# Harnessing Language Technologies in Multilingual Information Channeling Services

**Diman Karagiozov**
Tetracom IS Ltd.
diman@tetracom.com

## Abstract

Scientists and industry have put significant efforts in creating suitable tools to analyze information flows. However, up to now there are no successful solutions for 1) dynamic modeling of the user-defined interests and further personalization of the results, 2) effective cross-language information retrieval, and 3) processing of multilingual content. As a consequence, much of the potentially relevant and otherwise accessible data from the media stream may elude users' grasp.

We present a multilingual information channeling system, MediaTalk, which offers broad integration between language technologies and advanced data processing algorithms for annotation, analysis and classification of multilingual content. As a result, the system not only provides an all-in-one monitoring service that covers both traditional and social media, but also offers dynamic modeling of user profiles, personalization of obtained data and cross-language information retrieval. Bulgarian and English press clipping services relying on this system implement advanced functionalities such as identification of emerging topics, forecasting and trend prediction, all of which allow the users to monitor their standing reputation, events and relations. The architecture of the system is robust, extensible and adheres to the Big Data paradigm.

## 1. Introduction

In the last decade, the information available on the Internet has grown significantly[1] and has increased the demand for efficient monitoring and information extraction for the purposes of industry and research, including publishing, marketing, advertising, social research, etc. The problem is related not only to the vast volume of information but also to the dynamic nature of the information flow, the variety of sources, data formats, media types (printed, electronic, audio, multimedia) and languages. Monitoring applications for this purpose have complex structures implementing advanced data processing and language technologies.

Moreover, the processing and information extraction from multilingual and multimodal content is still an area of active research with no established solutions. The efficiency of methods is also essential since they need to have close to real-time performance and give high-quality results.

Section 2 describes existing media monitoring services which provide similar functionalities to the system presented in this paper. The key design and functional decisions, implementation and integration approaches are described in Section 3. Section 4 outlines the functionalities of an intelligent web application built with our system and the benefits of using it. Section 5 summarizes the main achievements of the system and suggests improvements and extensions.

[1] Digital Universe studies series, John F. Gantz et al., IDC 2007, 2008, 2009, 2010
http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm

## 2. Overview of Related Service

Media monitoring services have gained popularity in recent years and have focused on monolingual electronic content: web pages of broadcasting agencies, TV and radio channels, newspapers, governmental and non-governmental organizations. Some of the most popular companies offering these services are WebClipping[2], PressClipping[3], eReleases[4]. Typically, monitoring applications process monolingual information in structured form and employ methods based on keywords, text categorization, named entity recognition and monolingual information extraction.

Another branch of information services reside in the Google ecosystem. Public services like Google Reader[5] and Google Alerts allow the users to describe their fields of interests by providing search queries. Users are then notified of recent query results for the same terms via RSS feeds and emails.

The service provided by Prismatic[6] integrates with the social profiles of the users and tries to present a user-focused English language information stream. Companies such as CyberAlert[7] and CustomScoop[8] offer search queries in multiple languages and instant machine translation of extracted clippings along with comprehensive news coverage and minimization of irrelevant information. The service Mention.net[9] is able to process multilingual content; the user profile information is described as a set of keywords.

There are specialized broadcast monitoring services, such as Critical Mention[10] and TV Eyes[11], that combine real-time TV and radio broadcast monitoring with online and social media coverage. Such services capture text, audio and video content, analyze it to some extent and distribute it.

To the best of our knowledge, there is no system applying cross-language information retrieval in news monitoring and no existing unified approach with respect to dynamic user profiling and multilingual and multimodal content monitoring.

## 3. A Multilingual Information Channeling System

We present a system for multilingual information channeling which aims for the following key objectives: (a) relevant news coverage and adequate data analysis (implemented); (b) efficient dynamic modeling of user profiles (implemented); (c) a uniform approach to user profiling and multilingual content monitoring (implemented); (d) efficient cross-language information extraction (under development); and (e) a uniform approach to processing multimodal content (a future task).

