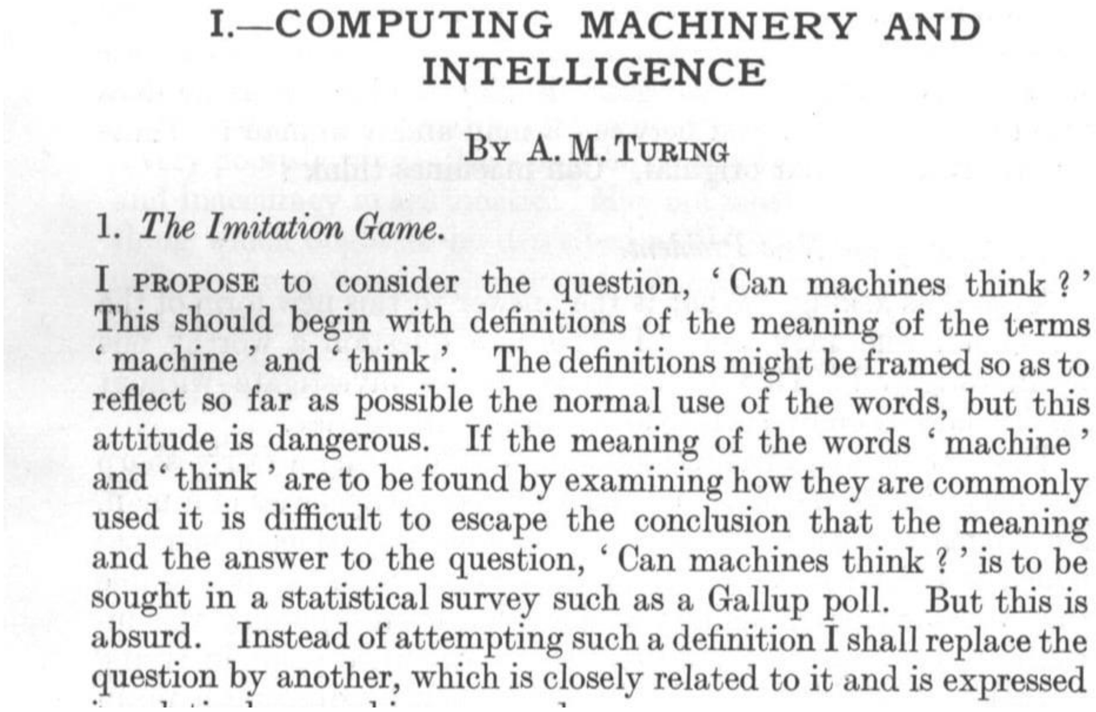
What is Natural Language Processing and what are its problems with different languages

> Natural language processing is a modern technology used in IT sphere to make applications and services that can work with various data (primarily text and voice data) in natural languages such as English, Spanish or Russian. But how does NLP work and what problems do companies face when developing NLP-powered apps and services?

What is Natural Language Processing

> First of all, let’s skin-deep discover how does NLP work and what steps do NLP algorithms perform to work with human voice and text data? We won’t dive too deep into Math and Programming of such NLP algorithms, but we will need basic overview to understand problems this sphere faces later.

Many hyped technologies move the IT industry forward nowadays. Some of them are: AI, Machine Learning, AR and VR, Cloud Computing and Streaming. Natural Language Processing is a field of computer science and linguistics that takes its roots in an article “Computing Machinery and Intelligence” published in 1950 by Alan Turing. In that very article, Alan Turing proposed the idea of a test nowadays known as the Turing test.



I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, ‘ Can machines think ? ’ This should begin with definitions of the meaning of the terms ‘ machine ’ and ‘ think ’. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words ‘ machine ’ and ‘ think ’ are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, ‘ Can machines think ? ’ is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed

NLP is all about giving computers the ability to understand text and spoken words in much the same way human beings can. It refers to processes of analysis, processing, and generation of statements in natural languages. That allows people to interact with computers with voice or text instead of using graphical interfaces or console commands.

