Julie Motooka
Jonathan Young
LIS 678/CIS 702: Personalized Information Delivery
Special Topics Paper
Dr. Luz M. Quiroga, Fall 2010

# Cross-Language Latent Semantic Indexing and Medical Applications

In cross-language information retrieval, searchers can make queries in their native language and retrieve relevant documents in a foreign language, without having to make any manual query translations (ie. look up in a dictionary). In an effective cross-language text retrieval system, rather than translating large document collections into the query language, documents are retrieved *before* translation, thus saving time (Oard, 1997). In this paper we explore one corpus-based approach—Latent Semantic Indexing (LSI)—to medical cross-language information access. Throughout, we refer to Cross-Language Information Retrieval as CLIR, which should not be confused with other organizations or concepts; this is also the accepted acronym in the field.

Great incentive exists for physicians and biomedical researchers to be able to find case reports and research in various languages. While the vast majority of biomedical research and information is written, distributed, and stored in English, often a comprehensive search for other medical cases similar to the one of interest is desired, and this is only truly possible by searching multilingual document collections.

Medical databases present somewhat unique problems to CLIR. First, medical terminology is a highly technical language that even native speakers will not understand without training. Second, medical databases such as MEDLINE, or clinical datasets, are often of large size and extended scope, making knowledge-based approaches difficult. Finally, the users interested in medical information span a range of knowledge and background, from patients to practitioners, exacerbating the multilingual problem.

Most other information retrieval methods “depend on exact matches between words in users’ queries and words in documents” and “treat words as if they are independent” (Littman, Dumais, & Landauer, 1998). Latent Semantic Indexing differs in that it attempts to generate semantic relationships between terms in the corpus. By using a mathematical transform similar to factor analysis—known as Singular Value Decomposition—a matrix of correlations between words is generated based on their context and coappearance in documents. When certain terms appear together in enough articles, the system will deem them semantically close and cluster them, so that a query on any one term will return relevant documents that, while not an exact match, contain terms semantically related to the query (Yu, Cuadrado, Ceglowski, & Payne, 2002).

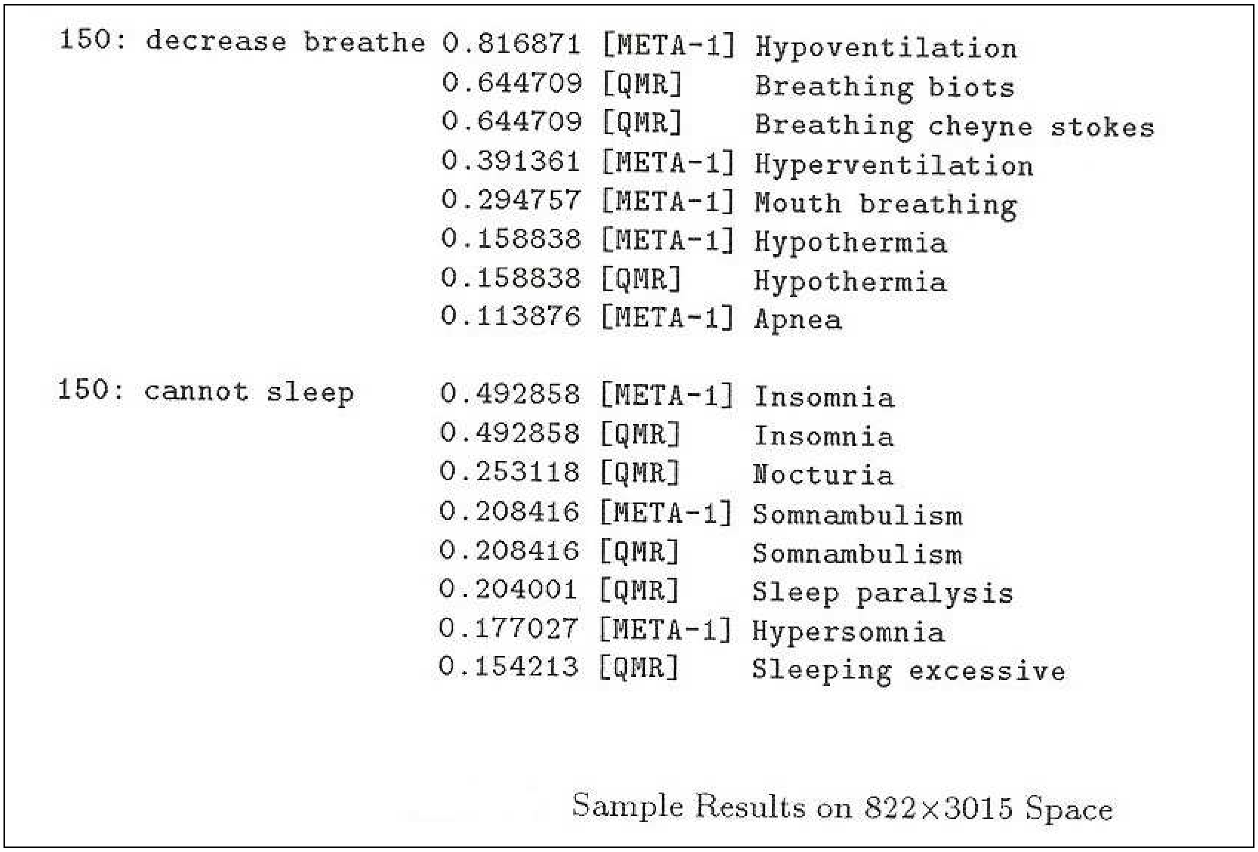
With LSI it is not necessary to use thesauri or dictionaries to determine word associations, since it is all done by the system, through numerical analysis of parallel texts (Dumais, Landauer, & Littman, 1997). First, a sample set of documents are translated in order to “train” the system, and then the rest of the monolingual documents are re-integrated back into this “dual-language semantic space” (Littman, Dumais, & Landauer, 1998).

