THE DESCRIPTION OF THE AUTISM SPECTRUM DISORDER QUESTION ANSWERING DATASET

Firsanova Victoria Igorevna (vifirsanova@gmail.com)

Saint Petersburg State University, Saint Petersburg, Russia

The study presents the Autism Spectrum Disorder Question Answering Dataset (ASD QA), a new Russian dataset based on the structure of the Stanford Question Answering Dataset (SQuAD), a machine reading comprehension dataset. The ASD QA dataset is a work in progress. The dataset version described in the paper consists of 1,134 question-answer pairs compiled by the author of the paper from the information website for individuals with autism spectrum disorders (ASD) and Asperger's syndrome and their parents. The paper also describes several question-answering models built to analyze the dataset.

Keywords: Dataset; Question Answering; Dialogue Systems; Machine Reading Comprehension

Introduction

Closed-domain question answering is a challenging Natural Language Understanding (NLU) task. Recent advances in Natural Language Processing (NLP) show that complex models trained on huge datasets that contain a large number parameters are usually more efficient than their small-scale counterparts. For example, Ilya Sutskever emphasized in his speech at the Matroid Scaled Machine Learning Conference 2019 (Matroid, 2019) that large-scale models are significantly more efficient in solving a wide range of tasks covering different domains, such as robotics, gaming and NLP, and are capable of zero-shot learning, a learning method allowing to solve a task without training on examples of that task.

Despite the fact that now we can use a large-scale pre-trained model to fine-tune it on a smaller dataset, to get better model performance on closed-domain question answering, we still need larger datasets allowing in-depth topic coverage and high-quality text generation. How to collect large question answering datasets when data resources are limited because it is not always possible to find ready-made question-answer pairs on a given topic and manual work imposes limits upon the speed of the dataset collecting and compiling? Which pieces of the work can be automated? How to find a balance between efficient automation and high quality?

Another important issue in building closed-domain question answering systems is user's questions processing. The formulation of the question depends on the user. For example, qualitative and quantitative differences in question asking patterns might be found in the speech of people of different age groups, social classes or cultures. Despite the fact that some differences might be more explicit among children, adults might have different preferences in choosing a question type whereas the question type influences the answer type (Kearsley, 1976), which becomes crucial in a question answering system mechanism.

NLU implies automatic decoding of meaning hiding in a given language sample. Understanding is made possible by different types of linguistic analysis, such as context analysis, the analysis of the utterance objectives, extraction of lexical features, for example, frequently used named entities, etc. (Canonico & Russis, 2018). New issues are emerging. How does high out-of-vocabulary rate (the percentage of unknown words in a language sample) influence a model's language understanding ability? How should a model respond to domain irrelevant questions?

Related Work

The ASD QA dataset presented in the paper is a tool for Natural Language Understanding tasks. Natural Language Understanding (NLU) and Natural Language Generation (NLG) are often considered as two major components of Natural Language Processing (NLP). Figure 1 shows a possible illustration of the correlation between NLU and NLG.

The function of NLU is to model the meaning of a given piece of somebody's text or speech. NLU models training usually implies such types of linguistic analysis as lexical analysis, syntactic analysis, semantic analysis, discourse analysis, and pragmatics analysis (Ovchinnikova, 2012). The function of NLG is the transformation of structured data into text or speech. The transformations remind the creative process in the human brain, which occurs when a person wants to express some formed ideas in a text form (Reiter & Dale, 2000). NLG models training implies such types of analysis as structural analysis, lexical analysis, analysis of structured data or images, discourse analysis, and data parsing.