Natural Language Processing usage

> Where can we see NLP algorithms? What impact does NLP have on our everyday lives and enterprise?

NLP stands behind computer programs that translate text from one language to another, respond to spoken commands, or summarize large volumes of text rapidly — even in real-time. Most probably, you’ve at least once interacted with NLP in the form of voice-operated GPS systems, digital assistants, speech-to-text dictation software or customer service chatbots.

The most known examples of Natural Language Processing technologies application to software are voice assistants such as Microsoft’s Cortana, Apple’s Siri, or Amazon’s Alexa. Virtual assistants and chatbots use speech recognition to recognize patterns in voice commands and natural language generation to respond with an appropriate action or helpful comments. Chatbots perform the same magic in response to typed text entries. The best of these are able to recognize contextual clues about human requests and use them to provide even better responses or options over time. The next enhancement for these applications is question answering, the ability to respond to our questions with relevant and helpful answers in their own words.

Many GPS applications such as Google Maps or Yandex Maps use NLP to generate speech directions they give to the driver.

Most of nowadays email services use NLP to filter spam in our inbox. The most known of them is Gmail. Companies use NLP algorithms to search for indicators of spam or phishing, such as overuse of financial terms, characteristic bad grammar, threatening language, inappropriate urgency, misspelled company names, and more [4]. Spam detection is one of a handful of NLP problems that experts consider 'mostly solved'.

NLP has become an essential business tool for uncovering hidden data insights from social media channels. Sentiment analysis can analyze language used in social media posts, comments and reviews to extract attitudes and emotions in response to products, promotions, and events–information companies can use in product designs, advertising campaigns, and more.

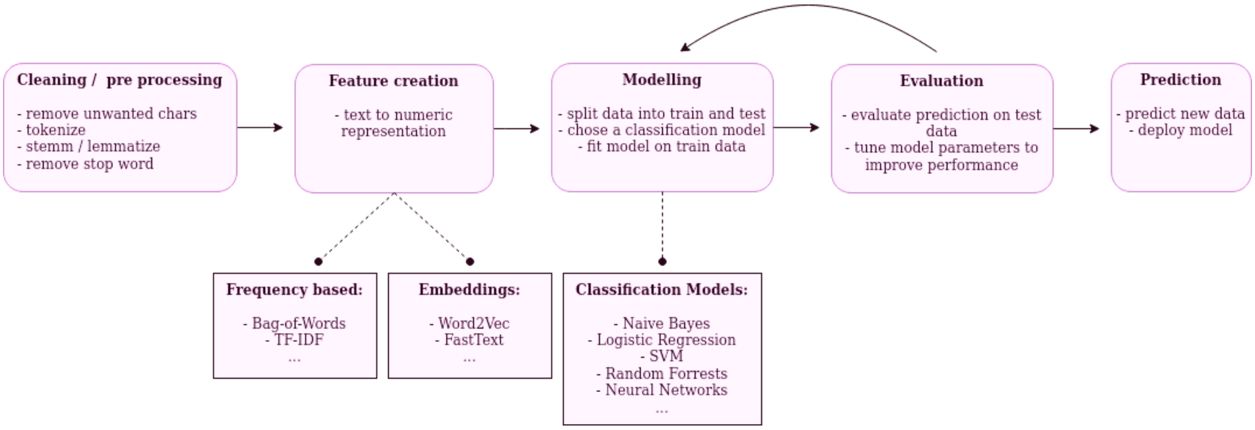
Text summarization uses NLP techniques to parse huge volumes of digital text and create summaries and synopses for indexes, research databases, or busy readers who do not have time to read full text. The best text summarization applications use semantic reasoning and natural language generation (NLG) to add useful context and conclusions to summaries.

NLP also plays a growing role in enterprise solutions that help streamline business operations, increase employee productivity, and simplify mission-critical business processes, what is not always seen from the outside.

