University of Ljubljana
Faculty of Computer and
Information Science

Paraphrasing sentences

Matej Kranjec, Domen Vilar, Timotej Petrič

Abstract

Paraphrasing sentences involves rephrasing them in a different way while keeping the same meaning. We present a methodology for generating paraphrases of Slovene sentences using the fine-tuned T5 and LLaMA-based model. To train our models, we needed to curate a dataset of input-output sentence pairs containing original and paraphrased sentences. For this purpose, we selected several parallel corpora and translated the original English sentences to Slovene using Nvidia NeMO NMT models trained specifically for Slovene. We used LLaMA to generate paraphrases of English sentences and then translated the generated LLaMA paraphrases to Slovene using the translation. To ensure the quality of our dataset, we applied various filtering techniques. We also conduct both manual and metric-based evaluations to assess the quality of the generated paraphrases. A modified version of the Parascore metric for automatic evaluation of paraphrases is proposed.

Keywords
Paraphrasing, Large Language Models, Seq2Seq Models, LLaMA, T5

Advisors: Slavko Žitnik

1. Introduction

The task of paraphrasing a sentence involves restating the meaning of the sentence in a different form, while preserving its original intent. Paraphrasing is an important task in natural language processing (NLP), with various applications in text simplification, summarization, and machine translation. In this paper, we propose a metric for automatic evaluation of paraphrases, a methodology for generating paraphrases of Slovene sentences using the fine-tuned *T5* and *LLaMA* based model, and evaluate the quality of the generated paraphrases both manually and using our metric.

2. Related Work

Sokolov and Filimonov [1] explored paraphrase generation for data augmentation. They proposed a machine translation inspired encoder-decoder deep recurrent neural network architecture for automatic paraphrase generation in English.

Witteveen and Andrews [2] proposed the use of a large language model, specifically GPT2, to generate paraphrases. They claim that their approach is capable of generating paraphrases even on paragraph level without needing to break down the text into smaller chunks.

Li et al. [3] proposed a transformer-based model that can generate paraphrases at different levels of sentence granularity. They further proposed an unsupervised method for domain adaptation.

All of the mentioned researchers were focusing on paraphrase generation in English. To our knowledge, no such tools have been developed for Slovene.

3. Dataset Methods

For training our models we needed to select and produce a dataset of input-output sentence pairs of original and paraphrase sentences. For this reason we selected a few parallel corpora and translated the original English sentences to Slovene. Another approach was to use LLaMA based Vicuna large language model to produce the paraphrases from the original English sentences and then translate this paraphrases to Slovene. In this section we describe both methods and our data filtering approaches.

3.1 Translating datasets

We used the *Europarl* parallel corpus [4] for obtaining pairs of sentences in Slovene language for the paraphrasing dataset. We took the Slovene-English parallel corpus variant and translated the English sentences using machine translation. To translate from English to Slovene, we used the *Nvidia NeMO NMT* models trained for the Slovene language [5]. The produced dataset contains approximately 600.000 paraphrase sentence pairs in Slovene language.

We replicated this approach to translate also other corpora which we uploaded to the shared NLP OneDrive folder. This way we produced multiple datasets: *Europarl600k_en_slx2* and *MaCoCu-sl-en-v1.0_slx2* datasets [6] (in collaboration with *The Sentence Shufflers* team). We also produced the *OpenSubtitles* dataset (again in collaboration with *The Sentence Shufflers*), but this was not uploaded to OneDrive, nor was the dataset used by us in the end. Further descriptions of the datasets can be found in the respective folders on the shared OneDrive folder.

3.2 LLaMA/Vicuna paraphrasing

Another approach we used is paraphrase generation using the *Vicuna* model [7]. This is a chatbot based on the *LLaMA* 13B model [8], fine-tuned on 70.000 conversations from ChatGPT. We used the *Vicuna* model to generate paraphrases of English sentences by prompting it to paraphrase a given sentence in English. After that we translated the generated *LLaMA* paraphrased sentence to Slovene using the translation approach described in section 3.1. By using the *Europarl* parallel corpus we then had two pairs of sentences in Slovene - the original Slovene sentence from *Europarl* corpus and the translated *LLaMA* paraphrased sentence from original English sentence from *Europarl*. This way we generated the *europarl-llama-slx3_enx2_600k*dataset, which was also uploaded to the shared OneDrive. *LLaMA*/*Vicuna* based approach of dataset generation is further explained in 5.2.

