ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет

«Высшая школа экономики»

Факультет гуманитарных наук Образовательная программа

Фундаментальная и компьютерная лингвистика

КУРСОВАЯ РАБОТА

На тему «Пробинг предобученных мультилингвальных моделей на материале Universal Dependencies»

Тема на английском “Probing of pretrained multilingual models on the knowledge of discourse on the material of Universal Dependencies”

Студентка 2 курса группы №212
Годунова Мария Эдуардовна

Научный руководитель
Сериков Олег Алексеевич
приглашенный преподаватель
Школы лингвистики ФГН

Научный консультант
Волошина Екатерина Юрьевна
ведущий инженер по разработке

Москва, 2023 г.

1. Introduction

While pretrained language models have shown remarkable performance on various language tasks, there is still much to be explored in terms of their ability to capture broader discourse in documents. Probing tasks can certainly offer means to analyze these capabilities, but it is important to design appropriate tasks that can capture the relevant aspects of discourse.

Although some studies [Koto et al., 2021; Chen et al., 2019; Nie et al., 2019] have looked at the ability of pretrained models to classify specific aspects of discourse, such as discourse markers and relations, there is still a need for more systematic investigations into their ability to model discourse structure as a whole. This can involve tasks such as identifying the relationships between sentences within a document, the role of one sentence in the document structure, investigating main topics, suitable ending and finding out whether one sentence belongs to the particular paragraph or not. Such tasks can help to shed light on the strengths and limitations of pretrained language models in capturing the nuances of discourse structure.

We frame almost all presented tasks as binary classification problems, but they involve different aspects of Rhetorical Structure Theory, models' understanding of which is being tested in this study. The peculiarity of our work is that we test multilingual, rather than basic, models on the corresponding data for ten languages, exploring only modern pre-trained transformer models.

Due to the Rhetorical Structure Theory, the use of relational propositions is crucial for creating coherence in multi-sentence texts. These propositions assert functional relations between different parts of the text, and a reader recognizes and attributes these assertions to the text in order to understand the overall structure and meaning. The tree-like connectivity of RST structures also ensures that removing any one relational proposition can disrupt the coherence of the entire text. RST alights units which are roughly equivalent to clauses, meanwhile our main focus is on relationships between selected sentences.

As was mentioned earlier, all models used in our research are transformers. We chose this architecture because this model has been shown to achieve state-of-the-art results on various natural language processing tasks, including machine translation, language modeling, and question answering. The Transformer model, introduced in (Vaswani et al., 2017) is a neural network architecture that relies on self-attention mechanisms to compute representations of its input and output sequences. This approach allows the model to capture long-range dependencies and relationships between elements in the sequence, without the need for recurrent neural networks (RNNs) that rely on sequence alignment. When updating the memory vector with recurrent neural networks there is no permanently stored information, so only the vector obtained from the previous layer is taken into account, but not all going before that. That is, all weights are updated for each layer. Therefore, it is difficult to accept long sequences of n-grams, and the meaning of the context is lost. For models with scaled dot-product attention a single convolutional layer whose kernel is smaller than the number of elements in the input does not connect all pairs of input and output positions, it has a scaled dot-product attention, meanwhile transformers provide us with multi-head attention, which allows the model to jointly attend to information from different representation subspaces at different positions. By using multiple attention heads, the model can learn more diverse and complex representations of the input which are processed by the function in parallel. Additionally, the linear projections provide the model with additional parameters to learn, allowing it to better capture complex relationships between the inputs. Overall, multi-head attention has been shown to improve the performance of various natural language processing tasks, such as machine translation and text classification.

The tasks we have compiled for different models have already been used in other probing studies. At the same time, we tried to make a collection of tasks from different previous works in order to compare the results obtained and identify new problems in the assimilation of the text discourse by neural networks. We are also interested in how differently the models perform depending on the language in which the text is written. Presumably, despite the fact that all the models we used are multilingual, the best accuracy will be revealed for tasks in English, regardless of the chosen model, since the vast majority of texts in the pretraining data are written in English.

In our study we tried to accumulate all kinds of tasks (collected from the materials of previous works, which will be described later, and adjusted in such a way as to cover as many aspects as possible in understanding the discourse). At the same time, as will be demonstrated in the next section, by limiting the task to predicting a smaller set of discourse relationships, the model can focus on key linguistic features that are more indicative of entailment and contradiction. This can result in more accurate predictions and better overall performance. Therefore, in our research we do not expect to get the same outstanding results.

2. Tasks

Several works are devoted to the study and prediction of a marker linking two sentences [Nie et al., 2019]. Discourse markers are important for understanding the relationships between sentences, and predicting them can be a useful task for training models to understand sentence meanings. This task falls somewhere between predicting entailment, which requires manual annotation, and predicting missing words in large corpora. Discourse marker prediction can provide a way to learn from naturally occurring data with relatively little need for manual annotation. We also use this task in our research, but since our work is devoted to testing multilingual models, the difference is that we do not set a fixed list of markers for model input (as was done in the article where English texts acted as a training dataset with sentence pairs from BookCorpus). The similarity with testing based on the DisSent task is that we also do not update the sentence embedding model's weights. The main difference between our approach and the described study of discourse markers is that the latter uses the BERT MLM model. In our work, on the other hand, the understanding of the structure of discourse by the model itself is checked, so the masking algorithm is not applied. Best results were achieved with multi-task training methods with LSTML between 78.6 and 94.0 accuracy level.

In [Chen et al., 2019] discourse probing is conducted via DiscoEval, which is the set of 7 tasks that covers a variety of domains, including Wikipedia, stories, dialogues, and scientific literature. Each task group is designed to test specific aspects of discourse understanding, such as coreference resolution, discourse relation classification, and sentence ordering. It was found that ELMo performs well on SentEval, while BERT performs best on DiscoEval. This suggests that BERT is better at understanding discourse than ELMo, while ELMo performs better with tasks on syntax. This result indicates that BERT as a whole may have a better understanding of the document structure than other models, but this needs to be checked on a wide range of tasks. The main difference between this research and our work is that we do not concatenate vectors for separate sentences, but use the sequence as an input for our models. From the described article we borrowed and adapted the following tasks making them suitable for multiple languages:

1. Sentence Position (SP)

This task was created to test the model's understanding of linearly-structured discourse. By randomly moving one of the 5 sentences to the first position, the model must be able to accurately predict the correct order within the discourse sequence based on the content of the sentences. This type of testing could be useful in evaluating the effectiveness of natural language processing models in tasks such as story comprehension or summarization.

2. Binary sentence ordering

This task is to identify the correct order between the two contextually codependent sentences. BSO could be useful in testing a model's ability to capture local discourse coherence and understand the relationships between adjacent sentences in a text. It could also be used in tasks such as question answering or information retrieval, where the order of sentences can be important for understanding the meaning of a text.

3. Discourse Coherence

Having a sequence of six sentences that form a coherent paragraph, we need to randomly replace one sentence from the coherent sequence with a sentence from another discourse. The replacement sentence can be at position 2, 3, 4, or 5 in the original sequence. A model needs to determine whether the resulting sequence of six sentences still forms a coherent document. To do so, we need to replace one of the six consecutive sentences in the original document with the sampled sentence from a new text.

For most of these tasks, except for DC, BERT showed best results.

We base our methodology on [Koto et al., 2021]. The peculiarity of this probing study is that the authors tried to trace differences of understanding of the discourse by models depending on the layer. It seems that RoBERTa and BART excel at capturing discourse information, particularly in the encoder layers. However, it was claimed that the BART decoder may not be effective enough at language understanding, as it focuses more on sequence generation, T5 performs in almost the same way. BERT performs well in deeper layers, while ELECTRA performs best in the last three layers. As for the tasks that we borrowed from this article, these are:

1. Next sentence prediction

This task is framed as a 4-way classification, with one positive candidate and three negative candidates for the next sentence. The preceding context consists of 2 to 8 sentences, while the candidates are always single sentences. Nevertheless, we adopted it as a binary classification task with mixing one of the sentences in a way that researchers in [Chen et al., 2019] did. Still we did not change the form of the task with its shape of the sequence.