Main algorithms of NLP

We have discussed what is NLP and how is it used in nowadays industry. But how do NLP algorithms work and what steps do they perform to understand our speech and text?

Human language is filled with irregularities that make it incredibly difficult to make a program that would accurately determine the intended meaning of text or voice data. There is a bunch of problematic cases that take humans years to learn. However, a program must understand them with a single circuit.



Since its creation, NLP had been developing as an emulation of natural language understanding with a collection of preset questions and answers aka rules. That type of NLP is known as **Symbolic NLP**. And it was the only way till the early 80s.

Starting in the late 80s, a revolution happened, caused by the introduction of machine learning algorithms for language processing. NLP that uses machine learning is now called **Statistical NLP**. Those NLPs allowed development of machine translation, used nowadays in law, diplomacy, and everyday life.

Nowadays, **Neural NLPs** are a thing. They use deep neural network-style machine learning to achieve state-of-the-art results in many tasks such as language modeling and parsing.

NLP can be broken down into several steps needed to break down human text and voice data into data that computer can understand. Some of the tasks include the following:

Speech recognition (also known as called speech-to-text), is the task of reliably converting voice data into text data. Speech-to-text conversion is necessary for any application that uses voice commands or answers spoken questions. The way people talk: quickly, slurring words together or with different accents and incorrect grammar makes task challenging.

Next step in NLP is **Speech tagging** (also known as grammatical tagging). It is the process of determining the part of speech of a particular word or piece of text depending on its use and context. This step identifies, for example, the word ‘make’ as a verb or as a noun.

When decided on part of speech for each word, the algorithm must select meaning of a word with multiple meaning using **Semantic analysis**. For example, word sense disambiguation helps distinguish the meaning of the verb 'make' as in ‘make the grade’ (achieve) or ‘make a bet’ (place).

After that program uses **Named entity recognition** algorithm to identify useful pieces of information, such as ‘Moscow’ as a location or ‘John’ as a man’s name.

Co-reference resolution is the process of identifying when two words refer to the same object. For example: when ‘she’ is ‘Mary’ from the previous sentence. This process also involves identifying metaphors or idioms.

Sentiment analysis tries to extract subjective qualities – attitudes, emotions, sarcasm, confusion, suspicion — from the text.

Natural language generation is the opposite of speech recognition or speech-to-text. It is the task of putting structured computer data into human language.