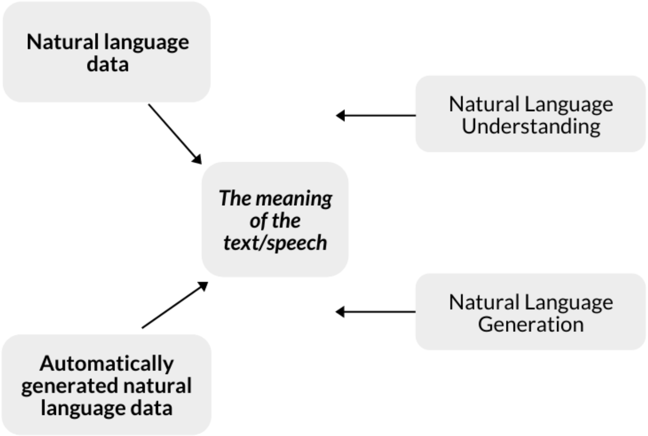

Natural language data
Natural Language Understanding
The meaning of the text/speech
Natural Language Generation
Automatically generated natural language data

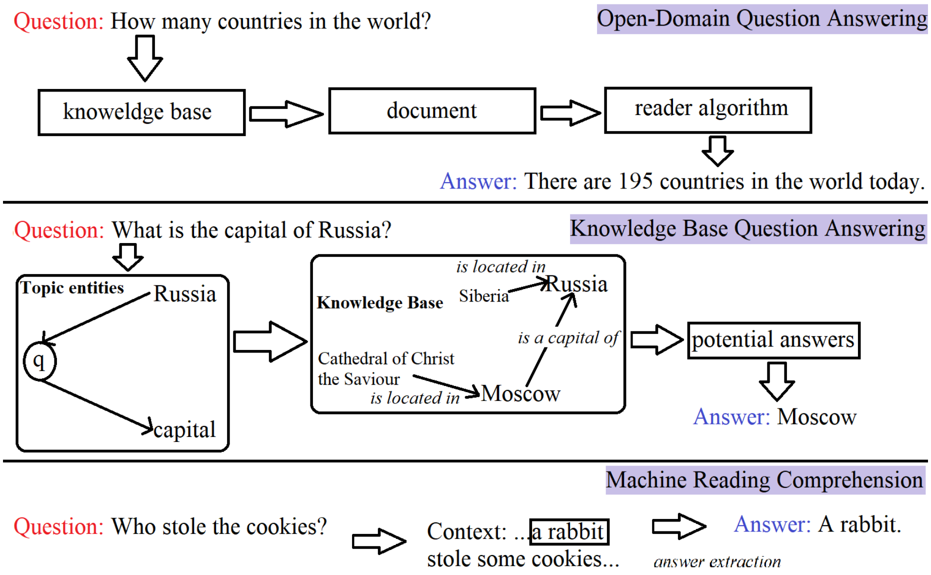
Figure 1. The correlation between NLU and NLG

The ASD QA dataset is designed to be used in training and evaluation question answering systems. One can distinguish different types of question answering datasets. For example, machine reading comprehension, open-domain question, and knowledge base question answering datasets are widespread. Figure 2 illustrates types of question answering tasks that use the listed question answering datasets.

Open-domain question answering (ODQA) models answer factoid questions, for example the ones about dates and locations, (Jurafsky & Martin, 2014) in natural language. For example, Quasar is an ODQA dataset that consists of cloze-style (fill-in-the-gap) queries constructed from the online question answering website for software developers Stack Overflow and trivia question-answer pairs obtained from online sources (Dhingra et al., 2017). Another example is Multilingual Knowledge Questions and Answers (MKQA), an ODQA dataset that consists of question-answer pairs aligned across 26 typologically diverse languages (Longpre et al., 2020).

Knowledge base question answering (KBQA) task is to give an answer in a natural language based on a knowledge graph to a question. For example, QALD-9 is a superset of previous versions of QALD based on DBpedia graph knowledge base and compiled by human experts (Usbeck et al., 2018). Another example is RuBQ, a Russian KBQA dataset. RuBQ consists of questions in Russian, their English machine translations, SPARQL queries to Wikidata, reference answers, and entities with Russian Wikidata labels (Korablinov & Braslavski, 2020).

Machine Reading Comprehension (MRC) task is to find an answer to a question in a given paragraph or document. For example, Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2018) used as a reference for ASD QA dataset described in this paper is an MRC dataset consisting of questions posed by crowdworkers on a set of Wikipedia articles. SberQuAD (Efimov et al., 2020) is a Russian analogue of the original SQuAD. The ASD QA dataset is also an MRC dataset tuned for training and evaluation of dialogue systems for inclusive education.



