Assessing the Difficulty of Spans Within Nested Text Markup Data Models

A Preprint

Andrey M. Tabachenkov
Department of Mathematical Methods of Forecasting
Moscow State University,
Machine Learning and Semantic Analysis
MSU Institute for Artificial Intelligence
a.tabachenkov@iai.msu.ru

Archil I. Maysuradze
Department of Mathematical Methods of Forecasting
Moscow State University,
Machine Learning and Semantic Analysis
MSU Institute for Artificial Intelligence
useraim@mail.ru

Abstract

While improving the performance of machine learning (ML) models, researchers assess the difficulty of objects in a dataset. There is no universally agreed-upon definition of what constitutes a difficult object. The most commonly used term is "difficulty"; although other terms such as "hardness" and "challenging cases" are also used. There are numerous methods for assessing object difficulty, ranging from traditional approaches to more recent techniques based on model training. These methods are applied in a variety of domains and to various data structures, helping to improve solutions by filtering out or re-labeling difficult objects. This paper extends these methods to the humanities, where researchers encounter inherently complex data schemas. Therefore, the use of conventional methods is complicated by the need to transform complex data into vector representations. In particular, we consider labeled text spans, for which the vector representation must take the entire context into account. To address this challenge, we study the use of language models to create vector representations of textual fragments that consider the context within the overall difficulty assessment pipeline. This yields a more accurate representation of a text's meaning and complexity, enabling a more precise estimation of the difficulty of each data instance.

Keywords Evaluation of difficulty for individual data instances · Text labeling · Digital humanities · Machine Learning · Natural language processing

1 Introduction

The evaluation of object complexity constitutes a critical component of data mining, as it provides insights into the performance of trained models. Identifying difficult objects is particularly valuable for the re-annotation of data and subsequent development of models. The methods employed for this assessment have evolved to encompass a broad spectrum of tasks, domains, and entities exhibiting varied characteristics.

In the realm of machine learning and data analysis, various methodologies exist for evaluating the difficulty associated with individual data instances. These methodologies range from traditional statistical approaches to contemporary techniques utilising neural networks. The evaluation process is further complicated by the

absence of a universally accepted definition of a "difficult object", which has led to diverse interpretations among researchers.

The terminology surrounding this concept remains contentious, with some scholars referring to such objects as "challenging" or "atypical", while others prefer the term "difficult". Nonetheless, "difficulty" appears to be the most widely endorsed designation within the literature.

The application of these machine learning techniques within the humanities is complicated by the need for precise knowledge formalisation. For instance, nested data are characterized by a hierarchical or multilevel structure, that is, are organized at more than one level. We consider following levels: the first and the second sub-levels of the level of spans and the level of elements, which are described in subsection 2.2.

In this article, we focus on the span level of a chain of nested data models, considering the text spans along with their corresponding tags as entities for difficulty assessment. The application of traditional methods to evaluate the difficulty of objects within a chain of nested data models encounters several challenges, including the phenomenon of multi-assessment and the necessity of data vectorization. This article will explore strategies to surmount these obstacles.

Firstly, we propose integrating assessor consistency into the loss function 2.3, thereby ensuring that the model is trained to yield reliable results across various assessors. Secondly, we advocate for the utilisation of language models to generate vector representations of text spans 2.4 and their contexts. This approach will aid in capturing the semantic essence of the text, facilitating easier comparisons between distinct objects.

The authors of considered articles about difficulty demonstrate varying interpretations of the concept of object difficulty. In this paper the difficulty of an object refers to the complexity encountered by a machine learning model when processing that object in a specific task. We focus exclusively on model-specific and task-specific methods since we consider text's span with single tag as object the current task is the span classification. Besides that, model specificity entails the need for training span classifier.

2 Related works

2.1 Difficulty assessement methods

Methods for assessing the difficulty have been addressed in a limited number of publications, such as the work Seedat et al. [2024]. We have drawn upon the mathematical and methodological foundations underlying these methods:

- Inclusion of distribution support estimation
- Inclusion of distribution density estimation for the object (or its features)
- Utilisation of reconstruction error as a measure of object difficulty
- Task-agnosticism (It is important to note that, in most instances, the task-agnostic nature of a method is directly related to the absence of a labelled or target feature for the object)
- Model-agnosticism
- Computation of statistics for a trained model (with respect to its layers; we focus here on model-specific approaches)
- Generation of (pseudo)difficult objects (for the purposes of training or validation).

