SLOmet – Slovenian Commonsense Description

Adrian Mladenic Grobelnik
Department for Artificial Intelligence,
Jozef Stefan Institute
Ljubljana Slovenia
adrian.m.grobelnik@ijs.si

Erik Novak
Department for Artificial Intelligence,
Jozef Stefan Institute,
Jozef Stefan International Postrgraduate School
Ljubljana Slovenia
erik.novak@ijs.si

Dunja Mladenic
Department for Artificial Intelligence,
Jozef Stefan Institute,
Ljubljana Slovenia
dunja.mladenic@ijs.si

Marko Grobelnik
Department for Artificial Intelligence,
Jozef Stefan Institute
Ljubljana Slovenia
marko.grobelnik@ijs.si

ABSTRACT

This paper presents Slovenian commonsense description models based on the COMET framework for English. Inspired by MultiCOMETs approach to multilingual commonsense description, we finetune two Slovenian GPT-2 language models. Experimental evaluation based on several performance metrics shows comparable performance to the original COMET GPT-2 model for English.

KEYWORDS

deep learning, commonsense reasoning, multilingual natural language processing, slovenian language model, gpt-2

1 Introduction

Recent research [1] into commonsense representation and reasoning in the field of natural language understanding has demonstrated promising results for automatic commonsense generation. Given a simple sentence or common entity, such technology can generate plausible commonsense descriptions relating to it. However, further testing on complex sentences, uncommon entities, or by increasing the quantity of requested commonsense descriptions usually gives nonsensical results.

Following the recent success on the automatic generation of commonsense descriptions proposed in COMET-ATOMIC 2020 [1], we focus on extending the COMET framework to the Slovenian language. We investigate the impact of different Slovenian language models on the overall performance of commonsense description generation. In our previous research [2], we expanded on an existing approach for automatic knowledge base construction in English [3] to work on different languages. We utilized the original ATOMIC dataset [4]. This was performed by finetuning the original English GPT model from COMET 2019 on automatically translated Slovenian data and evaluated based on exact overlap for the generated commonsense descriptions. Evaluations were performed on a small subset of 100 sentences. In this work we use the updated ATOMIC-2020 dataset [1] and finetune two Slovenian GPT-2 language models. We evaluate the models' performance using several performance metrics including BLEU, CIDEr, METEOR and ROUGE-L. The evaluation is performed on several thousand sentences and entities; we investigate how the predicted commonsense descriptions' performance relates to the language model used. Furthermore, given the complexity of the Slovenian language compared to English, we anticipate a noticeable drop in performance across all metrics for the Slovenian language models.

The main contributions of this paper are (1) the comparison of the performance of commonsense description models using different Slovenian language models and the English model, (2) a comprehensive evaluation using a variety of performance metrics. An additional contribution (3) is the Slovene ATOMIC-2020 dataset acquired by machine translation from the original English dataset [6].

The rest of this paper is organized as follows: Section 2 provides the data description. Section 3 describes the problem and the experimental setting. Section 4 exhibits our evaluation results. The paper concludes with discussion and directions for future work in Section 5.

2 Data Description

To train the Slovenian commonsense description models, we use data from the ATOMIC-2020 dataset, as proposed in the COMET framework for English. The ATOMIC-2020 dataset consists of English sentences and entities, labelled by up to 23 commonsense relation types describing their semantics.

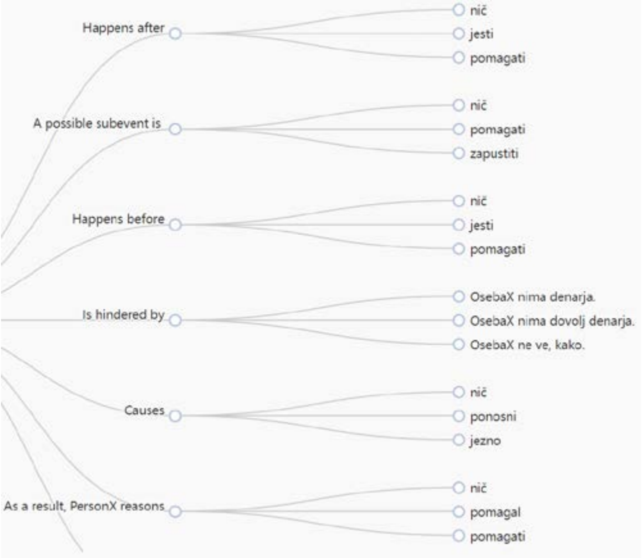

Figure 1 Close-up of "Event-Centered" descriptor values predicted for an example Slovene sentence "PersonX is sad" ("OsebaX je žalostna" in Slovenian)

We refer to them as descriptors, 9 of which are identical to those used in our previous research [2]. The 23 descriptors are organized into 3 categories: “Physical-Entity”, “Event-Centered”, and “Social-Interaction”. The “Physical-Entity” descriptors capture knowledge about the usage, location, content, and other properties of objects. The “Event-Centered” descriptors include IsAfter, Causes and other descriptors describing events. The “Social-Interaction” descriptors include xIntent, xNeed, oReact to distinguish between causes and effects in social settings. An example of a part of a labeled sentence is shown in Figure 1.