Question: How many countries in the world?
Open-Domain Question Answering
knowledge base
document
reader algorithm
Answer: There are 195 countries in the world today.
Question: What is the capital of Russia?
Knowledge Base Question Answering
Topic entities
Russia
q
capital
Knowledge Base
is located in
Siberia
Russia
is a capital of
Cathedral of Christ the Saviour
is located in
Moscow
potential answers
Answer: Moscow
Machine Reading Comprehension
Question: Who stole the cookies?
Context: ...a rabbit stole some cookies...
Answer: A rabbit.
answer extraction

Figure 2. Types of question answering (Firsanova, 2021)

Data Collection

The ASD QA dataset is in progress, which means that the number of question-answer pairs, topic coverage, and the number of resources will increase over time. We have chosen a reliable informational online-resource available at http://aspergers.ru/. The articles on the website are about autism spectrum disorder and Asperger syndrome, inclusive education, health of people with ASD, and interaction between neurotypical and atypical people. The data is being collected with the agreement of the website administration.

The data is being collected with an HTML parser built with Beautiful Soup 4 (Beautiful Soup Documentation) on Python 3.7 (Python). HTML content from the website pages was obtained with the *get* method from the Requests Python library. The data was parsed with *findAll* and *find* Beautiful Soup methods. The extracted texts were saved as text data for the dataset development. The code is given in the project GitHub repository, see section Reproducibility. The collected data comprises blog entries, messages to readers, and informational articles written in Russian or translated from foreign languages into Russian. Some of the texts were created by people with ASD. Figure 3 shows topic coverage in the collected data.

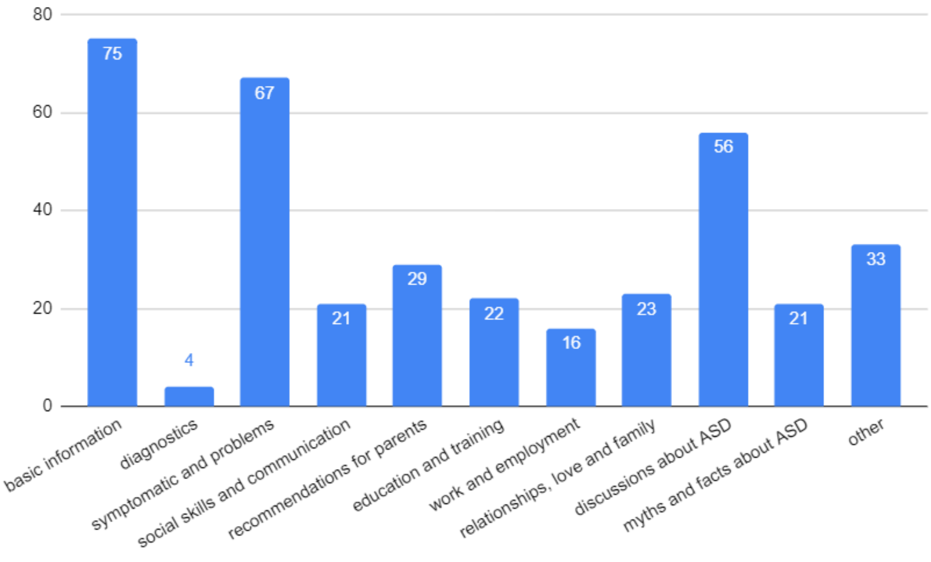


Figure 3. Topic coverage of the collected data in %

Dataset

The ASD QA dataset was designed as an MRC one, which means that ASD QA contains sets of question-answer pairs, as well as corresponding reading passages. An MRC model should learn to extract the answer from a given reading passage. Figure 4 where Q denotes a question, and A denotes an answer shows a question-answer pair for a sample passage in the ASD QA dataset. The answer is a segment of text from the passage.

