Natural Language Processing

and processing of

Google Search queries

Volodin Andrei

January 2020

ABSTRACT

The paper represents itself an attempt to research and understand the situation with the recently emerged phenomenon: Natural Languages Processing (NLP). Within the work on the project an investigation on use cases of application of NLP has been conducted. Works of leading scholars in the area have been analysed, and it was determined that in the western world Stanford University takes a leading position in the development of NLP methods. However, Microsoft Asia research has made even more significant breakthrough with the technology development, so that the magnitude of outcomes of eastern divisions of Microsoft Research dramatically underperforms outcomes of same corporation divisions in Asia.

Moreover, the research on NLP that was conducted within the context of the work on the essay writing includes an attempt to study corresponding background technologies which are beyond and utilised by NLP methods. Also an attempt was made to investigate how Google Search uses NLP for processing of queries. Researched subjects include but are not limited to the listed below: Machine Learning(ML), Mathematical methods, Artificial Intelligence (AI), Word2Vec, Seq2Seq, Stanford CoreNLP, Stanford Question Answering Dataset, Google BERT, Softmax, etc. Some practical implementations of NLP algorithms have been tested within Unix environment. And the research has shown that sophisticated designs of NLP algorithms can be used for the purpose of implementation of AI based journalism, as well as for the purpose of answering of search queries addressed to Google Assistant, and specifically in cases when speech-to-text conversion takes place, i.e. a voice search request is received by a device powered by Google and then the request is translated to a text form and a response is formed to address the request which can be either displayed on a device or pronounced aloud using the device's speakers.

The result of the research shows that together with Computer Vision AI, speech processing function will enable robots to operate in lieu of Journalists in dangerous areas, and in the near future it will become possible to survey and interview participants or observe events using aerial drones, wheeled and other mechanisms equipped with corresponding software and hardware.

Introduction

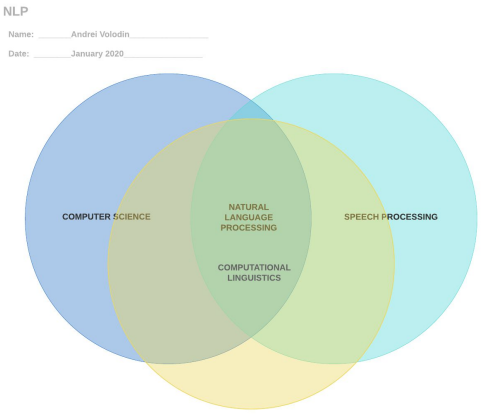
Originally, development of technologies that followed the technological revolution, allowed to accumulate significant resources that can be dedicated to this or that research and result in a breakthrough in this or that area. Corporations, institutions and individuals got connected via computer networks and that allowed to combine their efforts in resolving this or that issue almost instantaneously. It has led to a breakthrough in multiple areas of communications and computer science development. As a result of such breakthrough multiple new disciplines and scientific projects mushroomed and among them were projects developing computer speech processing, robotisation of computer operations, artificial intelligence, recognition and classification of data including audio-visual data, etc.

Since recent due to its success in processing queries and due to wide coverage with its services, Google managed to introduce not only worldwide services that can be used with Desktop computers, but also mobile operating system Android that supports sophisticated algorithms for Natural Language processing and can recognise voice commands from users and pass them to the Google Search engine with further delivery of the response in audio or visual form back to the user. Some of the queries can be answered offline and other require an internet connection in order to be addressed. In the essay I will try to investigate how the processing of requests from user happens and what mechanisms are involved in addressing the queries and constructing a response that will be delivered to users.

Definitions

Below is presented a short list of terms and definitions that will construct the context for understanding how speech recognition works and how Google Search requests are processed.

Natural Language Processing [NLP]: It has many corner terms that have similar meaning: Computational linguistics, Speech Processing, Computer Science. However, it is an interdisciplinary field that is aimed in making computers to understand and operate with human language. Its relation to its closest neighbours is depicted in Venn's diagram below.



