Probing of pretrained multilingual models on the knowledge of discourse

Mary Godunova[λ] and **Ekaterina Voloshina**[ϕ]
[λ]HSE University
[ϕ]Work done at AIRI

Abstract

With the raise of large language models (LLMs), different evaluation methods, including probing methods, are gaining more attention. Probing methods are meant to evaluate LLMs on their linguistic abilities. However, most of the studies are focused on morphology and syntax, leaving discourse research out of the scope. At the same time, understanding discourse and pragmatics is crucial to building up the conversational abilities of models.

In this paper, we address the problem of probing several models of discourse knowledge in 10 languages. We present an algorithm to automatically adapt existing discourse tasks to other languages based on the Universal Dependencies (UD) annotation. We find that models perform similarly on high- and low-resourced languages. However, the overall low performance of the models' quality shows that they do not acquire discourse well enough.

1 Introduction

Various methods of evaluating language models, including probing methods (Koto et al., 2021), have recently been popular. The probing methods help to shed light on the linguistic abilities of Large Language Models (LLMs), which could be later used to improve models' qualities (Saphra, 2021). However, probing studies were mainly conducted at such language levels as morphology and syntax (Kassner and Schütze, 2020; Marvin and Linzen, 2018). While pre-trained language models have shown remarkable performance on various language tasks, there is still much to be explored regarding their ability to capture broader discourse in documents. By *discourse*, we understand a language level that operates linguistic units bigger than sentences.

It involves organizing and connecting ideas to create coherent and cohesive communication.

In this paper, we are testing models' ability to capture different aspects of discourse knowledge. Discourse probing can involve tasks such as identifying the relations between sentences within a document or the role of one sentence in the document structure, investigating main topics, discovering a suitable ending, and finding out whether one sentence belongs to a particular paragraph or not (Koto et al., 2021; Chen et al., 2019a). Such tasks shed light on the strengths and limitations of pre-trained language models in capturing the nuances of discourse structure.

Our main contribution is a new suite of probing tasks on multilingual data from ten languages. Moreover, our method can be used for other languages with data available in Universal Dependencies format (De Marneffe et al., 2021). Overall, we state our contributions as follows:

- To bridge the gap in discourse probing research, the paper introduces a probing task to interpret the ability of pretrained LMs to capture discourse relations in 10 linguistically diverse languages;
- We present a tool to generate tasks for probing discourse in any language for which there is enough data in Universal Dependencies (UD) format[1];
- The study validates the findings across different models, languages, and discourse probing tasks, providing valuable insights into the limitations of current LMs in capturing discourse knowledge.

2 Related work

Probing tasks were first introduced in Conneau et al. (2018) and described as simple classification tasks that would reveal if a model contains any linguistic knowledge. Probing involves different methods, for instance, probing classifiers (Belinkov, 2022). After training a model on a specific task, we create representations using the model and then train a separate classifier to predict a particular attribute based on these representations. If the classifier demonstrates strong performance, we conclude that the model has acquired relevant information for the attribute. However, upon further examination, it becomes clear that additional complexities are at play. Also, probing methods involve prompting: transforming a set of probing tasks into question-answer pairs and directing the model to respond to the questions with a specific prefix (Li et al., 2021). This approach essentially serves as a probe that is independent of the model. By using prompting instead of a diagnostic probe, researchers can circumvent the challenge of distinguishing

[1]Our code is available at https://github.com/mashagodunova/discource_probing

Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), pages 78–90

between the content of the representations and what the probe learns. After all, one of the most developing fields in probing LLMs is task relevance which is aimed at investigating whether the information encoded in sentence representations, as discovered through a probe, is used by the model to perform its task. Task relevance is also our method of research, which will be discussed later.

Most probing studies focus on evaluating semantic knowledge, which focuses on the meaning of individual words and sentences. The probing methodology combining various annotated data is commonly used as the benchmark for language model comparison and evaluation of their generalizing ability (Conneau and Kiela, 2018). On the other hand, probing of discourse examines how linguistic units are organized and connected to form coherent texts, a crucial ability to generate long sequences. However, only some works investigate the ability of LLMs to understand discourse. Ettinger (2020) shows that BERT produces pragmatically incorrect outputs because it does not consider an extended context. Among other works, Nie et al. (2019a) evaluate models on discourse relations expressed with conjunctions. Chen et al. (2019b) propose a benchmark for model evaluation on different discourse tasks such as prediction of implicit discourse relations based on the Penn Discourse Treebank annotation (Prasad et al., 2008), discourse coherence, and others.

