
Contribution ID: **206** Type: **not specified**

# Natural Language Processing and Information Retrieval - Major Tasks and Applications

*Monday, 11 April 2011 14:00 (30 minutes)*

## Overview

In today's world the amount of data produced and stored keeps growing exponentially. Moreover, with the expansion of Internet and widespread use of text processing tools majority of that data is available in an unstructured text form. According to the International Data Corporation (IDC): "In organizations, unstructured data accounts for more than 80% of all information"(see http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf). It is now well understood that unstructured data contains significant business value and tools capable of effectively analyzing this type of data are becoming of paramount importance in everyday life of any organization. However, traditional Information Technology techniques are not sufficient to process and exploit the full potential of unstructured data. The recent advance in Information Retrieval (IR), Natural Language Processing (NLP) and related areas provides a number of instruments that are well suited to solve this problem. The other aspect of this problem is the volume of data that is being employed. This requires a huge storage and computing infrastructure. This problem is being addressed and today several solutions are offered: virtual supercomputers, grids, clouds and similar. To name one such solution, recently as a result of the Baltic Grid Second Phase project, an infrastructure for scientific research and business applications is created (see http://www.balticgrid.eu/).

## Impact

With the help of such data pool the user would benefit in saving operational costs when exploring web content. The service would benefit SME's, public and other organizations. To give a few examples:
• Content and news providers would use data pool to add functionality to their publications. They might be persuaded to share their content and get additional service on top of it in return.
• The tool would help governmental bodies to evaluate the feedback from the citizens after the decisions are made. These organizations might also be interested in sharing their content.
• Marketing/Public Relations companies would use event pool to evaluate outcomes of their campaigns.
• Predictive analytics companies would use event pool to create predictive models (e.g. portfolio manager). Some events change sentiment on the market which then moves asset price.
• Competitive intelligence companies would employ event pool to analyze market situation of the companies they are researching (e.g. corporate investor).

## Description of the work

To get an insight of what constitutes a text processing and how this applies in everyday life of an organization let us consider a working example of a system which deals with events as they are reported on the Web. The event is a segment of information describing who does what and when.

For example when the following news is reported:

"(NewsToday.com): Oceanic Airlines plans to lay off nine of its baggage workers at Western Airport in early April, according to a local union representative for the carrier.",

report meta-data and event components would be extracted:

Date –April, 2011
Provider –NewsToday.com
Organization –American Airlines, Western Airport
Industry –aviation
Event type –lay off
Features –nine of its baggage workers
Credibility [1...10] –( local union representative for the carrier / plans to) 8
Sentiment [-5...5] –(...) -3
Related events –(···)

On the one side a service provider would create a data pool of events. On the other side a user (SME, public or other organization) would get the query-relevant events with additional service on top of them: credibility –a measure of importance of the event, sentiment –opinions gathered and analyzed after the event was reported, associations –a network of related events for further exploration.
Implementation of a system like this would require knowledge and skills in several areas:
• Large-scale Web crawling. A Web crawler would read content published on the Web and harvest raw material for further analysis. Methods involved here rely on Web Crawling, Web Scraping, Indexing.
• Named Entity Retrieval. This would be the main component of the system extracting elements which constitute the event: organizations and persons, timing and actions, properties and features. Methods involved here rely on Natural Language Processing and Machine Learning.
• Sentiment Analysis. To truly tap into social-media content a measure of sentiment expressed in this media is needed. In this way original event description would be enriched with opinions and emotions expressed in relation to the event. Methods involved here rely on Natural Language Processing and Machine Learning.
• Event grouping. Connections between events are found. Methods involved here rely on Language Models and Machine Learning.
• Faceted Search. Named entities extracted and event categories defined would allow creating filters for fast and easy exploration of query-related events. Methods involved here rely on Information Retrieval.
• Distributed/Parallel Computing. In order to be valuable the system would need to process huge amount of data in very short time frame. This could be possible only when infrastructure of supercomputers or clusters of computers is employed. Also specific algorithms like Map-Reduce would be apt here.

## Conclusions

A justification for large-scale text processing is given and a model system is described which employs methods in Information Retrieval, Natural Language Processing, Machine Learning, Distributed Computing etc. Similar services that would use automated tools for the analysis of text data could also be and are developed. The main implication is that such systems are very complex and expensive in order to be implemented and maintained by separate users on their own. Thus only actors with the knowledge of advanced state-of-the-art scientific methodology and computing infrastructure are capable of creating such systems. On the other hand, users would identify the demand for specific functionalities in such a way determining "the shape"of a service.

**Presenter:** Dr ZUOKAS, Danas

**Session Classification:** General Workshops