The overview is presented in the Tables 1 and 2, respectively, for the articles under consideration and for the methods from scikit-learn. Additionally, we have also systematized the margin-based approach that is task- and model-specific.

All the methods discussed rely on vectorised objects as input, along with their accompanying labels.

Furthermore, when employing these methods directly, the assessment of difficulty is typically conducted on the entire dataset without exception, with the notable exception of the article Lee et al. [2018], which utilised a validation dataset containing previously known complex objects to refine hyperparameters (utilising objects from alternative datasets). In instances where pseudo-difficult objects (or pseudo-outliers) were employed, these were generated automatically using noise or an algorithm akin to that described in article Zhou and Wang [2024]. In other scenarios, the authors resorted to unsupervised learning methods or abstained from any training altogether, merely calculating statistics derived from pre-trained models.

Таблица 1: Considered articles

Article	Support estimation	Distribution estimation	Recon-struction	Task-agnostic	Model-agnostic	Statistics	Generation
Support vector data description Tax and Duin [2004]	✓			✓	✓		
Task-agnostic out-of-distribution detection using kernel density estimation Erdil et al. [2021]		✓		✓		✓	
Robust, deep and inductive anomaly detection Chalapathy et al. [2017]			✓	✓	✓		
A simple unified framework for detecting out-of-distribution samples and adversarial attacks Lee et al. [2018]		✓				✓	
Deep semi-supervised anomaly detection Ruff et al. [2019]	✓			✓	✓		
Rapp: Novelty detection with reconstruction along projection pathway Kim et al. [2019]			✓	✓		✓	
Unsupervised anomaly detection with generative adversarial networks to guide marker discovery Schlegl et al. [2017]				✓	✓		
Estimating example difficulty using variance of gradients Agarwal et al. [2022]						✓	
Grod: Enhancing generalization of transformer with out-of-distribution detection Zhou and Wang [2024]	✓					✓	✓

Таблица 2: Considered functions of sklearn

Function	Support estimation	Distribution estimation	Recon-struction	Task-agnostic	Model-agnostic	Statistics	Generation
One Class SVM	✓			✓	✓		
Elliptic envelope		✓		✓	✓		
Isolation forest				✓	✓		
Local outlier factor				✓	✓		

In contrast, when evaluating the efficacy of these algorithms, the authors relied on previously acknowledged difficult objects, either by employing alternative datasets or by designating one class as complex or noisy, subsequently excluding it during the training phase, or by using explicit annotations.

Since we consider model-specific and task-specific difficulty, this article will further examine the works Lee et al. [2018], Agarwal et al. [2022], and the margin-based approach.

2.2 Nested data models

In neuroscience synapses (level 1) are organized, or nested, in cells (level 2) Aarts et al. [2014]. The annotation of multiple spans in content analysis enhances the identification of human values in textual data Maysuradze et al. [2024], Rink et al. [2024], Vorontsov et al. [2025].

Within the framework of nested data models adopted for this study, analysis is confined to the span level and the element level, where elements comprise multiple spans. This framework was developed based on existing

datasets and comprises a sequence of text markup data models that are nested within one another, aligning with varying levels of markup complexity.

First sub-level of level of spans At this level, each document contains only a single markup, which is composed of spans. Each span is designated with only one tag, commonly referred to as a SpanTag. This structure aligns with many classical datasets characterised by simplicity. These datasets are predominantly utilised in Named Entity Recognition (NER) tasks. An example of such a dataset is the Kaggle NER Corpus Nadeau and Sekine [2009].

Second sub-level of level of spans At this level, a document may feature multiple annotations, introducing the concept of multi-assessority. Each span can be assigned zero or more tags; the absence of tags indicates that all spans within the dataset share the same tag, which is therefore omitted. Conversely, the presence of multiple tags corresponds to scenarios of multi-label classification or a complex hierarchical structure within the tagging system. For example, the RuSentNE dataset Golubev et al. [2023] assigns both Named Entity Recognition (NER) tags and sentiment tags to each span.

Level of elements At this level, each annotation comprises elements rather than spans, with an element being defined as a set of spans. The interpretation of elements may vary across different datasets. Each element is also assigned tags (ElementTag). Typical instances of elements include coreference clusters, which pertain to the coreference resolution task, as well as relations, frames, and multi-fragments Maysuradze et al. [2024]. Examples of datasets featuring elements include the Ruethics and RWSD tasks from the MERA project Fenogenova et al. [2024], as well as the ADE dataset Gurulingappa et al. [2012], NEREL Loukachevitch et al. [2023], RURED Gordeev et al. [2020], the RWSD task from RuSuperGlue, SCIERC Luan et al. [2018], and SemEval 2010 task 8 Hendrickx et al. [2019]. These datasets not only include individual spans but also incorporate the relationships between them and/or coreference clusters.

