Transformation-Based Information Extraction Using Learned Meta-Rules

Un Yong Nahm

Ask Jeeves, Inc.
1551 South Washington Avenue, Suite 400
Piscataway, NJ 08854
pebronia@acm.org

Abstract. *Information extraction* (IE) is a form of shallow text understanding that locates specific pieces of data in natural language documents. Although automated IE systems began to be developed using machine learning techniques recently, the performances of those IE systems still need to be improved. This paper describes an information extraction system based on transformation-based learning, which uses learned meta-rules on patterns for slots. We plan to empirically show these techniques improve the performance of the underlying information extraction system by running experiments on a corpus of IT resumé documents collected from Internet newsgroups.

1 Introduction

The goal of an information extraction system is to fill out the pre-determined template by finding relevant data in natural-language text [1]. In order to reduce human efforts to build an information extraction systems, automatic construction of complex IE systems began to be considered lately. One of the typical problems often found in existing information extraction systems is that the *recall* (percentage of correct slot fillers extracted) of an IE system is significantly lower than its *precision* (percentage of extracted slot fillers which are correct) [2]. In this paper, we present a method to boost the performance of given IE systems, especially recalls, by learning meta-rules by finding relationships between slots and applying these meta-rules on weakly-labeled data repeatedly. For example, we found that a pattern "`<degree>` in `<major>`" appears frequently in resumé postings on USENET newsgroups, e.g. "B.S. in Mathematics". This rule can be applied to partially-labeled data, such as "M.S. in `<major>`" or "`<degree>` in Mechanical Engineering", to extract additional fillers, e.g. `M.S.` for degree or `Mechanical Engineering` for major.

Since meta-rules only require target texts to be tagged and do not assume anything about the tagger, it becomes clear that meta-rules are not restricted to the initial weakly-labeled data tagged by the underlying IE system, but recursively to the data already labeled by meta-rules themselves. This notion of recursiveness resembles that of transformation-based learning which were found to be successful in some natural language processing tasks such as part-of-speech (POS) tagging [3]. The main advantage of our approach is that it is completely independent of the underlying information extraction systems and is therefore very flexible.