Machine Learning [ML]

This new field in computer science is aimed in an attempt to make computers behave as humans in order to perform certain tasks. The construct is based on training computes on datasets in a way computers with certain extent of probability will be able to recognise elements of this or that set. It is mostly implemented with the use of neural networks that are similar to binomial trees and graphs. With the utilization of mathematical statistics it becomes possible to implement predictions in a way an algorithm will output a predicted event and that will likely be an equivalent to an actually happening event. For example, weather forecasting based on previous data can be a task that can be implemented with ML. And looking at continuous statistics of development of weather conditions computer will be able to forecast further tendencies in weather and indicate e.g. what temperature is expected tomorrow or next week. Another and more complex example is people who are trying to predict development of situation in financial markets with the use of sophisticated mathematical algorithms and for a purpose to get profit from market fluctuations. Also what can be in focus of prediction and modelling in scientific area is distribution and development of epidemias of diseases, or natural and environmental processes and many types of events that can be modelled.

Crowdsourcing

Emergence of the Internet allowed to crowdsource multiple huge projects and dataset that were labeled by many workers. For example crowdsourcing mechanisms as Amazon MTurk and similar online platforms allowed to collect big datasets of labeled images that allowed to get a breakthrough in computer vision field. Based on the accumulated dataset it has become possible to run sophisticated algorithms on this data and get some prediction models trained in a way they could do labeling and recognition of images in a way similar to as a human does. Basically crowdsourcing allowed to get computer programs to operate as humans labeling images, and also crowdsourcing is a major factor for ML development

Artificial Intelligence [AI]

According to J. Russell (2016), the main theme in AI is an idea of intelligent agent. AI is still an emerging field which is still exists in a level of a toy example. Even the most sophisticated algorithms deployed to computer hardware have many failures and couldn't act nearly intelligent. On the other hand there are many functioning AI examples when it helps to assist people in this or that task: in cars equipped with voice navigator will speak to a driver to turn right or left and if the car has more sensors it may even drive autonomously for short intervals while driver's hands are out of the wheel. Another use case of AI in TV and especially in streaming platforms is that AI could give better recommendations to users according to their preferences and in a way suggested content will match users interests. For example, a voice command to AI assistant could select a movie for evening watch in a way the viewer will just say something like "pick up a comedy" or similar direction.

Neural Network [NN] are multidimensional trees or combinations of layers of neurons that serve as filters for inputs and are used for purposes of classifications of data. Otherwise a big algorithm can be considered to be a neural network in case it has many layers and can act in a way that the processing of a data will be adjusted on the fly with use of gradients and back propagation and will be passed to this or that layer for further categorisation or weighting. The predecessor of a neural network was a simple decision tree. But with continuous development

algorithms got depth structure so that it allows them to do a very sophisticated classification and identification of inputs as belonging to this or that set.

Transformers, "a novel neural network architecture based on a self-attention mechanism that we believe to be particularly well suited for language understanding." (Uszkoreit, 2017) According to Uszkoreit, Transformer performs a number of steps that are performed in a way a self-attention mechanism models relationships between words in a sentence. Establishing relations between words regardless of their position in a sentence algorithm will "know" that e.g. in a sentence "I arrived at the bank after crossing the river" the word bank relates to a word river and that the word bank doesn't refer to a financial institution.

Bidirectional Encoder Representations from Transformers [BERT] is said to be an innovative approach in pre training of language representations for NLP tasks and as Chang defines, it is also a language representation model that "is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications" (Chang et al. 2018). BERT code and pretrained models are available at https://github.com/google-research/bert . Pandu Nayak from Google in his blog points out that with BERT anyone can train a question-answering model.

