Verb Class Discovery from Rich Syntactic Data

Lin Sun[1], Anna Korhonen[1] and Yuval Krymolowski[2]

[1] Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
alk23@cam.ac.uk, ls418@cam.ac.uk
[2] Department of Computer Science, University of Haifa
31905, Haifa, Israel
yuvalkry@gmail.com

Abstract. Previous research has shown that syntactic features are the most informative features in automatic verb classification. We investigate their optimal characteristics by comparing a range of feature sets extracted from data where the proportion of verbal arguments and adjuncts is controlled. The data are obtained from different versions of VALEX [1] – a large SCF lexicon for English which was acquired automatically from several corpora and the Web. We evaluate the feature sets thoroughly using four supervised classifiers and one unsupervised method. The best performing feature set includes rich syntactic information about both arguments and adjuncts of verbs. When combined with our best performing classifier (a novel Gaussian classifier), it yields the promising accuracy of 64.2% in classifying 204 verbs to 17 Levin (1993) classes. We discuss the impact of our results on the state-or-art and propose avenues for future work.

1 Introduction

Recent research shows that it is possible, using current natural language processing (NLP) and machine learning technology, to automatically induce lexical classes from corpus data with promising accuracy [2–5]. This research is interesting, since lexical classifications, when tailored to the application and domain in question, can provide an effective means to deal with a number of important NLP tasks (e.g. parsing, word sense disambiguation, semantic role labeling), as well as enhance performance in applications, (e.g. information extraction, question-answering, machine translation) [6–10].

Lexical classes are useful for NLP because they capture generalizations over a range of (cross-)linguistic properties. Being defined in terms of similar meaning components and (morpho-)syntactic behaviour of words [11, 12] they generally incorporate a wider range of properties than e.g. classes defined solely on semantic grounds [13]. For example, verbs which share the meaning component of 'manner of motion' (such as *travel*, *run*, *walk*), behave similarly also in terms of subcategorization (*I traveled/ran/walked to London*) and usually have zero-related nominals (*a run*, *a walk*).

NLP systems can benefit from lexical classes in many ways. For example, such classes can be used i) to define a mapping from surface realization of arguments to predicate-argument structure, ii) as a means to abstract away from individual words when required, or (iii) to build a lexical organization which predicts much of the syntax and semantics of a new word by associating it with an appropriate class.