Graciela Rosemblat, Laurel Graham
National Library of Medicine, NIH, DHHS, Bethesda, MD, USA

Cross-Language Search in a Monolingual Health Information System: Flexible Designs and Lexical Processes

Abstract: The predominance of English-only online health information poses a serious challenge to non-English speakers. To overcome this barrier, we incorporated cross-language information retrieval (CLIR) techniques into a fully functional prototype. It supports Spanish language searches over an English data set using a Spanish-English bilingual term list (BTL). The modular design allows for system and BTL growth and takes advantage of English-system enhancements. Language-based design decisions and implications for integrating non-English components with the existing monolingual architecture are presented. Algorithmic and BTL improvements are used to bring CLIR retrieval scores in line with the monolingual values. After validating these changes, we conducted a failure analysis and error categorization for the worst performing queries. We conclude with a comprehensive discussion and directions for future work.

1 Background and Introduction

Online health information systems predominantly offer only English language support. This language barrier undermines non-English speakers' ability to access information. Cross-Language Information Retrieval (CLIR) techniques are often used to overcome this challenge by supporting searches in the users' native languages. Dynamic databases, such as ClinicalTrials.gov[1], add an extra layer of complexity to cross-language search, with time-sensitive information, such as protocol amendments and registration deadlines, which requires keeping cross-lingual retrieval synchronized with periodic, unanticipated changes.

Previous work (Rosemblat, Tse, & Gemoets, 2004) reported on two query-based approaches to Spanish-English CLIR at ClinicalTrials.gov, a health information system. Retrieval results with machine translation (MT) were compared against those from a then-newly developed Bilingual Term List (BTL). The BTL approach provided a transparent and controllable process in which the translation entries, corresponding to both medical and common vocabulary terms in English and Spanish, were obtained from publicly available sources. After a series of evaluations and subsequent improvements, BTL translation results were brought in line with the MT approach scores, through rudimentary normalization of Spanish-language query terms.

The current paper describes a fully functional prototype that supports Spanish search over the English-language ClinicalTrials.gov data set, and presents a design that is generalizable to other health information systems and languages. We stopped further MT evaluations and concentrated on comparing subsequent CLIR results via the BTL approach against an English monolingual standard. CLIR scores are now close to equivalent monolingual retrievals due to improvements in translation algorithms and the BTL. This paper focuses on 1) the prototype architecture and design, 2) strategies adopted in the BTL to render greatly improved retrieval, and 3) failure analysis from a random query sample, to categorize the worst performing queries.

2 Current Architecture

The project's goal is to provide cross-language search in a cost effective, generalizable method for monolingual systems. From the early design stages of the Spanish prototype,

we planned to use common software, hardware, and backend data (Figure 1) for the English and Spanish search, rather than maintaining completely separate systems. While this design required a greater initial effort, the savings in maintenance time and avoidance of synchronization errors were compelling. The resulting system:

- Shares the web application code, the backend code, and data for both language systems;
- Intermingles data sets: English/Spanish mixed tags in one XML document; and
- Displays Spanish or English data based on run-time language selection.

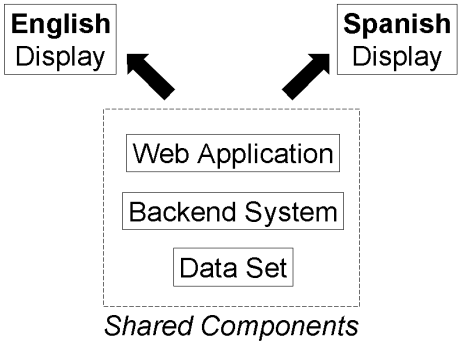

Figure 1: Prototype architecture overview

The efficiency of this mechanism notwithstanding, the recent increase in clinical trial registrations at ClinicalTrials.gov is forcing a re-evaluation of this approach. Large numbers of XML documents with mixed tags can offset the advantages, especially if more languages were to be introduced, increasing the information within each document and resulting in sluggish performance. Decoupling of the language specific data may be considered in the future as a means of optimizing performance while maintaining the advantages of the common backend system and web application.

