

RLPCS: Russian Language Processing Cloud Service

Andrey Sozykin, Mikhail Goldstein

Institute of Mathematics and Mechanics, Ural Branch of Russian Academy of Sciences, Yekaterinburg, Russia

Dmitry Ustalov, Denis Shirgin, Alexander Kudryavtsev

Institute of Physics and Technology
Ural Federal University, Yekaterinburg, Russia

Introduction

Natural language processing (NLP) is widely used in machine translation, information retrieval, knowledge retrieval and question answering systems. Despite the popularity and importance of these problems, Russian language processing has not achieved wide popularity due to the lack of open systems that perform NLP in Russian and the requirements of high-performance computers to run these systems.

There are many open source natural language processing systems (e.g. GATE, NLTK, Stanford NLP, LingPipe). These systems are mostly English language oriented and the support of Russian language is either absent or poor. In Russian Federation, open source NLP software had been developed at DIALING project, but it seems to be currently frozen. Therefore, there is no such production-ready open system for Russian language processing.

Natural language processing, like any Artificial Intelligence problem, requires a lot of computer resources. The creation of the Russian language processing cloud service is needed to allow Russian text processing for those who does not have enough resources to create their own hardware and software

Related work

Popular language processing systems (GATE, NLTK, Stanford NLP, LingPipe) have rich set of features, but they are mostly English oriented and lack of Russian support.

A tagset for Russian morphology has been developed as a part of MULTEXT-East project. This tagset is used in part-of-speech tagging systems like TreeTagger and TnT. Morphological analyzer for Russian language was developed within Showball project, but it can only stem the words.

There are many NLP software vendors in Russian Federation: Yandex, ABBYY, Mail.Ru Group, Rambler, RCO. Most of these systems are proprietary and created exclusively for vendor needs. Mystem morphological analyzer is an exception. It is free for non-commercial use and its algorithm has been published.

The most complete open source Russian language processing software is made within the DIALING machine translation system, which include graphematical, morphological, syntactic and semantic analyzers for Russian. Unfortunately, the development of DIALING had been stopped by 2008.

Therefore, currently there is no production-ready open natural language processing system that has analogy to English language processing systems.

RLPCS project

RLPCS objective is the creation the Russian language processing service that works in cloud. This processing include the following layers (Fig. 1.):

- 1) Graphematical analysis — text segmentation and token identification.
- 2) Morphological analysis — detection of words morphological interpretation: stemming, part of speech and corresponding grammemes set retrieval.
- 3) Syntactic analysis — syntactic parsing of sentence and building a parse tree.
- 4) Semantic analysis — building the graph of semantic categories and relations in sentence.

