Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Information Processing and Management 41 (2005) 217–242

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Information extraction with automatic knowledge expansion

Hanmin Jung [*], Eunji Yi, Dongseok Kim, Gary Geunbae Lee

Department of Computer Science and Engineering, Pohang University of Science and Technology,
San 31, Hyoja-dong, Nam-gu, Pohang, Kyungbuk 790-784, South Korea

Received 28 February 2003; accepted 15 July 2003
Available online 6 September 2003

Abstract

POSIE (**POS**TECH **I**nformation **E**xtraction System) is an information extraction system which uses multiple learning strategies, i.e., SmL, user-oriented learning, and separate-context learning, in a question answering framework. POSIE replaces laborious annotation with automatic instance extraction by the SmL from structured Web documents, and places the user at the end of the user-oriented learning cycle. Information extraction as question answering simplifies the extraction procedures for a set of slots. We introduce the techniques verified on the question answering framework, such as domain knowledge and instance rules, into an information extraction problem. To incrementally improve extraction performance, a sequence of the user-oriented learning and the separate-context learning produces context rules and generalizes them in both the learning and extraction phases. Experiments on the "continuing education" domain initially show that the $F1$-measure becomes 0.477 and recall 0.748 with no user training. However, as the size of the training documents grows, the $F1$-measure reaches beyond 0.75 with recall 0.772. We also obtain F-measure of about 0.9 for five out of seven slots on "job offering" domain.

Keywords: Information extraction; Question answering; User-oriented learning; Lexico-semantic pattern; Machine learning

1. Introduction

Information extraction is a process that takes unseen documents as input and produces a tabular structure as output. As Internet growth accelerates, information extraction is attracting

[*] Corresponding author. Tel.: +82-54-279-5581; fax: +82-54-279-2299.
E-mail addresses: jhm@postech.ac.kr (H. Jung), juicy@postech.ac.kr (E. Yi), dskim@postech.ac.kr (D. Kim), gblee@postech.ac.kr (G.G. Lee).

0306-4573/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0306-4573(03)00066-9

considerable attention from the Web intelligence community. Traditional information extraction tasks involve locating specific information from a plain text written in a natural language. Thus, the task biases an information extraction only as one area of natural language processing. Information extraction, as a fundamental front-end technique for knowledge discovery, data mining, and natural language interface to databases, on the Web has transformed a major Web application technology (Jung, Lee, Choi, Min, & Seo, 2003; Nahm, 2001; Nahm & Mooney, 2000).

One crucial challenge in information extraction as a Web application technology is to acquire domain portability. Since most previous systems require human annotated data to learn extraction rules or patterns, domain experts manually annotate the training data. Worse, when a new domain is added, a considerable portion of the time to graft for the domain is poured into laborious annotation. To circumvent this problem, recent research develops weakly supervised and unsupervised learning algorithms. However, the new techniques do not yet satisfy the back-end applications (Eikvil, 1999; Zechner, 1997).

Domain portability is greatly affected by the manner in which Web document types are used and by which point in time domain experts or users are involved. Thus, we propose two strategies: first, replacing laborious annotation with automatic knowledge extraction from structured Web documents, [1] and second, placing users at the end of a learning cycle in a deployment phase. To incrementally improve extraction performance, POSIE combines user-oriented learning to produce context rules with separate-context learning to generalize the rules.

The remainder of the paper is organized as follows. Section 2 reviews important related research on information extraction. Section 3 proposes our information extraction model based on a question answering framework. The detail architecture and the knowledge of the POSIE [2] are respectively described in Sections 4 and 5. Section 6 explains the techniques to expand this knowledge through user-oriented and separate-context machine learning. Section 7 analyzes experimental results for the practical "continuing education" domain. To conclude this paper, Section 8 discusses the functional characteristics of information extraction systems and future works.

2. Related research

Information extraction (IE) systems using an automatic training approach (Grishman, 1997; Sasaki, 1999; Yangarber & Grishman, 1998) have a common goal: to formulate effective rules to recognize relevant information. They achieve this goal in the manner of annotating training data and running a learning algorithm (Knoblock, Lerman, Minton, & Muslea, 2000; Riloff, 1996; Riloff & Jones, 1999; Sudo, Sekine, & Grishman, 2001).

Recent IE research concentrates on the development of trainable information extraction systems for the following reasons. First, annotating texts is simpler and faster than writing rules by hand. The rapid growth of the Web contents increases the need for a series of automatic processing steps. Second, automatic training ensures domain portability and full coverage of ex-

[1] Documents containing attributes which can be correctly extracted based on some uniform syntactic clues, for example, tables in the form of separated attributes and their contents.

[2] POSIE (**POS**TECH **I**nformation **E**xtraction System).