The current design provides a generic application program interface (API) between the NLM[2]-developed search engine "Essie" (previously "SE") (McCray, Ide, Loane, & Tse, 2004) and the Spanish prototype, unifying the search-and-retrieval process. We developed a quasi-translation module which implements the generic interface and is callable by search engines. The incorporation of this module into the search engine leverages information retrieval enhancements designed for the English monolingual system. These enhancements, now available for Spanish queries, include conceptual mapping, lexical variant generation (Divita, Browne, & Rindflesch, 1998), and a synonymy component via the UMLS® (Lindberg, Humphreys, & McCray, 1993). This allows for better CLIR retrieval scores with an economy of effort, as future improvements need only be made in the English system without a need to duplicate the changes in the Spanish. Thus cross-language search takes advantage of existing monolingual system enhancements.

3 The BTL Look-Up Process

The interchangeability of lexical databases such as glossaries or term lists dovetails well with the modular design of this prototype. While our BTL is focused on clinical trials, a different, new BTL or other lexical source may be used in the future, or combined with the current one, without redesigning the general architecture.

Before the Spanish-language queries are searched against the English-language corpus, they are converted into corresponding English queries in a stepwise process. First, the Spanish search page calls Essie with a flag indicating that the incoming query terms will be in Spanish. This triggers Essie to forward all incoming query terms to the translation module, which looks

up these terms in the BTL and returns the English translations. The term-look-up process consists of several stages: number and gender normalization, stripping of diacritics and conversion of all words to lowercase, and attempting to match the corresponding English-language expressions in the BTL. Multi-word expressions are initially considered as a single unit, then successively decomposed into smaller phrases and, ultimately, individual terms (McCray et al., 2004).

Spanish result lists will often exceed their English counterparts, for a variety of reasons:

- Multiple senses for a single Spanish expression (whether one-word or multi-word) may correspond to different English terms, a phenomenon known as polysemy or one-to-many relationship. For example, both English *dust* and *powder* translate to Spanish *polvo*; English nouns *drop* and *gout* translate to Spanish *gota*;
- Spanish translations in the BTL may include slight word-family or semantic variations from the corresponding English expression, to capture those cases where Spanish uses adjectives but English uses premodifying nouns, as in Spanish *punción pulmonar* (noun + adjective), English *lung puncture* (noun + noun); Spanish *tumor cerebral* (noun + adjective), English *brain tumor* (noun + noun);
- To optimize retrieval, expansion mechanisms (synonymy, lexical variants) apply to all putative translations, including any BTL context-independent alternate English translations for a single Spanish term, as in the polysemy cases outlined above; and
- More fields are searched in the Spanish searches because Spanish tags are searched in addition to English ones in the same XML document.

4 Improvements Implemented: Description and Validation

Our current prototype shows a 23% average increase[3] using the same unedited query sets and corpus (7,170 records) from the earlier project (Rosemblat et al., 2004). These queries came from two sources: English domain-specific queries from ClinicalTrials.gov, and Spanish general health queries from MedlinePlus en español[4]. External translators converted the original English queries into Spanish and the original Spanish ones into English, rendering two parallel query sets, one for each language. Retrieval results from the English set served as the monolingual standard for the CLIR results from the equivalent Spanish query set. CLIR performance was measured by F-values, which combine precision and recall into a single value (Van Rijsbergen, 1979).

Table 1 shows F-values for the test sets using the 10 document cut-off calculation (Rosemblat et al., 2004). The interim step displays improvements to the search and retrieval mechanism within Essie (outside the scope of this project), without improvements in translation algorithms and the BTL. The final step shows the total performance increase with both Essie and BTL improvements factored in.

	ClinicalTrials.gov (N = 488), F Factor			**MedlinePlus (N = 466), F Factor**		
Environment	Initial BTL Training Set	Modified BTL Training Set	Test Set	Initial BTL Training Set	Modified BTL Training Set	Test Set
Baseline[5]	0.398	0.460	0.481	0.443	0.489	0.487
BTL[5], Current Essie	0.534	0.539	0.516	0.546	0.543	0.526
Current BTL, Current Essie	0.68	0.688	0.672	0.606	0.606	0.607

Table 1: Comparing CLIR performance improvements

Lexical CLIR score improvements resulted from the following changes to the BTL:

- gender normalization (building on earlier algorithms for singular/plural variation[5]);
- increased vocabulary coverage, both domain-centered and data-focused;
- addition of stop words and punctuation; and
- removal of excessive context-dependent, semantic variations or "over-extended" English translations for a given Spanish entry.