To achieve these objectives, we integrate the appropriate language technologies with advanced data processing algorithms for annotation, analysis and classification of multilingual content. More particularly, semantic entities[12] are extracted, represented as time series and classified in order to obtain relevant news coverage and to provide adequate data analysis. Relations between semantic entities are represented as a semantic graph that enables users to track the related persons, dates, locations and the like. Data (both target information flow and user information) are provided with internal semantic links that preserve content integrity and allow information tracking. Our experiment towards crossing the language barrier utilizes a hybrid (example-based, statistical and dictionary-based) machine translation

[2] `http://www.webclipping.com`

[3] `http://www.pressclip.net`

[4] `http://www.ereleases.com`

[5] Google Reader service has been discontinued since 1st July 2013. Google Alerts service provides the relevant to the query information in email format which is not convenient for integration in 3rd party software systems.

[6] `http://getprismatic.com/`

[7] `http://www.cyberalert.com`

[8] `http://www.customscoop.com`

[9] `https://en.mention.net/`

[10] `http://www.criticalmention.com`

[11] `http://tveyes.com`

[12] We will refer to both named entities and noun phrases as semantic entities further on in this paper.

engine and a language-independent presentation of the semantic entities. The processing of multimodal content will be implemented as extensions to existing text and metadata extraction systems, such as Apache TIKA[13].

### 3.1. System Workflow

First, we present the processes of data acquisition, normalization, processing, indexing, analysis and visualization. After that we describe the novel approach, called *Lambda architecture*, adopted in solving the inevitable Big Data problem (Marz and Warren, 2013).

The general workflow in the system can be generalized in the following subsequent processes:

1. Harvesting the data – based on a wide collection of RSS feeds and user defined queries and resources from the social networks, a pool of information items is created.
2. Text extraction – each information item is transformed to a list of textual fragments.
3. Semantic excerpts – each textual fragment is processed so that the most "interesting" semantic excerpts are indexed.
4. Graph of semantically related excerpts – a graph of interrelated information items and semantic excerpts is created.
5. User perspective – the user "interests", described as another set of information items, are processed in a uniform way, and thus the harvested information items are "contextualized" for the user.
6. Statistical methods are applied on contextualized items to identify the emerging topics and trends.

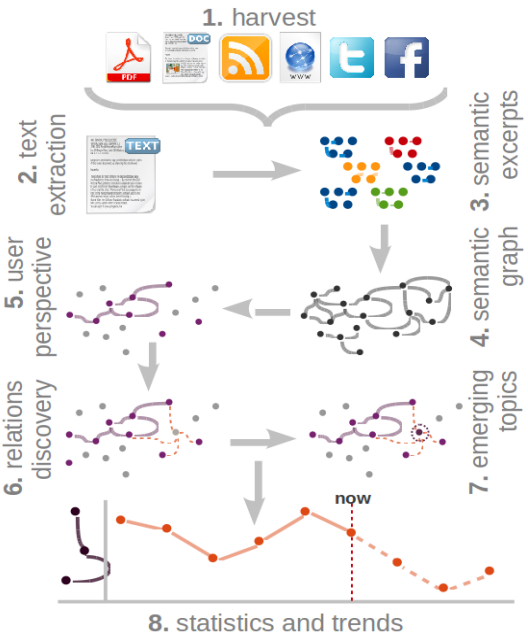


Diagram 1: Core system processes

Each of these processes are described in the following subsections.

#### 3.1.1. Harvesting and Text Extraction

A customizable data harvesting engine is designed and implemented to deal successfully with the various formats of data on the Internet and the almost random intervals of update, as well as to track simultaneously news and social media. Data providers (RSS feeds, Facebook pages and groups, search queries results, queries on Twitter, feeds, websites and document libraries) follow standard schema definitions (Ronallo, 2012), and their properties are stored in a content management system. The object-oriented description of the data provider types creates an abstraction layer between the actual provider's content and the harvesting and storage processes. New data providers can easily be added, and new data medium types, such as audio and video podcasts, can be supported.