2. Sentence ordering

This task is to determine whether the order of sentences in the document is correct. Texts in which from 3 to 7 sentences are mixed are presented as incorrect options. It is important to mention that in this case sentences are mixed not among all the documents, but within the same sequence.

3. Cloze story test

Data for this task consists of sequences with 4 sentences in each. Model needs to pick the best ending sentence for all documents. We made our adapted task binary, so for incorrect pairs ‘key:value’ we shuffle ending sentences within all documents.

All models perform well on the NSP task (about 0.8 accuracy) but struggle with sentence ordering (~ 0.4 accuracy). This suggests that these models may not be effective in modeling discourse over multiple sentences. The degradation of performance as the number of sentences to re-order increases further supports this conclusion. Despite the theoretical differences between the discourse connective data and RST, the models produced similar patterns in their output. BART seems to perform better than other models in layers 1-6 for the tasks of discourse connectives, RST nuclearity, and RST relation prediction, as well as NSP and sentence ordering. All models improve as we go deeper into the layers for the Cloze Story Test which suggests that the higher-level understanding of the story is captured deeper in the models. This is an important finding as it suggests that the models are able to capture the deeper meaning of the story as they process it.

Thus, it can be noted that probing studies in the field of discourse have already been conducted before, but they affected a small number of languages and focused on a limited number of tasks in terms of content: either predicting a discourse marker, or analyzing the model's understanding of the coherence of the entire text, or the connectivity between a certain number of sentences in a document. Therefore, it seems important to us to conduct a general study, having compiled tasks on various aspects of discourse and choosing different languages as a training sample, but at the same time not focusing heavily on the layers of the neural network, but analyzing the overall result.

3. Methods

3.1 Models

BERT multilingual base model (cased) was trained on a large corpus of text from Wikipedia in the top 104 languages [Devlin et al., 2019]. The MLM objective allows the model to learn contextual representations of words by predicting randomly masked tokens within the input text. This model is a powerful tool for natural language processing tasks in a wide range of languages, and its case sensitivity allows for even more accurate language modeling. One of the two BERT objectives is next sentence prediction. It means that during pretraining the multilingual BERT model concatenates two randomly selected sentences from a given corpus, with a [SEP] token between them, and masks some of the tokens in both sentences. The model is then trained to predict whether the two sentences were adjacent to each other in the original text or not. At the same time, by pretraining the model on this task, it can then be fine-tuned on a wide range of downstream NLP tasks with better performance. Therefore, we can suppose that multilingual BERT will perform well on the majority of the compiled tasks.

XLM-RoBERTa (base-sized model) was trained on 100 languages data (CommonCrawl corpus) [Conneau et al., 2020]. Unlike BERT, it was pretrained only with the Masked language modeling (MLM) objective. It means that by masking parts of the input sentence, the model is forced to learn contextual relationships between words and phrases, which helps it to generate more accurate and natural language responses. Therefore, this version of RoBERTa will precisely perform better on syntax tasks, than on discourse ones.

Multilingual GPT model was trained on 60 languages data using Wikipedia and Colossal Clean Crawled Corpus [Shilazhko et al., 2022]. The main feature of this neural network is that it uses sparse attention mechanism and, as a result, achieves greater scalability and performance across multiple domains than transformers with multi-head attention. In practice it is done with two-dimensional factorization of the attention matrix, doing this the network can use only two steps of the matrix to attend to all positions. The main difference between models like BERT, RoBERTa and GPT is that the first ones need a larger sequence for masking, that is, generating text, than the GPT.

mT5 model was trained on a Common Crawl-based dataset covering 101 languages. This model’s architecture takes parameters from encoder-decoder transformers, but replaces encoder-decoder attention with a full attention algorithm. By adding a sentinel token to represent consecutive masked tokens and training the model to predict the number of words missing in the blank, it can better understand the context and generate more coherent text – this is T5 advantage over BERT, for instance. Additionally, by splitting the input text and having the model auto-regressively regenerate the output, it can learn to generate text that is more similar to human writing. Presumably, taking into account this transformer’s masking qualities, it will perform better than other models on discourse connective prediction.

The typical distribution of performance observed for these models is that RoBERTa in most cases demonstrates the best accuracy with BERT being slightly worse, meanwhile mGPT performs the worst of all and mT5 being insignificantly better than mGPT.

3.2 Languages

All data for our probing tasks was taken from the UD framework. The table below shows the number of examples for each task and language that were extracted from treebanks:

	Russian	Bulgarian	Czech	Serbian	Catalan	French	Latin	English	Armenian	Turkish
Binary sentence ordering	15632	17354	1230	1389	1476	1468	1474	1823	2094	15203
Cloze story test	9385	67142	18437	6780	47852	1201	51867	21770	46209	12064
Discourse coherence	3450	33567	2143	2013	34701	1750	21602	3502	29436	3972

Next sentence prediction	12949	42781	13561	4998	21952	7620	13764	16067	49673	30166
Sentence ordering	5302	18579	9450	4356	1938	2395	1027	3750	19820	1960
Sentence position	2790	22152	7664	1732	9909	1042	1395	7438	10347	1704
Discourse connective prediction	14036	37620	2089	1503	7605	1201	3047	8993	28049	6775

Table 1: Number of examples for each language and task

Most of the languages in the sample belong to one language family – Indo-European. At the same time, it was found out that the sample (treebank) of the Turkish language is one of the largest in the database. In addition, Turkish is part of one of the largest language families, Altai, so we also included it in the training data. Among the groups of the Indo-European family, some of the most common were chosen: Romance, Germanic and Slavic. The sample also included the Armenian language, since data for this language was huge enough to parse it and no probing studies with its participation had been conducted before. The remaining languages were selected according to the following principle: they either often appeared in NLP works (English, French, Russian) or, conversely, were not or little studied using machine learning – this is the majority of the languages in the sample. This decision was made in order to compare the performance of one model with respect to different languages in order to understand how much the number of examples in a particular language (such as English, the examples in which the vast majority of models in the training sample) determines the understanding of several idioms by a multilingual transformer at once. All languages used in the work are presented in the table:

Indo-European	Slavic	Russian
		Bulgarian
		Czech
		Serbian
	Italic	Catalan
		French
		Latin
	Germanic	English
	Armenian	Armenian
Altaic	Turkic	Turkish

Table 2: Languages, their families and groups

3.3 Probing tasks

As it was mentioned earlier, the ideas for all the tasks were taken from the articles of the predecessors. At the same time, all the previous probing studies, the tasks from which we borrowed, were mainly conducted on the basis of English and they did not use multilingual models. Therefore, we needed to adapt all the borrowed tasks in such a way that they correspond to the treebanks of any language from the database. This is one of the reasons why in our study we did not test the models’ understanding of segmentation into clauses: in each language the division into clauses occurs differently, therefore, we are not allowed to implement a universal code for extracting EDU.

All tasks, except for the discourse connective prediction, are a binary classification problem. This approach was chosen to better evaluate the accuracy of the models. Taking into account that many of the languages in the sample are not very large and have not been studied sufficiently, their datasets are also small. As a result, if, for instance, in the case of a task for the order of sentences in a sequence, integer answers with an order were submitted to the input, not all numeric sequences would occur in the training sample. Therefore, given that all analyzed models have masking objects, correct and incorrect sequences should be generated by the models themselves. Thus, the correct sequences are marked as 1, the incorrect ones as 0.

First we will look at the task which stands out among other ones – prediction of discourse connective.