Alternate translations in the BTL for Spanish *trastorno* offer an example of over-extended translations: depending on word context (Table 2), *disease*, *disorder*, *condition*, and *disturbance* can all translate to *trastorno*:

Spanish	English
trastornos en la marcha	gait *disturbance*
trastorno de salud	adverse health *condition*
trastorno de los nervios periféricos	peripheral nerve *disorder*
trastorno de Tourette	Tourette's *disease*

Table 2: One Spanish source entry [trastorno/s] - many possible English translations

The prototype does not contain context-dependent rules to indicate which translation to select in the vicinity of other terms or collocations. Therefore, for Spanish searches, all the translations for a given Spanish expression are used against the English corpus, much as if they were joined by an OR operator. The original Spanish query term is searched along with the BTL translations. This results in Spanish retrievals often outnumbering those for the English monolingual search (gold standard), as each alternate translation contributes its own set of retrievals. The extra Spanish retrievals may not correspond to the original query searched. Thus, limiting the number of these over-extended translations is a critical part of the on-going BTL clean-up process, especially when including such translations hurts, rather than helps, precision values.

To evaluate how the bigger corpus affects retrieval in the prototype, we tested CLIR performance (Table 3) on the complete set of documents as of August 25, 2005 (15,064 records). The 10-document cut-off calculations were dropped because they require a frozen corpus for measurements to be comparable. Instead, calculations for precision at 10 now show the impact on the user, as this measure is independent from the data set(s) used in the search. Values for the prototype without lexical (translation) improvements are also shown, for comparison. Essie improvements are held constant in all rows.

	ClinicalTrials.gov (N = 483)				**MedlinePlus (N = 460)**			
	F Factor	Precision	Recall	Precision at 10	F Factor	Precision	Recall	Precision at 10
January 2004 BTL, Plural Normalization	0.7	0.838	0.601	0.856	0.683	0.853	0.57	0.87
Current BTL; Plural + Gender Normalization	0.83	0.836	0.823	0.868	0.846	0.843	0.848	0.875

Table 3: Comparing performance improvements on August 2005 clinical trials data set

The F-value increase reflects significant improvements in recall (43%) due to increased BTL coverage. While recall was prioritized over precision, considerable effort was given to ensure precision did not suffer. This increase was validated with a new random sample of 926 queries. For parallelism and consistency with the earlier study, the new sets of queries were extracted from the same sources, and underwent the same processing: external translators rendered two

parallel query sets, one for each language. Retrieval results from the English set served as the monolingual standard for the CLIR results from the equivalent Spanish query set.

	ClinicalTrials.gov (N = 470)				**MedlinePlus en español (N = 456)**			
	F Factor	Precision	Recall	Precision at 10	F Factor	Precision	Recall	Precision at 10
Current BTL, Number + Gender Normalization	0.861	0.888	0.844	0.905	0.879	0.843	0.918	0.861

Table 4: Validation of performance improvements with a new unedited test set

CLIR scores from the new query sets (Table 4) demonstrate that the improved F-values result from the strategies and changes implemented in the BTL, irrespective of the query sets used. Eliminating the 10-document cut-off computation will allow future performance comparisons as the data set grows, without having to recalculate measures at each point[6].

5 Failure Analysis

Failure analysis and subsequent categorization of the 200 worst performing queries[7] uncovered problem areas in the BTL and its interaction with lexical resources used by Essie, namely, the UMLS®. Two categories accounted for 69% of the worst performing queries:

- BTL Coverage: Missing translations or missing entries (46%), including Spanish-English pairs; and
- Semantic Coverage: (23%) Over-extended lexical variations and synonymy. Included are differences between BTL translations (too many or too few) and UMLS® entries and conceptual mappings used by Essie for each of the English translations.

Spanish *mareo* is an example of the latter category, with the following valid BTL translations: *dizziness*, *lightheadedness*, *airsickness*, *carsickness*, and *seasickness*. The UMLS® has no semantic mappings between these terms.

To illustrate how the differences between BTL translations and UMLS® entries affect precision values for the Spanish search, let's take Spanish *síndrome*, which has the following BTL translations: *syndrome*, *disease*, *condition*, and *disorder*. Since the UMLS® does not have a relationship for the last three terms, the English search will only include *syndrome*. For the Spanish search, however, the BTL look-up procedure will collect the four translations for Essie to search against the English corpus, along with the original Spanish query, *síndrome*. Thus more documents (not necessarily search-targeted) will be retrieved in the Spanish search than in the English. Alternatively, this may result in better Spanish returns. For example, until recently the UMLS® did not have *lung* as a synonym for *pulmonary*, while both terms were BTL translations for Spanish *pulmonar*. Since the English retrieval serves as the gold standard, the better Spanish results will be assigned low retrieval scores, as they indicate a mismatch between the English and Spanish results. This represents a weakness in our methodology for performance evaluation and validation.