ALBERT - is a lite version BERT for self-supervised learning of language representations. In 2019 it was found that BERT execution at GPU/TPU causes issues with memory limitations and BERT algorithm was reduced in a way that it will allow to avoid "out-of-the-bounds-of-memory" exceptions in the future. The new reduced model consumes less resources and works faster. However, it has less layers. (Chen, 2019) The code is available at https://github.com/google-research/ALBERT

Stanford Question Answering dataset [SQuAD]

It is a large reading comprehension dataset that consists of pairs of questions and answers collected with the use of crowdfunding. Question were posted to Wikipedia articles and answers to them were e.g. in a form of a corresponding reading passage. It also consists of unanswered questions approached by crowdworkers. As for today it has more than 100,000 pairs processed on more than 500 articles. (Rajpurkar, 2016). SQuAD can be used to train BERT.

WaveNet

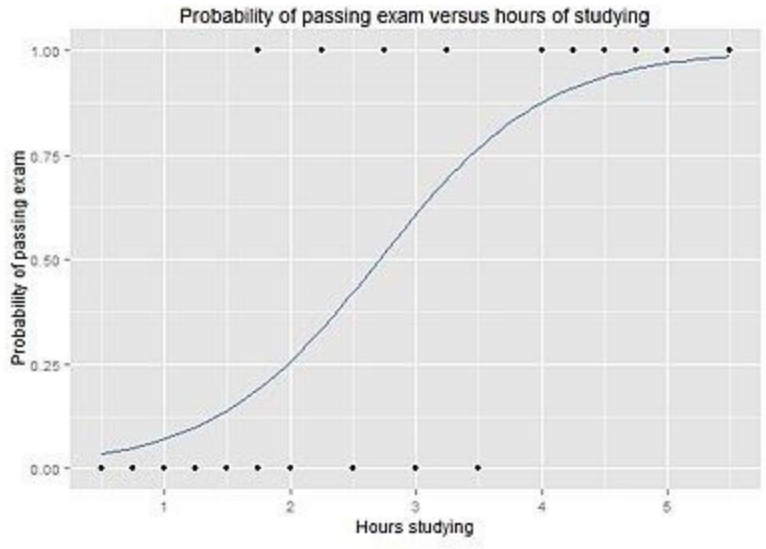
Oord in 2016 defines WaveNet as " a deep neural network for generating raw audio waveforms". It is based on PixelCNN which are generative models with trackable likelihoods.

WordNet

A thesaurus that has representations of many words with distribution of elements arranged according to the meaning of words in a way similar words are concentrated in certain locations.

Logistic Regression

"is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although" (Wikipedia). From the same Wikipedia article there is a reference graph representing it graphically.



Softmax Regression

Xue *(2018) underlines that a softmax regression is generated from logistic regression for multi-classification problems*

