Position Papers of the Federated Conference on Computer Science and Information Systems pp. 3–8
DOI: 10.15439/2020F212
ISSN 2300-5963 ACSIS, Vol. 22

From Machine Translated NLI Corpus to Universal Sentence Representations in Czech

Martin Víta
NLP Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno
Czech Republic
Email: info@martinvita.eu

***Abstract*—Natural language inference (NLI) is a sentence-pair classification task w.r.t. the entailment relation. As already shown, certain deep learning architectures for NLI task – INFERSENT in particular – may be exploited for obtaining (supervised) universal sentence embeddings. Although INFERSENT approach to sentence embeddings has been recently outperformed in different tasks by transformer-based architectures (like BERT and its derivatives), it still remains a useful tool in many NLP areas and it also serves as a strong baseline. One of the greatest advantages of this approach is its relative simplicity. Moreover, in contrast to other approaches, the training of INFERSENT models can be performed on a standard GPU within hours. Unfortunately, the majority of research on sentence embeddings in general is done in/for English, whereas other languages are apparently neglected. In order to fill this gab, we propose a methodology for obtaining universal sentence embeddings in another language – arising from training INFERSENT-based sentence encoders on *machine translated* NLI corpus and present a transfer learning use-case on semantic textual similarity in Czech.**

I. INTRODUCTION

NATURAL language inference (NLI) task, i.e., a sentence-pair classification task with respect to the entailment relation – usually into three classes (ENTAILMENT, NEUTRAL and CONTRADICTION) has been intensively studied in the last (approximately) fifteen years – formerly, this task was known as *recognizing textual entailment (RTE)*. The sentences forming the sentence pair to be classified are commonly known as premise and hypothesis.

The rapid development in NLI area was allowed, on one hand by strong progress in deep learning in NLP and, on the other hand, by releasing the first large volume annotated corpus for NLI in 2015 – well known Stanford NLI corpus (abbr. SNLI) [1], later followed by MultiNLI dataset [2] which covered wider range of topics and genres, both in English. Therefore, the majority of NLI research has been focused on NLI in English, other languages are still highly neglected. It is reminiscent of a "chicken-egg problem": research on languages different to English are neglected, since there are no suitable resources (annotated corpora), and, in the opposite direction, not "so strong research effort means lower pressure for development of relevant annotated corpora".

In [3], Conneau et al. shown, that NLI task is suitable for obtaining (supervised) universal sentence embeddings – these embeddings are produced by sentence encoders that form Siamese architecture called INFERSENT (two identical architectures are used for encoding both premises and hypotheses in the same manner). The entire classification architecture for NLI consists of these two encoders, a merging layer that combine these embeddings – the output of the merging layer is subsequently fed into a dense layer, followed by a final sigmoid layer. Sentences at the input are represented as sequences of word embeddings (like GloVe [4], word2vec or fastText. The INFERSENT authors trained this architecture on previously mentioned SNLI corpus, in some variants augmented by MultiNLI corpus. As we can observe, this work is again limited to English.

To fill this "language gab", we introduce a machine translated version of SNLI corpus into Czech. Subsequently, we have trained on this newly proposed dataset one of INFERSENT-based architectures. Alongside with this model for SNLI in Czech, we have obtained also sentence encoder for Czech. To demonstrate the capabilities of these Czech sentence embeddings, we used these sentence encoders for a task of *semantic textual similarity* in Czech.

This proposed process may be shift into a more general level – the process can be performed in the following steps:

1) The NLI corpus (e.g., SNLI) is machine translated to a selected target language (Czech for instance).
2) An INFERSENT is trained on the translated NLI dataset in the target language and sentence encoders are obtained.
3) Sentence encoders are used within models or other semantic oriented tasks in the target languages (transfer learning).

The requirements for this process are implicitly specified in the first two steps: this process relies on the availability of machine translation tools (or TranslationAPI) for a considered source-target language pair. The second requirement is the availability of suitable word embeddings. However, within MUSE project, FASTTEXT embeddings are available for more than one hundred languages.

In the following parts of this position paper we will elaborate on each step of the outlined process.

II. NLI Corpora and DNN Architectures

At first, we are going to summarize the key characteristics of the SNLI corpus and DNN architectures involved.