- **Discourse connective prediction**

Unlike previous approaches [Koto et al., 2021], we did not set a frequency threshold for accounting the connective due to the limited shapes of the data for some languages. Following the approach presented in [Malmi et al., 2017], we predict only connectives which occur in the beginning of the sentence, considering this as a base position for an explicit binding marker. This choice is explained by the fact that before testing the understanding of implicit connectives by a multilingual model, we must first pay attention to explicit ones.

S1	Obviously because I want to vote.
S2	If anyone else has voted, what did you guys vote for?
Discourse connective	And

Table 3: Examples of DCP task

- **Sentence position**

The position of a sentence within the text can provide context and help to understand the overall structure and purpose of the document. The opening sentences often provide an introduction to the topic, while the following sentences provide more detailed information and support the main idea. [Chen et al., 2019] discovered that in the SP task removing the surrounding sentences can make it more challenging to accurately predict the position of the target sentence, as the model has less information to work with. Due to the fact that the context plays a crucial role in a sentence position, we decided to take 5-sentence sequences for our dataset and swap the fourth of them with the other randomly chosen sentence in a sequence. This method was partly proposed by [Mostafazadeh et al., 2016], and, although in the described article researchers swap the forth sentence with the first one, we decided not to swap fixed elements of a text, and choose one of them randomly, so we complicated the task, because usually models demonstrate high results in this test.

Examples	Labels
Now, people wonder if Google can even survive. What they wonder is whether Google can be anything more than what it's always been -- a great search engine with some real grass-roots support, successful by the grace of simplicity. The problem is that customers attracted by a simple interface are among the least loyal you can find -- witness the fight-for-fewest-features between low-end camera companies. It's tough to make money branching out when your appeal is in your focus. If they continue to add features so they can justify their likely sky-high valuation, Google risks losing a huge chunk of their customer base to the next keep-it-simple search engine	1
The Greater New Orleans Fair Housing Action Center (GNOFHAC) filed a housing discrimination complaint against the Housing Authority of New Orleans (HANO) last week. The complaint, filed with the United States Department of Housing and	0

Urban Development, accuses HANO of violating a 2003 enforcement agreement entered into between former St. Thomas Housing Development residents, the City of New Orleans, HANO, and the US Department of Housing and Urban Development during the HOPE. VI redevelopment of St	

Table 4: Examples of SP task

- **Binary sentence ordering**

This task differs from SP in that a much smaller amount of context is supplied to the input, so this test allows us to evaluate the ability of the model to determine the relationship between the minimum context of two sentences.

Examples	**Labels**
Based on specific intelligence inputs, Army arrested Ghulam Mohiuddin Lone, a LeT man, from Doda district. During the preliminary interrogation, Lone 'confessed' his involvement in the blasts and gave several vital clues	1
Salon is clean and girls are nice. I didn't know what I was missing	0

Table 5: Examples of BSO task

- **Discourse coherence**

Capturing discourse coherence in sentence representations can be challenging because it involves identifying the relationships between words and phrases within a sentence and across multiple sentences. In order to evaluate the ability of a model to capture local discourse coherence, it would need to be able to capture characteristics of the entity being discussed or the topic of the sentence group, and perform inference across multiple sentences to determine the coherence of the discourse. This can be a non-trivial task, as it requires the model to have a deep understanding of the underlying meaning and context of the text being analyzed. Connectivity within the document, in accordance with our research and the previous work, is determined from 6 sentences. In our case, this number is fixed. Negative examples are created by replacing one of the sentences with a sentence from another text.

Examples	**Labels**
As we have already noted, many people today may also think of the universe in Genesis 1 as created from nothing. This idea may seem strange if they are familiar with the King James Version's translation: "In the beginning, God created the heaven and the earth." However, as we have seen, this translation is not correct. Even so, there might seem to be room for the idea of creation made from nothing. It might appear to readers that this idea of creation from nothing is expressed or symbolized in Genesis 1:2 by the mention of "void and vacuum". These two nouns, connected by a conjunction and forming a fixed, com pound phrase, would seem to describe precisely the kind of nothingness that facilitates the concept of creatioт ex nihilo	1
Genesis 1 envisions creation not simply as God making; it is as much as a process of "separation" and differentiation of elements from one another, as we will see in chapter 3. It involves a transformation from an unformed, wate1y mass into the world that sustains human existence with water.Creation is a process in which a deity makes the world, as it came to be. Psalm 33:6-7 nicely expresses this transformation. Let's consider this more closely.	0

Table 6: Examples of DC task

- **Next sentence prediction**

In the source paper there were 3 negative candidates and a single positive one for the next sentence, but we adopted it as a binary classification problem, therefore, for negative examples of sequences we shuffle the last sentence with the other sentence, but not within one document to sustain the text structure. This task is another step to predicting implicit discourse relations, since it gives to the model representation of a common next sentence [Shi, W., Demberg, V., 2019].

Examples	**Labels**
It was ok, nice management, they let us check in early, but the place was old. It was clean, but just a little dumpy. Hard to get into though because of road construction	1
Horrible customer service. I came in to get a nice gift for my wife. But thankfully there are other flowers shops around	0

Table 7: Examples of NSP task

- **Sentence ordering**

Originally this task was done by shuffling from 3 to 7 sentences, providing the model with the correct ordering and then predicting it. We reworked it by shuffling all the sentences for the incorrect sequences. This method allows the model to select the most consistent sequences in the dataset and further develop a coherency metric based on NLP analytics [Barzilay, Lapata, 2017].

Examples	**Labels**
This is unlike the situation last year in Asia when we evacuated US citizens from areas that were hit by the tsunami - a phenomenon that is much less predictable than the Hezbollah-provoked destruction that rained down on Lebanon. The American-Arab Discrimination Committee is suing Condoleeza Rice and Donald Rumsfeld, charging that they mismanaged the evacuation efforts	1
My favorite so far in Bellevue. They have good sushi for a good price	0

Table 8: Examples of SO task

- **Cloze story test**

As was described earlier, in this task the model receives a document containing 4 sentences as input and chooses the best completion for the text. We changed this task by making the answers binary and shuffling the last sentences in the sequence for negative samples. We also did not take into account text biases conducting stylistic feature analysis [Sharma et al., 2018] as it is harder to trace on a large language data. In [Mostafazadeh et al., 2016] it is claimed that cloze story test indeed helps to identify the model’s understanding of the text coherence. If a model performs well on this task, it suggests that it has some level of understanding of the story's narrative structure and can generate coherent and logical endings based on that understanding.

Examples	**Labels**
Heh, yep, I like to wear silk chemises. Also panties even stockings with garter belt .Later on I red somewhere that it's seakness	1

You've already asked this . Why would someone post the location of a dealer in a public place? Drop by my house, I can get you some real cheap. Give me an address or something please idk	0

Table 9: Examples of CST task

3.4 Probing methods

In our study we used all models without performing finetuning. Thus, all models were pre-trained – this is the basic probing approach. This was done because we aim to test the basic models in understanding the discourse. The diagnostic classifier logistic regression was used to assess the quality of the models' performance.

Also we use the [CLS] embedding to get the predictions from transformers following standard practice.