3 Tasks

3.1 General Description

All examples of the described tasks are presented in Appendix 7. We adapt tasks from DiscoEval (Chen et al., 2019a), a framework for discourse probing of language models. 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 paper, we borrowed and adapted the following tasks, making them suitable for multiple languages:

Sentence Position (SP): this task tests the model's understanding of linearly-structured discourse. By randomly moving one of the five 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.

Binary sentence ordering (BSO): 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.

Discourse coherence (DC): having a sequence of 6 sentences that form a coherent paragraph, we need to randomly replace one sentence from the coherent sequence with a sentence from another discourse. A model needs to determine whether the resulting sequence of 6 sentences still forms a coherent document. In the DC task, the models must determine the coherence of a document in which any of the five sentences could be replaced except for the first.

Besides that paper, we adapt several tasks from (Koto et al., 2021):

Next sentence prediction: The preceding context consists of 2 to 8 sentences, while the candidates (4 sentences) for prediction are always single sentences. Nevertheless, we adopted it as a binary classification task by mixing one of the sentences in a way that researchers in (Chen et al., 2019a) did.

Sentence ordering: This task is to determine whether the order of sentences in the document is correct. Texts from 3 to 7 sentences mixed within the same sequence are presented as incorrect options.

Cloze story test: Data for this task consists of sequences with four sentences in each. A model needs to pick the best-ending sentence for all documents. We adapted the task as binary, so for incorrect pairs 'key: value', we shuffle ending sentences within all documents.

Although probing studies (Koto et al., 2021; Chen et al., 2019a; Nie et al., 2019b) in the field of discourse have already been conducted, they included a small number of languages (mostly English). They focused on a limited number of tasks in terms of content: either predicting a discourse marker, 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 essential to conduct a general study, having compiled tasks on various aspects of discourse and choosing different languages as a training sample.

3.2 Tasks' theoretical background in terms of RST

As it was already mentioned, the main theoretical background for parsing UD documents was Rhetorical Structure Theory. We have not tried to consider individual types of relations, such as opposition or entailment. Instead, we focused on general patterns, called schemas in this theory, and the constraints they impose on the text.

There are 4 types of restrictions that must be observed in order not to violate the structure of the text:

- Completedness: The set contains one schema application that contains a set of text fragments that make up the entire text
- Connectedness: With the exception of the entire text in the form of a text fragment, each text fragment in the analysis is either a minimal unit or an integral part of another application of the analysis scheme.
- Uniqueness: Each schema application consists of a different set of area text, and within a multi-link schema, each link is applied to another a set of text areas.
- Adjacency: The text intervals of each schema application are equal to one text interval.

According to this classification, we divided all tasks into three groups. The first group included tasks in which the rules of coherence and contiguity were not observed at the same time. Among these tasks are Sentence Position and Binary sentence ordering. The difference between the tasks lies in the size of the sentences and the static part: in the first case, four out of five sentences remain static, while in the second one element moves relative to another. The similarity lies in the fact that in both tasks the order is disrupted by changing the adjacency relations, that is, the sentence changes its position in the general structure, but the new sentence, which was not originally in the discourse, is not involved.

Another group that we deduced was a group of examples in which the rules of completedness and uniqueness are violated: Discourse coherence, Next sentence prediction, Cloze story test. In this group, the desired element is removed from the discourse and replaced with an element from another discourse. Due to this general characteristic, tasks from this group can be characterized by two properties: loss of text integrity and the presence of elements that do not fit into the structure of the text.

The latter group is characterized by the absence of an important element (sentence or word form) necessary for the connectivity of the text (at the same time, nuclear part is not missing, therefore, in this sence the text is completed), therefore, only the rule of connectivity is violated in them. Among the tasks included in this category: Sentence ordering and Discourse connective prediction.

4 Methods

4.1 Data

All data for our probing tasks was taken from the UD framework (De Marneffe et al., 2021), which provides a standardized set of grammatical dependencies and syntactic relations for annotating treebanks in different languages (more than a hundred languages). One of the main tasks of our research was to create a parser that generates multilingual tasks for discourse on UD data automatically without the need for manual markup. As a result, we extracted .csv files as training samples from the UD data. The general format of such files consists of:

1. Answer in correctness rating format: 0 or 1. In this case 1 indicates that presented sentences (and discourse connective for DCP task) meet the criteria for the correctness of a specific task. For example, for the Binary sentence ordering task, two sentences will be presented; if they are in the correct order, there will be 1, otherwise - 0.