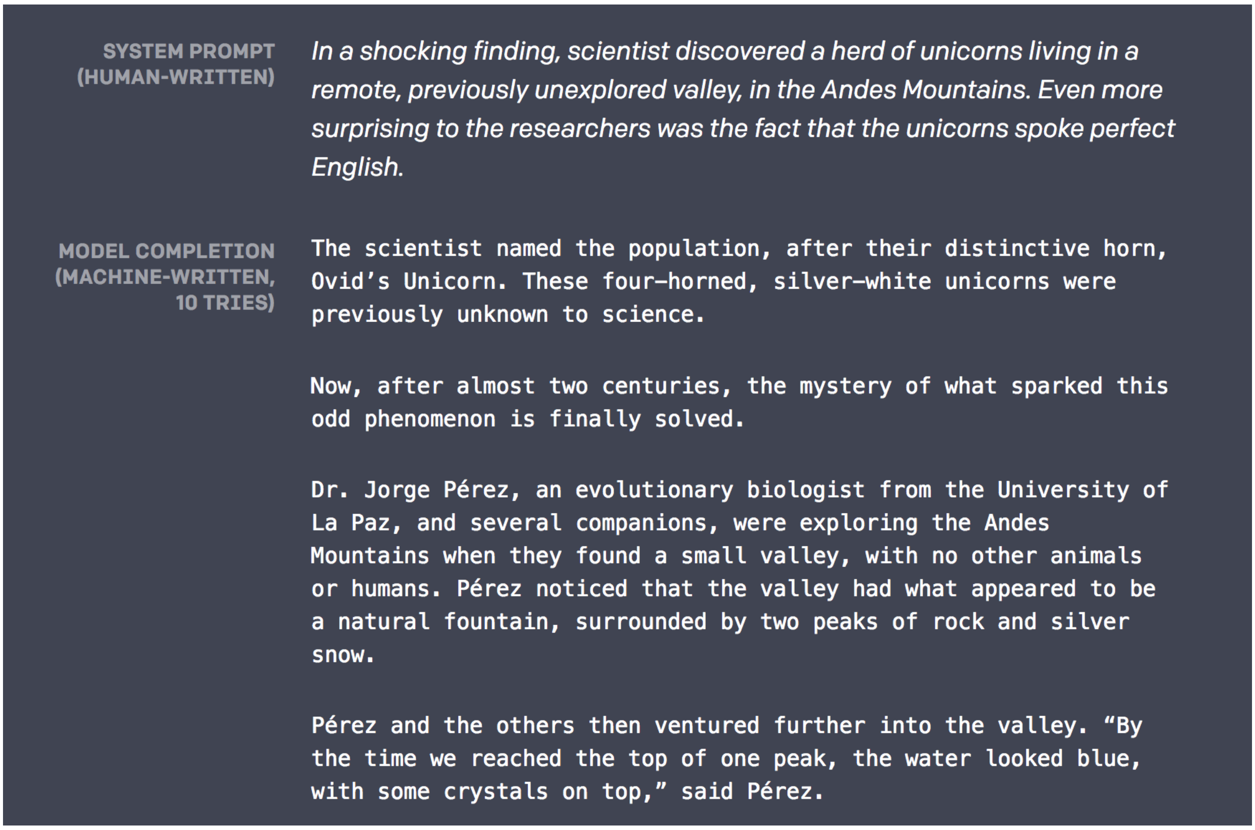
NLP problems with different languages

> For various reasons NLP haven’t been instantly integrated everywhere on Earth. What’s the reasoning for that? What problems do developers face when developing NLP-powered applications for markets outside of UK and US?

Italian

Even for languages that are as close to English as Italian is, time gap before integration of new models is enormous. In 2018 OpenAI presented their GPT-2 model for English language. GPT-2 has a whopping 1.5 billion parameters (10X more than the original GPT) and is trained on the text from 8 million websites.

It was the first NLP model that could generate stories about talking unicorns:



SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

```
The scientist named the population, after their distinctive horn,
Ovid’s Unicorn. These four-horned, silver-white unicorns were
previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this
odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of
La Paz, and several companions, were exploring the Andes
Mountains when they found a small valley, with no other animals
or humans. Pérez noticed that the valley had what appeared to be
a natural fountain, surrounded by two peaks of rock and silver
snow.

Pérez and the others then ventured further into the valley. “By
the time we reached the top of one peak, the water looked blue,
with some crystals on top,” said Pérez.
```

However, the first GPT-2 based model for Italian was released only in 2020 [7]:

> The availability of GePpeTto opens up substantial possibilities. In the same way that GPT-2 is changing the approach to several NLP English tasks, we can expect GePpeTto to serve a similar purpose in Italian language processing.