> Не каждый командный вид спорта требует навыков общения и взаимодействия на высоком уровне. Во многих из них отдельный спортсмен вполне может быть ценным членом команды. Плавание. Прекрасный вид спорта для большинства людей, включая детей с аутизмом. Бывает так, что аутичным детям сложно контролировать мяч, но они хорошо справляются с базовыми гребками и простой игрой на воде.
>
> Q: Подойдет ли плавание детям с аутизмом?
>
> A: Прекрасный вид спорта для большинства людей, включая детей с аутизмом.

Figure 4. A sample question-answer pair in the ASD QA in Russian

Figure 5 presents the ASD QA dataset structure where in brackets are given the types of the object (*str* is for string, *dict* is for dictionary, *bool* is for Boolean, *int* is for integer). It is a JavaScript Object Notation object that contains reading passages (*contexts*) and question-answer pairs (*qa-pairs*). Each answer contains also positional tags (*start* and *end*) denoting the numerical positions of the first and last symbols of the answer span in a reading passage. 5% of the questions in ASD QA are *unanswerable*, which means that corresponding reading passages do not contain any answers to them. We have made those questions deliberately irrelevant to the ASD QA topic coverage. For example, some questions ask a system to tell a joke or a fairy tale.

```
{
"qa_pairs":
{
"question": "Question?" (str),
"answer": {"text": "answer" (str), "start": n (int), "end": m (int)} (dict),
"is_unanswerable": TRUE/FALSE (bool)
} (dict),
"context": "Context." (str)
} (dict)
```

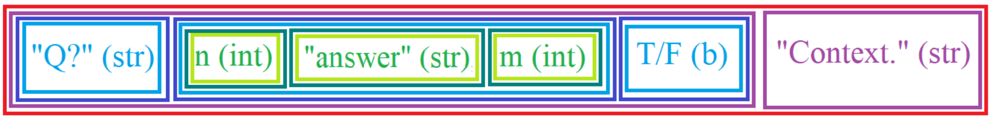


Figure 5. The ASD QA structure

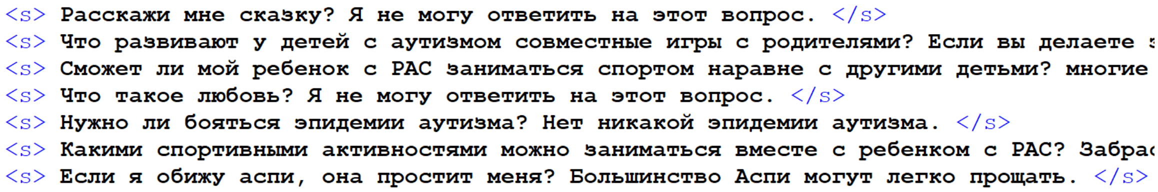
Table 1 presents the ASD QA dataset statistics. For the experiments, the dataset was shuffled and split with the Scikit-learn (Scikit-learn) *train_test_split* method. The training set contains 70% of the data, the validation set contains 15% of the data, and the test set contains 15% of the data. We have decided to train two types of question answering models, generative and extractive ones. The vocabulary size in extractive models is 30,522 tokens on a subword level, and the vocabulary size in generative models is 50,257 tokens on a subword level. Subword level was created with byte pair encoding (Gage, 1994). About 4.47% of the words were excluded from the analysis and replaced by <unk>'s (*unknown* tokens) as out-of-vocabulary tokens.

Parameter	Value
Number of reading passages	96
% of unanswerable questions	5
Number of question-answer pairs	1,134
Number of tokens (word-level) in questions	8,109
Maximum length of a question (word-level)	20
Minimum length of a question (word-level)	2
Number of tokens (word-level) in answers	18,160
Maximum length of an answer (word-level)	50
Minimum length of an answer (word-level)	2
Maximum length of a reading passage (word-level)	512