2.3 Multi-assessorship

In practice, it is not uncommon for situations to arise where the texts within a dataset are identical, yet their markup may vary. This phenomenon can be described as multi-assessorship without explicit information regarding the assessors, which corresponds to the second sub-level of the span level within a chain. Although this aspect is typically overlooked during the training of models and the assessment of object difficulty, we propose an alternative approach. Specifically, we recommend introducing an additional stage of multi-assessorship processing prior to the vectorisation of spans and the application of methods for assessing difficulty.

Firstly, we advocate for the utilisation of multiple markups of individual objects through the implementation of specialised loss functions during training, alongside various forms of consistency and consensus. An example of such an approach is articulated in article Le et al. [2023].

In that article, the authors examined the task of segmentation using bounding boxes and explored the classification of these boxes. For each image, they organised similar bounding boxes—derived from different markups of the images—into clusters based on the Intersection over Union (IoU) score. Subsequently, for each cluster, the authors computed an averaged bounding box, taking into account the reliability weights assigned to the assessors. For the computed bounding box and each class $k \in \{1, \ldots, K\}$, the authors calculated a confidence score denoted by:

$$c_k = c_0 \min(T, N),$$

where T represents the number of experts corresponding to the cluster in question, who labelled the boxes within that cluster with class k. This coefficient was then utilised as a weight in the loss function for the averaged bounding box. If T is sufficiently large and numerous assessors select very similar bounding boxes with the same label k, the resulting averaged bounding box assigned label k will exhibit a high confidence score.

According to Le et al. [2023], the incorporation of confidence scores into the loss function enhances the robustness and generalisation capability of the model.

2.4 Span vectorisation

There exists a variety of methods for deriving vector representations of spans through the use of language models. In the majority of instances involving spans, aggregation functions are utilised (which are used to aggregate vector representations of tokens in spans). Examples of such methodologies are detailed in the following studies: Joshi et al. [2020] (which discusses the application of boundary tokens and positional embeddings for spans' lengths), Eberts and Ulges [2020] (which explores max pooling along with positional embeddings of spans' lengths), and Toshniwal et al. [2020] (where the authors examined six distinct methods: average pooling, attention pooling, max pooling, endpoint, diff-sum, and coherent). We note, that in different tasks such methods may have different levels of success.

3 Problem statement

Let we have text $T = \{\tau_1, ..., \tau_l\}$, where τ_i is a token (tokenization corresponds to the considered tokenizer) and its text span $s = \{\tau_{begin}, ..., \tau_{end}\}$ with the SpanTag $y \in Y$. Thus, we consider triplet $(T, (begin, end), y)$.

It is required to estimate the difficulty of these spans with tags $DiffScore : X \to \mathbb{R}$, where X is the set of spans with tags.

If $D_{non-diff}$ is the set of nominal non-difficult objects and D_{diff} is the set of nominal difficult objects then the quality criterion is a ROC-AUC Fawcett [2006] applied to difficulty scores of objects, considering objects from $D_{non-diff}$ as negative and objects from D_{diff} as positive 2.1.

Let $ConfScore : X \to [0, 1]$ be the confidence score of pair span-tag 2.3. Then we also check hypothesises that $ConfScore$ correlates with $DiffScore$ and that objects with lower confidence scores have higher difficulty scores.

Список литературы

Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Dissecting sample hardness: A fine-grained analysis of hardness characterization methods for data-centric ai. In The Twelfth International Conference on Learning Representations, 2024.

David Tax and Robert Duin. Support vector data description. Machine Learning, 54:45–66, 01 2004. doi:10.1023/B:MACH.0000008084.60811.49.

Ertunc Erdil, Krishna Chaitanya, Neerav Karani, and Ender Konukoglu. Task-agnostic out-of-distribution detection using kernel density estimation. In Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3, pages 91–101. Springer, 2021.

Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Robust, deep and inductive anomaly detection. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10, pages 36–51. Springer, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. ArXiv, abs/1807.03888, 2018. URL `https://api.semanticscholar.org/CorpusID:49667948`.

Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. arXiv preprint arXiv:1906.02694, 2019.

Ki Hyun Kim, Sangwoo Shim, Yongsub Lim, Jongseob Jeon, Jeongwoo Choi, Byungchan Kim, and Andre S Yoon. Rapp: Novelty detection with reconstruction along projection pathway. In International Conference on Learning Representations, 2019.