4. Results

In the beginning we analyzed all averaged results depending on the type of tasks:

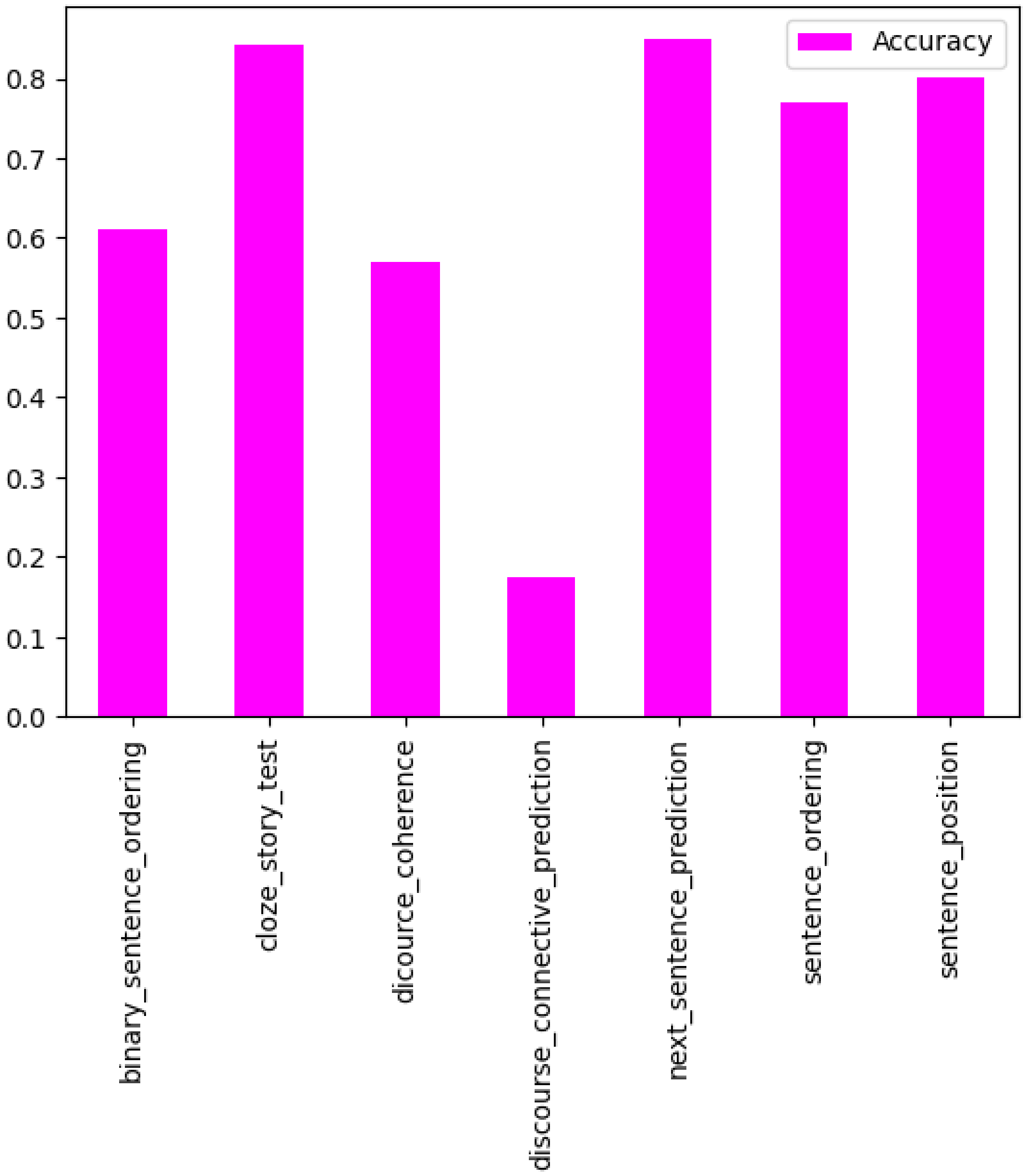

Figure 1: Averaged accuracy depending on the type of task

Transformers show the best performance on Cloze Story Test (0.842) and Next sentence prediction (0.849) tasks. These tasks are united by the fact that in both the focus of the prediction is the last sentence of the document. Due to the fact that an average accuracy for the Discourse coherence task is much lower (0.57), we cannot conclude that good models' results for these tests correlate with the number of sentences in the input. It can be attributed to the fact that for the DC test there was a sequence of 5 sentences provided, while for CST all documents consisted of 4 sentences. At the same time, DC task asked the models to determine the coherence of a document in which any of the five sentences could serve as a replaced sentence, except for the first. Meanwhile, in tasks in which the highest accuracy of the models' performance was recorded, the focus sentence is fixed (always the last, only the size of the sequence varied).

Nevertheless, context definitely affects the model’s performance on the task For example, models perform worse on a task in which it is required to determine the correctness of the order of sentences within a binary sequence (0.61) than on a task containing multiple sequences (0.77). Also in the context of this task (SO), it is important to mention that, unlike the work of our predecessors, we considered it as a binary classification problem, therefore, we cannot exactly know how much the model understands the real order, and not just reveals an irregularity in connectivity.

Quite unexpected and contrary to hypotheses results were obtained for the task Sentence Position. In the original paper, the BERT-Large accuracy for SP was 53.8%, while in our case we got an 80% accuracy. However, we changed the task in the way that we swapped not only the first sentence with the fourth sentence, but only the fourth sentence was fixed in our case, which could be swapped with any other sentence in the text. It means that with every iteration we swapped the fourth sentence of the sequence and randomly chose the other sentence. Such a difference in the results may indicate the importance of the first position in the sequence, the weight of which in the context of the multi-head attention method is the largest.

The worst average accuracy is for Discourse connective prediction (0.175). Our results for this task are much worse than in the source research. That can be explained by the fact that an average result is affected by too huge and too small values. The average value, unlike the median, may deviate from the actual most frequent value, so in some cases it cannot serve as an objective indicator. On the other hand, the median accuracy value (0.238) for this task is somehow closer to the results presented in the article. To sum up, the difference between results that we got and the ones received in the [Nie et al., 2019] can be explained by the fact that our data for particular languages is much smaller, therefore, models obtain less connective markers and, as a result, perform worse.

To find out whether certain languages distort the test results, we will compare the overall average accuracy obtained for all tasks for each language:

