# FactBook Semantic Technologies[1]

## Introduction

**FactBook** uses semantic technologies to understand users' current interests and collect appropriate facts on their behalf. The underlying deep structure learning algorithms presumably mimic the way children acquire their native language.[2] This Whitepaper presents the basics of our approach.

## Overview

Machine learning of natural languages was for a long time represented mostly by supervised learning. Maybe due to the vast bulk of knowledge already accumulated in linguistics. Learning made use of large linguistic corpuses, trying to find formal rules reproducing correct parsing provided by skilled linguists and using existing linguistic knowledge.

The resulting parsers were overcomplicated and could parse only several dozen sentences per second (~ 1 kB/s). Which was four orders of magnitude less than state-of-the-art search engines indexing speed (10 MB/s). Thus, practical semantic indexing at scale was prohibited. Situation changed recently, with parsing speed up to 100 and even 1000 sentences[3,4] per second.

However, most parsers still utilize supervised learning, with corpuses available only for some languages in a certain domains. Most scientific and blogosphere jargons are out of reach of such approach.[5] Thus, deep unsupervised learning without linguistic priors attract more and more attention in recent years.[6]

We seek to learn language from scratch in unsupervised fashion, without dictionaries or any other prior linguistic knowledge. The resulting "semantic processor" is capable of parsing syntactical and semantic structures of sentences as fast as 2000 sentences per second (~ 300 kB/s).[7]

Contrary to the mainstream approach based on distributed representations in deep and recurrent networks[8,9] our model uses extremely sparse representations, namely recurrent correlation analysis. Linguistic rules form a large (open) set of mutually competing linguistic patterns, represented as binary trees, based on correlation statistics. It is fully applicable to any language and any domain.

## Language model

Supervised approach describes language, using precompiled dictionaries and linguistic corpuses. We aim to extract the "language model" from scratch, making use of the abundance of raw texts.

---

[1] Sergey Shumsky, serge.shumsky@gmail.com

[2] S.A.Shumsky, "Brain and Language: How we Understand Speech", in "Approaches to Cognition Modeling", V.G.Red'ko (Editor), 2014, ISBN 978-5-9710-1049-4

[3] "Large-Scale Syntactic Processing: Parsing the Web", Final Report of the 2009 JHU CLSP Workshop

[4] Alexander Volokh and Günter Neumann (2012) Task-oriented Dependency Parsing Evaluation Methodology, IEEE 13th International Conference on Information Reuse and Integration, IEEE Systems, Man, and Cybernetics Society (SMC), 2012

[5] Say in biology various genes and proteins have a quite different biological "meaning".

[6] Schmidhuber J. "Deep Learning in Neural Networks: An Overview". arXiv:1404.7828 v3 [cs.NE], 2014

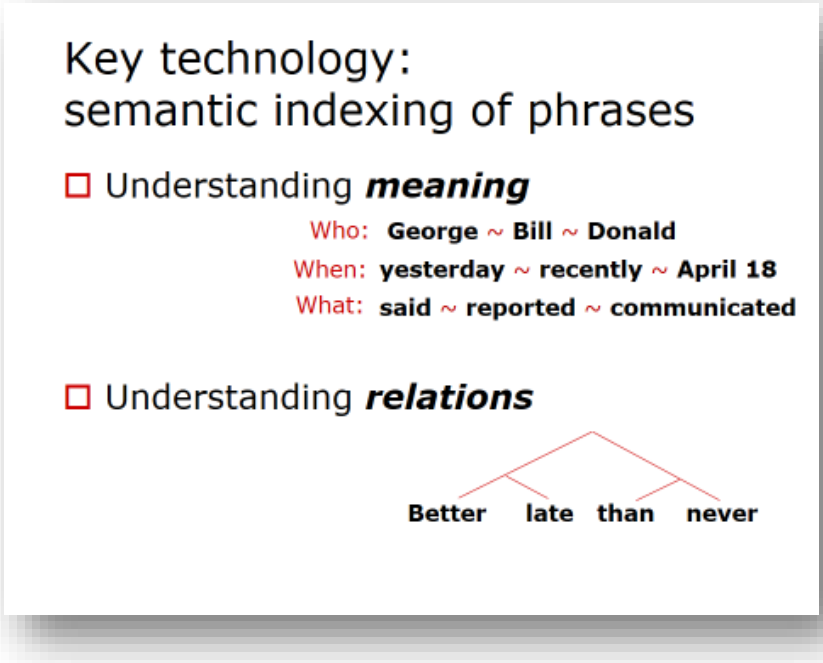
[7] Semantic indexing of English Wikipedia takes 15 hours on a single server.