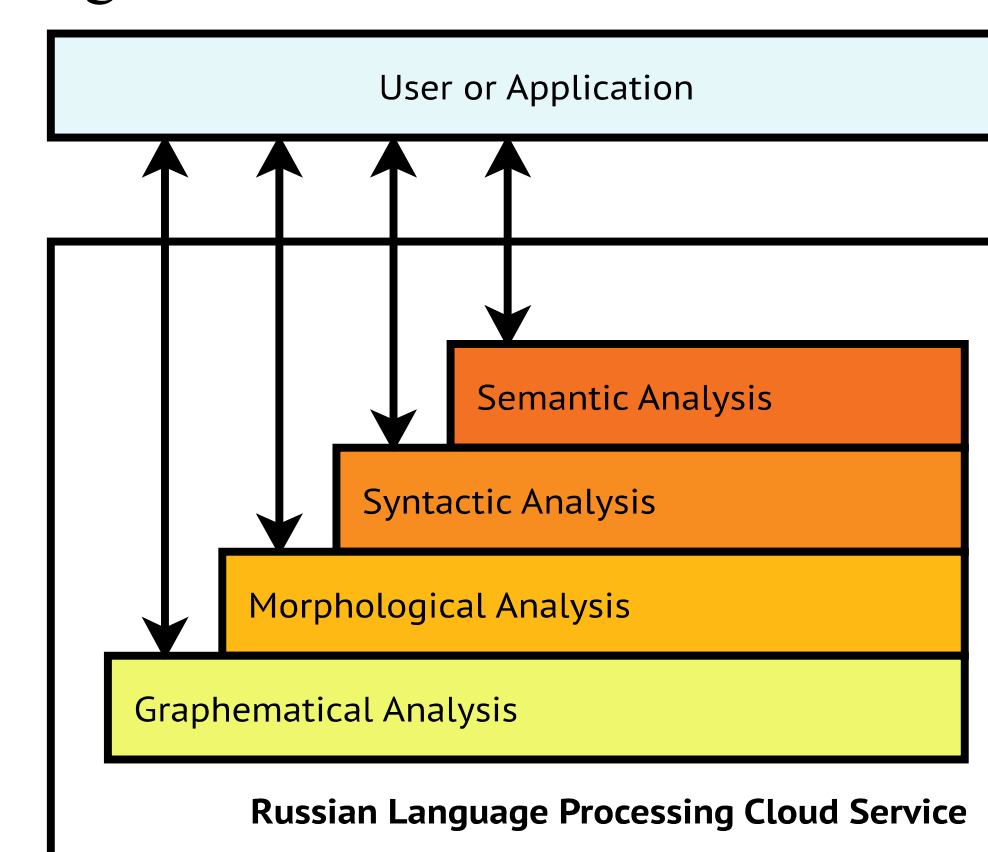


Figure 1. Levels of Russian language processing in RLPCS project.

Architecture

The RLPCS is provided in Software-as-a-Service model, when users work with application installed at service provider infrastructure using networking channels and standard protocols. RLPCS uses REST architecture style and JSON format (Fig. 2.).

