*: 488; MedlinePlus en español: 466).

4 Results

The BTL approach was found to be less effective than the MT approach when compared to the English monolingual system. In a detailed review of a random subset of the 954 queries, comprising 200 queries from MedlinePlus and 200 from *ClinicalTrials.gov*, we focused on those cases where BTL resulted in lower effectiveness than MT (Table 2), as measured by the F factor. Those queries where the BTL approach scored better or as good as MT were not considered. The failure analysis (Table 2) showed that a lack of a normalizing procedure for nouns/adjectives, missing terms, wrong translations in the BTL, and other categories, accounted for significantly lower effectiveness than MT:

- Lexical Variants: Only the canonical form of nouns and adjectives was found. (No stemmer or normalizing procedure was included).
- Missing Terms: Mostly quasi-technical expressions (*anaplastic*; *electroconvulsive*).
- Polysemy: Some Spanish terms have multiple meanings in English, due to synonymy (*alcohol*=alcohol, spirits, liquor), or homonymy (*gota*=arthritis, drip, drop, gout). Multi-listings interfere with the English search engine ability to match synonyms and phrases.
- Wrong Translations: Contextually erroneous (*seno*/ 'sinus, sine', should be 'breast')
- Part of Speech Ambiguity: For example: *crónica*/ 'chronicle' [noun] instead of 'chronic' [adjective]
- Search Procedure: Inability of Spanish search to handle Booleans at present.

Subsequently, we devised and applied a rudimentary normalizing strategy for nouns and adjectives, for singular/plural alternation only. The BTL sources for several mistranslations were identified and extracted from the BTL, and replaced with a different source (Table 1). We heuristically determined that this combination of changes would greatly improve results, as confirmed by a training set run with these changes implemented (Figure 1, Tables 2, 3: Modified BTL Training Set column). The non-parametric Wilcoxon Signed Rank Test showed the differences between F values in the two BTL training set runs (initial vs. modified) to be statistically significant ($p<0.0001$).

Category	N= Training Set Queries		Examples
	Initial BTL (N = 150)	Modified BTL (N = 118)	
Inflectional Variation	57 (38%)	17 (14%)	*drogas*/drugs (plural); *fibrosa*/fibrous (feminine)
Missing Term	40 (27%)	39 (33%)	*transplante*/transplantation; *inocuidad*/safety
Polysemy	23 (15%)	33 (28%)	*gota*/ (arthritis OR drip OR drop OR gout)
Wrong Translation	11 (7%)	7 (6%)	*ojos*/eyeglasses (should be 'eyes')
Part of Speech	3 (2%)	3 (3%)	*crónica* (adj)/chronicle (n) (should be 'chronic')
Search Procedure	16 (11%)	19 (16%)	Booleans: *gleevec, ovarian* vs. *gleevec ovarian*

Table 2. Categorization of problems in queries for BTL. Percentages are in relation to the number of queries identified in the failure analysis.

