IDP intellectual document processing

Max Talanov

April 25, 2008

Domain

80% of information is unstructured. UnQueriable - this means that you have this information in your computer but you can not create any analysis based on it. This information lies as sheets of paper in documents of various formats.

Technologies overview

- Natural language processing
  - **UIMA** IBM (Unstructured information management architecture)
  - **GATE** University of Sheffield (General architecture for text engineering)
  - **Minorthird** CMU (Toolkit for storing text, annotating text, and learning to extract entities and categorize text)
- Machine learning
  - **Weka** University of Waikato (Weka is a collection of machine learning algorithms for solving real-world data mining problems.)
  - **Rapid miner** rapid-i (Open-source data mining solution, built on base of Weka)

Let's consider simple example of structuring the job offer text in XML format.

Newsgroups: austin.jobs
Subject: COMPUTER TECHNITION NEEDED FOR RETAIL STORE 451-2489
Date: Thu, 28 Aug 1997 17:16:14 GMT
Organization: Jump Point Communications, Inc.
Message-ID: 3405a5f4.92436695@NEWS.JUMPNET.COM>
The computer and photo industries have merged into CompuImage "where photos & computers meet". This exciting concept holds tremendous growth opportunitines. We are looking for a customer friendly person expienced with the Internet, Networking, Windows 95, Windows NT, Foxpro, general troubleshooting and product titles. Call Clifford @ 451-2489 or E-mail cliff@compu-image.com.

We are going to get following structure of XML document:

root for XML wellformness

header The supplemental information section

newsgroup austin.jobs

subject COMPUTER TECHNITION NEEDED FOR RETAIL STORE 451-2489

post date 28 Aug 1997

id 3405a5f4.92436 ...

body Main part of an offer

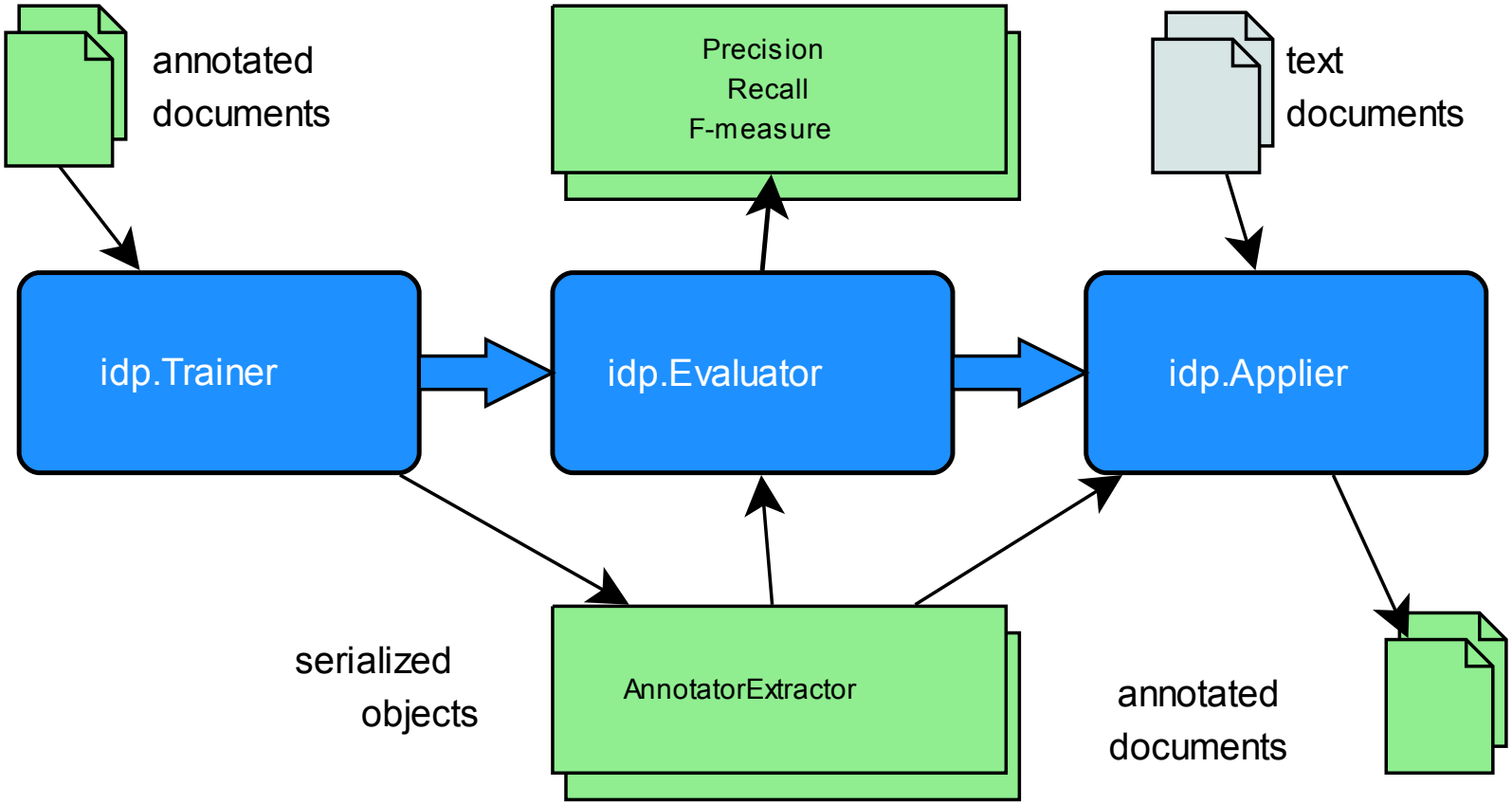
Information Extraction Description

We are going to use Information Extraction approach to structure the text document in XML.
There are tree parts of information extraction task:

- Train (Create annotators)
- Test (Evaluate)
- Apply (Most important for us)

First of all we will train the annotators (programs that can mark proper parts of the text) on positive examples, then we are going to let them annotate plain text examples and put results in XML file.

Application structure



Application structure description

idp.Trainer	input is Annotated (XML) documents - training set
idp.Trainer	output is Annotators - serialized Java objects, that are learned to annotate
idp.Evaluator	input is Annotators, output is Precision, Recall, F-measure parameters
idp.Applier	input is Annotators and text to structure - test set
idp.Applier	output is structured documents

<root>
<_predicted_i_header>
Newsgroups:
<_predicted_i_newsgroup>austin.jobs</_predicted_i_newsgroup>
Subject: <_predicted_i_subject>COMPUTER TECHNITION
NEEDED FOR RETAIL STORE 451-2489</_predicted_i_subject>
Date: Thu, <_predicted_i_post_date>28 Aug
1997</_predicted_i_post_date>17:16:14 GMT
Message-ID: <_predicted_i_id>3405a5f4...</_predicted_i_id>
</_predicted_i_header>
<_predicted_i_body>
The computer and photo industries have merged into CompuImage
"where photos & computers meet". This exciting concept holds
tremendous growth opportunitines. ...
</_predicted_i_body>
</root>

This looks much better, we can form analytical report about number of specialist are required by company, skills that are required average salary that is suggested during 1997 year. We can track trends of average salary fluctuation for different skill set through 1997 year. All this information could be retrieved by XQuery for example.

Annotation Process of marking up some parts of the text according to some rule(learned in case of Machine learning)

Annotator Some program that makes annotation