To simplify reading, we will use *LLaMA* abbreviation when referring to the *Vicuna* model from here on forward.

3.3 Filtering sentences in dataset

We filtered the resulting dataset by considering the amount of digits in a pair of sentences, the amount of special characters, and the sentence lengths. Furthermore, we employed a metric similar to Parascore to remove similar sentences. The Parascore metric acknowledges that paraphrase evaluation is inherently different from the evaluation of other tasks, such as machine translation and summarization. A good paraphrase must satisfy two criteria: semantic similarity and lexical divergence. Semantic similarity requires that the paraphrase remain similar meaning to the original sentence. Lexical divergence requires that the paraphrase has syntactic or lexical differences from the original sentence. Most metrics used in previous paraphrase generation research described in section 2 are designed for tasks other than paraphrase evaluation, which can lead to suboptimal results. To determine the appropriate threshold for filtering out sentences, we examined the distribution of scores. By visualizing the distribution of scores in figure 1, we were able to make the decision where to set the threshold. This enabled us to avoid cutting out too many sentences, which could result in a loss of training data. We filtered the dataset by looking at our modified version of the Parascore [9] (described in section 4.1.2) of pairs of sentences from the dataset, and removing pairs if the score was too low.

We also removed sentences that did not start or end with an alphanumeric symbol. When reviewing the list of removed sentences, we noticed that many of them had started with arbitrary sequences of characters, such as (SI), (AU), (HU), and others. After further investigation, we realized that these were actually identifiers, and that the sentences were otherwise appropriate for training. To identify these sentences and retain them in our dataset, we applied a regular expression filter to the start of each sentence.

We used 80% of the data for training and the remaining 20% for validation. As we are generating paraphrases and comparing the generated sentences to the prompts, we do not need a labeled dataset for testing, we will use a unlabeled corpus instead.

The info for 3 different filtered datasets we produced are listed below. This datasets were used for training the T5 model described in section 5.1.

- *filtered Europarl* dataset
  - based on *Europarl600k_en_slx2*
  - contains 104003 of original 598872 sentences
- *filtered LLaMA Europarl Paraphrases* dataset
  - based on *europarl-llama-slx3_enx2_600k*
  - contains 243129 of original 598866 sentences
- *filtered MaCoCu-LLaMA Europarl* dataset
  - based on *europarl-llama-slx3_enx2_600k* and *MaCoCu-sl-en-v1.0_slx2*
  - contains 966401 of original 3483328 sentences

4. Evaluation Methods

To evaluate the quality of the generated paraphrases, we will use a combination of automatic and manual evaluation techniques.

4.1 Automatic Evaluation

For automatic evaluation, we explored existing techniques for scoring paraphrases. We modified an existing metric as the original version did not separate good paraphrases from bad ones in our datasets.

4.1.1 Parascore, 2022

The paper [9] explores the effectiveness of automatic metrics for paraphrase evaluation and proposes a new metric called *Parascore*. The authors found that reference-free metrics perform better than reference-based metrics and that commonly used metrics do not align well with human annotation. We noticed that this metric gives a high score for identical sentences or sentences with modified word order.

4.1.2 Our metric

We propose a modified version of the *Parascore* metric. Instead of taking into account lexical divergence, we use the *BLEU* score [10] to penalize too similar or even identical pair of sentences. We compute the semantic similarity using *BERTScore* [11] with the *SloBERTa* model [12]. For diversity, we use $1-BLEU_4$. Both of the values are between 0 and 1, and we use their geometric mean as the final score. However, we correct for completely different sentences by setting their score to 0. We decide to do so based on observation that pairs of sentences with a Slovenian reference and English candidate sentence scored too high despite not being paraphrases. The final metric is presented in equation 1.