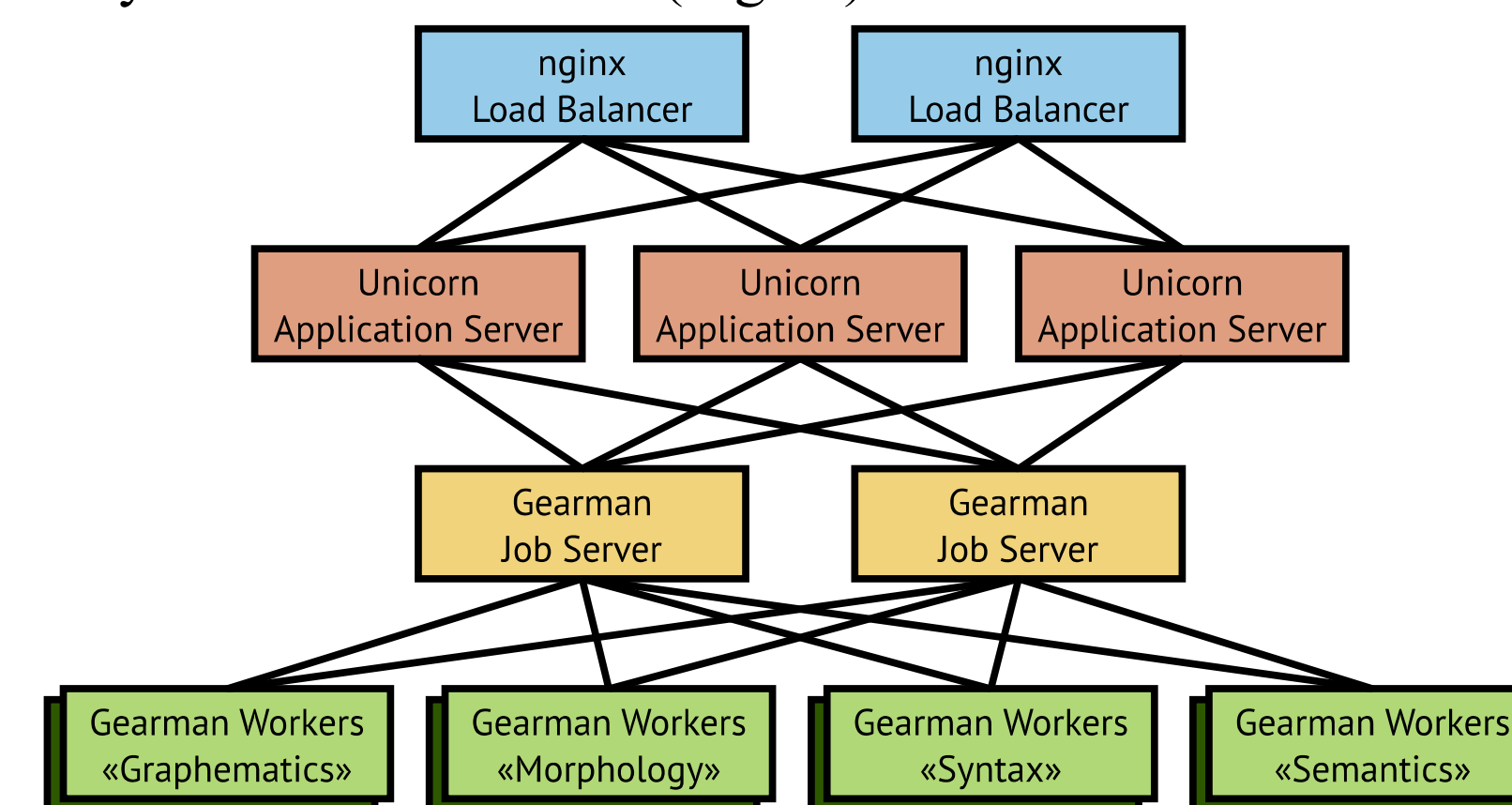


Figure 2. Architecture of RLPCS.

To achieve high scalability and to protect from hardware failures we use the cluster of independent servers of standard hardware architecture, which work under the GNU/Linux operating system.

Text processing actions are organized by Gearman application framework to perform parallel evaluation, load balancing and procedure calling in different programming languages.

Realization

Current implementation includes Graphematical and Morphological analyzers.

Graphematical analyzer is based on finite state machine (FSM). Input alphabet of this FSM is a set of Russian letters in UTF-8 encoding, Arabic digits, separators, in-sentence punctuation marks, punctuation marks and End-of-Line/End-of-File signs. The result of graphematical analysis is a graph representing the input text structure elements and relations of them.

Morphological analyzer in RLPCS is based on Mystem algorithm and the structure of morphological dictionary is based on morphological modules from the DIALING project.

Performance Evaluation

Performance evaluation of morphological analyzers was made on cluster with 4 nodes (CPU AMD Opteron 2218 Dual Core 2.6 GHz, 8 GB memory, Ubuntu Linux) connected by Gigabit Ethernet. We compare performance of our morphological analyzer “myaso” with Mystem from Yandex and Snowball stemmer (Fig. 3 and 4.).

