

# Extracting Information about Outbreaks of Infectious Epidemics

**Roman Yangarber**

Department of Computer Science

University of Helsinki, Finland

first.last@cs.helsinki.fi

**Lauri Jokipii**

**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 evaluated based on carefully annotated corpora, at most a few hundred documents.<sup>1</sup> The goal of the work presented here is to explore the IE process *in the large*: the system integrates a number of off-line and on-line components around the core IE engine, and serves as a base for research on a wide range of problems.

The system is applied to a large dynamic collection of documents in the epidemiological domain, containing tens of thousands of documents. The topic is outbreaks of infectious epidemics, affecting humans, animals and plants. To our knowledge, this is the first large-scale IE database in the epidemiological domain publicly accessible on-line.<sup>2</sup>

<sup>1</sup>Cf., e.g., the MUC and ACE IE evaluation programmes.

<sup>2</sup>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<sup>3</sup> is shown in Fig. 1. The core IE Engine (center) is implemented as a sequence, or “pipeline,” of stages:

- Layout analysis, tokenisation, lexical analysis;
- Name recognition and classification;
- Shallow syntactic analysis;
- 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.<sup>4</sup>

The IE engine is based in part on earlier work, (Grishman et al., 2003). 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 Fig. 1), and automatic post-validation of extracted facts for detecting and reducing errors (upper right). Novel features include the notion of confidence,<sup>5</sup> and aggregation of separate facts into outbreaks across multiple reports, based on confidence.

Operating in the large is essential, because the learning components in the system rely on the availability of large amounts of data. Knowledge

<sup>3</sup>PLUS: Pattern-based Learning and Understanding System.

<sup>4</sup>ProMED, [www.promedmail.org](http://www.promedmail.org), is the Program for Monitoring Emerging Diseases, of the International Society for Infectious Diseases. It is one of the most comprehensive sources of reports about the spread of infectious epidemics around the world, collected for over 10 years.

<sup>5</sup>Confidence for individual fields of extracted facts, and for entire facts, is based on document-local and global information.

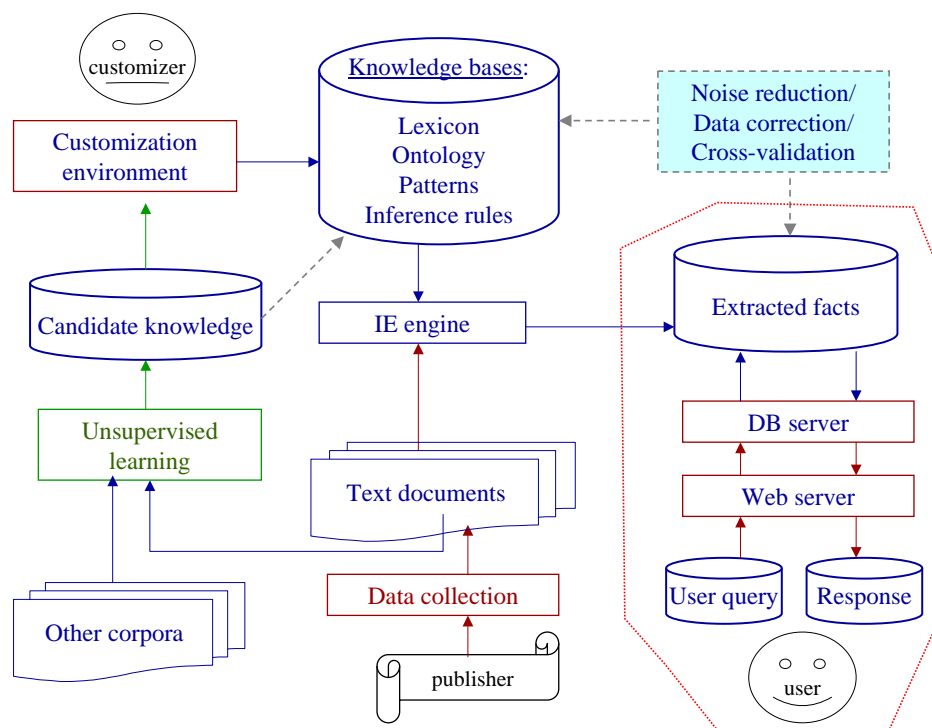


Figure 1: System architecture of ProMED-PLUS

acquisition, (Yangarber et al., 2002; Yangarber, 2003) requires a large corpus of 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, (Yangarber and Jokipii, 2005). For this kind of approach to work, it is necessary to aggregate over a large body of extracted records.

The interface to the DB is accessible on-line at [doremi.cs.helsinki.fi/plus/](http://doremi.cs.helsinki.fi/plus/) (lower-right of Fig. 1). It allows the user to view, select and sort the extracted outbreaks, as well as the individual incidents that make up the aggregated outbreaks. All facts in the database are linked back to the original reports from which they were extracted. The distribution of the outbreaks may also be plotted and queried through the *Geographic Map* view.

## References

- 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 unification of information extraction and data mining using conditional-probability, relational models. In *IJ-CAI'03 Workshop on Learning Statistical Models from Relational Data*.
- U. Y. Nahm and R. Mooney. 2000. A mutually beneficial integration of data mining and information extraction. In *AAAI-2000*, Austin, TX.
- R. Yangarber and L. Jokipii. 2005. Redundancy-based correction of automatically extracted facts. In *Proc. HLT-EMNLP 2005*, Vancouver, Canada.
- 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.