2. Data type marker: training or test

3. Sentences - each sentence is displayed in a separate column

4. Present only in the Discourse connective prediction task - discourse connective itself

This parser can be used on treebanks for any language. We frame almost all presented tasks as binary classification problems, and they involve different aspects of Rhetorical Structure Theory[2], models' understanding of which is being tested in this study. More information about the generation of tasks is presented in section A.

4.2 Models

In our study, we probe several multilingual LLMs of different architectures: mBERT (Devlin et al., 2019), XLM-XLM-RoBERTa (Yinhan et al., 2019), mGPT, and mT5. We do not fine-tune models since we aim to test the basic models in understanding the discourse. Instead, we extract [CLS] embeddings and train a Logistic regression on these representations to assess the quality of the models' performance.

4.3 Languages

Most of the languages in the sample belong to the Indo-European language family (limited to the most common language groups – Romance, Germanic, and Slavic); as for our experiments, the dataset size was essential. We also included Turkish, which treebank is one of the largest in the Universal Dependencies. In addition, Turkish is part of one of the largest language families, Altai. The sample also included the Armenian language since data for this language was massive enough to parse it, and it has never been included in any previous probing studies. The table below shows the number of examples for each task and language that were extracted from treebanks:

Most of the languages in our sample were chosen as they have been mentioned little to no in previous works. However, we also include languages often appearing in Natural Language Processing works, such as English, French, and Russian, to make our results comparable to other works.

Moreover, the difference in corpora sizes shows how models perform in best (high-resourced languages) and worst cases (low-resourced languages). It allows us to investigate further how the number of examples in a particular language determines a multilingual transformer's understanding of several idioms at once.

5 Results

5.1 Results by languages

Overall, models show some understanding of discourse structures, especially in high-resourced languages.

As for the differences in performance on different languages, as Figure 1 shows, models show better quality

[2]Rhetorical Structure Theory (Forsbom, 2005) is a framework for analyzing and understanding how texts are organized and constructed rhetorically. It focuses on the patterns and relationships between different text elements, such as the primary point or argument, supporting evidence, and rhetorical devices used

Language	BSO	CST	DC	NSP	SO	SP	DCP
Russian	15632	9385	3450	12949	5302	2790	14036
Bulgarian	17354	67142	33567	42781	18579	22152	37620
Czech	1230	18437	2143	13561	9450	7664	2089
Serbian	1389	6780	2013	4998	4356	1732	1503
Catalan	1476	47852	34701	21952	1938	9909	7605
French	1468	1201	1750	7620	2395	1042	1201
Latin	1474	51867	21602	13764	1027	1395	3047
English	1823	21770	3502	16067	3750	7438	8993
Armenian	2094	46209	29436	49673	19820	10347	28049
Turkish	15203	12064	3972	30166	1960	1704	6775

Table 1: Number of examples in each treebank. *BSO*: Binary Sentence Ordering, *CST*: Cloze Story Test, *DC*: Discource Coherence, *NSP*: Next Sentence Prediction, *SO*: Sentence Ordering, *SP*: Sentence Position, *DCP*: Discourse Connective Prediction

in the languages better presented in the training set. As can be seen, a writing system does not appear to be an essential factor, as models show better performance in Armenian than in Turkish or even French in some cases.

Armenian XLM-RoBERTa performs best in this language, although mBERT and mT5 demonstrate almost identical results. Although there are practically no studies devoted to the structure of discourse in the Armenian language, and this language is considered underresourced, it is surprising that models show results similar to results in English.

Bulgarian In this case, there is a distribution common to most tasks (and obtained by averaging the results for both tasks and languages), in which XLM-RoBERTa demonstrates the highest accuracy, mBERT performs slightly worse, followed by mT5, and the worst results are observed for mGPT.

English Results demonstrated by models for English may show the actual distribution of ratings because this language always has the largest number of examples in the training sample. We can assume that mBERT potentially has more knowledge about discourse, but it is more difficult to cope with longer sequences, or it has a smaller multilingual base.