[8] R.Collobert, et.al. "Natural Language Processing (Almost) from Scratch". J. of Machine Learning Research 12 (2011) 2493-2537

[9] T.Mikolov, I.Sutskever, K.Chen, G.Corrado, J.Dean "Distributed Representations of Words and Phrases and their Compositionality" (NIPS 2013)

Namely, we seek to:

- Understanding **meanings**. Words and phrases have close meanings, if they are used in a similar way (e.g. names, surnames, dates, cities, countries, etc.). To that end, we exploit statistics of usage of linguistic variables in that or another context.
- Understanding **relations** - binding of words and phrases in a sentence. (E.g. name and surname, subject and its action, etc.). Each successive binding in our model increases the probability of similar bindings in the future.



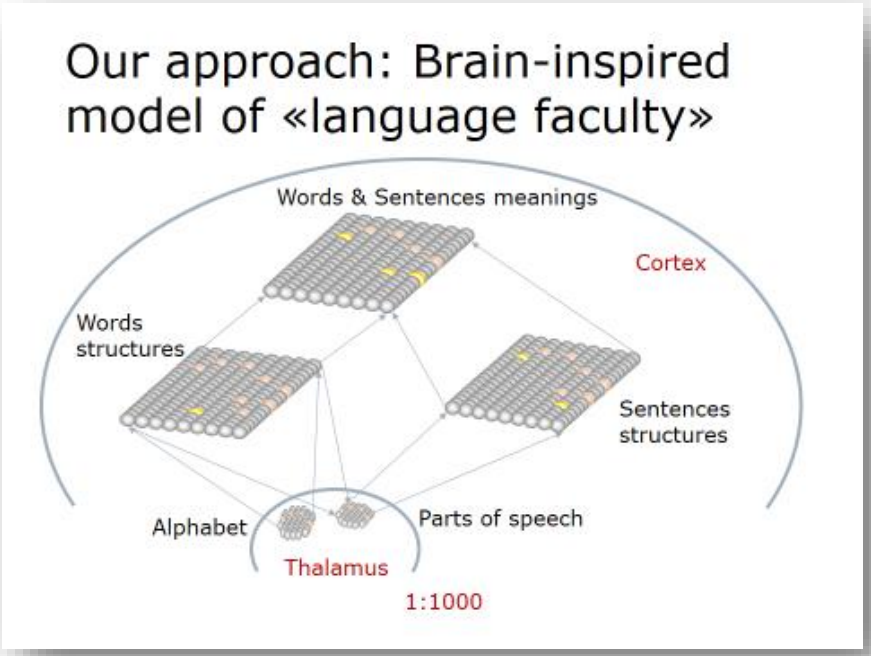

## Golem's language model

Children can easily acquire any language, since all human languages have adapted themselves to our mental capabilities. Thus, it is natural to construct language model based on what we already know about the algorithms of human brain.[10]

In particular, we know that all our experience is stored in the neocortex in a set of interconnecting 2D self-organizing feature maps. We built our model of human "language faculty" having in mind recursive self-organizing maps, as possible correlation analysis machinery in the neocortex. We also supposed different parts of neocortex to process morphological, syntactical and semantic information.

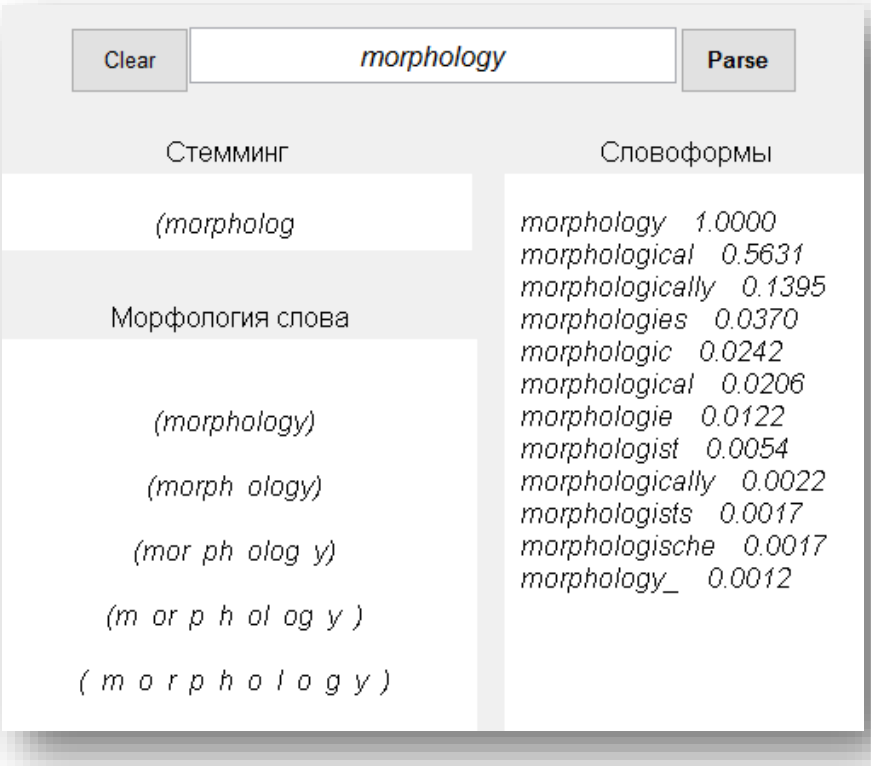
The resulting language model "Golem" is to our best knowledge the first biologically viable model of unsupervised language learning "from scratch". It has several modules, using the same set of basic learning algorithms in a hierarchical manner - to find out and encode typical linguistic patterns.