Sentences and entities were manually labelled by human workers on Amazon Turk; they were assigned open-text values for 23 commonsense descriptors, reflecting the workers' subjective commonsense knowledge. For instance, when workers were given the sentence “PersonX chases the rabbit” and asked to label it for the “xWant” descriptor, one wrote “catch the rabbit” and another wrote “cook the rabbit for dinner”. A more detailed explanation can be found in the ATOMIC-2020 paper. There are 1.33 million (possibly repeating) descriptor values. The distribution of data across the descriptors is depicted in [1].

To finetune our Slovenian language models, we have automatically translated the sentences, entities, and descriptor values from the ATOMIC-2020 dataset from English to Slovenian. The translation was done using DeepL’s Translate API [7]. We have found that while the majority of inspected translations were of good quality, there were also incorrect translations due to word disambiguation problems. Nevertheless, we conclude that the dataset is of good enough quality to be used for our experiments. The translated dataset is publicly available [6].

3 Problem Description and Experimental Setting

The addressed problem is predicting the most likely values for each descriptor in the Slovene-translated ATOMIC-2020 dataset, given a Slovenian input sentence or entity. We take inspiration from the approach proposed in MultiCOMET [2].

To compare the performance of the models, we utilize a variety of performance metrics described below. Each performance metric is a value between 0 and 1 indicating the quality of a generated commonsense descriptor value. Values closer to 1 represent higher quality descriptions.

BLEU — Bilingual Evaluation Understudy was first used to evaluate the quality of machine translated text by examining the overlap of candidate text n-grams in the reference text. BLEU-1 only uses 1-grams in the evaluation, while BLEU-4 only considers 4-grams. [8]

CIDEr — Consensus-based Image Description Evaluation was originally used to measure image description quality. It first transforms all n-grams to their root form, then calculates the average cosine similarity between the candidate and reference TF-IDF vectors. [9]

METEOR — Metric for Evaluation of Translation with Explicit Ordering is a metric initially used for evaluating machine translation input. The metric is based on the harmonic mean of unigram precision and recall with other features such as stemming and synonymy matching. [10]

ROUGE-L — Recall-Oriented Understudy for Gisting Evaluation is a metric used for evaluating machine produced summaries or translations against a set of human-produced references. The score is calculated using Longest Common Subsequence based statistics, which involves finding the longest subsequence common to all sequences in a set. [11]

Comparison of the Slovene commonsense models was performed by finetuning two state-of-the-art Slovene GPT-2 language models: macedonizer/sl-gpt2 [12], gpt-janez [13]. As a reference model, we used the original COMET-2020 GPT2-XL English language model [1]. Moving forward, we will refer to our Slovenian finetuned models as “COMET sl-gpt2” and “COMET gpt-janez”.

4 Experimental Results

We performed a train, test, and development split on the ATOMIC-2020 dataset identical to the split used in COMET-2020. Our evaluation split consisted of over 150,000 descriptor values with their corresponding sentences and entities.

We finetuned our Slovene commonsense models on our training set consisting of over 1 million descriptor values. Both models were trained for 3 epochs under the same parameters; the maximum input length was set to 50, the maximum output length was set to 80; the training was performed using a train batch size of 64. The model updates were performed using the weighted adam optimizer [14] with the starting learning rate set to 10^{-5}. The experiment’s implementation can be found on our GitHub repository [5].

Model	Language	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L
COMET sl-gpt2	Slovene	0.297	0.150	0.086	0.058	0.487	0.207	0.383
COMET gpt-janez	Slovene	**0.324**	**0.174**	**0.108**	**0.076**	**0.508**	**0.225**	**0.397**
COMET (GPT2-XL)	English	0.407	0.248	0.171	0.124	0.653	0.292	0.485

Table 1: Comparison of the two Slovene commonsense models with the English model at the bottom.

Experimental results show performance comparable to the original COMET-2020 English model. Both Slovene models were comparable to the English model across all metrics, “COMET gpt-janez” consistently outperformed “COMET sl-gpt2” achieving a METEOR score of 0.225 compared to 0.207. The performance gap was smallest for BLEU-4, as all models struggled to produce generations whose 4-grams overlapped with those in the reference set. The gap in performance between the Slovene and English models could be attributed to multiple factors. The English model from COMET-2020 was trained for longer on more capable hardware and is larger. Moreover, the machine translation done to acquire our dataset can be erroneous at times.