LSI has the great advantage for medical databases of being able to do both cross language and subject language correlations. All that is necessary is a large enough corpus that contains related terms in context. This can be provided by translating a sample set of documents, as mentioned above. Evans (1998) demonstrated the power of LSI for medical information retrieval by mapping a 3,000-term space in Spanish, English, and medical terminology.

Because the knowledge needed to list synonyms is small, Evans was able to complete this list “by one person in approximately three hours of work.” This example demonstrates the potential for LSI to do automated CLIR. As shown in Figure 1 and Figure 2, the LSI method gives high correlations for both standard English to medical terms, and Spanish queries to medical terms.

In conclusion, the potential benefits of CLIR for medical applications are great, both for patients and medical practitioners. LSI offers one powerful, mostly automated method to accomplish CLIR in the medical domain. Future work needs to expand on Evans' pilot studies and demonstrate that LSI is practical on a full-sized medical database such as MEDLINE.

## Figures



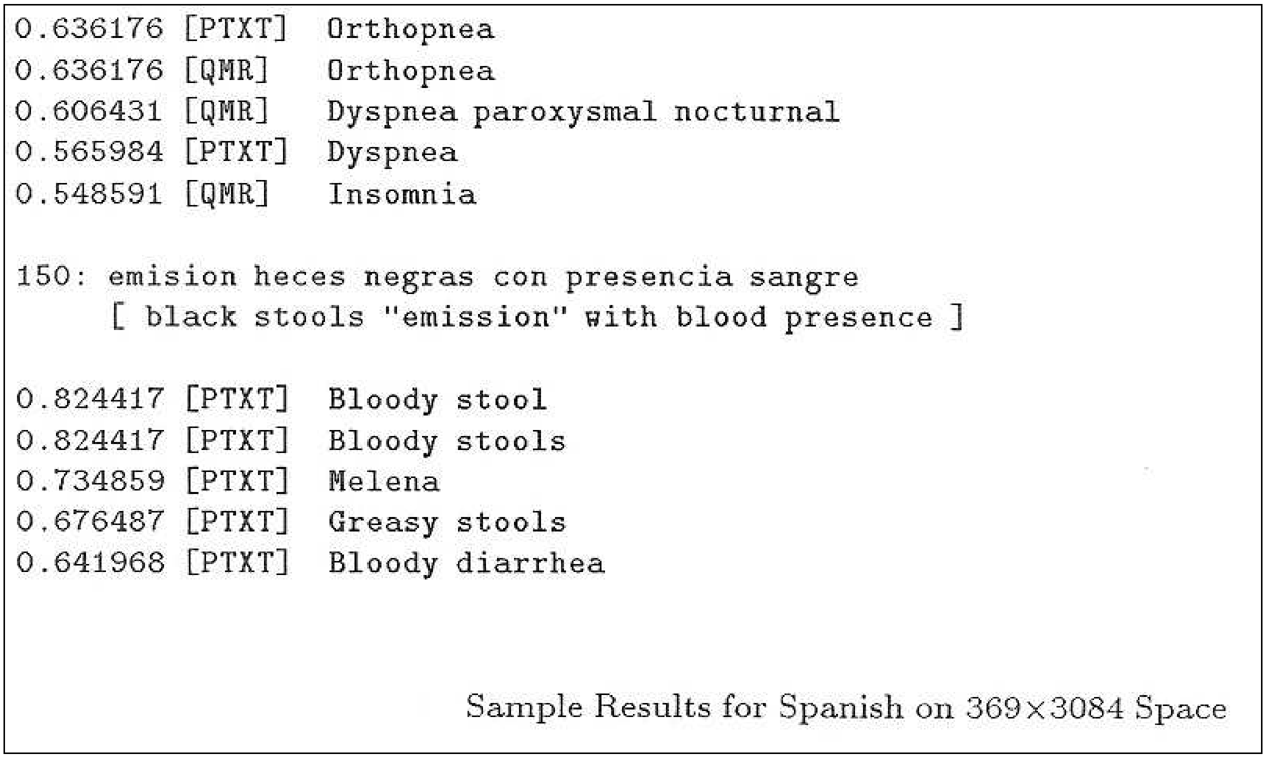
```
150: decrease breathe 0.816871 [META-1] Hypoventilation
                      0.644709 [QMR]    Breathing biots
                      0.644709 [QMR]    Breathing cheyne stokes
                      0.391361 [META-1] Hyperventilation
                      0.294757 [META-1] Mouth breathing
                      0.158838 [META-1] Hypothermia
                      0.158838 [QMR]    Hypothermia
                      0.113876 [META-1] Apnea

150: cannot sleep     0.492858 [META-1] Insomnia
                      0.492858 [QMR]    Insomnia
                      0.253118 [QMR]    Nocturia
                      0.208416 [META-1] Somnambulism
                      0.208416 [QMR]    Somnambulism
                      0.204001 [QMR]    Sleep paralysis
                      0.177027 [META-1] Hypersomnia
                      0.154213 [QMR]    Sleeping excessive
```