[10] S.A.Shumsky, "Reverse Engineering the Brain Architecture: Role and Co-operation of the Main Subsystems", in: Lections on Neuroinformatics, pp.13-45. (Neuroinformatics-2015), Moscow, 2015.




## Golem's morphology module

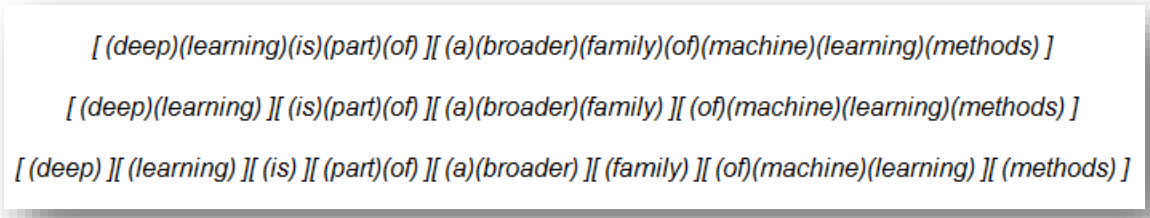
Morphology module learns to merge letters iteratively into morphemes and words. The resulting morphological structure of the words is used by syntactic and semantic modules, e.g. for stemming and part of speech analysis.

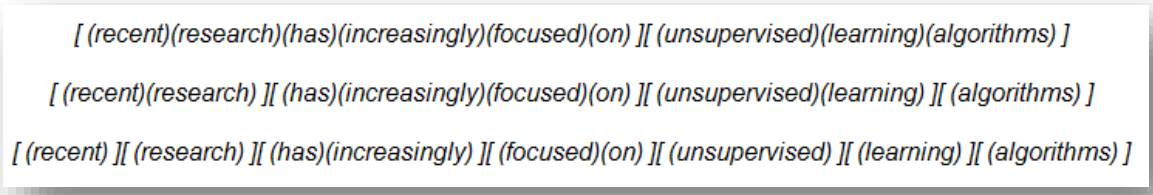


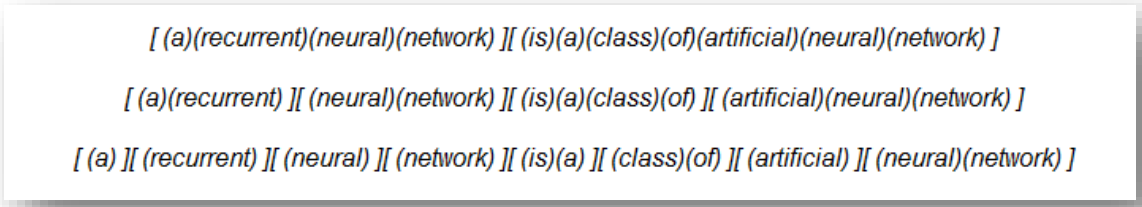
The above figure illustrates the merging procedure, stemming and various word forms of the word '*morphology'*, learned in unsupervised manner.

## Golem's syntactic module

In a similar fashion, the syntactic module learns various patterns in parts of speech usage in sentences. The figures below illustrate the iterative procedure of corresponding bottom up sentences parsing.