A. Original SNLI Corpus in English

Nowadays, SNLI (Stanford NLI corpus) is probably the best known corpus for NLI task. The entire corpus contains of 570K labeled sentence pairs split in a TRAIN (550K), DEV (10K) and TEST (10K) sets. These pairs were generated by annotators (crowdworkers) based on image captions mostly of the FLICKR30K dataset [5] and a minor part of the TRAIN set (4K) on captions that were taken from the VisualGenome dataset [6]. The annotators were asked, given a textual caption (without the original photo), to create a three other sentences (i.e., alternative captions) that satisfy the following conditions [1]:

- one is "definitely a true description of the photo",
- one "might be a true description of the photo",
- one is "definitely a false description of the photo."

The original sentence given to annotators was taken as premise, the three sentences produced by annotators were taken as hypotheses. These sentence pairs were labeled according to the conditions as ENTAILMENT, NEUTRAL and CONTRADICTION, respectively. Subsequently, 56,941 samples were validated by four additional judgments showing a high annotation agreement. The details about the corpus development process is provided in the original paper [1].

B. Machine Translated SNLI Czech Version of SNLI Corpus

In order to obtain Czech NLI annotated corpus, we chose a (machine) translation approach. Since the inference is a semantic phenomenon (and hence "invariant to translation", i.e., the entailment relation between a premise and a hypothesis expressed by the label, is the same in both original/source and target language), we can simply use the original labels.

In recent years, the machine translation (MT) approach was utilized for German in a task of *contradiction detection*, see [7]. However, in this case, the authors took only a part of SNLI corpus (110,000 items in particular) and translated it subsequently using DeepL service[1]). No analysis of the German counterpart was performed.

In our case, the Czech MT version of SNLI was created using translation LINDAT Translation API [2] – we have translated the entire SNLI corpus sentence-by-sentence. The TRAIN/DEV/TEST splits remain unchanged. This process relies on the implicit assumption MT system produces translations in a sufficient quality. This assumption is supported by the fact that image captions that form the "premises" part of the corpus are usually short and do not have a complicated dependency structure, thus we may expect reasonable results of machine translation process. However, this quality assumption will be analyzed in the further text.

[1]https://www.deepl.com/translator

[2]https://lindat.mff.cuni.cz/services/translation

TABLE I
EXAMPLE OF ITEMS IN CZECH MT VERSION OF SNLI CORPUS

Premise: *Přes řeku právě projíždí terénní vůz.* (orig.: *A land rover is being driven across a river.*)
Hypothesis: *Vozidlo přejíždí řeku.* (orig.: *A vehicle is crossing a river.*)
Label: ENTAILMENT
Premise: *Muž v černé košili se dívá na kolo v dílně.* (orig.: *A man in a black shirt is looking at a bike in a workshop.*)
Hypothesis: *Muž se rozhoduje, které kolo si koupí.* (orig.: *A man is deciding which bike to buy.*)
Label: NEUTRAL
Premise: *Holky jdou po ulici.* (orig.: *The girls walk down the street.*)
Hypothesis: *Dívky se usadily na ulici.* (orig.: *Girls set down in the street.*)
Label: CONTRADICTION

TABLE II
1- TO 4-GRAMS BLEU SCORES

Type	1-gram	2-gram	3-gram	4-gram
Score	80.35	62.18	50.92	42.38

To provide a better idea about the corpus, we selected three sentence pairs from Czech MT version of SNLI corpus – from the TEST subset in particular (one sentence pair for each label), see Table I. This table also shows the original source sentences, hence it provides also the examples of original sentence pairs of SNLI corpus.

The Czech MT version of SNLI corpus is freely available for download[3].

C. Selected Characteristics of Czech MT version of SNLI Corpus

As we have already mentioned, we are going to present an evidence that justify our the MT approach. At first, we have computed a "traditional" MT evaluation metric: BLEU score [8], [9]. We have prepared a sample of 100 randomly selected hypotheses from the TEST set and translate them manually from English to Czech. This manual translation was done by two independently working Czech native speakers. Then we have computed BLEU score w.r.t. machine translation and this human translated (reference) sentences using Interactive BLEU score evaluator[4]. The results for 1- to 4-grams are summarized in Table II.

