Authors
Petr Fanta
Duc Tam Hoang
Adam Huječek
Václav Pernička
Jakub Vlček

Supervisor
Ondřej Bojar

Version
0.1

Date
July 25, 2014

Contents

1. Introduction

1.1 Introduction to the project

Status: A dummy documentation - Text Analyser

Generally, Police often generate some documents with regards to a specific person (criminal records, conflict, ...). In the documents, there is a lot of information such as person name, date of birth, name of other persons, his/her location, other persons' location. They need an application to manage such documents in a convenient way and extract the important details without much human effort. This is the topics of our software project, call Text Analyser (TextAn).

TextAn is a client-server application for analysing text. It is to manage documents and exploit the information out of the documents. Given a collection of written texts (typically Police's reports) in Czech language, TextAn automatically breaks down the unstructured text to exploit the structured information.

Furthermore, The application allows user to add/remove/adjust the structured information. They can change the relation, correct mistake or provide further information. The structured information could be seen in graph-view, which is a fancy way that user could look at his/her profile.

Implementation of TextAn is recognised in Java 1.8 with Spring Framework. Information is stored with mysql database.

The developers's team consists of 5 students:

- Adam
- Jakub
- Peter
- Tam
- Vecca

Status: An even dummer introduction - Text Analyser

Textan (Text Analyser)

Generally, there is a huge amount of unstructured written texts which consist of structure information. It ranges from newspapers, medical bills to even personal letters. In every written documents, there are entities like names, address, dates, etc. These entities do not stand alone, there are relationship between them. The relationship could facilitate the problem of assigning an entity from new text to an existing objects. Such problem of extracting structure information out of a general documents has recently gained attention from both research and industry.

Out of all unstructured sources which contains structured information, Czech police reports is typical. They are the descriptions with regard to a specific person, his/her profile and actions, criminal records and so on. Such information by no doubt is the most valuable data in the whole documents. Currently, the task of extracting it is done manually. In other words, there is a person who read through all documents, look at the database and match the description in the document with existing data. This is not-surprisingly time-consuming and labour-intensive.

Project Textan is devoted to build a pleasant and effective tool for extracting structured data in Czech police reports. Our goal is to provide an effective extraction of structured data with user involvement. It consists of three main contributions. First, an efficient design of database for structured data is implemented. Second, we aim to maximize the performance of automatic detection of entities. Third, Textan is designed to provide a friendly method for users to confirm or adjust the system's suggestions. At first, the second task may achieve low accuracy, but the magic of machine learning could improve the performance through confirmation by users.

1.2 Work on the project

1.3 Adopted solutions and decisions

1.4 Related work

Mining structured information from unstructured text documents has recently gained attention. Current approaches are still far from an fully automated and completely accurate processing. However, the challenge has led to a number of

research and applications. Each works focused on solving a particular problem in the big pictures with regard to some specific languages.

This section provides an insight into other contributions to the mining problem. Both research-oriented works and industry applications are taken into account.

Knowtator is a general-purpose text annotation tool. Developed by scientists at Division of Biomedical Informatics, it uses Protégé knowledge-base as the database. As an early development of tool for annotating data, Knowtator has a number of limitations (OS dependent, no client-server, no automatic detection). The remarkable feature of Knowtator is the ability to relate annotations to each other via the slot *reference*. The tool has been implemented as a Protégé plug-in for wider-spread usage.

Anafora is an open source web-based text annotation tool, developed by scientists of University of Colorado at Boulder. It distinguishes itself from previous contribution in the field of OS platform. Before the introduction of Anafora, old tools were written as a local application in a local machine under the threat of data fragmentation. Anafora is a web-based tool with client-server structure.

