Mining the Semantics of Text Via Counter-Training

Roman Yangarber

Department of Computer Science, University of Helsinki, Finland

Abstract. We report on a set of experiments in text mining, specifically, finding semantic patterns given only a few keywords. The experiments employ the Counter-training framework for discovery of semantic knowledge from raw text in a weakly supervised fashion. The experiments indicate that the framework is suitable for efficient acquisition of semantic word classes and collocation patterns, which may be used for Information Extraction.

1 Introduction

In this paper we describe experiments with automatic, minimally supervised acquisition of semantic knowledge for Information Extraction (IE). In particular, our experiments employ the Counter-training framework for automatic acquisition of semantic patterns, and introduce variations on the framework in an attempt to broaden its applicability.

Counter-training has been successfully deployed previously for acquisition of semantic resources. For example, [1] shows how it can be used to acquire classes of related terms in various domains, starting with a very small set of "seed" terms. [2] presents acquisition of semantic collocation patterns, which are then used for building IE systems, again starting with a small seed. Semantic term classes and patterns are essential components in a (pattern-based) IE system, and both types of resource are notoriously difficult to develop manually. As a parallel development, a variety of supervised methods have been investigated for acquisition of semantic resources. Supervised training requires an investment in the manual labor of tagging training corpora, which can be substantial. Our goal is to reduce the amount of supervision necessary for building semantic resources, even if we trade off some of the precision of the resulting resources for a reduction in the amount of supervision required. The rationale is twofold: first, it has been shown (e.g., in [3]) that the time needed to verify the resulting resources is usually shorter than the time necessary for tagging corpora for supervised training. Second, an unsupervised, as opposed to supervised, method will allow us to leverage a much larger training corpus to improve recall.

In our experiments we used a general corpus of news articles, and smaller subsets of documents extracted from the general corpus by means of a simple keyword-based IR query. We have then processed our corpus with tools designed to acquire word co-occurrence patterns and word classes from text. These resources are then used for proposing candidate IE templates, along with example members of the word classes for filling slots in these templates. The patterns describe of the major events/facts, and were found to cover (i.e., match) a substantial portion (50–60%) of the documents retrieved by the queries—so that they may be considered representative of the retrieved collections.

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 647–657, 2005.

The following sections describe the data used in the experiments (2), the methods (3), and the results (4).

2 Data

The base corpus consists of a collection of about 35,000 English-language articles, from the Associated Press (AP), from 1989 and 2000.

In each experiment, a keyword-based query was issued against this corpus, and the top 1000 documents returned were used as a "reference" corpus.[1] In preparation for the experiments the corpus was further pre-processed, as described below. The goal of these experiments was to apply unsupervised learning to cover a representative portion of the reference corpus. We describe an experiment in which a simple query consisted of the names of four African countries: Nigeria, Kenya, Ethiopia and Uganda.

The key idea we are exploring is summarized as follows: ideally, in IR, one hopes that the sub-collection retrieved in response to a query somehow constitute a unified "semantic whole." It should be possible to demonstrate this, by capturing the semantics in concise form, namely, as a set of patterns which describe the semantics (and further, which enable us to build an IE system to extract events from text). To use the weakly supervised pattern discovery methods mentioned above, one must provide seed patterns; however, to provide the seed patterns one must have some prior knowledge of the domain(s) in the retrieved sub-collection(s). We'd like to help obtain this knowledge automatically.

3 Methods

Several methods have been proposed for learning patterns relevant to a specific scenario.[2] This includes supervised learning algorithms (e.g., [4,5]), and weakly supervised algorithms ([6,3]), which start from a small set of seed patterns, and iteratively bootstrap a larger set of relevant patterns.

The intent of the present work is to investigate what happens when *no* specific scenario is given *a priori*. Rather the input to the learning algorithm is the retrieved reference corpus—a set of documents, which likely cover multiple, different topics. We would like to induce the topics themselves automatically, without reading through any portion of the 1000-document corpus, which may be time-consuming, and may not yield a clear result.

We planned to experiment with the ExDisco algorithm, [6], in this setting, but found that it is not well suited for two reasons. First, the target scenarios are not known in advance. Second, even when a scenario is given manually, ExDisco tends to overgeneralize after a few dozen initial iterations: the learner begins to acquire patterns that are too general (with respect to the target scenario). At that point, ExDisco also acquires documents which do not belong to the scenario and the entire learning process diverges.

[1] In section 4, we discuss some of the problems encountered during data preparation, which reduced the number of usable documents below 1000.

[2] The term *scenario* is used in its traditional IE sense: the topic of interest in which events are to be extracted, e.g., "company mergers and acquisitions," or "terrorist attacks."