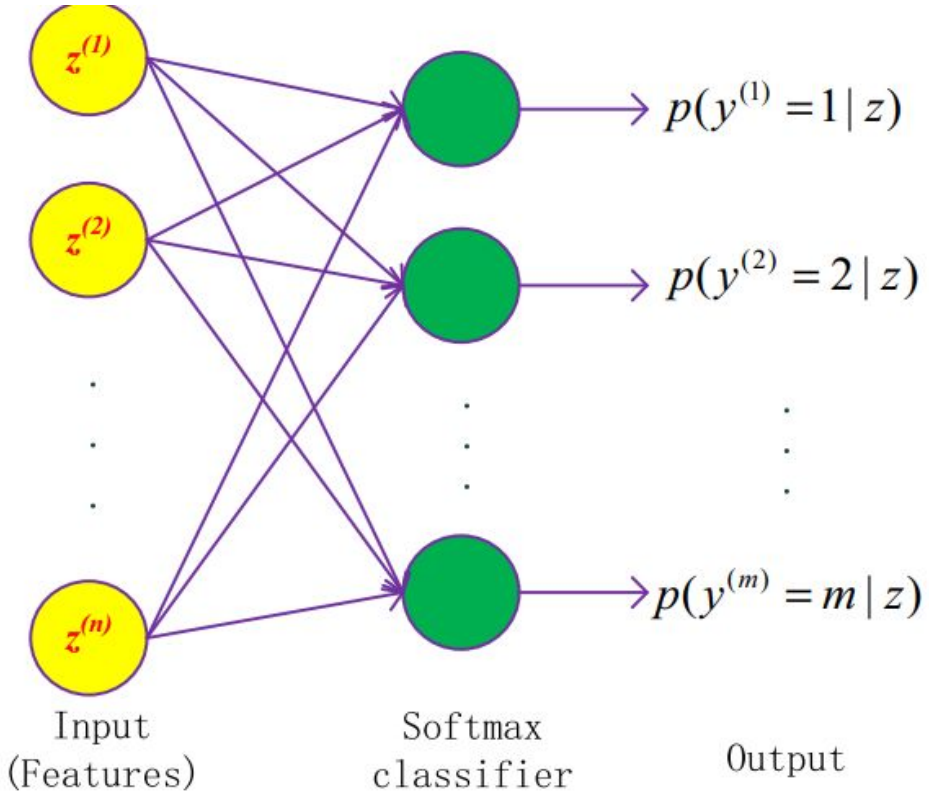


Figure 1: softmax regression model

As it is seen from two images above, softmax regression has more complexity and that allows to use more robust functions to define interdependencies among elements.

Model Training

Consists of assigning of an initial state to a system of elements with further check if a prediction based on given inputs worked or not. Different initial states are assigned to the same system of elements [nodes of a graph] in order to establish which initial parameters will impact on increase and reduce of predictability of outcomes of the system.

Google Assistant. It is an application that is being actively developed and constantly enhanced with features. In the near future it is expected that it will allow even more operations

and scheduled events executions than is implemented today. Today among many other implementation of assistants e.g. Croatia, Alexia & Echo Echo it allows people to talk and command to computer devices and IoT devices with voice or other inputs.

Word2Vec

An example of Word2Vec is representation of words by a number like zip or postal code. That way presentation of a word denoting a city is recorded using mathematical expressions rather than symbols. However, naturally a word is represented by multidimensional vector that could have up to 100 to 1000 dimensions that define its position within a semantic picture or a text. Words are assigned with weights like in graphs theory and that allows to operate and select somehow them using their weights and defined patterns.

Seq2Seq

For a translation of sequences from one language to another language a seq2seq method is used. It allows to index words in one language and process them by an algorithm and translate them to another language based on their weights or somehow otherwise due to the defined method. Normally, seq2seq is based on sophisticated algorithms that process text translations automatically.

Stanford CoreNLP

"a Java (or at least JVM-based) annotation pipeline framework, which provides most of the common core natural language processing (NLP) steps, from tokenization through to coreference resolution" (Manning, 2014).

Hate Speech

"Numerous kinds of insulting user-created content addressed in the individual works" (Schmidt & Wiegand, 2017). It is often used in cyberbullying and can be known as hostile or abusive messages and flames.

Use cases

For example, an algorithm can be executed to process a large amount of data sourced from the Internet, e.g. on a social media platforms conversations can be collected and corresponding neural networks can be trained on them and for example that will result in a

possibility to create an AI that will use this or that approach to conduct conversation and that will be constructed on the top of conversations from social media data. Another approach which is similar to it was used for SQuAD gathering when questions and extended answers were matched and collected from Wikipedia in order to construct the database that further can be used for training of various models. Moreover, many social media websites already have some implementations of AI's in a form of bots and text processing algorithms that can address generic user questions and generate responses based on written text enquiries.

Semantic Analysis of Texts

Trained models and toolkits like Stanford CoreNLP can be used to understand the connotation of texts and determine either a sentence has a positive meaning or a negative meaning. Social media crawlers could be used by intelligent services in order to diffuse or predict dangerous events. Models can be trained and adjusted to recognise hate speech at social websites and signal or automatically mute users. For example having a database with texts representing examples of hate speech it is possible to make a model to analyse any text and map matches of analysed text with the database. Malmasi & Zampieri (2017) in the paper titled *Detecting Hate Speech in Social Media* underline the importance of thorough analysis of classes in a dataset in order to achieve more precise details. In their work they point out that they "applied text classification methods to distinguish between hate speech, profanity, and other texts." They "applied standard lexical features and a linear SVM classifier to establish a baseline for this task. The best result was obtained by a character 4-gram model achieving 78% accuracy." Their results showed that distinguishing profanity from hate speech is a very challenging task. Badjatiya et al. (2017) argues : "Hate speech detection on Twitter is critical for applications like controversial event extraction, building AI chatterbots, content recommendation, and sentiment analysis. We define this task as being able to classify a tweet as racist, sexist or neither." In their work titled *Deep Learning for Hate Speech Detection in Tweet*s, they describe their experiments with Twitter inputs and processing them by various approaches: Logistic Regression, Random Forest, SVMs, Gradient Boosted Decision Trees (GBDTs) and