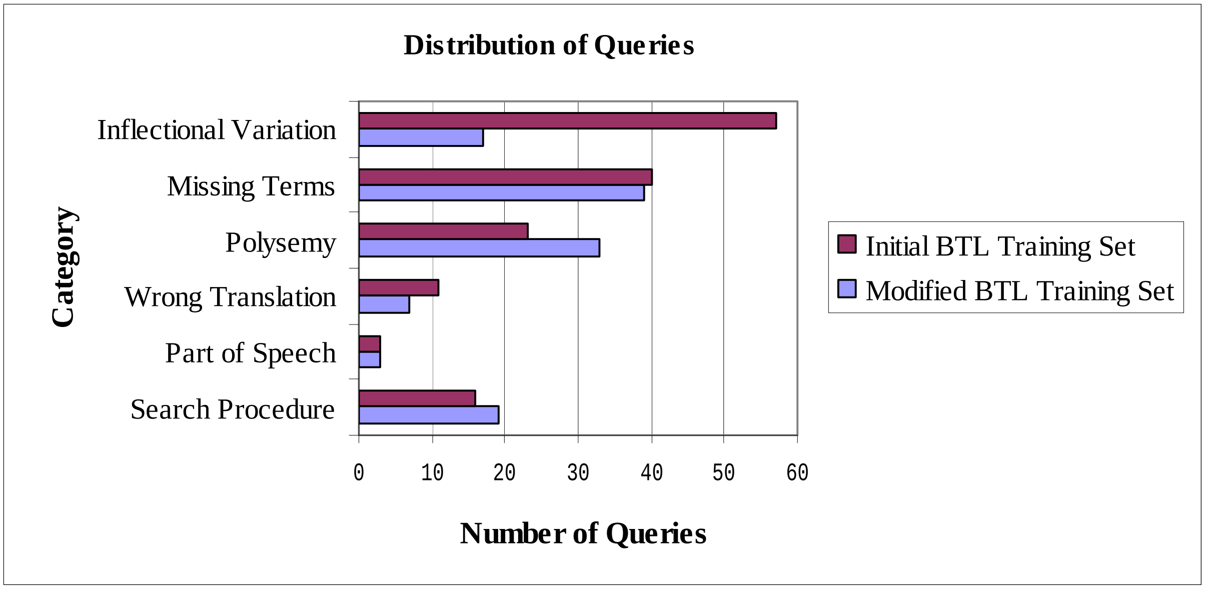

Figure 1. Graphical representation of the distribution of categories in the failure analysis.

A test set of 488 new queries from *ClinicalTrials.gov* (14 January 2003), and another 466 from MedlinePlus en español (March 6, 2003) against the modified BTL validated the changes applied (Table 3, Test Set column):

	***ClinicalTrials.gov* (N=488)**			**MedlinePlus (N=466)**		
Approach	Initial BTL Training Set	Modified BTL Training Set	Test Set	Initial BTL Training Set	Modified BTL Training Set	Test Set
BTL	0.398	0.460	0.481	0.443	0.489	0.487
MT	0.561	0.551	0.585	0.596	0.595	0.588

Table 3. Effectiveness (F Factor) of two CLIR approaches:
Three iterations each of BTL and MT.

In the BTL modified training set, the F factor for MT also changed, as the modifications to the BTL method resulted in different queries retrieving zero documents in both CLIR approaches. Following the procedure described earlier, these queries were discarded for the purposes of calculating retrieval effectiveness.

5 Discussion

As expected, the changes to the BTL improved overall retrieval results and effectively decreased the total number of queries with lower F values than in the MT approach for some categories (Figure 1), such as Inflectional Variation and Wrong Translation. Not surprising, resolving these issues led to an increase in problems further along ("downstream") in the CLIR process: polysemy and search procedure. Thus, inflectional variation heuristics are clearly not enough for improving F values in some individual queries.

The lower effectiveness of the BTL-based query translation CLIR confirms the findings of previous studies, which suggest that, for general language documents, simple automated BTL-based query translation approaches 60% of monolingual retrieval under optimal conditions (Ballesteros & Croft, 1997). The consumer health domain requires both general and technical words and phrases. To improve coverage, a BTL should include both Spanish and English lay terms: common expressions for non-specialists for medical concepts, such as "*pain killer*" for "*analgesic*" in English or "*calmante para el dolor*" for "*analgésico*" in Spanish.

Additional research is needed to explore what size BTL is optimal or sufficient to reach critical mass for the consumer health domain. Demner-Fushman and Oard (2003) reported that, for the print news genre, term lists with at least 30,000 general vocabulary entries in the query language provided the best average mean precision. Larger dictionaries had marginal effects because additional terms rarely appeared in queries. However, in a Finnish-English CLIR study in the medical domain, Pirkola (1998) utilized a general dictionary of over 65,000 terms, and a medical dictionary of 67,000 Finnish-English terms. In our own study, the 38,654 general and technical language entries in the modified BTL do not provide sufficient coverage for even a subset of the consumer health domain, namely clinical research. Quality of the source entries is equally important and should be considered.

