# Information Extraction from Epidemiological Reports

**Roman Yangarber** **Lauri Jokipii**
Department of Computer Science
University of Helsinki, Finland
first.last@cs.helsinki.fi

**Antti Rauramo**
Index, Oy
Helsinki, Finland

**Silja Huttunen**
Department of Linguistics
University of Helsinki, Finland

## Abstract

This work demonstrates the ProMED-PLUS Epidemiological Fact Base. The facts are automatically extracted from plain-text reports about outbreaks of infectious epidemics around the world. The system collects new reports, extracts new facts, and updates the database, in real time. The extracted database is available on-line through a Web server.

## 1 Introduction

Information Extraction (IE) is a technology for finding facts in plain text, and coding them in a logical representation, such as a relational database.

Much published work on IE reports on "closed" experiments; systems are built and evaluations are conducted based on carefully annotated training and test corpora, at most a few hundred documents.[1]

The main goal of the work presented here is to explore the IE process *in the large*: the system exhibits the integration of a number of off-line and on-line components around the IE engine proper. Further, the system is applied to a large dynamic collection of documents in the epidemiological domain, currently containing tens of thousands of documents.

The topic is outbreaks of epidemics of infectious diseases, affecting humans, animals and plants.

To the best of our knowledge, this is the first large-scale IE database, especially in the epidemiological domain,[2] publicly accessible on-line.

[1] Cf., e.g., the MUC and ACE IE evaluation programmes.

[2] On-line IE databases do exist, e.g., CiteSeer, but none that extract multi-argument events from plain natural-language text.

## 2 System Description

The architecture of the ProMED-PLUS system[3] is shown in figure 1. The core IE engine (center of the figure) is viewed and implemented as a sequence, or a "pipeline", of stages of processing:

- Layout analysis, tokenisation, lexical analysis;
- Name recognition and classification;
- Shallow syntactic parsing;
- Resolution of co-reference among entities;
- Pattern-based event matching and role mapping;
- Normalisation and output generation

The database (DB) contains facts extracted from ProMED-Mail, a mailing list about epidemic outbreaks. ProMED[4] is one of the most comprehensive sources of reports about the spread of infectious diseases around the world, collected for over 10 years.

The core IE engine is based on earlier work, described in (Grishman et al., 2003; Grishman et al., 2002). Novel components use machine learning at several stages to enhance the performance of the system and the quality of the extracted data: acquisition of domain knowledge for populating the knowledge bases (left side in the architecture diagram), and automatic post-validation of extracted facts for detecting and reducing errors (upper right). Novel features include the notion of confidence,[5] and aggregation of separate facts into outbreaks, across multiple reports, based on confidence, (submitted for review).

[3] **PLUS**: Pattern-based Learning and Understanding System.

[4] www.promed.org is the Program for Monitoring Emerging Diseases, of the International Society for Infectious Diseases.

[5] Confidence for individual fields of extracted facts, and for entire facts, is based on document-local and global information.

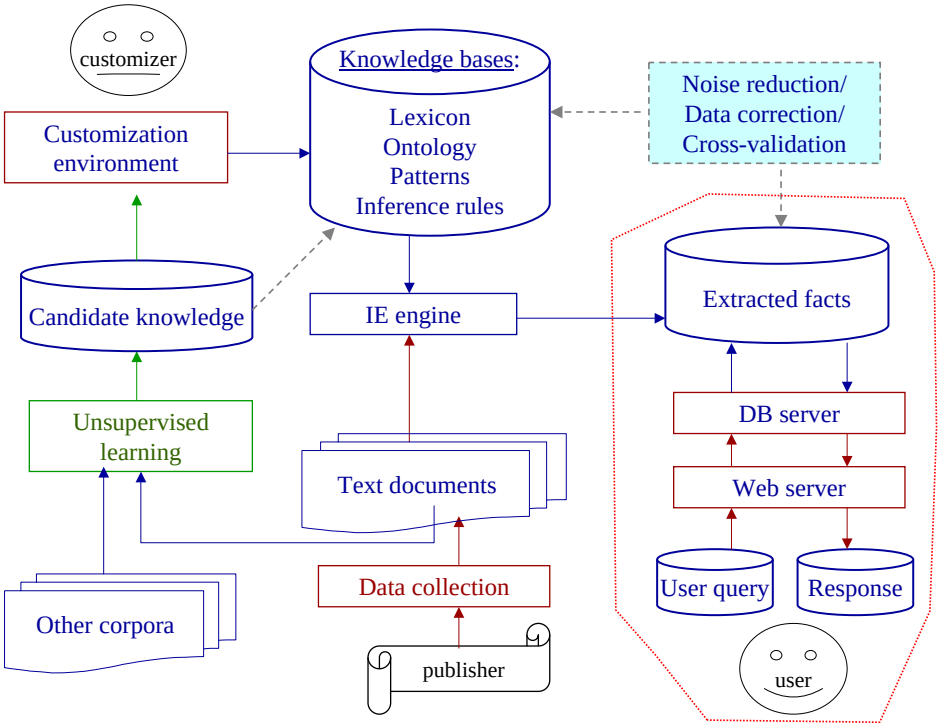


Figure 1: System architecture of ProMED-PLUS

Operating in the large is essential, because the learning components rely on the availability of large amounts of data. Knowledge acquisition, (Yangarber et al., 2002; Yangarber, 2003) requires a large amount of raw material, both domain-specific and general-topic texts. On the other hand, automatic error reduction requires a critical mass of extracted facts. Tighter integration between IE and KDD components, for mutual benefit, is advocated in recent related research, e.g., (Nahm and Mooney, 2000; McCallum and Jensen, 2003). In this system we have demonstrated that redundancy in the extracted data (despite the noise) can be leveraged to improve quality, by analyzing global trends and correcting erroneous fills which are due to local mis-analysis. For this kind of approach to work, it is necessary to aggregate over a large body of extracted records.[6]

Accessible on-line at doremi.cs.helsinki.fi/plus is the interface to the DB (lower-right of the diagram). Records may be viewed, selected and sorted, querying for the aggregated outbreaks, as well as for individual incidents. Distribution of outbreaks may also be plotted and queried through the *Geographic Map*.

[6]The system's knowledge bases are deliberately primed with a moderate amount of *a priori* information, without large gazetteers or disease lists, to make the learning problem interesting.

## References

R. Grishman, S. Huttunen, and R. Yangarber. 2002. Event extraction for infectious disease outbreaks. In *Proc. 2nd Human Language Technology Conf. (HLT 2002)*, San Diego, CA.

R. Grishman, S. Huttunen, and R. Yangarber. 2003. Information extraction for enhanced access to disease outbreak reports. *J. of Biomed. Informatics*, **35**(4).

A. McCallum and D. Jensen. 2003. A note on the unifi cation of information extraction and data mining using conditional-probability, relational models. In *IJCAI'03 Workshop on Learning Statistical Models from Relational Data*.

U. Y. Nahm and R. Mooney. 2000. A mutually benefi cial integration of data mining and information extraction. In *AAAI-2000*, Austin, TX.

R. Yangarber, W. Lin, and R. Grishman. 2002. Unsupervised learning of generalized names. In *Proc. COLING-2002*, Taipei, Taiwan.

R. Yangarber. 2003. Counter-training in discovery of semantic patterns. In *Proc. ACL-2003*, Sapporo, Japan.