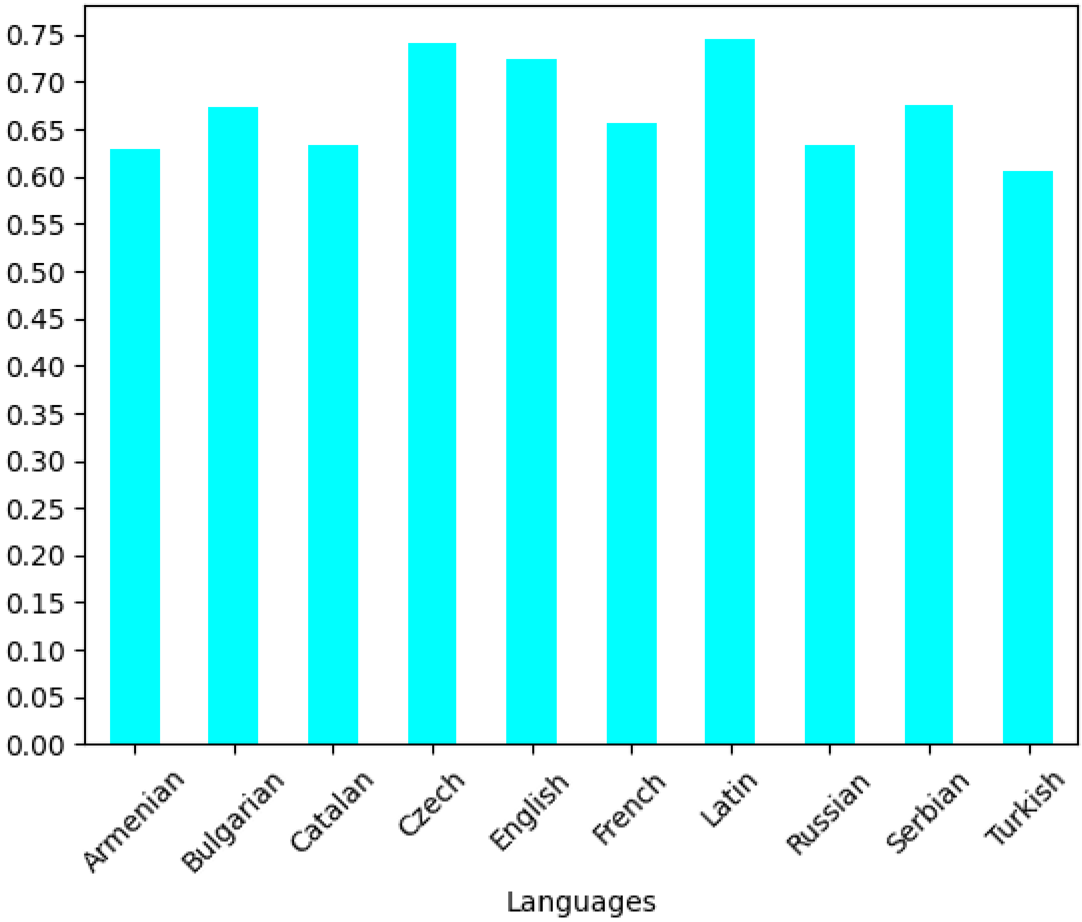

Figure 2: Average accuracy for each language

It can be seen that languages do not vary drastically in their average results. Contrary to expectations, the best results were obtained not for English, but for Latin and Czech, supposedly, due to the fact of bigger data for particular tasks.

To better assess the correlation between the structure of a particular language and the performance of transformers, consider the accuracy averaged over various tasks for each model and each idiom.

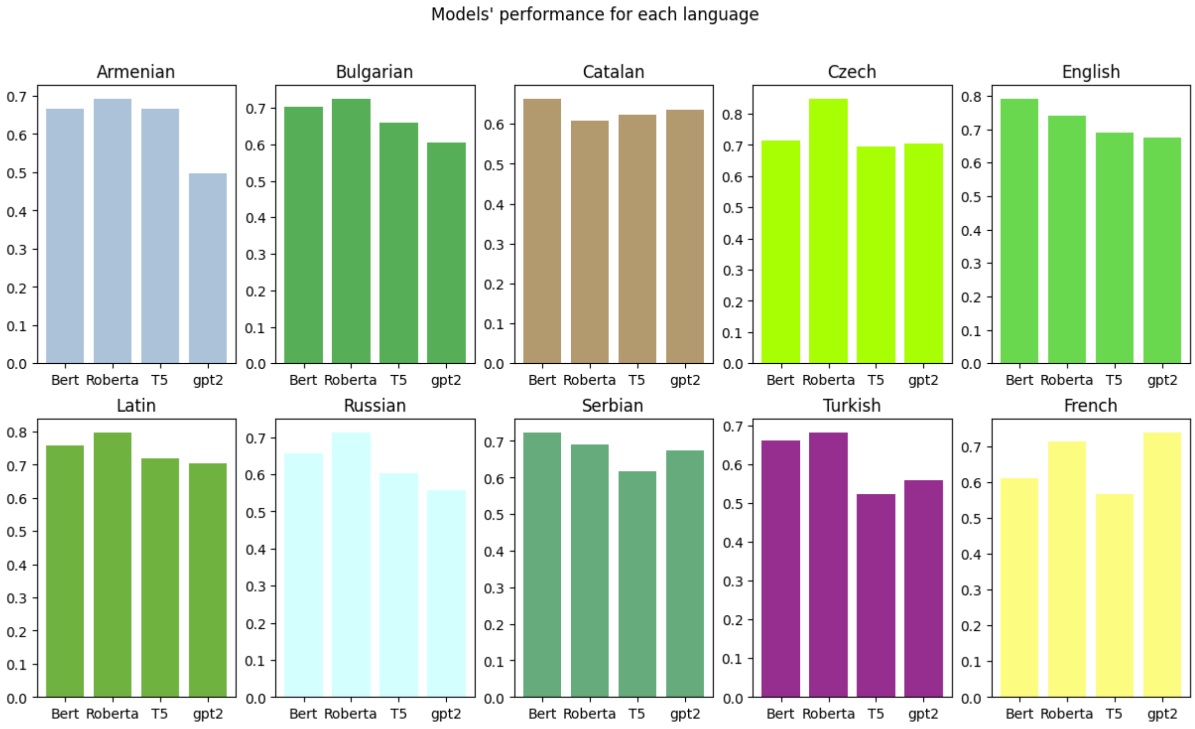

Figure 3: Average accuracy depending on the language and type of model

1. Armenian

The average results for Armenian (see Figure 3.1) demonstrate the trend that is the most common in accuracy distribution for these models: RoBERTa shows the best accuracy. The result is unexpected as BERT and T5 demonstrate almost completely identical results. Since there are practically no studies devoted to the structure of discourse in the Armenian language, the only feature that we can rely on in analyzing the result obtained is the lack of resources of this idiolect.

BERT and T5 use a bidirectional Transformer, while mGPT uses a left-to-right Transformer. This difference in architecture provides the advantage of the first two models over mGPT, since they can directly access all positions in the sequence simultaneously (parallel receipt of vectors in the encoder), which, as has been proven, provides better accuracy in low-resource languages. The fact that the Armenian language was not in the mGPT training sample also plays a crucial role in further results.

2. Bulgarian

In this case (see Figure 3.2), there is a distribution common to most tasks (and obtained by averaging the results for both tasks and languages), in which RoBERTa demonstrates the highest accuracy, BERT performs slightly worse, followed by T5, and the worst results are observed for mGPT.

According to the study [Dagnev et al., 2019] conducted on the basis of the corpus of English and Bulgarian articles, there are the so-called ‘complicated simple sentences’. This term is used to denote differences between the structure of utterance in English and Bulgarian, which may reduce the accuracy of speech for transformers with unidirectional attention. If the model generates the weights of each input based on the previous context, but does not take into account the subsequent context with the help of recalculation, the features of the utterance can serve as a barrier to good results. Those sentences generate the heavy complementation and use of compound and complex sentences in Bulgarian, and that is the main difference between Bulgarian and English rhetorical structure. It can be that the model borrows discourse patterns from the language that prevails in the training sample. Thus, presumably, the fewer languages in the model and the greater the presence of English, the greater the accuracy in those languages whose discursive patterns are similar to patterns in English.

3. English

Results (see Figure 3.5) demonstrated by models for English may show the real distribution of ratings, due to the fact that this language always has the largest number of examples in the training sample. It can be seen that BERT, instead of RoBERTa (as in most other cases), performed best, RoBERTa has lower average accuracy, meanwhile T5 and mGPT demonstrate almost the same results with T5 slightly outperforming mGPT. Therefore, we can assume that BERT potentially has more knowledge about discourse, but either it is more difficult for it to cope with longer sequences, or it has a smaller multilingual base, that is, we are no longer talking about the differences between the parameters of BERT and RoBERTa, but about the differences between their multilingual counterparts, bearing in mind the inequality in the training sample and lack of language embeddings in case of RoBERTa which provides its better dealing with code-switching.

4. Catalan

For Catalan we observe extremely unexpected results (see Figure 3.3), as BERT demonstrates the best accuracy (while still lower than the average value for other languages), and mGPT is in second place. A slightly lower average accuracy was demonstrated by T5, and RoBERTa performs the worst. These results strongly contradict intuition and can be explained by particularly clear rules for constructing a text in Catalan, namely right-branching (right-dislocation constructions). Therefore, we may assume that RoBERTa does not cope well with languages that have a rhetorical structure, which is very different from the structure of discourse in English. As for the high accuracy obtained when testing mGPT, this fact can be explained by the results and hypotheses proposed by [mGPT: Few-Shot Learners Go Multilingual]. For a few-shot method increasing the number of examples in a sample leads to the decreasing of accuracy, therefore, having no examples at all in training data may be beneficial in case of this approach.