To illustrate the performance of the models, we investigate their generated descriptor values on the same inputs. Table 2 shows a side-by-side example comparison of the descriptor values generated by our three models, given the same input sentence in their respective language. Table 3 compares the models on an example entity. For the example sentence “Marko went to the shop”, the descriptor “oWant” indicates what the others want as a result of the event. “COMET gpt-janez” generates a valid output “None” but fails to provide alternatives. The other two models agree on the most likely descriptor value being “None” (“nič” in Slovenian) and provide plausible alternatives. The “IsBefore” descriptor relates to possible events following the input event. In our case, “COMET gpt-janez” gives the most plausible output of “Buys something”. The other two models provide still plausible outputs including “Is in the pet store” and “PersonX buys a new car”.

Marko je šel v trgovino (Marko went to the shop)

Descriptor	COMET sl-gpt2	COMET gpt-janez	COMET (GPT2-XL)
oWant	Nič	Nič	None
	Se zahvaliti osebiX	Nič	To give him a receipt
	se zahvaliti	Nič	To give him a discount
IsBefore	Zaslužiti denar	Kupiti nekaj	PersonX buys a new car
	V trgovino za hišne ljubljenčke	Kupiti nekaj	PersonX takes the car back home
	V trgovino z živili	Kupiti nekaj	PersonX buys a new one

Table 2: Illustrative example comparing the output of the three models on the same input sentence across two descriptors.

For our example entity “car”, the descriptor “ObjectUse” describes possible usages for that entity. Table 3 shows all models are capable of generating plausible descriptor values for such common entities. Nevertheless, the descriptor “HasProperty” proves challenging for the Slovenian models, suggesting a car is “crazy” and is “found in the car”. The English model gives reasonable outputs such as “Found in parking lot”.

Avto (car)

Descriptor	COMET sl-gpt2	COMET gpt-janez	COMET (GPT2-XL)
ObjectUse	Vožnja do trgovine	Priti do hiše	Drive to the store
	Vožnja do hiše	Priti do hiše	Get to the store
	Vožnja do cilja	Priti do hiše	Drive to the restaurant
HasProperty	Noro	Najden v avtomobilu	Found in parking lot
	Vrata	Najden v avtomobilu	Found on road
	Pohištvo	Najden v avtomobilu	Found in car dealership

Table 3: Illustrative example comparing the output of the three models on the same input entity across two descriptors.

In our example sentence and entity, COMET gpt-janez returns the same output when different commonsense descriptors are requested. We have observed this for all input sentences and entities thus far. We presume such results are due to the trained parameters in the original gpt-janez model, as macedonizer/sl-gpt2 was finetuned using the same workflow and returns different descriptor values. While unsure of the exact cause, we reason it could be due to an insufficient vocabulary or unoptimized choice of parameters during training.

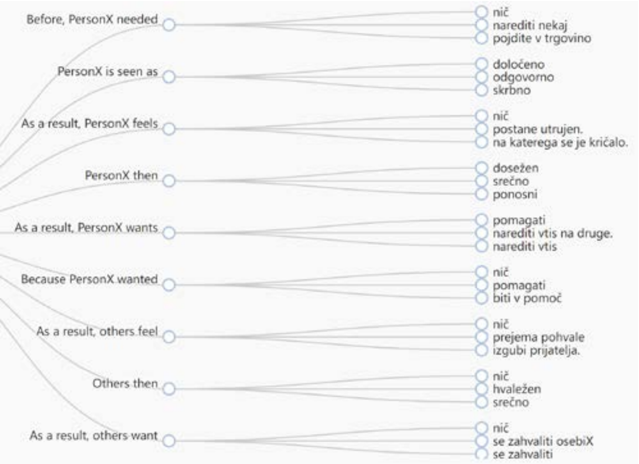

Figure 2 Close-up of “Social-Interaction” descriptor values predicted for an example Slovene sentence “John is very important” (“Janez je zelo pomemben” in Slovenian)

Figures 1, 2 and 3 show the outputs generated by “COMET sl-gpt2” for three different inputs. Figure 2 visualizes the output for the sentence “John is very important”. Outputs include “PersonX is then accomplished, happy, proud” and “As a result, others want none, to thank PersonX”. We can see that for many descriptors the highest ranked output is “None” (“nič” in Slovenian), indicating no commonsense inference can be made.

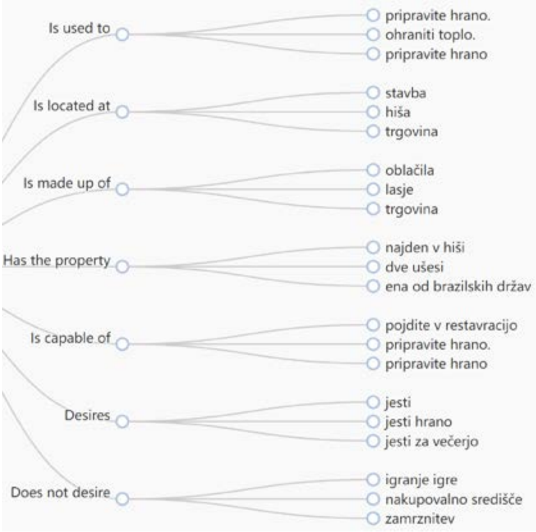

Figure 3 Close-up of "Physical-Entity" descriptor values predicted for an example Slovene entity "banana"

Figure 3 exhibits the output for the entity "banana", the model claims the banana can be used to prepare food, is located in a building or shop, desires to be eaten for dinner and does not desire to be frozen. On the other hand, the model claims the banana is made up of clothes and is capable of going to a restaurant. This is likely due to the overall significantly lower number of physical-entity descriptor values provided in the ATOMIC-2020 dataset.

In Figure 1 we can see the "Event-Centered" descriptors for the sentence "PersonX is sad". Top descriptor values are again "None", but the model also claims it is more difficult for PersonX to be sad, if PersonX has no money.

5 Discussion

This paper applied an existing approach to multilingual commonsense description to the Slovene language. To implement our approach, we machine translated the ATOMIC-2020 dataset to Slovene and finetuned two Slovene commonsense models. We compared our models to the original English commonsense model from COMET-2020 and achieved comparable experimental results across multiple performance metrics. Among others, our models achieved a 0.487 CIDEr score, a 0.383 ROUGE-L score, and a BLEU-1 score of 0.297.

Through examination of individual examples, we observed that while "COMET gpt-janez" has the highest performance scores on the Slovene language, it fails to provide alternative descriptor values. "COMET sl-gpt" provides multiple values for the same descriptor, but in average has lower performance. It is important to emphasize the models were trained on subjective commonsense knowledge provided by individual humans. For example, workers labelled the sentence "PersonX digs holes" with the descriptor values "PersonX plants a garden" and "PersonX places fence posts in the holes" for the "IsBefore" descriptor. While both labels are plausible for some context, they are not necessarily true.

Possible directions for future work include evaluating the models' performance for individual descriptors, as there are drastic differences in quantity of training data and lengths of values across them. After achieving results comparable to the original English commonsense model COMET-2020 GPT2-XL, we intend to finetune and evaluate models for other languages.

ACKNOWLEDGMENTS

The research described in this paper was supported by the Slovenian research agency under the project J2-1736 Causalify, the RSDO project funded by the Development of Slovene in a Digital Environment project, and the Humane AI Net European Unions Horizon 2020 project under grant agreement No 952026.

REFERENCES

[1] Hwang, J.D., Bhagavatula, C., Le Bras, R., Da, J., Sakaguchi, K., Bosselut, A., & Choi, Y. (2021). COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. AAAI.
[2] Mladenic Grobelnik, A., Mladenić, D., & Grobelnik, M. (2020). MultiCOMET - Multilingual Commonsense Description. In Proc. SiKDD 2020, Ljubljana, Slovenia (pp. 37–40).
[3] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, Yejin Choi. (2019). COMET: Commonsense Transformers for Automatic Knowledge Graph Construction.
[4] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, Yejin Choi. (2019). ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA. Allen Institute for Artificial Intelligence, Seattle, USA.
[5] SLOmet-ATOMIC 2020 Github https://github.com/eriknovak/RSDO-SLOmet-atomic-2020#slomet-atomic-2020-on-symbolic-and-neural-commonsense-knowledge-graphs-in-slovenian-language Accessed 30.08.2022
[6] ATOMIC-2020 Slovene Machine Translated Data https://www.dropbox.com/sh/gs8iqcwpwkaqkuf/AAAmnCqG89JOz_umtq42MMxxa?dl=0 Accessed 30.08.2022
[7] DeepL Translate API https://www.deepl.com/pro-api Accessed 30.08.2022
[8] Papineni, Kishore & Roukos, Salim & Ward, Todd & Zhu, Wei Jing. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation.
[9] Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).
[10] Lavie, Alon & Denkowski, Michael. (2009). The METEOR metric for automatic evaluation of Machine Translation. Machine Translation. 23. 105-115.
[11] Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out.
[12] Documentation page for "macedonizer/sl-gpt2" on HuggingFace https://huggingface.co/macedonizer/sl-gpt2 Accessed 1.09.2022
[13] gpt-janez supporting project: RSDO https://www.cjvt.si/rsdo/en/project/ Accessed 30.08.2022
[14] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 201