Making Senses:
Bootstrapping Sense-tagged Lists of
Semantically-Related Words

Nancy Ide

Department of Computer Science, Vassar College, Poughkeepsie, New York USA
`ide@cs.vassar.edu`

Abstract. The work described in this paper was originally motivated by the need to map verbs associated with FrameNet 2.0 frames to appropriate Word-Net 2.0 senses. As the work evolved, it became apparent that the developed method was applicable for a number of other tasks, including assignment of WordNet senses to word lists used in attitude and opinion analysis, and collapsing WordNet senses into coarser-grained groupings. We describe the method for mapping FrameNet lexical units to WordNet senses and demonstrate its applicability to these additional tasks. We conclude with a general discussion of the viability of using this method with automatically sense-tagged data.

1 Introduction

Lists of semantically-related words and phrases are heavily used in many automatic language processing tasks. A common use of such lists in recent work is in attitude or opinion analysis, where words indicative of a given semantic orientation—often, "positive" or negative" polarity—are detected to classify documents such as movie and product reviews as more or less favorable ([1], [2], [3]). Approaches include simple term counting [4] as well as training machine learning algorithms to classify documents. In machine learning approaches, semantically-related words and phrases are often used as a part of the feature set (e.g., [2], [3], [5]. NLP tasks such as event recognition also typically rely on lists of semantically-related verbs coupled with frames or patterns that are used to identify participants, etc. (e.g., [6] [7]).

Largely due to the recent upsurge in work on attitude and opinion analysis, numerous lists of semantically-related words have been made available within the language processing community. The lists are compiled using a variety of means, including extraction from existing resources such as lexicons, thesauri, and pre-compiled content category lists such as the General Inquirer [8]; automated extraction [2] [3]; and manual production; and often include hundreds or even thousands of words.

Whatever the source, available lists of semantically-related words do not identify the sense of the included items, despite the fact that many of the words are highly