Catalan For Catalan we observe extremely unexpected results exceeding XLM-RoBERTa, as mBERT demonstrates the best accuracy (while still lower than the average value for other languages), and mGPT is in second place. mT5 demonstrated a slightly lower average accuracy, and XLM-RoBERTa performed the worst.

Czech XLM-RoBERTa's absolute superiority may stem from the fact that the compilers of the treebank for the Czech language emphasized long-distance discourse relations in accordance with (Poláková et al., 2020), meaning that to capture a core sense of the sentence you need to 'parse' it from the beginning to an end and keep in mind all the details. As proven, one of the main advantages of XLM-RoBERTa is the ability to analyze large text sequences (Conneau et al., 2020).

French The utterance in Romance languages (compared to the linear structure of utterance in English) is distinguished by ornateness. The main idea is usually expressed at the beginning and at the end. In this vein, the accuracy of mGPT can be explained by the sparse attention mechanism, which allows each output position to focus on only a subset of input positions, selected based on predefined patterns or rules (Martins et al., 2020).

Russian For Russian, we observe the same distribution that has already been described for Bulgarian. Since the distribution was almost the same for the Czech language (the difference is that the mGPT showed slightly higher accuracy than mT5), it can be assumed that such similarity in the results is explained by the affiliation of the above languages to the same language group.

Latin In (Kroon, 2009), it is established that the structure of discourse in Latin is characterized by solid 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.

Serbian 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 mBERT's superiority over XLM-RoBERTa, it can be assumed that differences in the token masking procedure explain it - in the case of mBERT, it is always a fixed set of tokens when the model is working, which may help in working with low-resource languages.

Turkish XLM-RoBERTa achieved the highest performance, surpassing mBERT, mT5, and mGPT. However, mBERT still performed better than mGPT and mT5; mT5 showed the lowest accuracy among the four models.

5.2 Results by tasks

Now, we will examine the correlation between each model's understanding of discourse and different types

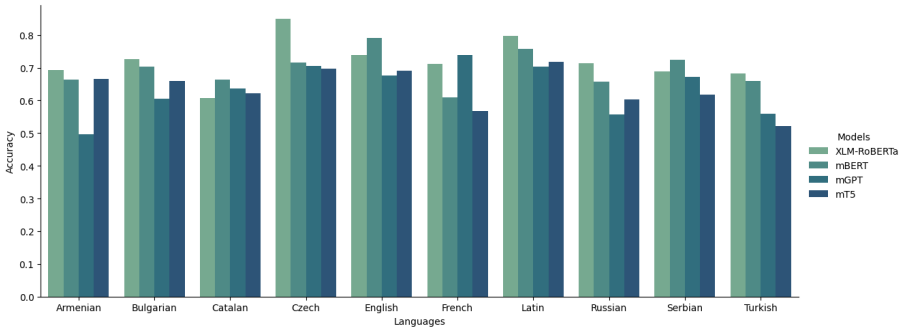

Figure 1: Average accuracy depending on the language and type of model

of tasks. As seen from Figure 5.2, the models show the best performance on *Cloze Story Test (CST)* and *Next sentence prediction* tasks. In both tasks, the focus of the prediction is the last sentence of the document. However, the accuracy on a similar task, the *Discourse coherence (DC)* task, is much lower. We can conclude that the number of sentences is not a crucial factor, as for the DC test, there was a sequence of 5 sentences provided, while for CST, all documents consisted of 4 sentences. However, the position of a shuffled sentence appears to be important.

Binary sentence ordering is the only task where mGPT copes with it best, but in all other tasks, it demonstrates the lowest accuracy rates due to obvious issues like the lack of some investigated languages in the mGPT's training data.

Cloze story test In this task XLM-RoBERTa shows the best performance. Our results replicate the results by Conneau et al. (2020) where they show that XLM-RoBERTa surpasses mBERT on cross-lingual classification, but specifically with low-resource languages used in training data. XLM-RoBERTa's superiority over mBERT can be explained not only by its overall better accuracy in most tasks but also by the phenomenon called the "generalization gap", which occurs when a language model's ability to perform well on downstream tasks exceeds its performance on the validation set during training.

Discourse coherence Even though one of the two main mBERT's objectives is Next sentence prediction, we should remember that the DC task provides the model with not two but several sentences as input to determine whether they are coherent. As shown by the results, XLM-RoBERTa copes better with long sequences because compared to mBERT, more extensive training data with lengthier sequence segments is trained. Results for this task indicate that the model's architecture type does not play a crucial role in this case. Although mBERT and XLM-RoBERTa are encoders, mGPT is a decoder, and mT5 is an encoder-decoder transformer, we can see that mT5 and mGPT-2 have shown almost the same results, which are relatively close to mBERT's accuracy.