We plan to explore other techniques reported in the literature to improve BTL retrieval:

- Local-Feedback Technique: Query expansion with terms extracted from highly relevant documents (pre-, post-query translation, or both) – PubMed "related articles" idea (Adriani & Croft, 1997);
- Local Context Analysis: Query expansion using terms extracted from local and global document context analysis (Xu & Croft, 1996);
- Curating our bilingual term list to remove poor translations and/or questionable entries;
- Implementing a Spanish lexical variant system to handle gender for adjectives; and
- Better handling of stopwords, Boolean operators, and hyphenated terms.

6 Conclusion

This study describes one approach to supporting access to consumer health documents not written in the information seeker's native language. Overall, research on improving CLIR effectiveness to rival monolingual information retrieval systems is needed. In particular, quality and coverage of entries in a BTL for specialized (medicine) and "hybrid" domains (e.g., consumer health) where technical information is intended to be accessible to non-specialists require further investigation. Thus, for consumer health information, the "language barrier" is compounded by the vocabulary problem – explaining complex medical concepts to laypersons.

CLIR is only part of the solution to information access; once users retrieve documents in another language, they need translation to understand their contents. English retrieval for a Spanish query is not enough and more work is needed in this area.

Notes

[1] MeSH® is developed by the NLM (http://www.nlm.nih.gov/mesh/meshhome.html)

[2] SPECIALIST Lexicon is developed by the NLM (http://www.nlm.nih.gov/research/umls)

[3] We acknowledge Paula Kitendaugh for generously providing MedlinePlus en español queries, Olivier Bodenreider for his knowledge on statistical issues, and Dina Demner-Fushman for useful insights and comments.

References

Adriani, Mirna & Croft, W. Bruce. (1997). The effectiveness of a dictionary-based technique for Indonesia-English cross-language text retrieval. CLIR Technical Report IR-170. University of Massachusetts, Amherst.

Ballesteros, Lisa A. & Croft, W. Bruce. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the ACM-SIGIR Conference*. pp. 64-71.

Berland, G.K., Elliott, M.N., Morales, L.S., Algazy, J.I., Kravitz, R.L., et al. (2001). Health information on the Internet: accessibility, quality, and readability in English and Spanish. *Journal of the American Medical Association*. 285(20) May: 2612-2621.

Demner-Fushman, Dina & Oard, Douglas W. (2003). The effect of bilingual term list size on dictionary-based cross-language information retrieval. In *Proceedings of the Annual Hawaii International Conference on Systems Sciences (HICSS-36)*. p. 108b.

Divita, Guy, Browne, Allen C. & Rindflesch Thomas C. (1998) Evaluating lexical variant generation to improve information retrieval. In *Proceedings of the American Medical Informatics Association Annual Symposium*. pp. 775-779.

McCray, Alexa T., Ide, Nicholas C., Loane, Russell F., & Tse, Tony. 2004. Strategies for supporting consumer health information seeking. Medinfo. In press.

NLM. (2003). Unified Medical Language System® (UMLS®). [http://www.nlm.nih.gov/research/umls/] Accessed 12/01/03.

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*.

Ramirez, Roberto R. & de la Cruz, G. Patricia. 2002. The Hispanic population in the United States: March 2002. Current Population Reports. P20-545. U.S. Census Bureau, Washington, D.C. [http://www.census.gov/prod/2003pubs/p20-545.pdf] Accessed 2/03/04.

Rosemblat, Graciela, Gemoets, Darren, Browne, Allen C., & Tse, Tony. (2003). Machine translation-supported cross language information retrieval for a consumer health resource. In *Proceedings of the American Medical Informatics Association Annual Symposium*. Washington, D.C., November 2003. pp. 564-568.

Pirkola, Ari. (1998). The effects of query structure and dictonary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM Sigir Conference on Research and Development in Information Retrieval*. Melbourne, Australia, pp. 55-63.

Xu, Jinxi & Croft, W. Bruce. (1996). Querying expansion using local and global document analysis. In *Proceedings of the International Conference on Research and Development in Information Retrieval*. pp. 4-11.