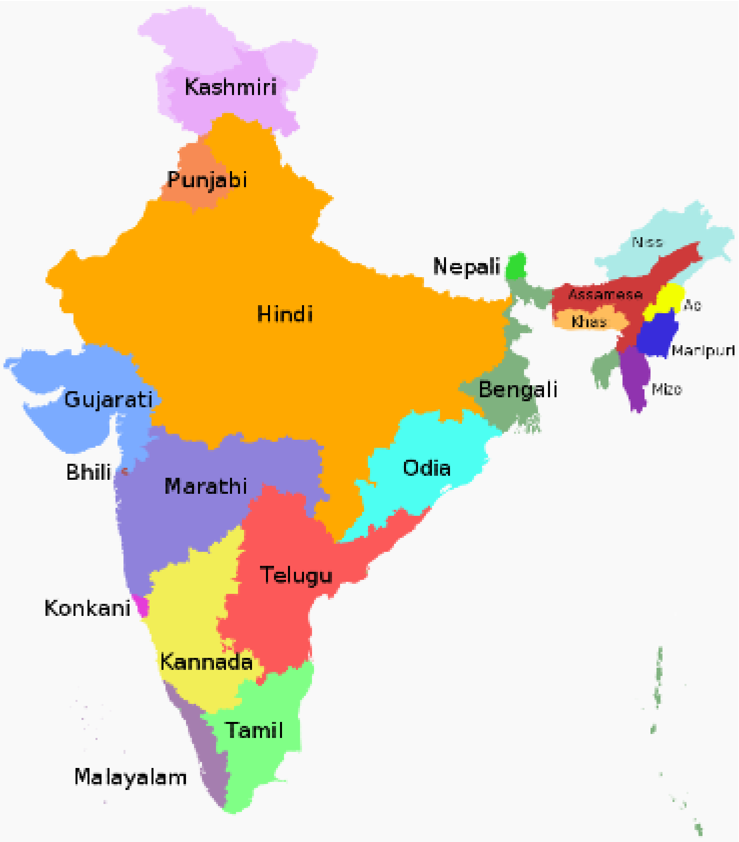
Another important aspect to understand is costs for businesses to use different NLP models for each language. Even though well-optimized English corpus model can understand French, Deutsch or Spanish, languages from Slavic group or Arabic would suffer greatly.

Overall, these are the problems with NLP in Italy and other smaller first-world countries:

- Variety in languages as theirs vocabulary, grammar and other rules;
- Small to medium population often not comparable to that of English-speaking world. For example, Polish is native language for only 42 million people [1] compared to 379 million for English.

Those listed problems are only obstacles that cause extra costs for developers. Overall, Europe and smaller first-world countries are profitable markets for developers. Nevertheless, companies tend to save money and directly translate models developed for English into other languages, such as Italian, and that often causes bad user experience.

Indian languages



Only 10% of Indian population speaks English. Most Indians have Hindi as their first language, followed by Marathi, Telugu, Punjabi, etc. Moreover, each one of them has many different dialects.

The most known problems for NLP in Indian languages are:

1. Ambiguity at different levels — syntactic, semantic, phonological ambiguity, etc.
2. Dealing with idioms and metaphors.
3. Dealing with emotions.
4. Finding referents of anaphora and cataphora.
5. Understanding discourse and challenges in pragmatics.

Issues in dealing with Hindi:

1. Scarcity of annotated corpora and tools.
2. Lack of education and training institutes.
3. Large set of morphological variants.

Many datasets available at authentic sources like Kaggle, UCI machine learning repository, etc. are available only in English. Limited datasets, corpora, and other resources are a huge hurdle. For Hindi is only one list of stop words having 184 words in total, released for the public back in 2016.

Overall, these are the problems with NLP in India:

- Many different and difficult languages;
- Poor population.

Therefore, it's unprofitable for big companies to invest into NLP for this market. Those factors can also be the reasons for the famous "indian programmist" as 77% of Indian IT market works for export [5].

Russian

Despite recent advances in machine learning, developing Russian natural language processing (NLP) systems remains a big challenge for researchers. Russian, like other Slavic languages, is a morphologically rich language with free word order and inflection. These linguistic factors make it difficult to collect enough relevant multilingual data to train machine-learning models.