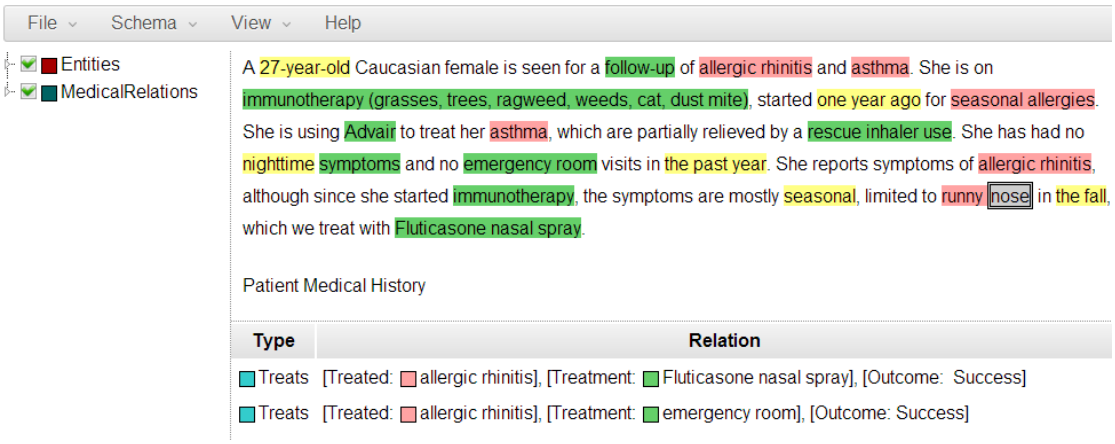

Figure 1.1: Anafora graphics user interface

Anafora is designed for the medical domain with a medical named entity tags, not a general domain. It is developed with a server in Django (a Python framework) and a client in jQuery (a JavaScript library). Besides, it stores annotations in each single XML files. By doing so, authors claim that the tool is agile and flexible. The project hierachy is designed to be the file/directory structures. The application provides an automatic detection of named entities. Users could change/add details to the annotation by mouse or keyboard. Language choice is English.

The limitation of Anafora lies in the data structure. The complicated definition like relations (or relation properties) among entities are not supported. It may be the trade-off when authors want to develop a light tools based on a simple data structure such as XML.

BRAT rapid annotation tool is a web-based tool for text annotation, most useful to add note to existing text documents (taken from the introduction of BRAT). It is developed by Japanese. This tool does support the linking relation between two entities and even the link from entities to a definition outside documents (such as wikipedia).

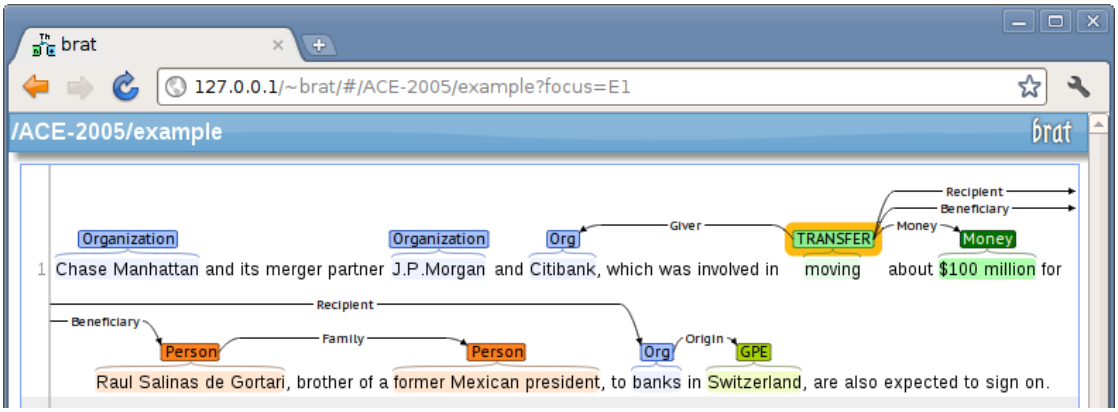

Figure 1.2: BRAT graphics user interface

The strong point of BRAT is the simplicity of usage. The tool offers a comprehensive visualization with mouse-based activities, a linking method to an external resources (such as Freebase, Wikipedia, and Open Biomedical Ontologies). A number of other useful functions are *save, export from standoff format, search*. The brat standoff format is a simple format developed for the visualization of the tool, in which data is stored in two files: a text file *.txt* and an annotation file *.ann*.

The weak point of the tool is that it does not support any automatic recognition of entities. All annotations have to be done manually. This leads to a features that the tool could be applied on a wide range of languages. However, the limitation leads to a fact that BRAT is merely an editor for annotating.

1.5 Conventions

1.6 Glossary

2. Requirement specification

3. Architecture

4. Developer notes

5. Possibilities of future extensions