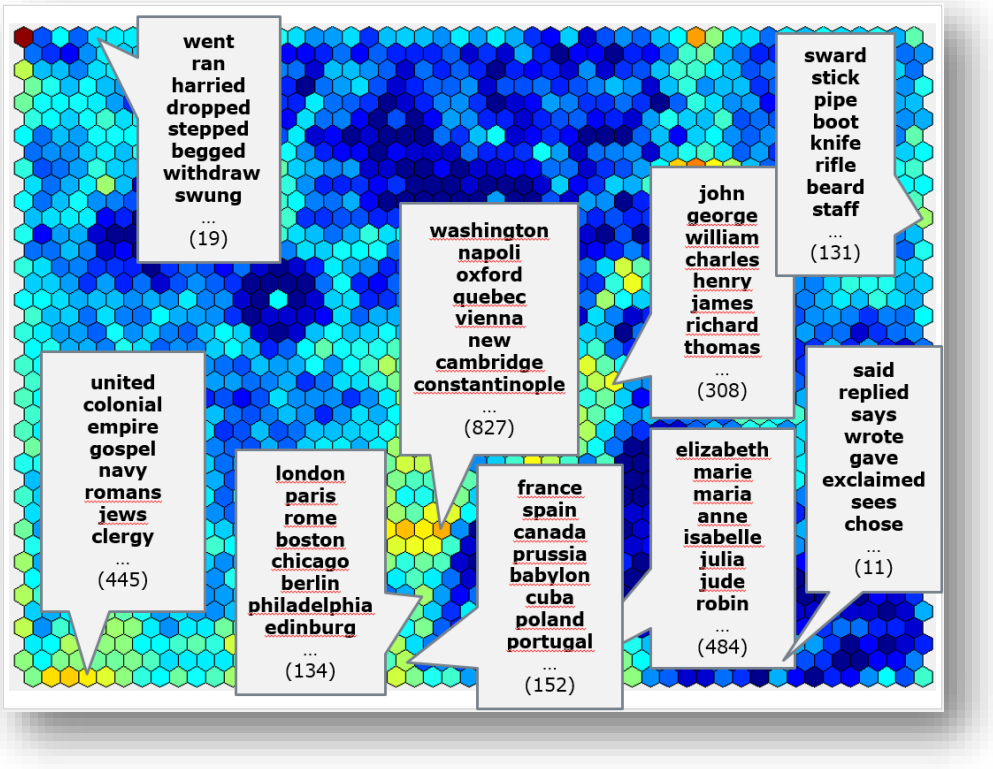




## Golem's semantic module

Semantic module encodes the meanings of the words, phrases and sentences. The starting point is the semantic space for words, illustrated in the next figure. Each «semantic cell» on that map represents one of the basic meanings. For some cells the words with the closest meanings are shown. (The words in Golem may have several meanings, i.e. belong to several cells.)

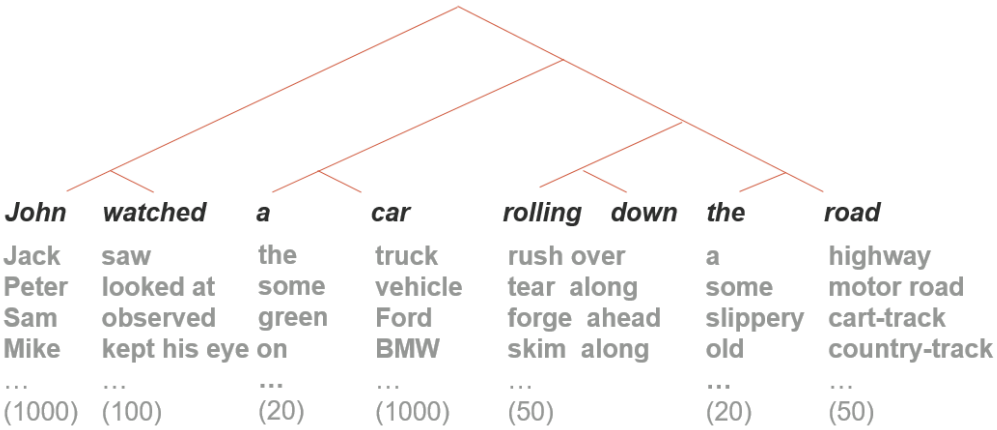
To represent the meaning of phrases and sentences we extend this initial semantic space, adding new meanings in the course of learning as most frequently used combinations of already known ones.

Since new meanings don't change the existing ones, such unsupervised semantic learning may proceed in the course of semantic indexing. Thus, Golem gradually increases its semantic knowledge, reading new volumes of documents in various domains.




## Semantic search

Semantic indexing may be used, e.g. for semantic search, which takes into account deep linguistic structures, describing the general 'scenarios': who makes what with whom, etc. There are zillions of ways to express one and the same deep structure using different words. Golem provides compact sparse representation for the meaning of sentences, allowing fast semantic search, using inverted semantic index.