Other category areas that hurt retrieval (31% combined) were:

- Query Translation used: Lack of context may cause a variance in query interpretation, and there are often several ways to translate a given query. Professional renderings may vary slightly from commonly used translations, resulting in zero matches;
- Search Procedure: Failure caused by bugs or limitations in the search; and
- Language Differences: Polysemy, false cognates or general language differences.

Figure 2 shows the distribution[8] of categories in percentage of total failed queries:

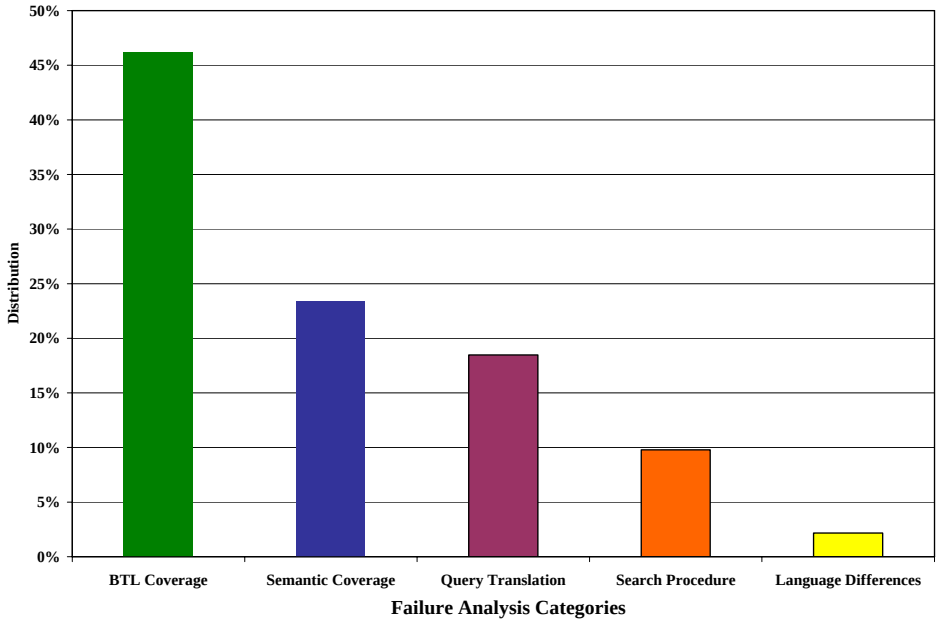

Figure 2. Graphical representation of the failure analysis distribution

Some problems from the initial failure analysis have since been resolved. For example, several queries failed due to punctuation being treated as a search term in the Spanish search. The equivalent punctuation was skipped for English as it was part of the stop word list. This led to the inclusion of punctuation as stop words in the BTL, rendering the Spanish stop word list more equivalent to the English, and unifying search results for the previously failed queries.

6 Discussion

Table 1 shows a 23% improvement in the F-values of the Spanish retrieval results against their English counterparts. This performance increase (from Rosemblat et al., 2004), especially in recall values, resulted entirely from the normalization of Spanish terms and additions and clean up of the BTL. These changes were validated with a new, unedited query set (Table 4). The high F-values obtained in our tests (0.860 – 0.900) attest to the viability of dictionary-based CLIR, despite the known pitfalls of term-lookup lexical-based systems.

Once some simple problems were resolved, the analysis highlighted a fundamental weakness with dictionary-lookup systems not implementing context-based translation rules: translation gaps. BTL gaps accounted for 46% overall in the failure analysis, which pointed to a need for automated entry generation. Roughly 72% of the gaps were caused by English-Spanish pairs missing entirely, as opposed to specific translations for certain Spanish expressions. One potential solution would be to use the lexical resources of the monolingual system to ensure entries in these resources map to translations in the BTL. This would take care of those missing BTL entries that are present in the lexical resources, such as Spanish synsets in the UMLS®, and could be used to increase coverage. In addition, Spanish queries from medical websites, such as MedlinePlus en español may be mined to locate potential lexical candidates for addition in the BTL.