Deep Neural Networks(DNNs). As baselines they compared char n-grams , Term Frequency - Inverse Document Frequency [TF-IDF] vectors, and Bag of Words vectors (BoWV).

NLP in the discourse of Google Search Requests processing

Nayak (2019) writes that the emergence of a new hardware in the form of Google Cloud Tensor Processing Units [TPU] allowed them to apply BERT to ranking and matching of user queries and that allowed to better understand 1 in 10 searches in the US. He points out the following example: “Here’s a search for “2019 brazil traveler to usa need a visa.” The word “to” and its relationship to the other words in the query are particularly important to understanding the meaning. It’s about a Brazilian traveling to the U.S., and not the other way around. Previously, our algorithms wouldn't understand the importance of this connection, and we returned results about U.S. citizens traveling to Brazil. With BERT, Search is able to grasp this nuance and know that the very common word “to” actually matters a lot here, and we can provide a much more relevant result for this query.”

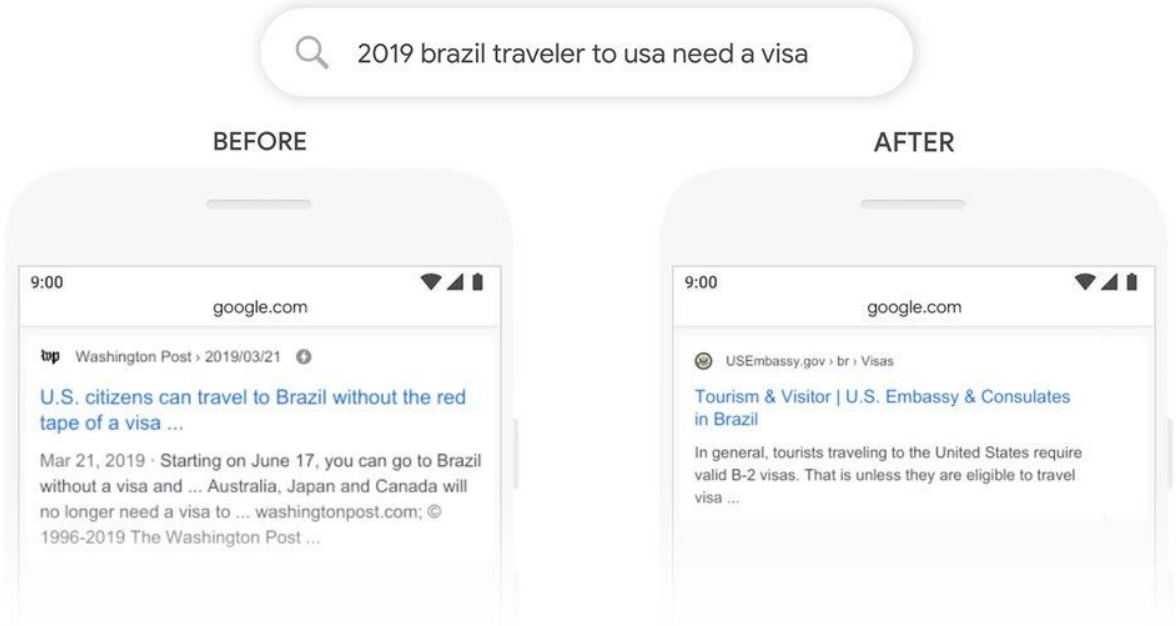


Figure 2. BERT in Google Search.