***1 deep structure***: 1000 × 100 × 20 × 1000 × 50 × 20 × 50 = **$10^{12}$ surface variants**

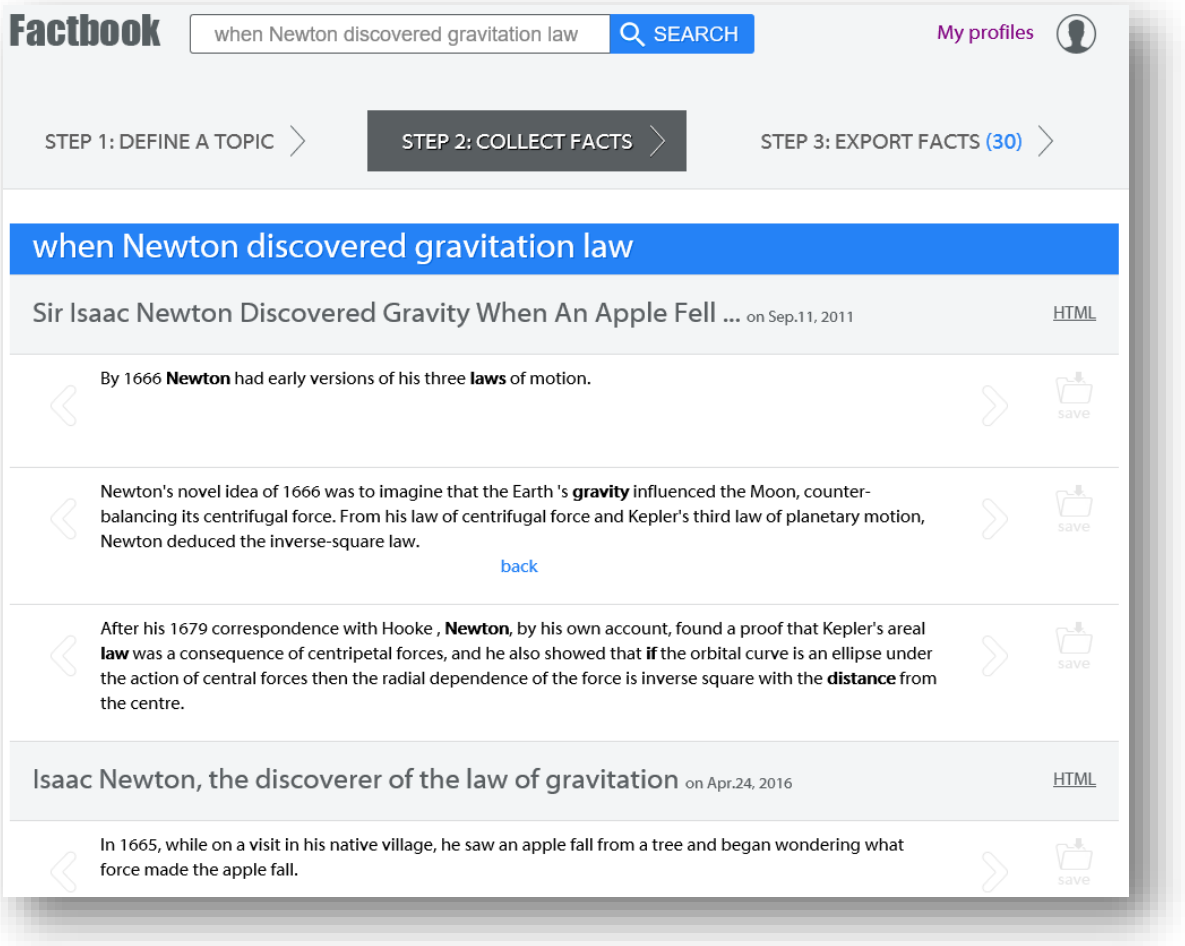
In a way, semantics describe typical 'scenarios', where each 'role' may be played by that or another word. Our inverted index contains both words and their roles in scenarios.

## Collecting facts with semantic search

FactBook uses semantic search (together with the keyword search) to collect facts on a given topic from numerous documents. The topic of interest is automatically detected, based on the facts, already selected by the user.

Understanding the meaning of facts helps the system to grasp the interest profile from a handful of examples. The more facts user saves in his archive, the better are the new facts, extracted from the documents collection.

The following figure illustrates the facts, selected for a user, interested in the history of Newton's great discovery.




FactBook thus provides easy to use interactive tool for collecting facts from the Web or other large documents collections.

## Conclusion

In conclusion, we developed biologically inspired unsupervised learning algorithms, capable of learning language model from scratch. This model is nonparametric, i.e. it can continuously learn, extending its semantic knowledge, and improving the quality of semantic search.