An interrelated issue relates to conflicts between the BTL and existing lexical resources. Resolving this problem would require aligning and/or mapping the existing lexical resources and the term list, which goes against the philosophy of a pluggable, modular design. One mechanism to address this would be using the UMLS® to cover existing gaps. The UMLS® contains multilingual entries that could be used to extend the BTL and align the two resources. But the UMLS® is ever evolving with constant updates and extensions, like any lexical

resource. Thus keeping alignments or mappings in synch between two lexical resources could entail replacing one set of problems with another.

Occasionally, English synonyms in the UMLS® coincide with semantically unrelated Spanish terms, a phenomenon known as false cognate. These synonyms, missing in the BTL, could have a negative impact on precision for a Spanish search. For example, Spanish *herpes* has two BTL translations: *herpes* and *shingles*. The UMLS® offers several semantically related expressions for English *herpes*, such as *zona* among others, a bona-fide English synonym (Dorland's Medical Dictionary, 2000). However, z*ona* is also an unrelated Spanish term, meaning *zone* or *area*. Since English synonymy is included in the Spanish search, the Spanish query *herpes* will retrieve not only pertinent documents on this condition, but also some not pertinent ones on *marginal zone lymphomas* for example, translated in Spanish as *linfomas de la zona marginal*. Alternatively, the English search for *herpes*, which will include a search for English *zona*, will only retrieve pertinent records, as this term has no other English semantic senses and only English fields will be included in the search.

Thus, adding all potential alternate translations in the BTL for a single Spanish expression magnifies retrieval in the Spanish search, hurting precision. Conversely, deleting some of the alternate translations may hurt both precision and recall for Spanish searches, as key documents may be missed altogether. Possible approaches include disambiguating translations based on frequency of usage and commonality of terms, or including context-based rules, or a combination of both. These strategies could be used either during translation or in a post-translation disambiguation module, but will require extending the BTL design to include frequency or context-based information for each translation entry. Further research is required.

7 Future Developments

We have just completed a consumer-centered usability study to assess whether the Spanish prototype provides accessible, readable content that encourages Spanish-speaking users to read clinical trials information, learn about key health opportunities, and make appropriate decisions based on their own situations. The analysis and subsequent categorization of the different types of errors point the way to problems that arise when using a BTL approach to CLIR, and working with ontologies in general.

The next phase of the project requires extensive tools and manual labor to ensure the consistency of the BTL and reduce the number of gaps and over-extended translations. Automated methods for identifying and adding lexical entries for translation will extend the BTL and cover missing entries. Aligning it with other lexical resources should be seen as part of a larger project to curate and validate it. Once curated and validated, we will be able to provide the BTL as a free resource.

This prototype offers immediate extensibility to monolingual health information systems. It can also be applied to creating controlled vocabulary translations of key documents in different languages, and extended to other websites, within and outside the health domain.

Notes

[1] Available: http://www.clinicaltrials.gov/
[2] U.S. National Library of Medicine
[3] Enhancements to Essie, which further increase performance percentages, are not included.
[4] Available: http://medlineplus.gov/spanish/
[5] Rosemblat, Tse, & Gemoets (2004)
[6] The 10-document cut-off required a constant clinical trials corpus for the values to be comparable.
[7] In terms of zero results and low retrieval scores, against the English monolingual standard.

[8] In the two instances in which the queries displayed problems that fell into multiple analysis categories, we coded the queries once for each applicable category.

[9] Acknowledgements: We are largely indebted to Tony Tse for significant feedback and valuable contribution to earlier versions of this manuscript. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine/Lister Hill National Center for Biomedical Communications.

References

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.

Dorland's Illustrated Medical Dictionary. 29th Edition (2000) W.B.Saunders Company.

Lindberg, Donald A., Humphreys, Betsy L., McCray, Alexa T. (1993) The Unified Medical Language System. Methods Inf Med 1993;32:281-91; Unified Medical Language System® (UMLS®) Available: (http://www.nlm.nih.gov/research/umls)

McCray, Alexa T., Ide, Nicholas C., Loane, Russell F., Tse, Tony. (2004) Strategies for supporting consumer health information seeking. Medinfo. 2004:1152-6.

Rosemblat, Graciela, Tse, Tony, Gemoets, Darren. (2004) Adapting a monolingual consumer health system for Spanish cross-language information retrieval. Proceedings of the 8th International ISKO Conference. London, England. pp. 315-321.

Van Rijsbergen, C.J. (1979) *Information Retrieval, 2nd edition*. Department of Computer Science, University of Glasgow.