We evaluate the translated *Europarl* corpus, described in section 3.1, using the proposed modified *Parascore* metric metric and display the distribution in figure 1. Due to the metric punishing identical and completely different (due to different languages etc.) pairs of sentences, the amount of sentences with score 0 stands out.

$$Sim = BERTScore(Candidate, Reference)$$
$$Div = 1 - BLEU_4(Candidate, Reference)$$
$$Score = \begin{cases} \sqrt{Sim * Div} & Div \neq 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

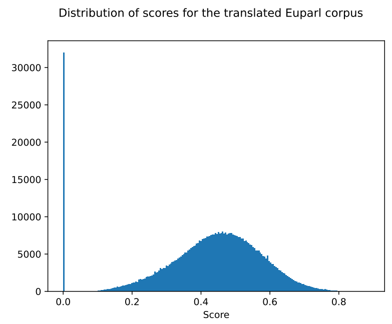


Figure 1. Distribution of scores on the translated *Europarl* dataset.

4.2 Manual evaluation

We will also evaluate a smaller subset of the test dataset manually. We will consider grammar and syntax, lexical divergence, meaning preservation, fluency and naturalness.

4.2.1 Grammar and Syntax

In the evaluation of sentences, we first assessed the correctness of their grammar and syntax. We assessed whether the sentences have proper sentence structure, accurate word order and lack grammatical errors.

- 1: Very poor - The sentence has lots of grammar mistakes, incorrect word order.
- 2: Poor - The sentence has some grammar errors and could be structured better.
- 3: Average - The sentence has acceptable grammar and syntax, with few errors in word order and sentence structure.
- 4: Good - The sentence has good grammar and reads well, with only a few small errors.
- 5: Excellent - The sentence has excellent grammar and syntax, with flawless word order and proper sentence structure.

4.2.2 Lexical Divergence

Focus of this task is to look at lexical variations in paraphrases. We assessed whether the paraphrased sentences utilized different words compared to the original sentence. Sentences that have more varied vocabulary are more favorable. This criterion tries to evaluate the ability to introduce new words while still preserving the meaning and context of the original sentence.

- 1: Very poor - The paraphrased sentence uses almost the same words as the original, lacking variety.
- 2: Poor - The paraphrased sentence has only a few different words compared to the original.
- 3: Average - The paraphrased sentence introduces some different words, but there's room for improvement.
- 4: Good - The paraphrased sentence shows good word choice, with noticeable differences from the original.
- 5: Excellent - The paraphrased sentence uses a wide range of different words while preserving meaning and context.

4.2.3 Meaning Preservation

In the case of paraphrased sentences, the meaning preservation criterion is very important. We focus on how well the paraphrased sentence preserves the original sentence's meaning. By comparing the two, we determined whether the paraphrase accurately presented the intended meaning.

- 1: Very poor - The paraphrased sentence completely changes the original sentence's meaning.
- 2: Poor - The paraphrased sentence has significantly different meaning from the original.
- 3: Average - The paraphrased sentence generally keeps the main message and context, but with a few mistakes.
- 4: Good - The paraphrased sentence retains the original meaning with minor mistakes.
- 5: Excellent - The paraphrased sentence accurately retains the original meaning.

4.2.4 Fluency and Naturalness

The fluency and naturalness criterion considered the overall flow and readability of the sentences. We assessed whether the sentences read smoothly and sounded natural in Slovene. We pay attention to the clarity of the sentences, ensuring that they were easily understood without any ambiguity. Additionally, we examined the word order and sentence structure to determine if they followed the conventions of natural Slovene language usage.

- 1: Very poor - The sentence reads awkwardly, doesn't flow well, and sounds unnatural in Slovene.
- 2: Poor - The sentence has issues with flow and readability, with some awkwardness and unnatural phrasing.
- 3: Average - The sentence reads smoothly, but there is occasional awkwardness or unnatural phrasing.
- 4: Good - The sentence reads well and sounds natural in the language.
- 5: Excellent - The sentence is perfectly fluent, reads really well and sounds completely natural.

5. Paraphrasing Methods

We tested a few different methods for generating paraphrases.

5.1 T5

We fine-tuned multiple sequence to sequence T5 models for the task of paraphrase generation. We used the *Slovene T5* model [13] - the smaller variant *t5-sl-small* - and further trained it using the different datasets which we described in section 3.

We fine-tuned about 60 trained models and multiple checkpoints for each. This model checkpoints were also trained on mainly 3 different dataset combinations which are described at the end of section 3.3. We mainly fine-tuned the models for 2 whole epochs.

From this fine-tuned models we then generated 5 sentences with each of the checkpoints based on the same input sentence "*Evropski policijski urad Europol je konec marca posvaril pred zlorabami tovrstnih besedilnih generatorjev.*". We read the generated paraphrases at the end empirically chose the model that was trained for 2 whole epochs on the *filtered Europarl* dataset, because it produced the best sentence paraphrases. Examples of generated paraphrases from the selected T5 model can be read from table 3 in appendix. Example generated paraphrases from the *filtered LLaMA Europarl Paraphrases* datasettrained T5 model can be seen in table 4.

To our surprise, compared to fine-tuning on *filtered Europarl* dataset, fine-tuning on *filtered LLaMA Europarl Paraphrases* dataset did not improve the models ability to generate good paraphrases and the model still managed to negate what was being said, change the meaning or introduce new unrelated words to the generated paraphrases. Using a 10 times bigger *filtered MaCoCu-LLaMA Europarl* dataset also did not improve the generated paraphrases. Possible reasons will be discussed in section 7.

Automatic metric	Average Score T5	Average Score LLaMA
First candidate	0.50	0.51
Best candidate	0.62	0.63

Table 1. Average scores comparison between T5 and LLaMA for automatic evalution.

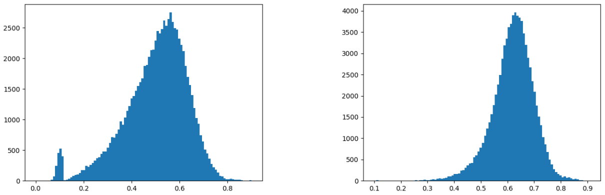
5.2 LLaMA/Vicuna paraphrasing

Another approach we used is to generate the *LLaMA* paraphrases directly. Because the large language model (LLM) is trained mainly on English corpora [8], we opted for the following pipeline: User inputs the sentence in Slovene, the sentence is translated to English using the machine translation approach described in section 3.1, the *LLaMA* based model paraphrases the English sentence and then the paraphrased sentence is translated back to Slovene. The prompt that was used to generate the paraphrases: `Please paraphrase the given text. Write nothing else, just the paraphrased text. "English sentence".` Where English sentence is then replaced with the original English one that we want to paraphrase.

A few examples of direct *LLaMA* paraphrasing are presented in table 5.

6. Results

We conducted a comparison between the first paraphrase generated by the model and the best paraphrase identified using our automatic scoring system. The average score of the best-selected paraphrase was considerably higher than that of the first generated paraphrase.



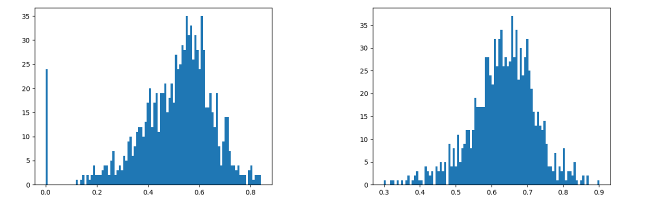
(a) First selected score value of the generated paraphrases.

(b) Maximum selected score value of the generated paraphrases.

Figure 2. Difference between maximum and first selected score of the generated paraphrases with T5 model trained on *filtered Europarl* dataset.

6.1 Comparison of automatic metric scores: T5 vs. LLaMA

We evaluated our approaches automatically on 1000 sentences. We noticed that both approaches perform similarly on our



(a) First selected score value of the generated paraphrases.

(b) Maximum selected score value of the generated paraphrases.

Figure 3. Difference between maximum and first selected score of the generated paraphrases with *LLaMA* model on 1000 sentences from *MaCoCu* dataset.

Manual metric	Average Score T5	Average Score LLaMA
Grammar	3.55	4.5
Lexical divergence	2.28	4.14
Meaning preservation	3.02	3.77
Fluency	2.91	4.14

Table 2. Average scores comparison between T5 and LLaMA for manual evaluation.

automatic evaluation metric. Furthermore, we notice that the first generated paraphrase candidate is not always scored the best and the metric increases by about 0.1 when taking into the account the best score rather than the first one. The distribution of scores is displayed in figure 2 for the T5 model and in figure 3 for the LLaMA model. The average scores are presented in table 1.

6.2 Comparison of manual metric scores: T5 vs. LLaMA

We conducted manual evaluation on 90 sentences from the *MaCoCU* dataset, comparing the average scores and distribution of scores for the T5 and *LLaMA* models across four key metrics: Grammar, Lexical Divergence, Meaning Preservation, and Fluency. The results, shown in Table 2 and Figure 4, indicate that *LLaMA* consistently outperforms T5, achieving higher average scores in all metrics.

7. Discussion

The chosen T5 model that was fine-tuned on *filtered Europarl* dataset, on the smallest of the 3 proposed datasets in section 3.3. The *filtered LLaMA Europarl Paraphrases* dataset may have been insufficient in size and because of its higher diversity the dataset proved to be too different in structure and word order, compared to using the normal *filtered Europarl* dataset. As a consequence, we observed instances where the model mistakenly replaced certain words with others and introduced new words that were not contextually related to the original paraphrase. This suggests that the model's understanding of word meaning and contextual preservation was limited due to the limitations of the dataset.

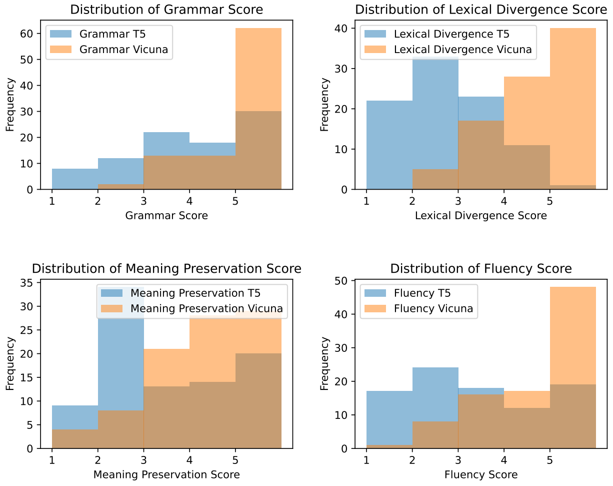


Figure 4. Distribution of Evaluation Scores for T5 and LLaMA Models

The *filtered MaCoCu-LLaMA Europarl* dataset also proved to not improve the generated paraphrases. This may be because the original sentences in the dataset are of too low quality since they were automatically scraped from ".si" internet top level domains, where there is many times less importance given to the English translations of original Slovene pages.

We noticed that our automatic metric, unlike our manual metrics, failed to differentiate between the models we tested. However, it still proved to be useful in filtering the training datasets.

One drawback of this method is that some information is lost when translating from Slovene to English and then back to Slovene. But because the *LLaMA* large language model (LLM) already contains the learned knowledge from pre-training on large corpora of 1.3T tokens, this is not a big issue - the model generates coherent sentences with good fluency and naturalness. The last translation from the *LLaMA* English paraphrase to Slovene still introduces some word artifacts, but generally we observed that this happens a lot less frequently than with the T5 fine-tuned model.

This approach would probably work even better if the input text was longer than one sentence, because the more context the LLMs get put into, less hallucinations occur.

8. Conclusion

In conclusion, our paper focuses on the task of paraphrasing sentences in Slovene. We propose a metric for automatic evaluation of paraphrases and present a methodology for generating paraphrases using the fine-tuned T5 and LLaMA-based models. We evaluate the quality of the generated paraphrases both manually and using our metric. Our findings demonstrate the effectiveness of LLaMA and translation in generating paraphrases in Slovene.

References

[1] Alex Sokolov and Denis Filimonov. Neural machine translation for paraphrase generation, 2020.

[2] Sam Witteveen and Martin Andrews. Paraphrasing with large language models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 215–220, Hong Kong, November 2019. Association for Computational Linguistics.

[3] Zichao Li, Xin Jiang, Lifeng Shang, and Qun Liu. Decomposable neural paraphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3403–3414, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit*, 2005.

