Building a Medical Spanish Lexicon

Guy Divita, MS, Graciela Rosembalt, PhD, Allen C. Browne, MS
National Library of Medicine, Bethesda, Maryland

Abstract

We are developing a freely available Spanish medical syntactic lexicon, initially populated with medical terms from a bilingual list, and then from corpus based term discovery. The lexical records are a simplification of the SPECIALIST English lexicon. Lexical variant generation and normalization tools will be provided along with the lexicon.

Background

A freely available Spanish medical syntactic lexicon is an important underpinning to our effort toward Spanish lexical tools to deal with lexical variation in Spanish. The Specialist Spanish lexicon will facilitate indexing and retrieval of Spanish medical terms in the UMLS. Although commercial Spanish syntactic lexicons exist, to date, none are known to cover the biomedical domain, nor have been found to be freely distributable. The lexicon is to support a variety of lexical resources including a text normalization tool, a text inflectional variant generation tool, a part-of-speech (POS) tagger. When ready, these resources will be distributed from SPECIALIST.nlm.nih.gov. There is a rich tradition in corpus based lexicon development. The techniques employed here are most similar to what is being done to create the Morph-It! Italian syntactic lexicon.

Structure of the Spanish Lexicon. Attributes

As with the English Lexicon, lexemes may be multi-word, for example, vitamin A. The lexical record for a lexeme includes a set of inflectional variants with appropriate attributes. Gender is included, as well as spelling variants, acronyms and their expansions. As nouns agree with their modifiers (determiners and adjectives, if present), the gender attribute for each noun in the corpus can be inferred and added to the lexical record from the determiner preceding it.

Methodology

This lexicon is being built incrementally, starting with the adaptation of the dTagger POS tagger. The tagger's lexicon is the basis for lexical record lemma discovery as well as for POS assignment. Subsequent term discovery continues first by augmenting the tagger lexicon, then reviewing the tagger lexicon additions to be folded into the record based lexicon. Periodically, the record based lexicon is used to replace the core tagger lexicon.

Initial Tagger Lexicon

The initial lexicon was created from 10,000 inflected terms drawn from a Spanish/English bilingual term list developed in house, while working on a Spanish prototype of ClinicalTrials.gov. POS's to Spanish forms were assigned based on the POS of their English equivalent as assigned by an English POS tagger. The results were corrected and became the core tagger lexicon. The Spanish tagger was induced with a very small set of manually annotated medical texts and a much larger set of annotated medical texts taken from Spanish medical websites. Additional medical texts were tagged, manually corrected, and used to retrain the tagger. Terms from the tagged training set became candidates for inclusion back into the tagger’s lexicon. This process is iterative.

Lexemes from the Tagger Lexicon

An initial set of lexemes can be induced by clustering terms that differ by regular suffix endings from the same POS. This is done via a trie algorithm that annotates the category, inflection labels, a canonical form indicator, and lexeme identifiers to each inflected form upon the final pass through the trie structure.

Variant Generation and Normalization

An effort was made several years ago to generalize to Spanish from the English lexical variant generation (LVG) programs. This included both an inflection and an un-inflection tool. These tools, together with the lexicon, and the methods borrowed from the English LVG will be the basis for a Spanish normalization tool.

References

1. UMLS Knowledge Sources. Documentation. NLM