Finding Instance Names and Alternative Glosses on the Web: WordNet Reloaded

Marius Paşca

Google Inc.
1600 Amphitheatre Parkway
Mountain View, California 94043
mars@google.com

Abstract. This paper presents an approach to extending existing lexical resources with instance names and alternative definitions acquired from textual documents. The experiments involve WordNet and approximately 300 million Web documents, but the method is more generally applicable. We leverage formally-structured, human-validated resources, on one hand, and data-driven instance names and definitions on the other, which opens the path to new applications of the reloaded resources.

1 Motivation and Goals

Large-scale lexical, hierarchical resources have a broad range of applications in computational linguistics, information extraction and information retrieval. When manually building such resources, the focus is justifiably on selecting and organizing words into hierarchies of conceptual entries, with manual selection of an ideal, single definition for each entry. For example, by grouping together English words with the same meaning (e.g., *lawyer* and *attorney*) into sets of synonyms (or *synsets*, such as {*lawyer, attorney*}) associated with a single definition (or *gloss*), WordNet [1] became a de-facto standard for lexical resources. Its uses span word sense disambiguation [2], information extraction [3] and machine translation [4], to name only a few.

Hierarchical resources organize noun synsets along *IsA/InstanceOf* relations. The conceptual coverage of WordNet is impressive, with more than 150,000 English words encoded in over 115,000 synset entries or lexical concepts - more than half of which are nouns. However, WordNet and other resources are not necessarily complete for obvious practical reasons. This particularly applies to the lower-level hierarchies, where the more specific concepts occur, in the form of both missing specialized concepts and missing instance names. WordNet does not contain *telecom company* or *meta search engine* under *company* and *search engine* respectively; similarly, there are no instance names such as *Google* under *search engine*, or *Ferrari* under *car company*. Only a fraction of the encoded concepts are accompanied by corresponding instances; the number of such instances embedded under a given concept is usually small. For instance, 600 instance names exist under *city*; comparatively, there are eight instance names