5. Czech

With Czech (see Figure 3.4) RoBERTa performs best and other models' results are much lower being almost the same in comparison to each other. This result may stem from the fact that the compilers of the treebank for the Czech language emphasized long-distance discourse relations in accordance with [Polakova, Mırovsky, 2019]. As it has been proven, one of the main advantages of RoBERTa is the ability to analyze large sequences of text [Conneau et al., 2020].

6. French

French (see Figure 3.10) seems to be the only language for which mGPT outperforms other models with RoBERTa still having high accuracy value, BERT performing significantly worse and T5 occupying the last place in the results. These results can be partially explained by the differences in the structure of discourse of French and English, presented in [Kaplan, 1966]. Kaplan states that the utterance in Romance languages (in comparison with the linear structure of utterance in English (see Figure 4)) is distinguished by ornateness, length and constant deviation from the topic, therefore, in order to understand the sequence of sentences, you need to constantly keep in mind the initial information and have a complete picture, since the main idea is usually expressed at the beginning and at the end.

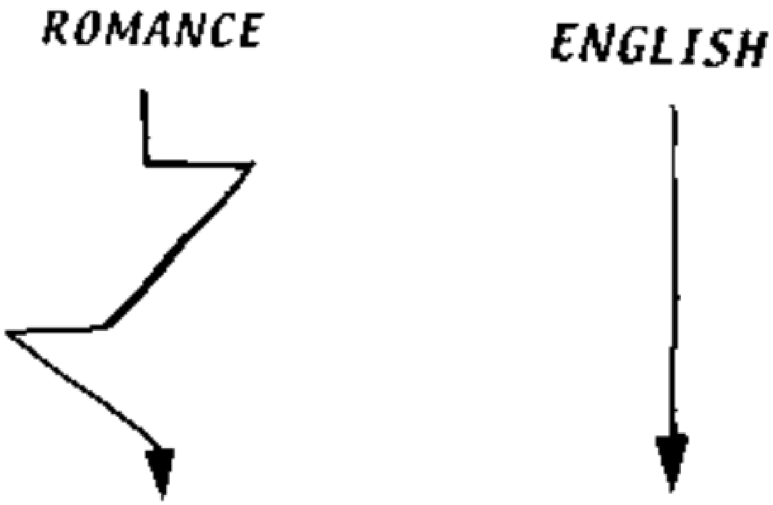

Figure 4: Comparison of discourse structure in Romance and English from Kaplan (1966)

In this vein, the high accuracy obtained by RoBERTa is easily explained, but it is more difficult to justify an even higher accuracy of mGPT. We assume that a similar result was achieved thanks to a sparse attention mechanism in which each output position only attends to a subset of input positions, and the subset is typically chosen based on some predefined pattern or rule. By limiting the number of input positions that each output position attends to, the computational cost of attention can be significantly reduced, especially for long input sequences. In such cases, the attention mechanism can focus on the relevant input tokens and ignore the irrelevant ones [Martins et al., 2020].

7. Russian

For Russian (see Figure 3.7) we observe exactly the same distribution that has already been described for Bulgarian: RoBERTa shows the best performance, BERT performs slightly worse, T5 is in third place and mGPT shows the worst accuracy. Since the distribution was almost the same for the Czech language (the difference is that the mGPT showed slightly higher accuracy than T5), it can be assumed that such similarity in the results is explained by the affiliation of the above languages to the same language group. This assumption is contradicted by the distribution of model accuracy obtained for Serbian, but in this case it seems appropriate to refer to the lack of resources of this language.

It also seems important to note that we can hardly consider that the sparse attention mechanism applied for mGPT helps to cope best with long sequences, rather it turns out to be the best in the case when the main topic of the utterance is concentrated at the beginning and end of the text. At the same time, for Russian Kaplan establishes a structure characterized by situationality, instability of discourse patterns and a constant change of focus of text, which, although in some sense similar to the ornate rhetorical structure in French (both are nonlinear with respect to discourse in English), differs in the lack of integrity according to Kaplan:

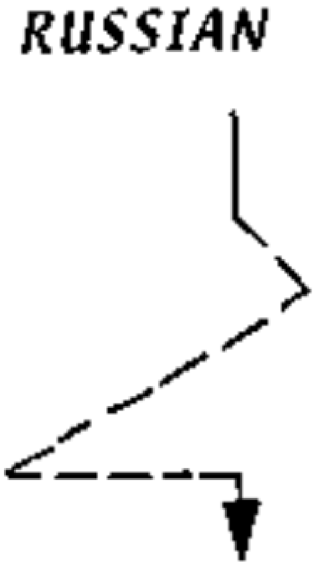

Figure 5: Discourse structure in Russian from Kaplan (1966)

It can be assumed that this difference is the reason for the strong decrease in the accuracy of the mGPT for Russian compared to French.

8. Latin

The hypothesis that the models act equally (in relation to each other) for languages belonging to the same language group and therefore having common discourse patterns is confirmed by the example of French and Latin (see Figure 3.6). RoBERTa achieved the highest performance compared to BERT, T5, and GPT. BERT, on the other hand, showed slightly lower performance than RoBERTa but still outperformed T5 and mGPT. T5 was ranked third in terms of accuracy, and mGPT showed the lowest accuracy of the four models. At the same time, this is still a hypothesis, since such a distribution seems to be universal in most cases and has also been recorded for most languages of the Slavic group.

ALso in [Kroon, 2009] it is established that the structure of discourse in Latin is characterized by strong fragmentation in the sense of the distance between discursive units united by various word forms, which are also polysemic. Thus, the high average accuracy of most models in tasks with the Latin language reflects the ability to build non-trivial connections within the text and understand the general meaning.

9. Serbian

As was mentioned earlier, Serbian (see Figure 3.8) stands out among the other studied languages of the Slavic group in that for him BERT turns out to be better in accuracy than RoBERTa, and mGPT surpasses T5 in average test results. Since Serbian discourse has not been sufficiently studied before, the only factor by which we can explain such a distribution of model performances is the small amount of data for the language under study. Regarding BERT's superiority over RoBERTa, it can be assumed that it is explained by differences in the token masking procedure - in the case of BERT, it is always a fixed set of tokens when the model is working, which may help in working with low-resource languages. mGPT is supposedly superior to T5 in accuracy due to zero-shot and few-shot learning approaches, which do not require a large array of input data.

10. Turkish

RoBERTa achieved the highest performance (see Figure 3.9), surpassing BERT, T5, and mGPT. However, BERT still performed better than mGPT and T5, T5 showed the lowest accuracy among the four models. To find out the cause of these results, we should carefully analyze the discourse of Turkish. Distinctive features in this case are shared arguments and properly contained arguments which can both contribute to the complexity of discourse structure [Demirsahin, 2015]. Shared arguments occur when two distinct discourse connectives use the same text span as their argument. This can create ambiguity or confusion for the reader or listener, as it may not be immediately clear which connective is governing the argument.

Properly contained arguments occur when a larger text span that is the argument of one connective contains a smaller text span that is the argument of another connective. This can also create ambiguity or confusion, as the reader or listener may need to mentally sort out which connective is governing which argument.

Both of these phenomena can increase the cognitive load required to process and understand the discourse, making it more complex and potentially more difficult to comprehend.

As was claimed in [Shliazhko et al., 2022], sometimes the variety and adversarial nature of examples leads to the deterioration of accuracy, and again we refer to the specific methods of learning used for mGPT. For RoBERTa the complexity of text may be potentially overcome via dynamic masking, as in this case the number of potentially different masked versions of each sentence is not bounded like in BERT, therefore the probability of understanding complicated structures gets bigger.

The other reason which is mutual for RoBERTa and mGPT as opposed to BERT and T5 respectively is an improved text generation, which may affect the quality of performance for languages with a more complex discourse structure than in English, since the model learns to "guess" the subsequent context without relying on already learned patterns.

Now we are going to examine the correlation between each model's understanding of discourse and different types of tasks.