Nayak underlines that language understanding remains an ongoing challenge. And that Search by Google is being actively developed and have many aspects to improve.
Previous Google attempt in 2015 was to use Rankbrain which was used in less than 15% of queries. It allowed to establish for a given unknown to the algorithm word a close synonym from known words and this way return relevant results to a user query.

However, in addition to text based requests Google also uses Natural Language Processing for voice queries that could come from a computer equipped with a microphone or from a mobile device. There are many registered by Google patents that define how complex software and hardware systems can take an input from end users and through a chain of servers and network services deliver them for processing to a speech-to-text processing algorithm in a way given input will be recognised and post processed as if it was provided in text form. Reverse text-to-speech processing has also many implementations and today even a tiny app installed at smartphone can use synthetic voices to read aloud. Excerpt below is from a US Patent US 2012/0271625A1 which defines "a natural language query system and method for processing and analyzing multimod ally-originated queries, including voice and proximity-based queries." The complex process defined in the paper got slightly obsolete due to technological progress but reflects all major characteristics and steps of alike systems. " *The natural language query system includes a Web enabled device including a speech input module for receiving a voice-based query in natural language form from a user and a location/proximity module for receiving location/proximity information from a location/proximity device. The query system also includes a speech conversion module for converting the Voice-based query in natural language form to text in natural language form and a natural language processing module for converting the text in natural language form to text in searchable form. The query system further includes a semantic engine module for converting the text in searchable form to a formal database query and a database-look-up module for using the formal database query to obtain a result related to the Voice-based query in natural language form from a database*."

On the other hand, the major difference with widely used today devices is that most of the components of speech-to text and back text-to speech processing are implemented locally at mobile devices and only the search database requires Internet connection while other elementary functions e.g. a request to Android device what is the time or date could be performed offline without Internet connections. Speech queries to a device can be matched with corresponding values in internal sql database and further processing will take them as inputs that will be processed according to the values of parameters in the database tables cells (Ghosh et al., 2014).

In his paper Speech Recognition for Mobile Devices at Google Mike Shuster in 2010 points out that the Speech Group started in 2005 and managed to developed many services for US and other countries.

In 2006 GOOG-411, a speech recognition driven directory assistance service which works from any phone, was launched. Which could be called in order to get an address or contacts of a business or in order to be connected with them. As a backend Google Maps information was used.

In 2008 Google Search by Voice was introduced to the public. It uses relatively standard voice recognition methods :

"Front-End and Acoustic Model. For the front-end we use 39-dimensional PLP features with LDA. The acoustic models are ML and MMI trained, triphone decision-tree tied 3-state HMMs with currently up to 10k states total. The state distributions are modeled by 50-300k diagonal covariance Gaussians with STC. We use a time-synchronous finite-state transducer (FST) decoder with Gaussian selection for speedy likelihood calculation."

For dictionary they use "between 200k and 1.5M words in the dictionary, which are automatically extracted from the web-based query stream. The pronunciations for these words are mostly generated by an automatic system with special treatment for numbers, abbreviations and other exceptions"

For language model they mostly use 3-grams or 5-grams with Katz backoff trained on months or years of query data. Models are pruned in a way they could fit server's memory.

For acoustic data they use more than 250k collected acoustic queries from specifically designed application at android devices. "Several hundred speakers read queries off a screen and

the corresponding voice samples are recorded". As most queries are spoken without errors they don't have to manually transcribe these queries. (Shuster, 2010) Shuster points out that they want to optimize user experience and that traditionally speech recognition systems are targeted in reducing error rate. They use WebScore as a metric indicating users experience. He also underlines that some phones have very few buttons and some languages would be very difficult to type with fingers due to complexity in alphabet. (Shuster, 2010). For example Mandarin Chinese is much more complex than English and Wu uses n n-gram language model over Mandarin words. In their experiments In the experiments observed in the work titled Search by Voice in Mandarin Chinese and published in 2010 they used Katz smoothing and entropy-based pruning as well as and count-cutoff thresholds of 0, 0, 0, 1 and 2 which were used for n-gram units. (Wu et al., 2010)

Ballinger et al. in the paper published in 2010 ad titled On-Demand Language Model Interpolation for Mobile Speech Input states that for android operating system they provide search by voice, voice input in any text and an API for application developers (Ballinger et al, 2010). He points out that Voice input is used 49%, Search by Voice 44% and API by 7%.