Table 1. The ASD QA statistics

Experiments

We have implemented two approaches, generative and extractive using Transformer architecture. Transformer based decoder GPT-2 (Radford et al., 2019) that was pre-trained to solve traditional language modeling task (predicting a probability distribution over token sequences) was used for the generative approach. The ASD QA dataset was transformed. Question-answer pairs without meta-data were retrieved from the MRC dataset and supplemented with start- (<*s*>) and end-of-sentence (<*/s*>) tags (see Figure 6). Unanswerable questions were provided with a plausible answer translated from Russian into English as "*I cannot answer this question*". We have tested the dataset on small and medium GPT-2 models. A small one has 117 million parameters, and a larger one has 774 million parameters.



```
<s> Расскажи мне сказку? Я не могу ответить на этот вопрос. </s>
<s> Что развивают у детей с аутизмом совместные игры с родителями? Если вы делаете :
<s> Сможет ли мой ребенок с РАС заниматься спортом наравне с другими детьми? многие
<s> Что такое любовь? Я не могу ответить на этот вопрос. </s>
<s> Нужно ли бояться эпидемии аутизма? Нет никакой эпидемии аутизма. </s>
<s> Какими спортивными активностями можно заниматься вместе с ребенком с РАС? Забрас
<s> Если я обижу аспи, она простит меня? Большинство Аспи могут легко прощать. </s>
```

Figure 6. The ASD QA dataset transformed for generative question answering

We have implemented the extractive approach based on machine reading comprehension with the Transformer encoder based model BERT (we have used a base version pre-trained on 104 languages and a version fine-tuned for Russian by Geotrend (Geotrend)) (Devlin et al., 2019) pre-trained to solve masked language modeling (fill-the-gap task) and other BERT based architectures, such as cross-lingual model XLM-RoBERTa (Conneau et al., 2020), a distilled (knowledge distillation allows transferring knowledge from a large model to a smaller one) version of BERT (also pre-trained on 104 languages) (Sanh et al., 2019). Table 2 shows the models' performance obtained on the ASD QA dataset.

Base model	F1-Score	Exact Match
XLM-RoBERTa	0.48	0.39
Multilingual BERT	0.40	0.29
Multilingual DistilBERT	0.42	0.32
Russian BERT (Geotrend)	0.39	0.30
774M GPT-2	0.63	0.52
117M GPT-2	0.53	0.43

Table 2. Results obtained on the ASD QA dataset

The training was performed on Google Colab with Nvidia Tesla T4 graphics processing unit (GPU). During the hyperparameters optimization, we have tuned batch size, learning rate, and the number of epochs. Each model was retrained nine or ten times with different parameters. All the models used GELU activation function and 12 attention heads. All the models had 12 hidden layers except for DistilBERT, which had 6 layers. Each BERT-based model has trained from 10 to 20 minutes. 117M GPT-2 has trained 30 minutes, and 774M GPT-2 has trained 1 hour 30 minutes. The optimum learning rate for GPT-2, XLM-R and BERT was 3e-5, and the optimum rate for DistilBERT was 1e-5. The optimum batch size for BERT based models was 1, and for GPT-2 was 16.

Conclusion

The paper introduces a new Autism Spectrum Disorder Question Answering Dataset, a dataset based on the structure of a machine reading comprehension dataset Stanford Question Answering Dataset (Rajpurkar et al., 2018). The ASD QA dataset uses the data from the informational online resource about autism spectrum disorder (ASD) and Asperger's syndrome. Articles from the website highlight different aspects of life of children and adults with special needs, give advice for parents of children with ASD and neurotypical people who communicate and interact with atypical individuals. The dataset is in progress, the number of question-answer pairs and topic coverage will continue to grow. The dataset is an open-source, see the link given in the section Reproducibility.

The best performance of the extractive question answering models is 0.48 F1-Score and 0.39 Exact Match. The best of the generative question answering models is 0.63 F1-Score and 0.52 Exact Match. The performance suggests opportunities for improvement. We can explore other, more complex algorithms for solving question answering task on the ASD QA dataset, new methods of dataset collection and compilation, try different approaches towards dataset and question answering models testing, like usability testing, showing the products to test audiences, focus groups engagement, and evaluation of the human performance on the dataset.