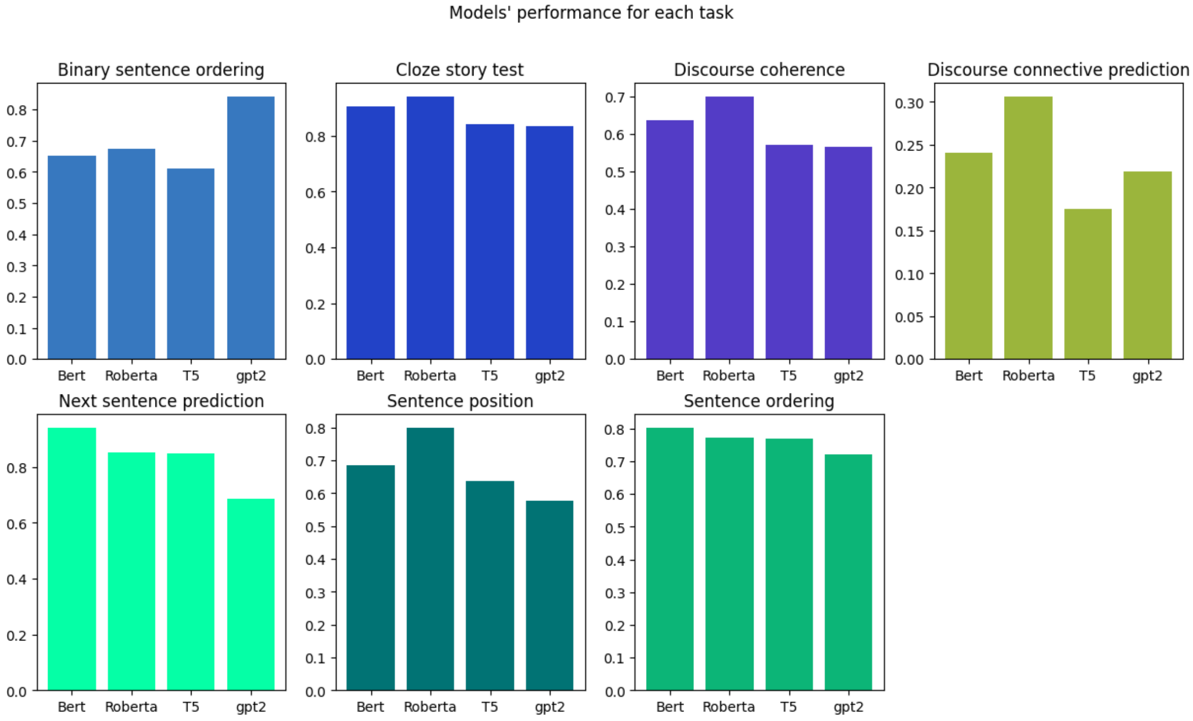

Figure 6: Average accuracy depending on the task and type model

1. Binary Sentence Ordering

The uniqueness of this task lies in the fact that mGPT copes with it best (see Figure 6.1), which in all other tasks demonstrates the lowest accuracy rates due to some obvious issues like the lack of some investigated languages in the mGPT's training data (still this issue does not cause too much problems as multilingual transformers tend to adjust to the new idioms using available data).

Other models perform with almost the same results and set a general trend, according to which T5 is the worst in the remaining three transformers to cope with most tasks, BERT and RoBERTa show the best results, while most often, as in this case, RoBERTa demonstrates the best results.

2. Cloze story test

In this case (see Figure 6.2) RoBERTa shows the best performance. As mentioned above, we have used some low-resource languages in our research, and it was proved in [Conneau et al., 2020] that RoBERTa surpasses BERT on cross-lingual classification, but specifically with low-resource languages being used for training data. RoBERTa's superiority over BERT can be explained by the phenomenon called the "generalization gap", which occurs when a language model's ability to perform well on downstream tasks (such as text classification or machine translation) exceeds its performance on the validation set during training.

One possible explanation for the generalization gap is that the validation set may not be representative of the true distribution of data that the model will encounter in the real world. Another possibility is that the model may be overfitting to the training set, and thus not generalizing well on new data. RoBERTa deals with this problem by the continuation of the performance even after validation perplexity has plateaued. The main reason explaining such a difference in the performance of BERT and RoBERTa is that RoBERTa was probed on tasks from the XNLI corpus. Researchers and practitioners in machine learning use the XNLI dataset to develop and test new models for cross-lingual NLU tasks, such as text classification, sentiment analysis, and language translation. The performance of these models on the XNLI dataset provides valuable insights into the state-of-the-art in cross-lingual NLU, and helps to advance the field by identifying areas for improvement and innovation. Thus, such tasks in perspective allow us to give the model a greater understanding of the whole text, which gives an advantage in tests where a document containing more than two sentences is submitted for input.

3. Discourse coherence

Even though one of the two main BERT's objectives is Next sentence prediction, we should keep in mind that the DC task (see Figure 6.3) provides the model with not two, but several sentences as input. And in this case, the best results obtained by RoBERTa are explained by the fact that this model copes better with long sequences, due to the fact that compared to BERT more extensive training data with lengthier sequence segments is trained.

Results for this task indicate that the type of architecture of the model does not play a crucial role in this case. Although BERT and RoBERTa are encoders, mGPT is a decoder and T5 is an encoder-decoder transformer, we can see that T5 and mGPT-2 have shown almost the same results, which are relatively close to BERT's accuracy. Decoder-only models' main focus is on the bonds between an output vector and the next target vector, but not between an input and an output corresponding to it. In encoder-only models there is an opposite situation, meanwhile encoder-decoder transformers are aimed to perform both tasks [Platen, 2020]. The difference also concerns their self-attention layer, as encoder-only transformers have bidirectional self-attention layer and decoder-only have unidirectional ones. Therefore, for discourse coherence we can conclude that type of attention in this case does not mean a lot for prediction results, consequently, it is hard to make a conclusion about which type of model is best suited (presumably the one which deals best with long token-sequences).

4. Next sentence prediction

The fact that BERT performed best (see Figure 6.5) on this task can be attributed to its NST objective, which was later removed from RoBERTa fine-tuning process. Quite unexpected was the result obtained for the T5 model, which performed almost as well as BERT. Perhaps this can be explained by the fact that in mT5 the decoder typically produces two additional tokens: the class label and an end-of-sequence token, which can contribute to a better understanding of the connectivity of the final element of the sequence and the previous elements. This hypothesis can be applied to all results in which T5 exceeds mGPT in accuracy.

NST is a task of the type for which we expect high accuracy of predictions from a model whose main specificity is text generation (mGPT). Hypothetically, in this case, bidirectional self-attention is not required and it is enough to predict the output based only on the previous context. To understand why mGPT still performs the worst, and T5 shows the same results as RoBERTa (thereby neutralizing the importance of having a decoder in the architecture), we need to take into account the differences between generating the next sentence and a single token. Presumably, for the accurate recognition of the next sentence, the context of both the previous and the subsequent sentences plays a decisive role, the full understanding of which is impossible without the encoder (due to the mechanism of bidirectional attention).

5. Sentence ordering

Again, an unexpected result with BERT performing better than RoBERTa (see Figure 6.7), which may have been caused by differences in the masking procedures for RoBERTa and BERT. In RoBERTa, the masking of 15% of tokens is dynamic and changes for each pre-training epoch. This helps to prevent the model from memorizing specific patterns or sequences of masked tokens and encourages it to learn more general representations of language. By changing the masking patterns for each epoch, the model is exposed to a wider range of training examples and is forced to learn more robust and flexible representations. At the same time, this change could lead to a deterioration in the quality of the model's performance in this case, since the SO task assumes that for incorrect examples all sentences in a sequence are being shuffled. Accordingly, in this case, masking the fixed part of the input can serve as an advantage of BERT. Besides, in [Rothe et al., 2020] it was demonstrated that BERT performs best with sequence-splitting tasks, indicating that its understanding of sentence ordering exceeds RoBERTa’s one.