Thomas Schlegl, Philipp Seeböck, Sebastian Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. pages 146–157, 03 2017. ISBN 978-3-319-59049-3. doi:10.1007/978-3-319-59050-9_12.

Chirag Agarwal, Daniel D'souza, and Sara Hooker. Estimating example difficulty using variance of gradients. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10368–10378, 2022.

Yijin Zhou and Yuguang Wang. Grod: Enhancing generalization of transformer with out-of-distribution detection. arXiv preprint arXiv:2406.12915, 2024.

Emmeke Aarts, Matthijs Verhage, Jesse V Veenvliet, Conor V Dolan, and Sophie Van Der Sluis. A solution to dependency: using multilevel analysis to accommodate nested data. Nature neuroscience, 17(4):491–496, 2014.

Archil Maysuradze, Olga Rink, Artem Fedorov, Andrey Tabachenkov, and Konstantin Vorontsov. Does annotating multi-spans improve classification in considerable text collections? In 2024 10th International Conference on Systems and Informatics (ICSAI), pages 1–6, 2024. doi:10.1109/ICSAI65059.2024.10893762.

Olga Rink, Viktor Lobachev, and Konstantin Vorontsov. Detecting human values and sentiments in large text collections with a context-dependent information markup: A methodology and math. In International Conference on Human-Computer Interaction, pages 372–383. Springer, 2024.

K. Vorontsov, I. Gladchenko, V. Lobachev, A. Mamontova, O. Rink, and N. Shabelskaya. Developing an open interdisciplinary classifier of human values by means of annotating multi-fragments. Vestnik of Saint Petersburg University. International Relations, 18:38–62, 2025. doi:10.21638/spbu06.2025.103.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. In Named Entities: Recognition, classification and use, pages 3–28. John Benjamins publishing company, 2009. URL `https://www.kaggle.com/datasets/naseralqaydeh/named-entity-recognition-ner-corpus`.

Anton Golubev, Nicolay Rusnachenko, and Natalia Loukachevitch. RuSentNE-2023: Evaluating entity-oriented sentiment analysis on russian news texts. In Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue" (arxiv:2305.17679, 2023. URL `https://github.com/dialogue-evaluation/RuSentNE-evaluation`.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. MERA: A comprehensive LLM evaluation in Russian. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9920–9948, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.acl-long.534. URL `https://aclanthology.org/2024.acl-long.534`.

Harsha Gurulingappa, Abdul Mateen-Rajpu, and Luca Toldo. Extraction of potential adverse drug events from medical case reports. Journal of biomedical semantics, 3:1–10, 2012. URL `https://huggingface.co/datasets/ade-benchmark-corpus/ade_corpus_v2`.

Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, Igor Rozhkov, Artem Shelmanov, Elena Tutubalina, et al. Nerel: a russian information extraction dataset with rich annotation for nested entities, relations, and wikidata entity links. Language Resources and Evaluation, pages 1–37, 2023. URL `https://github.com/nerel-ds/NEREL`.

Denis Gordeev, Adis Davletov, Alexey Rey, Galiya Akzhigitova, and Georgiy Geymbukh. Relation extraction dataset for the russian language. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog"[Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]. Russian State University For The Humanities Moscow, Russia, 2020. URL `https://github.com/denis-gordeev/rured2`.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreferencefor scientific knowledge graph construction. In Proc. Conf. Empirical Methods Natural Language Process. (EMNLP), 2018. URL `https://nlp.cs.washington.edu/sciIE/`.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. arXiv preprint arXiv:1911.10422, 2019. URL `https://huggingface.co/datasets/sem_eval_2010_task_8`.

Khiem H Le, Tuan V Tran, Hieu H Pham, Hieu T Nguyen, Tung T Le, and Ha Q Nguyen. Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis. IEEE Access, 11: 14105–14114, 2023.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. Transactions of the association for computational linguistics, 8:64–77, 2020.

Markus Eberts and Adrian Ulges. Span-based joint entity and relation extraction with transformer pre-training. In ECAI 2020, pages 2006–2013. IOS Press, 2020.

Shubham Toshniwal, Haoyue Shi, Bowen Shi, Lingyu Gao, Karen Livescu, and Kevin Gimpel. A cross-task analysis of text span representations. In Proceedings of the 5th Workshop on Representation Learning for NLP, pages 166–176, 2020.

Tom Fawcett. An introduction to roc analysis. Pattern recognition letters, 27(8):861–874, 2006.