The ASD QA dataset is available online to contribute and explore more question answering models. At the moment, the ASD QA dataset forms the basis of an eponymous project that supports the inclusion of people with special needs. The project is supported by the Open Data Science community. See the following link for more information: https://ods.ai/projects/asd_qa. Supposing that lack of autism spectrum disorder awareness might cause social problems and create obstacles to the development of inclusive and tolerant society, we believe that tools and projects like ASD QA are small steps to address this problem.

Reproducibility

The dataset for this paper is available on FigShare:
https://figshare.com/articles/dataset/Autism_Spectrum_Disorder_and_Asperger_Syndrome_Question_Answering_Dataset_1_0/13295831. The code for this paper is available on GitHub:
https://github.com/vifirsanova/ASD-QA.

References

1. Matroid. (2019). *Ilya Sutskever - GPT-2. YouTube.* https://www.youtube.com/watch?v=T0I88NhR_9M&amp;ab_channel=Matroid.
2. Kearsley, G. P. (1976). Questions and question asking in verbal discourse: A cross-disciplinary review. *Journal of Psycholinguistic Research*, *5*(4), 355–375. https://doi.org/10.1007/bf01079934
3. Canonico, M., Russis, L.D. (2018) A Comparison and Critique of Natural Language Understanding Tools. *CLOUD COMPUTING 2018: The Ninth International Conference on Cloud Computing, GRIDs, and Virtualization*, 110–115.
4. Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing*. Prentice Hall, Pearson Education International.
5. Dhingra, B., Rivard, K., & Cohen, W. (2017). Quasar: Datasets for Question Answering by Search and Reading. *ArXiv Preprint ArXiv:1707.03904*.
6. Longpre, S., Lu, Y., & Daiber, J. (2020). MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. *ArXiv Preprint ArXiv:2007.15207*.
7. Usbeck, R., Gusmita, R. H., Ngomo, A.-C. N., & Saleem, M. (2018). 9th Challenge on Question Answering over Linked Data (QALD-9). *Joint Proceedings of the 4th Workshop on Semantic Deep Learning (SemDeep-4) and NLIWoD4: Natural Language Interfaces for the Web of Data (NLIWOD-4) and 9th Question Answering over Linked Data Challenge (QALD-9) Co-Located with 17th International Semantic Web Conference (ISWC 2018)*, 58–64.
8. Korablinov, V., & Braslavski, P. (2020). RuBQ: A Russian Dataset for Question Answering over Wikidata. *Lecture Notes in Computer Science*, 97–110. https://doi.org/10.1007/978-3-030-62466-8_7
9. Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don’t Know: Unanswerable Questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. https://doi.org/10.18653/v1/p18-2124
10. Efimov, P., Chertok, A., Boytsov, L., & Braslavski, P. (2020). SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis. *Lecture Notes in Computer Science*, 3–15. https://doi.org/10.1007/978-3-030-58219-7_1
11. Firsanova, V. (2021). The Evolution of Chatbots: Psychoanalytics, Business Partners, Soul Mates (Jevoljucija chat-botov: Psihoanalitiki, biznes-partnery, sputniki zhizni). https://github.com/vifirsanova/NLP-Discussion-Group/blob/master/NLP_slides/Chatbots.pdf.
12. *Beautiful Soup Documentation*. Beautiful Soup Documentation - Beautiful Soup 4.9.0 documentation. (n.d.). https://www.crummy.com/software/BeautifulSoup/bs4/doc/.
13. *Python*. Python.org. (n.d.). https://www.python.org/.
14. *Scikit-learn*. scikit-learn: machine learning in Python. (n.d.). https://scikit-learn.org/stable/ .
15. Gage, P. (1994). A New Algorithm for Data Compression. *The C User Journal*. https://doi.org/10.5555/177910.177914
16. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI.*
17. *Scikit-learn*. scikit-learn: machine learning in Python. (n.d.). https://scikit-learn.org/stable/ .
18. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the Orth American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/n19-1423
19. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.747
20. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *5th Workshop on Energy Efficient Machine Learning and*

Cognitive Computing - NeurIPS 2019.

21. *Geotrend*. Geotrend - Revealing the world's connections. (n.d.). https://www.geotrend.fr/en/ .