The harvester automatically collects structured and unstructured information from the Internet and stores it in an easy-to-process format while omitting irrelevant information such as navigational elements, templates, advertisements, etc. The collected raw data are converted to textual content and metadata using either Apache TIKA, when the content appears to be in a binary format, or a boilerpipe detection library (Kohlschütter et al., 2010), when the content is an (x)HTML page. The text content and the extracted metadata are stored in the content management system as immutable objects.

[13] http://tika.apache.org/

### 3.1.2. Linguistic Annotation

The system is designed for the processing of multilingual content. At the moment Bulgarian and English are implemented and the technology can be applied to other languages as well. We utilized the ATLAS linguistic framework (Ogrodniczuk and Karagiozov, 2011) as it provides multilingual light-weight automatic text annotation functionalities through a multilingual UIMA-based (Ferrucci and Lally, 2004) framework. The ATLAS framework is capable of segmenting the text and extracting named entities and noun phrases. Furthermore, ATLAS provides text extractive summarization, automatic categorization and cross-language information retrieval (CLIR) modules. The framework is extendable in terms of new annotations (e.g., semantic relations) and new languages.

Language-independent names similarity measure (Steinberger et al., 2011) is implemented in order to automatically link translation equivalents of named entities and noun phrases in a multilingual content, thus facilitating the CLIR-based analysis.

### 3.1.3. Integrating Language Technologies in Data Processing Algorithms

The dimensionality of the data targeted for analysis is reduced by tf.idf weighting as the top-ranked semantic entities are represented as time series. We use classification algorithms (Ratanamahatana et al., 2010) on these series in order to find cylinders (sharp raise, plateau and sharp drop), funnels (sudden increase and gradual decrease) and bells patterns (gradual increase and sharp drop). As such patterns indicate a significant change, they are used for the identification of emerging topics. Signal analysis of all possible bonds between the identified patterns reveals hidden semantic relations. A coherent signal alerts the user of themes and topics that constitute trends and provides knowledge for further actions. The signal route represents the evolution of processes and events within the series, allowing for identification of relevant data and generation of recommendations and conclusions.

The most relevant content items encompassing the identified semantic entities are clustered. Most text clustering algorithms (Aggarwal and Zhai, 2012) can easily group texts into clusters but provide synthetic labels (if any labels at all) which are far from meaningful. Instead, we have adopted the Lingo3 clustering algorithm (Osinski et al., 2004), which decides on cluster labels prior to the clustering. Such clusters are used for showing the user what is happening at the moment, or what has happened several hours ago, yesterday, last week or last month.

Furthermore, the user is able to track the relations between different semantic entities. Performing deep semantic analysis in a multilingual environment is a complex task and requires a lot of language-specific resources (Navigli and Ponzetto, 2012). Thus we assume that two semantic entities (concepts, people, locations, etc.) are related if they both appear in a sentence. Additionally, we implement anaphora resolution in order to replace pronouns with their antecedents if the latter are recognized as semantic entities. After that the semantic entities are indexed as bi- and tri- grams in a language-independent SOLR[14] core. Each n-gram is considered to represent a semantic relation, and a semantic relations graph is built on top of a SOLR index.

### 3.1.4. User Profiles

The system provides only those fragments of the information flow that are most relevant to the users' interests. To achieve this, the user profile is dynamically built using a daily analysis of the user's sources. The profile serves as a pattern against which multilingual textual content from digital media sources and social networks is screened and rendered.

User interests are not described with static keywords but are derived from data provided by or related to the user – websites, documents, news items, etc. The same process of harvesting and linguistic processing is applied to the user data, after which we cluster the user content and formulate the user's interests. Further on, we create supervised models which are later used for automatic categorization (channeling) of items in the information stream towards the user profile. We apply chi2 feature reduction (Manning et al., 2008), subsequently building a smoothed naïve Bayesian model (Chen and Goodman, 1996).

[14] `http://lucene.apache.org/solr/`

The user profile, represented as a graph of semantic relations, factors the full semantic graph using a graph pattern-matching algorithm (Gallagher, 2006). All other analysis – summarization, relations tracking, identification of emerging topics, suggestions for further evolution of the user profile – are based on the user-factored semantic graph.

## 3.2. Lambda Architecture

There are several important non-functional requirements behind the system that are necessitated by the complexity of the technologies: (a) robustness and fault tolerance in distributed environment, which address the random changes in machine and human behavior; (b) low latency in reads and updates needed for modeling near-real-time data analysis; (c) ad-hoc queries that support business optimization and new applications of the data and the system; (d) scalability, needed to address the increasing data volume and system load.

Traditional software architectures only partially provide the above-listed non-functional requirements. Consequently, we have adopted a novel approach called *Lambda architecture* (Marz and Warren, 2013). Lambda architecture enables the execution of arbitrary functions on arbitrary data in real-time by decomposing the problem into three layers: batch layer, serving layer and speed layer.

Our implementation of the Lambda architecture is as follows: (a) the master data store is maintained through the content management system ATLAS CMS,[15] which seamlessly integrates a linguistic processing framework into the processes of content management; the semantic excerpts – people, places, organizations, salient noun phrases, relations between them, summary of information items and categorization labels – are appended to the master datastore; (b) the batch layer and views are based on a relational database (PostrgeSQL[16]) and a set of SOLR[17] cores: both solutions provide horizontal scalability by replication and shading, which guarantees optimal performance and stability; (c) the serving layer is integrated in the front-end component; as it facilitates the merging process, the same data model is used in the master datastore and the batch.

## 4. Press-Clipping Case Study

The described system provides an all-in-one monitoring service that simultaneously tracks traditional and social media. In this way, essential information encompassing what is coming from the official media and what is said by people is captured. Hence, a press clipping service implemented on top of our system allows the users to effectively monitor their reputation, particular products or practices and to make timely and well-informed decisions. The service can successfully meet the needs of large international corporations and organizations, publishing houses, news and PR agencies, political entities, etc. Two press-clipping services have been built as a proof-of-concept:

- Press clippings in English:[18] – the service monitors 30 international news agencies and an average of 1,500 news items are processed daily.
- Press clippings in Bulgarian:[19] – the service monitors 44 Bulgarian news sources and blogging sites, and an average of 4,000 news items are processed daily.

Both services feature 12 public profiles mapping the top-level news topics such as World, Business, Entertainment, Sports, Technologies, Health, etc. The English press-clipping service provides special topic-oriented profiles for Terrorism and Security. The Bulgarian press-clipping service provides special topic-oriented profiles for emerging local events, such as Elections 2013, Economical crisis, Judicial system, Crime rate, Caretaker government, etc.

[15] `http://www.atlasproject.eu/`

[16] `http://www.postgresql.org/`

[17] `http://lucene.apache.org/solr/`

[18] The service address is `http://en.mtalk.eu/login`.

[19] The service address is `http://mtalk.eu/login`.

The general functionalities of the services can be accessed through Google or Facebook accounts or by registering as a news user with username and password.

In addition we have created several enterprise-oriented profiles in order to test the user-focused capabilities of the system. As an example we have set up four profiles of interest – gas & oil drilling, biofuels and renewable energy, Arctic shelf and Gulf of Mexico, for the British Petroleum company[20].

Figure 1 shows the main screen of the English-language press-clipping service. The screenshot shows the list of latest news, the news clusters (highlights), people, place and organizations which are currently in the news focus, and the merging topics.

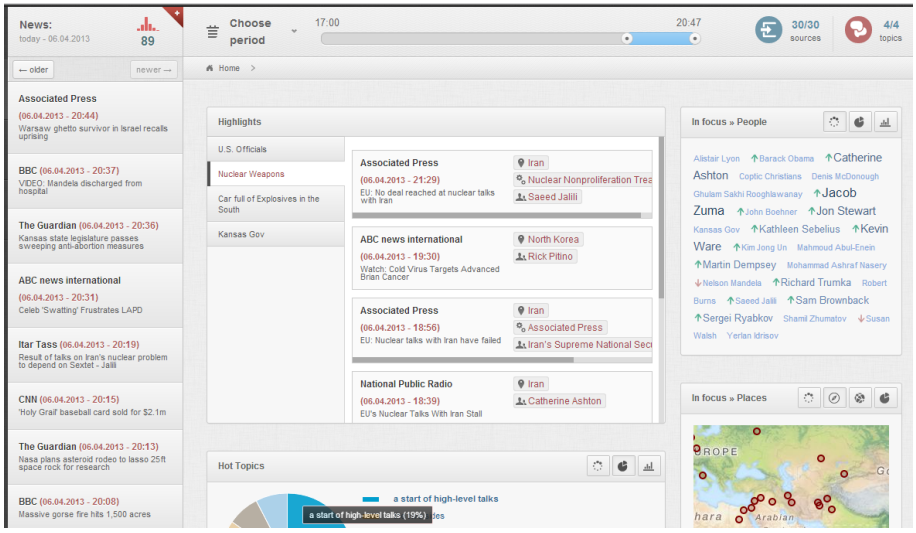


Figure 1: English press-clipping main screen

Figure 2 shows the news filtered through a user profile. In addition to the widgets shown in Figure 1, the screenshot includes statistics of news related to the particular user profile.

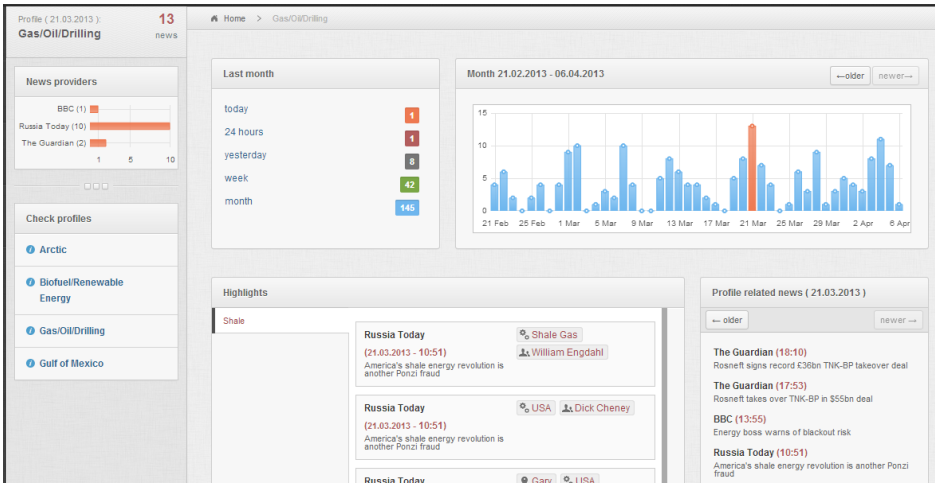


Figure 2: A user profile screen

[20] The press-clipping service for BP can be accessed at: `http://en.mtalk.eu/login` using `bp@en.malk.eu` as username and "bptest" (without the quote marks) as password.

## 5. Conclusion and Further Work

We have presented a multilingual information channeling system that can analyze formally structured and unstructured data representing users' interests, build user profiles dynamically and channel the information flows through these user profiles. The integration of language technologies and algorithms for analysis of time series utilizes functionalities such as identification of emerging topics, forecasting, and trend predictions. The fusion between an extensible linguistic processing framework and a multilingual content management system allows deeper semantic analysis and support of further languages. The architecture of the system is robust and adheres to the Big Data paradigm.