Sample Results on 822×3015 Space

**Figure 1: Correlations for Standard English to Medical Terms (Source: Evans, Handerson, Monarch, Pereiro, Delon, & Hersh, 1998).**

**Shown are the terms retrieved by LSI when using standard English queries shown at left. Center column are correlation strengths, brackets indicate the vocabulary the term is drawn from.**



```
0.636176 [PTXT]  Orthopnea
0.636176 [QMR]   Orthopnea
0.606431 [QMR]   Dyspnea paroxysmal nocturnal
0.565984 [PTXT]  Dyspnea
0.548591 [QMR]   Insomnia

150: emision heces negras con presencia sangre
     [ black stools "emission" with blood presence ]

0.824417 [PTXT]  Bloody stool
0.824417 [PTXT]  Bloody stools
0.734859 [PTXT]  Melena
0.676487 [PTXT]  Greasy stools
0.641968 [PTXT]  Bloody diarrhea
```

Sample Results for Spanish on 369×3084 Space

**Figure 2: Correlations for Spanish Queries to Medical Terms (Source: Evans, Handerson, Monarch, Pereiro, Delon, & Hersh, 1998).**

**Shown are the terms retrieved by LSI when using Spanish language queries shown at top. Left column are correlation strengths, brackets indicate the vocabulary the term is drawn from.**

## Sources Cited

Dumais, S. T., Landauer, T. K., & Littman, M. L. (1997). Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing. *AAAI Technical Report*, 15-21.

Evans, D. A., Handerson, S. K., Monarch, I. A., Pereiro, J., Delon, L., & Hersh, W. R. (1998). Mapping Vocabularies Using Latent Semantics. In G. Grefenstette, *Cross-Language Information Retrieval* (pp. 63-80). Boston: Kluwer Academic Publishers.

Littman, M. L., Dumais, S. T., & Landauer, T. K. (1998). Automatic Cross-Language Information Retrieval Using Latent Semantic Indexing. In G. Grefenstette (Ed.), *Cross-Language Information Retrieval* (pp. 51-62). Boston: Kluwer Academic Publishers.

Oard, D.W. (1997). Alternative Approaches for Cross-Language Text Retrieval. *AAAI Technical Report*, 154-162.

Yu, C., & Cuadrado, J., Ceglowski M., & Payne J.S. (2002). *Patterns in Unstructured Data: Discovery, Aggregation, and Visualization.* Retrieved Oct. 22, 2010, from National Institute for Technology and Liberal Education, Georgetown, Texas. Web site: http://knowledgesearch.org/lsi/cover_page.htm.).

## Keywords

cross-language information retrieval, latent semantic indexing, medical terms