6. Sentence position

In this case (see Figure 6.6), Roberta demonstrates the best results. This task is somewhat similar to the previous one, the difference is that not all proposals are mixed in the SO, but only 4 and another randomly selected, while in the SP all proposals for incorrect options occupy new randomly selected positions. Perhaps these differences explain the difference in the performance of the models for the two tasks. Presumably, in this case, RoBERTa's superiority is explained by the fact that RoBERTa was trained on a much larger corpus of text data than BERT, which allowed it to learn more complex and nuanced patterns in language. Additionally, RoBERTa was trained for longer than BERT, with a larger batch size and more training steps. Thus, RoBERTa has a greater understanding of the language as a whole, rather than a specific sequence of sentences in the document, which helps to establish the wrong order again and again without fixed changes.

7. Discourse connective prediction

RoBERTa unsurprisingly demonstrates the best results (see Figure 6.4), nevertheless, it is interesting for us why mGPT has shown higher accuracy than T5.

In [Liu et al., 2019] it was found out that input format used in BERT (SEGMENT-PAIR+NSP LOSS) deteriorated the overall level of accuracy, therefore, for RoBERTa NSP loss was removed and whole input was replaced with full sentences. For this task, this is one of the key changes, since the prediction of the union in most cases can be provided only when the context of both surrounding sentences is known.

An obvious problem with mGPT and T5 in solving these kinds of tasks is their generative objective. In order to perform classification, these transformers will by themselves generate the masking token, consequently, if the answer is not in a generated sequence, it will be regarded as an incorrect choice. The problem is that the training sample of models and the sample used for fine-tuning may lack the necessary connectives, in which case the correct answer simply cannot be generated by the model by definition and will eventually be read as incorrect.

5. Conclusion

Our work is devoted to the study of the degree of discourse assimilation by various multilingual models. Despite the fact that many tasks and hypotheses were built on the materials of their predecessors, our research differs from them in that it involves several languages in discourse probing at once and combines completely different tasks that ultimately somehow test the understanding of the model of the whole text. Also, some of our results do not correspond to the conclusions of other researchers and add new information about the understanding of the language by individual models.

For instance, in the [Koto et al., 2021] it was found out that BERT and RoBERTa are generally better at understanding discourse tasks. But this assumption contradicts the fact that the mGPT copes best with the task in which it is required to determine the order for two sentences of the document, which is explained by the peculiarities of studying the training of this transformer.

Moreover, we have come to a conclusion that models, on average, perform equally in low-resource and conventional (popular) languages with binary-classification tasks. This result may indicate the presence of certain trends associated with the assimilation of the document structure by models, which apply to all idiolects.

As already mentioned, the type of task strongly determines the performance of the model due to its individual parameters, which may respond to the task in varying degrees. For example, mGPT shows a poor performance on classification tasks with many classes and sequence generation tasks, therefore, we have seen that more sentences in a document give less accuracy and vice-versa.

We also identified some characteristics of tasks and training samples that affect the performance of the model, such as the size of the sequence, the number of sentences involved in shuffle, the focus of prediction (the last sentence is often easier to predict than the first) – and this factor is stronger than the significance of the size of the context. The more randomness there is in choosing proposals that will change the position in the document, the better the performance of some models, for example, RoBERTa, since its main principle is masking an unfixed set of tokens.

Consequently, we have identified certain aspects of tasks that models generally do worse with, such as predicting the connective marker when there is a limited amount of resources, as well as those factors of individual model's architecture that worsen the results. We also compared the results obtained with the accuracy of the predictions of monolingual models and did not reveal a significant deterioration in the quality of transformers.

As a practical result of this research, we present a new publicly available repository with discourse parsers and results of our tests[1].

[1] https://github.com/mashagodunova/discource_probing

References

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019, 4171–4186*

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman,´F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale.

Shliazhko, O., Mikhailov, V., Fenogenova, A., Kozlova, A., Tikhonova, M., Shavrina, T. (2022). mGPT: Few-Shot Learners Go Multilingual

Xue, L., Constant, N., Roberts, A., Kale M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer

Koto F., Lau, J., H., Baldwin, T. (2021). Discourse Probing of Pretrained Language Models, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 3849–3864*

Shi, W., Demberg, V. (2019). Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 5790–5796*

Chen, M., Chu, Z., Gimpel, G. (2019). Evaluation results and Learning Criteria for Discourse-Aware Sentence Representations

Martins, A., F., T., Farinhas, A., Treviso, M., Niculae, V., Aguiar P., M., Q.,, Figueiredo A., T., M. (2020). Sparse and Continuous Attention Mechanisms. *4th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.*

Nie, I., Bennett, E., D., Goodman, N., D. (2019). DisSent: Learning Sentence Representations from Explicit Discourse Relations.

Child, R.,, Gray, S., Radford, A., Ilya Sutskever, I. (2019). Generating Long Sequences with Sparse Transformers

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., N., Kaiser Ł., Polosukhin, I. (2017). Attention Is All You Need

Hiroko Kitano. (1990). Cross-Cultural Differences in Written Discourse Patterns: A Study of Acceptability of Japanese Expository Compositions in American Universities

Pires, T., Schlinger, E., Garrette, D. (2019). How multilingual is Multilingual BERT?

Kaplan, R. (1966) ‘Cultural thought patterns in inter-cultural education’, Language Learning, 16, 1-20. *: Information from lecture delivered by M. R. Montaño-Harmon, Ph. D., Professor Emeritus, California State University, Fullerton, June, 2001, based on (1) doctoral dissertation research and (2) ongoing research in four states in the United States.*

Malmi, E., Pighin, D., Krause, S., Kozhevnikov, M. (2017). Automatic Prediction of Discourse Connectives

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., Allen, J. (2016). Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories

Sharma, R., Allen, J., F., Bakhshandeh, O., Mostafazadeh, N. (2018). Tackling the Story Ending Biases in The Story Cloze Test

Rothe, S., Narayan, S., Severyn, A. (2020). Leveraging Pre-trained Checkpoints for Sequence Generation Tasks

Poláková, L., Mírovský, J., Nedoluzhko, A.,Jínová, P., Zikánová, Š., Hajičová, E. (2013). Introducing the Prague Discourse Treebank 1.0.

Procházka, M., Malá, M., Šaldová, P. (2010). The Prague School and Theories of Structure

Polakova, L., Mırovsky, J. (2019). Anaphoric Connectives and Long-Distance Discourse Relations in Czech

Villalba, X. (2013). Right-dislocation in Catalan: tails, polarity and activation

Kroon, C. (2009). Latin Linguistics between Grammar and Discourse. Units of Analysis, Levels of Analysis

Pal, A., Balasubramanian, V., N. (2019). Zero-Shot Task Transfer

Platen ,V., P. (2020). Transformers-based Encoder-Decoder Models. *https://huggingface.co/blog/encoder-decoder*

Hernault, H., Bollegala, D., Ishizuka, M. (2010). Towards Semi-Supervised Classification of Discourse Relations using Feature Correlations

Feng, V., W., Hirst, G. (2014). A Linear-Time Bottom-Up Discourse Parser with Constraints and Post-Editing

Ji, Y., Eisenstein, J. (2014). Representation Learning for Text-level Discourse Parsing

Koto, F., Lau, J., H., Baldwin, T. (2021). Top-down Discourse Parsing via Sequence Labelling

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., Allen, J. (2016). A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories

Barzilay, R., Lapata, M. (2017). Modeling Local Coherence: An Entity-Based Approach

Dagnev, I., Saykova, M., Yaneva, M. (2019). Discourse and linguistic characteristics of RMA introduction sections – a Bulgarian-English comparative study

Demirsahin, I. (2015). The discourse structure of Turkish

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach