MultiCOMET – Multilingual Commonsense Description

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

Dunja Mladenic
Artificial Intelligence Laboratory
Jozef Stefan Institute
Ljubljana Slovenia
dunja.mladenic@ijs.si

Marko Grobelnik
Artificial Intelligence Laboratory
Jozef Stefan Institute
Ljubljana Slovenia
marko.grobelnik@ijs.si

ABSTRACT

This paper presents an approach to generating multilingual commonsense descriptions of sentences provided in natural language. We have expanded on an existing approach to automatic knowledge base construction in English to work on different languages. The proposed approach has been utilized to develop MultiCOMET, a publicly available online service for generating multilingual commonsense descriptions. Our experimental results show that the proposed approach is suitable for generating commonsense description for natural languages with Latin script. Comparing performance on Slovenian sentences to the English original, we have achieved precision as high as 0.7 for certain types of descriptors.

CCS CONCEPTS

•CCS Information systems Information retrieval Document representation Content analysis and feature selection

KEYWORDS

deep learning, commonsense reasoning, multilingual natural language processing

1 Introduction

As artificial intelligence systems are becoming better at performing highly specialized tasks, sometimes outperforming humans, they are unable to understand a simple children's fairy tale due to their inability to make commonsense inferences from simple events. With recent breakthroughs in the area of deep learning and overall increases in computing power, it has enabled us to model commonsense inferences with deep learning models. In our research, we expand on the approach to automatic generation of commonsense descriptors proposed in COMET [1] by applying their deep learning models to languages other than English.

The approach presented in COMET tackles automatic commonsense completion with the development of generative models of commonsense knowledge, and commonsense transformers that learn to generate diverse commonsense descriptions in natural language [1].

Our research hypothesis is that the approach proposed by COMET [1] can be expanded to Latin script languages other than English. To test this claim, we have trained our own deep learning model on the original training data, and another model on the data translated into another natural language.

The main contributions of this paper are (1) a new multilingual approach to annotating natural language sentences with commonsense descriptors, (2) implementation of the proposed approach that is made publicly available as an online service MultiCOMET http://multicomet.ijs.si/ (illustrated in Figure 4), (3) evaluation of the proposed approach on the Slovenian language. An additional contribution is the publicly available source code [3] allowing users to train their own models for other natural languages.

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

2 Data Description

One might say the only way for AI to learn to perform commonsense reasoning, is to learn from humans. Following the approach proposed by COMET [1], we used data from the ATOMIC [2] dataset. The ATOMIC dataset consists of over 24,000 sentences containing common phrases manually labelled by workers on Amazon Turk. For each sentence the workers were asked to assign open-text values to nine descriptors which capture nine if-then relation types to distinguish causes vs. effects, agents vs. themes, voluntary vs. involuntary events and actions vs. mental states [2] as described in ATOMIC.

The following are the nine descriptors and their explanations:

xIntent – Because PersonX **wanted…**

xNeed – Before, PersonX **needed…**

xAttr – PersonX is **seen as…**

xReact – As a result, PersonX **feels…**

xWant – As a result, PersonX **wants…**

xEffect – PersonX **then…**

oReact – As a result, others **feel…**

oWant – As a result, others **want…**

oEffect – Others **then…**

The dataset contains almost 300,000 unique descriptor values for the listed nine descriptors. An example of a labeled sentence is shown in Figure 3.

In order to test the proposed approach, we implemented it for the Slovene language. We have translated the sentences from the ATOMIC dataset to Slovene, keeping the descriptor values in English. The translation was done using Google Cloud's Translation API [4].

3 Problem Description and Algorithm

The problem we are solving is predicting the most likely values for each tag in the ATOMIC [1] dataset, given an input sentence in a Latin script language. Following the proposal in COMET, we are addressing the following problem:

Given a training knowledge base of natural tuples in the $\{s, r, d\}$ format, where s is the sentence, r is the relation type and d represents the relation values. The task is to generate d given s and r as inputs.

Figure 1 depicts our approach to solving this problem. The system takes labelled sentences as input, translates them to the targeted Latin language and trains a deep learning model capable of labelling previously unseen sentences with values for nine descriptors capturing the nine predefined relation types as described in Section 2.

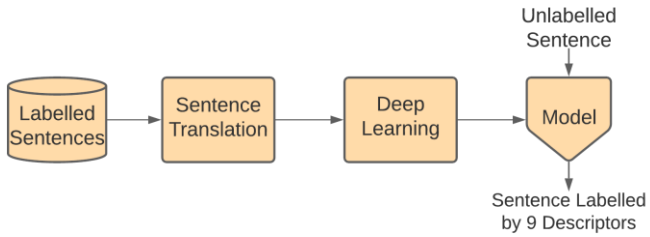

Figure 1: Architecture of the proposed approach

4 Experimental Results

Prior to training the model, we split the ATOMIC dataset into train, test and development sets identical to those used in COMET [1]. In our evaluation we used 100 sentences from the test set.

Our deep learning models are trained on the ATOMIC [2] dataset. We have trained one model on the original dataset in English, and another model on an automatically translated dataset to Slovene. Both models were trained under the same parameter settings: batch size=6, iterations=50000, maximum number of input features = 50.

To evaluate the performance of the proposed approach, we compared the predictions of the model trained on Slovene sentences with the predictions of the English model. As the performance metrics, we took the top 5 predicted values for each descriptor and checked their overlap. By taking the English predictions as the ground truth, we are measuring the precision of our model by the number of identical descriptor values. Note that we were strict in our comparisons, for instance "to stay away from people" and "to get away from others" do not count in overlap.

Experimental results show there is considerable difference in performance between the nine descriptors. The best performing descriptor was xReact, where precision@5 was 0.716, followed by oReact and oWant with precisions@5 of 0.706 and 0.468 respectively. The worst performing descriptor was xWant, with a precision@5 of 0.21 (see Table 1).

Descriptor	Precision
xIntent	0.324
xNeed	0.352
xAttr	0.438
xReact	0.716
xWant	0.210
xEffect	0.456
oReact	0.706
oWant	0.468
oEffect	0.310
Average	0.442

Table 1: Experimental results on the nine descriptors, showing precision of the top 5 predictions.

The best performing descriptor was xReact (representing the relation: As a result, PersonX feels). This was likely due to the fact that most predicted values were only one word long for both models, making it considerably easier for their predictions to overlap.

The worst performing descriptor was xWant (representing the relation: As a result, PersonX wants), this could be attributed to the fact that the most predicted values were at least 3-4 words in length, greatly decreasing the likelihood of overlap. Another reason for such low precision could be our strict overlap comparisons.

	Original	Translated/Predicted
Sentence	PersonX looks PersonY ___ in the face	PersonX izgleda PersonY ___ v obraz
xReact Values	nervous	**satisfied**
	happy	**happy**
	satisfied	attractive
	powerful	proud
	confident	angry

Table 2: One of the worst performing test sentences for xReact

Table 2 shows the predicted values of one of the worst performing sentences for the xReact descriptor. Note the sentence "PersonX looks PersonY ___ in the face" can refer to "Bob looks Mary slowly in the face" or "Adrian looks Anna kindly in the face" or something

else. The columns in Table 2 and Table 3 labelled “Original” show the original English sentence and its predicted descriptor values. The columns labelled “Translated/Predicted” show the sentence translated into Slovene and its predicted descriptor values.

Table 3 shows the predicted values of one of the worst performing sentences for the xWant descriptor. We can see that there are no common predictions between the two models. Note the sentence “PersonX avoids every ___” can refer to “Marko avoids every car on the road” or “Dunja avoids every boring event” or something else.

	Original	Translated/Predicted
Sentence	PersonX avoids every ___	PersonX se izogiba vsakemu ___
xWant Values	to stay away from people	to get away from others
	to avoid trouble	to make sure they are ok
	to stay away	to get away from the situation
	to not get caught	to be alone
	to not be noticed	to make a decision

Table 3: One of the worst performing test sentences for xWant

While Tables 2 and 3 show the model’s outputs for a single descriptor, Figure 3 shows the full output of the model, given an example sentence “Mojca je pojedla odličen sendvič” (Mary ate an excellent sandwich). Figure 2 shows a close-up of the output of Figure 3. The images in Figures 2 and 3 were taken directly from the interface of our online service MultiCOMET [5].

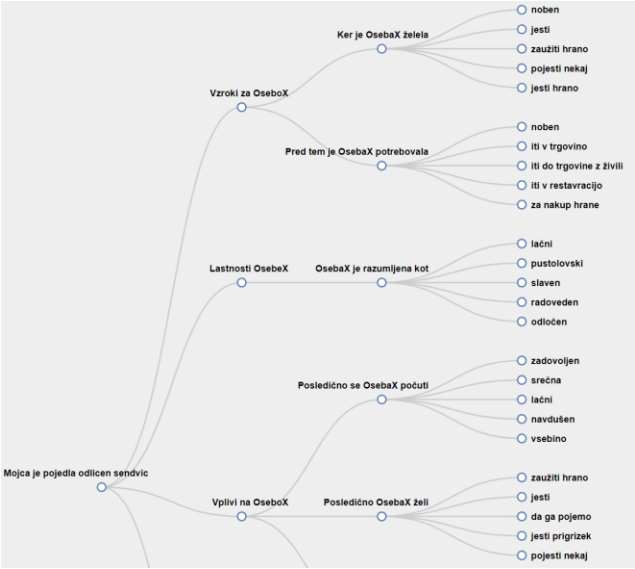

Figure 2: Close-up of predicted descriptor values generated for an example Slovene sentence

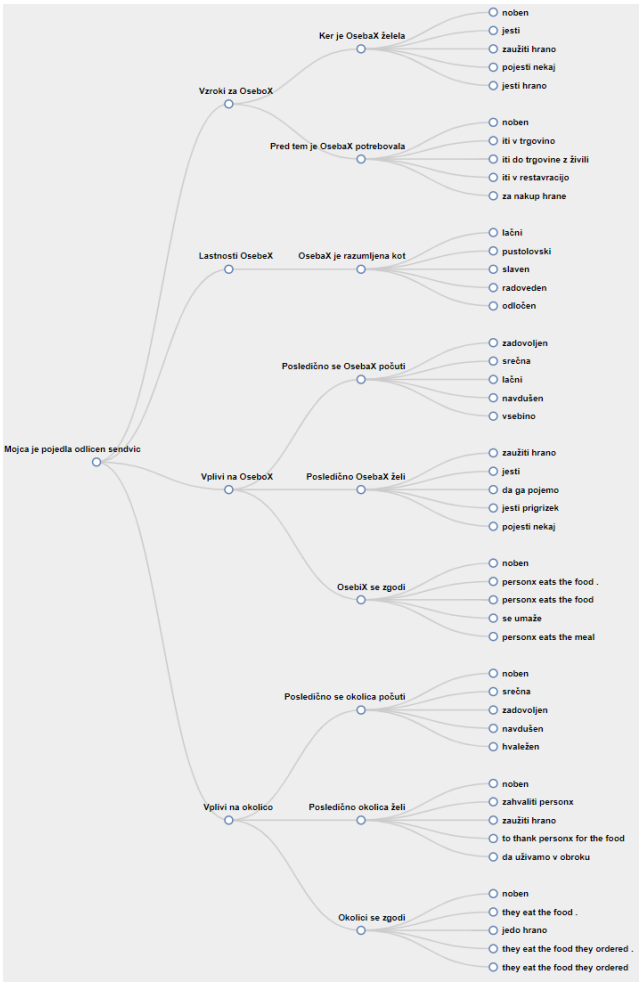

Figure 3: Full tree of predicted descriptor values generated for an example Slovene sentence

For the sentence “Mojca je pojedla odličen sendvič” (Mary ate an excellent sandwich) depicted in Figures 2 and 3, here is a potential English interpretation of the Slovenian output of the model:

Mary was hungry (xAttr) and wanted to eat food (xIntent). To do that, she needed to go to the restaurant (xNeed). At the restaurant, other people were also eating food (oEffect). As a consequence of eating the sandwich, Mary’s clothes got dirty (xEffect). Mary feels impressed (xReact) and wants to eat something else (xWant). The restaurant is grateful (oReact) for Mary’s visit and wants to thank Mary (oWant).

The MultiCOMET online service is a publicly available implementation of our proposed approach, shown in Figure 4. At the time of writing, MultiCOMET only supports English and Slovene.

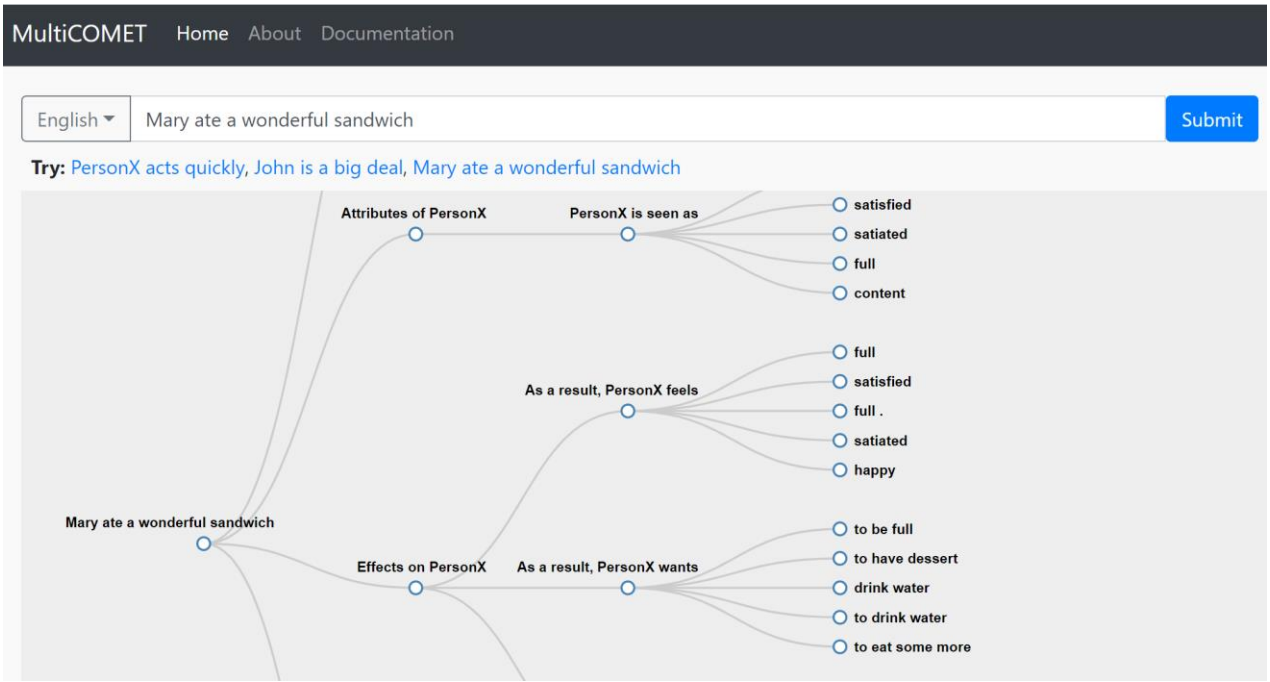

Figure 4: Illustrative example of MultiCOMET after submitting a query "Mary ate a wonderful sandwich."

5 Discussion

In our research we expanded on an existing monolingual approach and proposed a new approach to generating multilingual commonsense descriptions from natural language. In order to implement our approach, we built on an existing library, implementing the approach proposed by COMET [1]. Our experimental results show that we are getting meaningful values for the descriptors. Experimental comparison of the predicted descriptor values of the Slovene and English models show an average precision of 0.44, given our strict comparison methodology. We noted the precision values ranged from 0.716 to 0.210 across different descriptors.

Based on our literature review (September 2020), none of the articles citing the original COMET [1] paper expanded their approach to include other languages. The most similar work we found in the literature combining commonsense and multilinguality was [6] where the authors were extending the SemEval Task 4 solution using machine translation.

The possible direction for future work includes improving the quality of the translated sentences from ATOMIC by manual translation to improve the precision of the models. Another possible direction would be to evaluate the performance of our models on a larger number of sentences to increase the reliability of the results.

After testing the proposed multilingual approach on the Slovene language, we intend to expand our coverage to other Latin script languages including Croatian, Italian and French.

ACKNOWLEDGMENTS

The research described in this paper was supported by the Slovenian research agency under the project J2-1736 Causalify and co-financed by the Republic of Slovenia and the European Union under the European Regional Development Fund. The operation is carried out under the Operational Programme for the Implementation of the EU Cohesion Policy 2014–2020.

REFERENCES

[1] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, Yejin Choi. (2019). COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. Allen Institute for Artificial Intelligence, Seattle, WA, USA. Paul G. Allen School of Computer Science & Engineering, Seattle, WA, USA. Microsoft Research, Redmond, WA, USA.

[2] 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.

[3] MultiCOMET GitHub https://github.com/AMGrobelnik/MultiCOMET Accessed 31.08.2020

[4] Google Cloud's Translation API Basic https://cloud.google.com/translate Accessed 31.08.2020

[5] MultiCOMET http://multicomet.ijs.si/ Accessed 31.08.2020

[6] Josef Jon, Martin Fajcik, Martin Docekal, Pavel Smrz. (2020). BUT-FIT at SemEval-2020 Task 4: Multilingual commonsense. arXiv. https://arxiv.org/pdf/2008.07259.pdf