Next sentence prediction NSP 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, bidirectional self-attention is not required in this case, and it is enough to predict the output based only on the previous context. To understand why mGPT still performs the worst and mT5 shows the same results as XLM-RoBERTa (thereby neutralizing the importance of having a decoder in the architecture), we must consider 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 complete understanding of which is impossible without the encoder (due to the mechanism of bidirectional attention).

Sentence ordering Unexpectedly, mBERT performs better than XLM-RoBERTa, which differences in the masking procedures for XLM-RoBERTa and mBERT may have caused. In XLM-RoBERTa, the masking of 0.15 of tokens is dynamic and changes for each pretraining epoch. Our results correlate with (Rothe et al., 2020) where the authors demonstrated that mBERT performs best with sequence-splitting tasks, indicating that its understanding of sentence ordering exceeds XLM-RoBERTa's.

Sentence position In this case, XLM-RoBERTa demonstrates the best results. This task is similar to the previous one, the difference is that in SO not all proposals are mixed, but only four and another randomly selected. In contrast, in the SP all proposals for incorrect options occupy new randomly selected positions. Presumably, in this case, XLM-RoBERTa's superiority is explained by the fact that XLM-RoBERTa was trained on a much larger corpus of text data than mBERT, which allowed it to learn more complex and nuanced patterns in language. Additionally, XLM-RoBERTa was trained for longer than mBERT.

Discourse connective prediction For this task where the input consists of two sentences and transformers must predict correct connective XLM-RoBERTa unsurprisingly demonstrates the best results. This result can be attributed to the NSP loss being removed in XLM-RoBERTa's architecture and the whole input being replaced with full sentences. An obvious problem

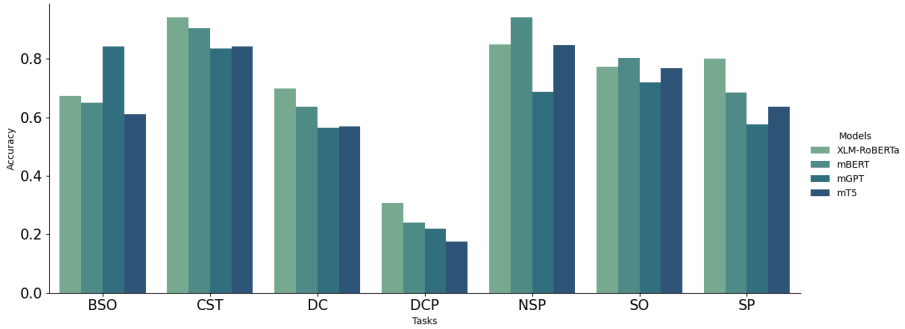

Figure 2: Average accuracy depending on the task and type model
BSO: Binary Sentence Ordering, *CST*: Cloze Story Test, *DC*: Discource Coherence, *NSP*: Next Sentence Prediction, *SO*: Sentence Ordering, *SP*: Sentence Position, *DCP*: Discourse Connective Prediction

with mGPT and mT5 in solving these kinds of tasks is their generative objective since 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.

6 Discussion

The influence of the discourse structure in English The so-called 'complicated simple sentences' (Dagnev et al., 2019) in Bulgarian generate heavy complementation, 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. The results obtained for Catalan, which is also structurally significantly different from English, display the same trend and can be explained by right-branching (right-dislocation constructions), which is not often found in English.

mGPT's sparse attention mechanism Due to mGPT's performance for French and Russian we can hardly consider that the sparse attention mechanism applied for mGPT helps to cope best with long sequences found in Russian, 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 (as in French). At the same time, for Russian Kaplan (Kaplan, 2006) 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 non-linear with respect to discourse in English), differs in the lack of integrity according to Kaplan. 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.

Models performing similarly with languages belonging to the same group 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. 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. 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.

Advantages of dynamic masking procedure In the case of Turkish, we were talking about shared arguments that 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 governs 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. For XLM-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 mBERT, therefore the probability of understanding complicated structures gets bigger. At the same time, we can see that dynamic masking procedure benefits only in cases where the complicated structure of the text does not change drastically. For instance, in SO task 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 mBERT.

mT5's superiority over mGPT In the NSP task we can assume that the results obtained 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 mT5 exceeds mGPT in accuracy.