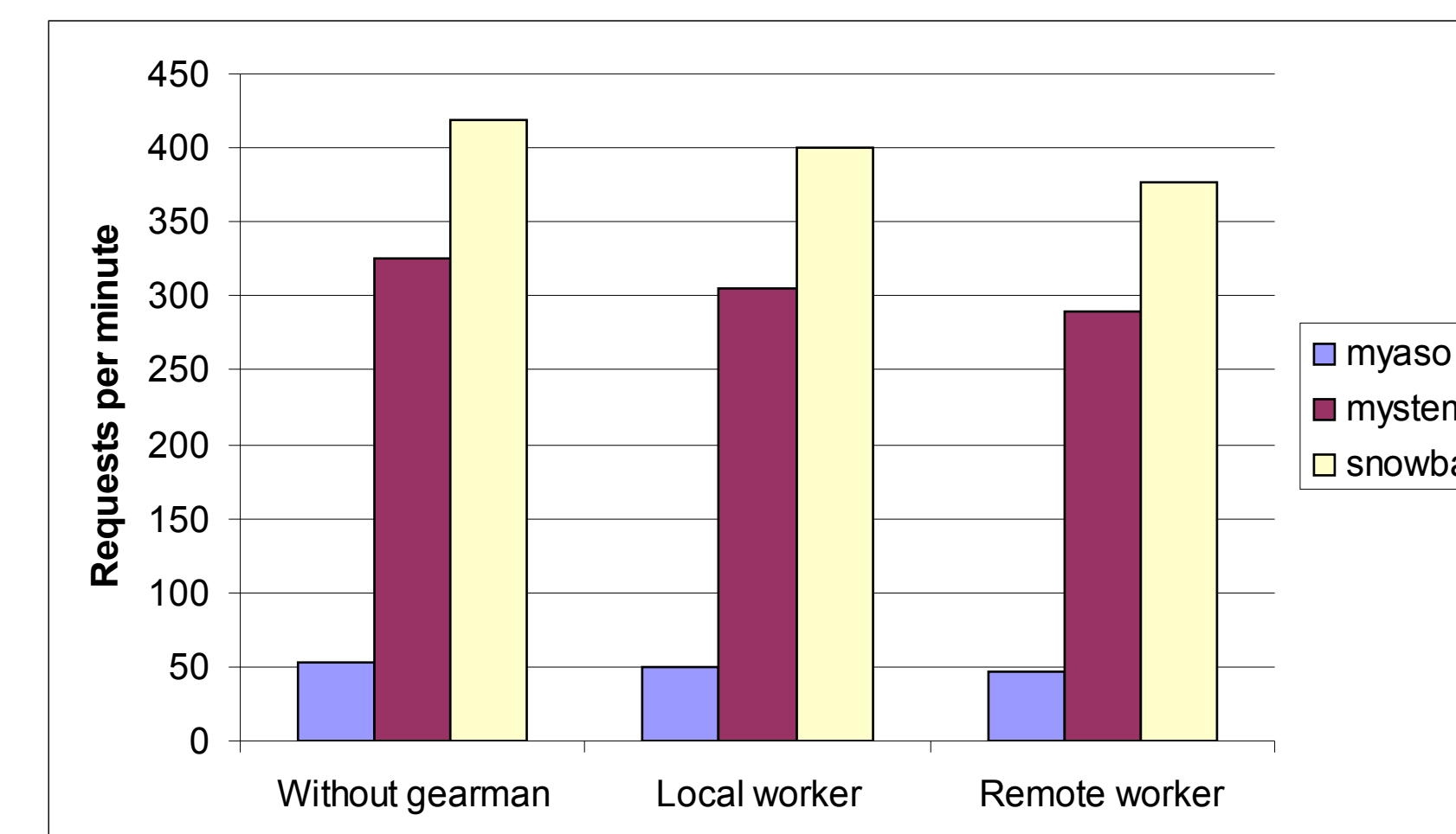


Figure 3. Gearman overhead estimation.



Figure 4. Morphological analyzers scalability evaluation.

The performance of non-parallel morphological analyzer working on only one CPU core was near 50 requests per minute. Cluster morphological analyzer implementation, based on Gearman, provides approximately 710 requests per minute. Testing shows that Gearman framework has relatively low overhead: 5% with local worker and 12% with distributed worker. As a result, Gearman provide effective scaling of morphological analyzer in a cluster configuration.

Summary and future work

We have described the RLPCS, a Russian Language Processing Cloud Service: presented features and architecture of Web-service, explained the current implementations of graphematical and morphological analyzers.

The demonstration of RLPCS is available at URL: <http://rlp.imm.uran.ru/>

Areas of further development are:

- 1) Implementation of syntactic and semantic analyzers.
- 2) Increasing the quality of graphematical analyzer by using better algorithms and methods.
- 3) Architectural improvements in parallel text processing in cluster configurations.

The advantage of RLPCS is the possibility of processing Russian language texts without developing own software and hardware infrastructure, suitable to that processing, which may significantly decrease the material and time costs.

Literature cited

- I. Segalovich, “A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine,” in *Proceedings of the International Conference on Machine Learning: Models, Technologies and Applications. MLMTA '03*.
- A. Sokirko, “Semantic Dictionaries in Machine Translation System (by materials of the DIALING system),” Ph.D. thesis, Russian State University for the Humanities, 2001.
- I. Nozhov, “Implementation of Automatic Syntactic Segmentation of Russian Language Sentence,” Ph.D. thesis, Russian State University for the Humanities, 2003.

For further information

Please contact dau@dpt.ustu.ru or avs@imm.uran.ru. More information on this and related projects can be obtained at rlp.imm.uran.ru