The system will be gradually extended in the directions of:

- Quality evaluation: Although techniques for evaluation of each individual component in the system (e.g. text categorization, text summarization, machine translation, information extraction) do exist, they are not applicable to the system as a whole. Furthermore, there is no annotated corpus suitable for the evaluation purposes of the multilingual information channeling system. Thus, a set of criteria will be developed in order to evaluate the quality of the implemented workflows and the usability of the system. The envisioned approach consists, not exclusively, of gathering a focus test group of individuals, developing test scenarios and questionnaires and building a manually annotated test corpus.
- Competitive intelligence: In addition to the news streams we will focus on defining, gathering, analyzing and distributing intelligence about products, customers and competitors. A foreseen challenge is the vast volume of potentially interesting information as well as the variety of source media types and data formats (printed, electronic, audio, multimedia) and languages.
- Cross-lingual information retrieval: The system processes information in six languages (English, Bulgarian, Greek, Polish, Romanian and German) and can easily be extended to support other languages. Future research and development will be focused on heavier utilization of CLIR in order to increase the added value of the provided analysis.
- Multi-document summarization: A logical extension of the current topical clustering mechanism is to create a summary report on the set of information items in each cluster. Algorithms as Lexrank (Erkan and Radev, 2004) and semantic graphs (Plaza and Díaz, 2011) can be effectively employed. The harmonization of the CLIR, MT and multi-document summarization components is still an open field for research and development.

## References

Aggarwal, C. and Zhai, C. (2012). A Survey of Text Clustering Algorithms. In Aggarwal, C. and Zhai, C. (eds.), *Mining Text Data*, pages 77–128. Springer.

Chen, S. F. and Goodman, J. (1996). *An Empirical Study of Smoothing Techniques for Language Modeling*, pages 310–318. Association for Computational Linguistics.

Erkan, G. and Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence (JAIR)*, 22:457–479.

Ferrucci, D. and Lally, A. (2004). UIMA: an Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10:327–348.

Gallagher, B. (2006). Matching Structure and Semantics: A Survey on Graph-Based Pattern Matching. *In AAAI FS '06: Papers from the 2006 AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection*, pages 45–53.

Kohlschütter, C., Fankhauser, P., and Nejdl, W. (2010). Boilerplate Detection Using Shallow Text Features. *In Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 441–450. ACM.

Manning, C. D., Raghavan, P. and Schutze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Marz, N. and Warren, J. (2013). *Big Data: Principles and Best Practices of Scalable Realtime Data Systems.* Manning Publications Company.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Ogrodniczuk, M. and Karagiozov, D. (2011). ATLAS Multilingual *Language* Processing Platform. *Procesamiento del Lenguaje Natural*, 47:241–248.

Osinski, S., Stefanowski, J., and Weiss, D. (2004). Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. *Intelligent Information Processing and Web Mining Advances in Soft Computing, 25*:359–368.

Plaza, L. and Díaz, A. (2011). Using Semantic Graphs and Word Sense Disambiguation Techniques to Improve Text Summarization. *Procesamiento del Lenguaje Natural*, 47:97–105.

Ratanamahatana, C. A., Lin, J., Gunopulos, D., Keogh, E. J., Vlachos, M., and Das, G. (2010). *Mining Time Series Data*, chapter Data Mining and Knowledge Discovery Handbook, pages 1049–1077. Springer.

Ronallo, J. (2012). HTML5 Microdata and Schema.org. *The Code4Lib Journal*, 16 (2012-02-03).

Steinberger, R., Pouliquen, B., Kabadjov, M. A., Belyaeva, J., and der Goot, E. V. (2011). JRC-NAMES: A Freely Available, Highly Multilingual Named Entity Resource. *In Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, Hissar, Bulgaria,* pages 104–110.