Multilingual Noun Phrase Extraction Using a Part-of-Speech Tagger

John O. Woods[1], Ryan Richardson[2]
Edward A. Fox[3], Department of Computer Science, Virginia Tech

Accurate noun phrase extraction for large texts in multiple languages is useful for building indexes for corpora, particularly for phrase-based searching. Extraction of noun phrases is also useful for translation of concept maps (Richardson & Fox, 2005).

A method for the extraction of commonly appearing phrases using heuristics and statistical analysis was developed within the last ten years (Evans & Zhai, 1996). More recently, a system was developed which was capable of identifying noun phrases in captions using comparable Spanish and English corpora (López-Ostenero et al., 2004). Part-of-speech tagging has been used since the 1960s for English, and with the development of the Viterbi algorithm for solving hidden Markov models (HMM) in the 1980s, English-language taggers have reached nearly 98% accuracy.

Our work implements a more robust and accurate version of López-Ostenero's idea, designed for use on large texts instead of captions, based around Dekang Lin's HMM part of speech tagger and Evans' method for identifying significant phrases. We developed a tag set for Spanish and re-trained Lin's tagger on various Spanish-language *Wikipedia* and *Wikinews* articles, resulting in a bilingual noun phrase extractor. Furthermore, our system can be just as easily expanded to work for nearly any language that favors word order over inflection. We intend to compare the results of our extractor with a baseline of manual noun phrase identification.

REFERENCES

1. Evans, D. A. & Zhai, C. (1996). Noun-Phrase Analysis in Unrestricted Text for Information Retrieval. In *Proceedings of ACL 1996*.
2. Richardson, R. & Fox, E. A. (2005). Evaluating concept maps as a cross-language knowledge discovery tool for NDLTD. In *Proceedings ETD2005*, 8th International Symposium on Electronic Theses and Dissertations, Sydney, Sept. 28-30, 2005, to be given.
3. Lin, D. (2003). A C++ Implementation of a Hidden Markov Model. [Cited 9 September 2005]. Available from World Wide Web: (http://www.cs.ualberta.ca/ ~lindek/hmm.htm).
4. López-Ostenero, F., Gonzalo, J., & Verdejo, F. (2004). Noun Phrases as building blocks for Cross-Language Search Assistance. *Information Process and Management, 41*, 549-568.

[1] Undergraduate in Computer Science, Virginia Tech, presenter, involved in undergraduate research and an Honors thesis related to this Abstract

[2] Ph.D. Candidate, Virginia Tech, applying results of this to dissertation studies

[3] Professor of Computer Science, Virginia Tech, advisor for Woods and Richardson