Conclusion

Speech recognition and Natural Languages Processing are developing fields. And development and research in Machine Learning, crowdsourcing and computer science allows continuously to introduce new features which results in optimization and increase of performance and reduce the number of errors in processing users requests. Most progress was made due to accumulation of large collection of speech samples, which made possible various enhancements including but not limited to Deep Neural Networks algorithms that allows to improve understanding of search queries. Often such features are still of experimental nature. Moreover, wide scale accessibility of voice recognition in devices allowed to provide services to users with sight issues and to users who have no option to type requests to devices due to the certain limitations like disability, complexity of their alphabets or due to other reasons. However, there are certain issues with usability of devices, with their failures, freezes, glitches, bugs, and

unstable updates which can prevent users from benefiting from the introduced innovative technology in certain cases. On the other hand, 5G networks technology could provide more usability and stable services that will include not only processing based on voice directives, but also based on video streams analysis, e.g. gestures, mimics, etc. which will become possible with fast networks.

References

1. Bernard. (Oct. 25, 2012). Patent. Multimodal natural language query system for processing and analyzing, voice and proximity based queries. Retrieved from https://patentimages.storage.googleapis.com/17/c4/4e/18d0d8996f47d2/US20120271625A1.pdf
2. Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
4. Room, C. (2019). BERT (Language Model). Retrieved January 2020 from https://www.blog.google/products/search/search-language-understanding-bert/
5. Uszkoreit, J. (2017). Transformer: A novel neural network architecture for language understanding. *Google AI Blog*, *31*.
6. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
7. Zhang, J., Zhu, X., Chen, Q., Dai, L., Wei, S., & Jiang, H. (2017). Exploring question understanding and adaptation in neural-network-based question answering. *arXiv preprint arXiv:1703.04617*.
8. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
9. Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.
10. Figure 1. Softmax regression model. Jiang, M., Liang, Y., Feng, X., Fan, X., Pei, Z., Xue, Y., & Guan, R. (2018). Text classification based on deep belief network and softmax regression. *Neural Computing and Applications*, *29*(1), 61-70.
11. Wikipedia. Logistic regression. Retrieved January 2020 from https://en.wikipedia.org/wiki/Logistic_regression

12. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60).
13. Malmasi, S., & Zampieri, M. (2017). Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*.
14. Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1-10).
15. P. Nayak. (2019). Understanding searches better than ever before. Retrieved January 2020 from https://www.blog.google/products/search/search-language-understanding-bert/
16. Figure 2. BERT in Google search. Retrieved January 2020 from https://www.blog.google/products/search/search-language-understanding-bert/
17. Wikipedia. *RankBrain*. Retrieved January 2020 from https://en.wikipedia.org/wiki/RankBrain
18. Bernard. Patent US 2012/0271625A1. Retrieved from https://patentimages.storage.googleapis.com/17/c4/4e/18d0d8996f47d2/US20120271625A1.pdf
19. Schuster, M. (2010, August). Speech recognition for mobile devices at Google. In *Pacific Rim International Conference on Artificial Intelligence* (pp. 8-10). Springer, Berlin, Heidelberg.
20. Ballinger, B., Allauzen, C., Gruenstein, A., & Schalkwyk, J. (2010). On-demand language model interpolation for mobile speech input. In *Eleventh Annual Conference of the International Speech Communication Association*.
21. Shan, J., Wu, G., Hu, Z., Tang, X., Jansche, M., & Moreno, P. J. (2010). Search by voice in mandarin chinese. In *Eleventh Annual Conference of the International Speech Communication Association*.