How context and focus sentence position affects models' performance In tasks in which the highest accuracy of the models' performance was recorded, the focus sentence for prediction 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, 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 0.538, while in our case we got an 0.8 accuracy. 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.

7 Conclusion

Our work is devoted to the study of the degree of discourse acquisition 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 which analyzed English and other few languages (Chinese in most cases) and add new information about the understanding of the language by individual models. 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. 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, XLM-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.

References

Barzilay, R. and Lapata, M. (2017). Modeling local coherence: An entity-based approach.

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Chen, M., Chu, Z., and Gimpel, K. (2019a). Evaluation benchmarks and learning criteria for discourse-aware sentence representations.

Chen, M., Chu, Z., and Gimpel, K. (2019b). Evaluation benchmarks and learning criteria for discourse-aware sentence representations. *arXiv preprint arXiv:1909.00142*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale.

Conneau, A. and Kiela, D. (2018). SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Dagnev, I., Mariya, Saykova, and Yaneva, M. (2019). Discourse and linguistic characteristics of rma introduction sections – a bulgarian-english comparative study.

De Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal dependencies. *Computational linguistics*, 47(2):255–308.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Forsbom, E. (2005). Rhetorical structure theory in natural language generation.

Kaplan, R. B. (2006). Cultural thought patterns in intercultural education. *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*, pages 1–20.

Kassner, N. and Schütze, H. (2020). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818.

Koto, F., Lau, J. H., and 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*, pages 3849–3864.

Kroon, C. H. M. (2009). Latin linguistics between grammar and discourse: Units of analysis, levels of analysis.

Li, L., Ma, C., Yue, Y., and Hu, D. (2021). Improving encoder by auxiliary supervision tasks for table-to-text generation. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5979–5989, Online. Association for Computational Linguistics.

Malmi, E., Pighin, D., Krause, S., and Kozhevnikov, M. (2017). Automatic prediction of discourse connectives.

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

Marvin, R. and Linzen, T. (2018). Targeted syntactic evaluation of language models. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Mostafazadeh, N., Chambers, N., He, X., Devi Parikh, D. B., Vanderwende, L., Kohli, P., and Allen, J. (2016). A corpus and evaluation framework for deeper understanding of commonsense stories.

Nie, A., Bennett, E., and Goodman, N. (2019a). DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

Nie, A., Bennett, E. D., and Goodman, N. D. (2019b). Dissent: Learning sentence representations from explicit discourse relations.

Poláková, L., Mírovský, J., Nedoluzhko, A., Jínová, P., Šárka Zikánová, and Hajičová, E. (2020). Introducing the prague discourse treebank 1.0.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). The Penn Discourse TreeBank 2.0. In *LREC*.

Rishi Sharma, James Allen, O. B. N. M. (2018). Tackling the story ending biases in the story cloze test. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) Authors:*.

Rothe, S., Narayan, S., and Severyn, A. (2020). Leveraging pre-trained checkpoints for sequence generation tasks.

Saphra, N. (2021). Training dynamics of neural language models.

Yinhan, L., Myle, O., Naman, G., Jingfei, D., Mandar, J., Danqi, C., Omer, L., Mike, L., Luke, Z., and Veselin, S. (2019). Roberta: A robustly optimized bert pretraining approach.

A Examples of tasks' generation

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.

A.1 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.

$Sent_1$	$Sent_2$	Discourse Connective
Obviously because I want to vote	If anyone else has voted, what did you guys vote for?	And

Table 2: Example of a discourse connective prediction task

A.2 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., 2019a) 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
The problem is that customers can find 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 th,ey can justify their likely sky-high valuation.	1
The Greater New Orleans Fair Housing Action Center (GNOFHAC) filed a housing discrimination ast week. The complaint, filed with the United States Department of Housing and Urban Development. Thomas Housing Development residents, the City of New Orleans. VI redevelopment of St	0

Table 3: Example of a sentence position task

A.3 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.

A.4 Discourse coherence

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

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 4: Example of a binary sentence ordering task

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
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 creation 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 5: Example of a discourse coherence task

A.5 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.

Examples	Labels
It was ok, but the place was old. It was clean, but just a little dumpy. Hard to get into, though.	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 6: Example of a Next sentence prediction task

A.6 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 and Lapata, 2017).