Different companies try to put together an exhaustive list of the best Russian datasets available on the web, covering everything from social media data to natural speech:

> The RU-EVAL 2014 has brought together a number of IT companies and academic groups that work on Russian NLP tasks (pos-tagging, parsing, anaphora and coreference resolution), and made it possible to assess the state-of-the-art in the field (so far, mostly in Russia). The forum has shown that there are competitive teams that develop NLP components on a considerably high level. However, these tasks have some peculiarities and complications due to high inflectional and fusional properties of Russian language, its free word order and the absence of overt definiteness markers for NP. The absence of free semantic resources such as WordNet and freely distributed syntactic parsers make the task even more difficult for newly organized NLP small teams. However, the event was the challenge for those teams that conduct the experiments on various machine-learning techniques.

However, Russian local IT industry is capable of developing NLP algorithms and apps for the market. One of such companies is Yandex that spends $325M on R&D in AI annually. Voice assistant "Alice" is their biggest project in AI. Yandex's main competitor - Mail.ru Group - has developed their voice assistant called "Marusya". Company has $5B capitalization [6]. Many banks such as Sberbank or Tinkoff are using AI to futher promote their products and increase user experience in banking apps.

These companies tend to use datasets collected from their own services in other fields. For example, Yandex uses data from their Yandex.Maps or Toloka services [3]. From official websites you can download datasets such as “Toloka Aggregation Features”. It is described as:

> This dataset includes 60,000 crowdsourcing evaluations collected in Toloka for 1000 tasks, including correct answers for most tasks. The task was to classify websites into five categories based on whether they contain NSFW content. Additionally, for each task dataset includes 52 indicators that can be used for category prediction.

Overall, these are the problems with NLP in Russia:

- Difficult language;
- Limited datasets for models training.

There are some obvious setback in developing NLP for Russia. But they mostly open an opportunity for local companies such as Yandex or Mail.ru to outcompete international IT giants.

One of many problems with NLP is an abundance of different languages used by humanity. English is one of them. This issue is especially important for developing countries or nationalities with small population talking a language. Other common problem is lack of sufficient online data to design such NLP algorithms. Development of NLP for a particular language largely depends on local IT industry capital.

Conclusion

After all, NLP is one of the many frontiers that move technology forward. One of the bigger problems with this field of computer science is differences between languages, which make it near impossible to adapt existing models to new languages. Another one is its dependency on the society and the economy of any given country. Nationalities with small and poor populations will not see NLP in everyday use in near future.

References

S. Toldova, O. Lyashevskaya, A. Bonch-Osmolovskaya and M. Ionov Evaluation for morphologically rich language: Russian NLP. Int'l Conf. Artificial Intelligence, ICAI'15, pp 300-306 ICAcombined.pdf (worldcomp-proceedings.com)

Primer | Russian Natural Language Processing

NLP for Indian Languages. Join me in exploring the need, the… | by Navkiran Singh | Towards Data Science

[7] GePpeTto Carves Italian into a Language Model paper_46.pdf (ceur-ws.org)

NLP (Natural Language Processing) Tutorial: What is NLP & how it works (mygreatlearning.com)

GPT-2: Understanding Language Generation through Visualization | by Jesse Vig | Towards Data Science

India Languages - Demographics (indexmundi.com)

The Problem With The English Language In India (forbes.com)

14 Best Russian Language Datasets for Machine Learning | Lionbridge AI

Red Hen Lab - Russian NLP

Using CoreNLP on other human languages - CoreNLP (stanfordnlp.github.io)

What is Natural Language Processing? | IBM

Natural Language Processing (NLP) – What Is NLP & How Does it Work? (monkeylearn.com)

Плавное введение в Natural Language Processing (NLP) (datastart.ru)

Основы Natural Language Processing для текста / Блог компании Voximplant / Хабр (habr.com)

[1] Estimate of the number of native Polish speakers (jakubmarian.com)

[2] The Most Spoken Languages in The World in 2020 - Speakt.com

[3] https://toloka.ai/ru/datasets

[4] Machine learning for email spam filtering: review, approaches and open research problems - ScienceDirect

[5] IT & BPM Industry in India: Market Size, Opportunities, Growth, Report | IBEF

[6] AI Report - Review