[5] Iztok Lebar Bajec, Andraž Repar, Marko Bajec, Žan Bajec, and Mitja Rizvič. NeMo neural machine translation service RSDO-DS4-NMT-API 1.0, 2022. Slovenian language resource repository CLARIN.SI.

[6] Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. Slovene web corpus MaCoCu-sl 1.0, 2022. Slovenian language resource repository CLARIN.SI.

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[9] Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. On the evaluation metrics for paraphrase generation. In *The 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[11] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[12] Matej Ulčar and Marko Robnik-Šikonja. Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0, 2021. Slovenian language resource repository CLARIN.SI.

[13] Matej Ulčar and Marko Robnik-Šikonja. Sequence to sequence pretraining for a less-resourced slovenian language, 2023.

Appendix 1: Example sentences

Table 3. Example generations from T5 model fine-tuned with the *filtered Europarl* dataset.

Input	Evropski policijski urad Europol je konec marca posvaril pred zlorabami tovrstnih besedilnih generatorjev.
Out1	Evropski policijski urad Europol je konec marca opozoril na zlorabo takšne besedilne Generacije zaradi zlorabe besedilnih generatorjev.
Out2	Evropski policijski urad Europol je marca obvestil Europolove stranke, da so zlorabile omenjena besedilna generatorja.
Out3	Evropski policijski urad Europol je na koncu marca posvaril pred uporabo teh besedilnih generatorjev.
Out4	evropski policijski urad Europol je konec marca posvaril pred zlorabami besedilnih generatorjev.
Out5	Evropski policijski urad Europol je marca objavil nagovor zlorabam teh besedilnih generatorjev.

Table 4. Example generations from T5 model fine-tuned with the *filtered LLaMA Europarl Paraphrases* dataset.

Input	Evropski policijski urad Europol je konec marca posvaril pred zlorabami tovrstnih besedilnih generatorjev.
Out1	Evropski policijski urad Europol je v marcu priznal uporabo besedilnih generatorjev, ki so v javni uporabi.
Out2	Evropska komisija je v začetku marca svarila pred uporabo tiskalnih generatorjev.
Out3	Europol je na koncu marca razglasil, da zlorabe besedilnih generatorjev.
Out4	Evropski policijski urad Europol je v začetku marca obvestil Evropske policijske organe, da lahko ponarejeni besedilni generatorji uporabijo svoje opise.
Out5	Evropski policijski urad Europol je v začetku marca opozoril, da obstaja možnost zlorabe različnih besedilnih generatorjev v elektronski obliki.

Table 5. Examples of paraphrase generation with LLaMA and machine translation.

Original	**Paraphrase**
menim, da je konferenca na Baliju odlična priložnost za vse nas, da se zberemo in razmislimo o potrebnih ukrepih glede podnebne politike po letu 2012.	Konferenca na Baliju je primerna priložnost za skupno preučitev potrebnih ukrepov za podnebno politiko po letu 2012.
To je najbolj prazno upanje, kar jih je.	Ta izjava izraža najbolj žalostna pričakovanja.
Vse te čudovite lastnosti pa se skrivajo v skoraj neuničljivi leči, ki v primerjavi s prejšnjo različico kar 25% hitreje reagira na spreminjajočo svetlobo.	Objektiv ima izjemne lastnosti, ki so skrite v strukturi, ki je zelo odporna na poškodbe. Ta objektiv je sposoben reagirati do 25% hitreje na spreminjanje ravni svetlobe v primerjavi s prejšnjim modelom.
Opozoriti želim tudi na problem velikokrat netransparentnega lastništva in nejasnega financiranja, morda celo s tajkunskim denarjem, nekaterih slovenskih medijev, na kar je v svojem nedavnem poročilu o stanju pravne države opozorila tudi Evropska komisija.	Problematiko nepreglednega lastništva in financiranja v nekaterih slovenskih medijih, pri katerih gre lahko tudi za tajkunski denar, je v svojem poročilu o stanju pravne države izpostavila tudi Evropska komisija.
Geox je postala ena vodilnih svetovnih blagovnih znamk, znana po visoki kakovosti.	Geox je pridobil priznanje kot ena izmed najboljših svetovnih blagovnih znamk zaradi svoje vrhunske kakovosti.