A.7 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 (Rishi Sharma, 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
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 7: Example of a sentence ordering task

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 8: Example of a cloze story test task

B Detailed results

Task	Language	Models			
		mBERT	XLM-RoBERTa	mGPT	mT5
Cloze story test	Bulgarian	**1.0**	0.924	0.899	0.9
	Catalan	0.947	0.9	**0.948**	0.934
	English	0.838	**0.892**	0.865	0.784
	French	0.875	**0.889**	0.625	0.633
	Armenian	0.8	**0.943**	0.829	0.8
	Latin	0.906	0.969	0.903	0.906
	Russian	0.875	0.884	0.625	0.75
	Czech	1.0	1.0	0.909	0.879
	Turkish	0.833	0.917	0.708	0.792
	Serbian	0.971	1.0	0.941	0.941
Binary sentence ordering	Bulgarian	0.517	0.724	0.759	0.621
	Catalan	0.577	0.615	0.808	0.615
	English	0.759	0.552	0.793	0.586
	French	0.514	0.6	0.943	0.429
	Armenian	0.8	0.8	0.6	1.0
	Latin	0.762	0.78	0.75	0.75
	Russian	0.515	0.697	1.0	0.455
	Czech	0.77	**1.0**	0.97	0.75
	Turkish	0.529	0.588	0.971	0.5
	Serbian	0.8	0.4	0.8	0.453
Discourse coherence	Bulgarian	0.75	0.719	0.594	0.594
	Catalan	0.548	0.645	0.546	0.677
	English	0.875	0.833	0.75	0.708
	French	0.333	0.667	0.998	0.667
	Armenian	0.615	0.769	0.462	0.615
	Latin	0.75	0.75	0.45	0.55
	Russian	0.667	0.689	0.333	0.667
	Czech	0.571	**1.0**	0.857	0.571
	Turkish	0.75	0.25	0.25	0.25
	Serbian	0.5	0.7	0.4	0.4
Discourse connective prediction	Bulgarian	0.226	0.29	0.161	0.258
	Catalan	0.313	0.313	0.375	0.125
	English	0.4	0.35	0.4	0.45
	French	0.429	0.429	0.429	0.286
	Armenian	0.184	0.026	0.158	0.105
	Latin	0.077	0.154	0.031	0.077
	Russian	0.357	0.214	0.286	0.286
	Czech	0.167	0.292	0.125	0.167
	Turkish	0.051	**0.999**	0.051	0.063
	Serbian	0.25	0.03	0.25	0.033
Next sentence prediction	Bulgarian	0.758	0.788	0.576	0.727
	Catalan	0.968	0.563	0.688	0.625
	English	0.981	0.939	0.697	0.758
	French	0.936	0.733	0.7	0.733
	Armenian	0.957	0.967	0.5	0.9
	Latin	1.0	0.998	0.97	1.0
	Russian	0.922	0.742	0.452	0.903
	Czech	0.958	1.0	0.783	0.99
	Turkish	0.94	0.774	0.677	0.839
	Serbian	1.0	**0.986**	0.833	1.0

Table 9: Overall results of different models on each task in each language

Task	Language	Models			
		mBERT	XLM-RoBERTa	mGPT	mT5
Sentence ordering	Bulgarian	0.759	0.793	0.62	0.586
	Catalan	0.531	0.563	0.656	0.5
	English	0.917	0.792	0.75	0.625
	French	0.682	0.682	0.727	0.729
	Armenian	0.629	0.63	0.519	0.593
	Latin	**1.0**	0.91	0.893	1.0
	Russian	0.867	0.8	0.767	0.811
	Czech	0.923	0.934	0.962	0.808
	Turkish	0.897	0.689	0.828	0.862
	Serbian	0.833	0.867	0.852	0.7
Sentence position	Bulgarian	0.912	0.765	0.797	0.559
	Catalan	0.761	0.61	0.71	0.585
	English	0.775	0.815	0.8	0.6
	French	0.52	**1.0**	0.47	0.75
	Armenian	0.667	0.714	0.703	0.333
	Latin	0.815	0.963	0.74	852
	Russian	0.4	0.92	0.42	0.4
	Czech	0.636	0.727	0.545	0.455
	Turkish	0.667	0.556	0.444	0.431
	Serbian	0.714	0.857	0.688	0.786

Table 10: Overall results of different models on each task in each language