In unigram setting, we have obtained a value exceeding 80%. This suggest a sufficient quality of translation. At this point we should notice that our primary aim is not to focus on "translation quality" and its assessment, but on the quality of the NLI corpus being developed. (And, we should take into account that the "wrong" translation does not necessarily lead to incorrect entailment labels. It may be obvious mainly in case of sentence pairs labeled as NEUTRAL: if the sentences forming a pair in NLI corpus are translated incorrectly, then the label is regardless most likely correct.). Nevertheless, we performed an experiment that elucidate the question of quality of labels in the Czech MT version of SNLI corpus.

[3]https://github.com/martinvita/CZinferSent

[4]https://www.letsmt.eu/Bleu.aspx

D. EN-CZ Label Transfer and its Quality

To estimate the quality of the entailment labels in the target language corpus, we assess the entailment labels manually again by two independently working Czech native speakers. The task was stated as follows: given (only) sentence pairs in Czech (machine translated) accompanied with transferred labels, the annotators were asked to check the correctness of the label (in a binary way) without the knowledge of the original sentence pairs (in the source language). This experiment was done on a random sample of 500 sentence pairs from the Czech TEST dataset with the following results:

- 454 items were marked as correct, i.e., the label corresponding to Czech premise-hypothesis sentence pair was correct.
- 46 (i.e., 9.2% of 500) items were revealed as incorrect.

However, in the further (human) analysis it was found that in majority of the incorrect cases, the incorrectness of the labels was contained already in the source SNLI corpus.

E. INFERSENT Architecture

Nowadays, we can observe a huge number of deep learning approaches to NLI in general. A comprehensive overview of the architectures involved in SNLI task can be found on the SNLI dashboard[5]. However, for the purposes of this paper, it is not necessary to provide a survey of these approaches, we only divide the deep learning into two major classes:

- Architectures encoding premise and hypothesis separately (usually using Siamese architectures), there is no mutual "interaction" between premise and hypothesis within the "encoding phase". Premise and hypotheses embeddings are subsequently merged and the final decision is made usually using fully connected layers.
- Architectures encoding the problem into a "joint embedding" using based on cross-sentence features constructed by various attention mechanisms between premise and hypothesis.

From our perspective (i.e., development of sentence embeddings) the first class of approaches is a keystone. The general architecture of such approaches / architectures is depicted in Figure II-E. (it is a generalization of a scheme on Figure 1 in [3]): premise and hypothesis embeddings u, v (obtained from GRUs or LSTMs for instance) are merged using a function f, that may be a simple concatenation of u, v, i.e., $f(u, v) = (u, v)$, or enriched representation dealing with pointwise absolute value of difference of u, v, and their pointwise product, i.e., $f(u, v) = (u, v, |u - v|, u * v)$ – this "enriched" approach is the utilized in [3] and also in this work. The final decision is made by a dense layer(s) and a 3-way softmax.

The INFERSENT approach is basically a collection of similar architectures corresponding to scheme in Figure II-E with different encoders, including LSTM [10], GRU [11] and their bidirectional variants, self-attention architecture, hierarchical convolutional networks and others. Their detailed description is provided in [3].

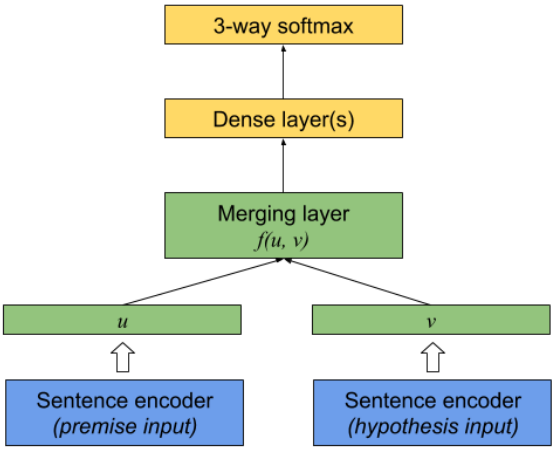

Fig. 1. General architecture of the first class of approaches (no attention between premise and hypothesis)

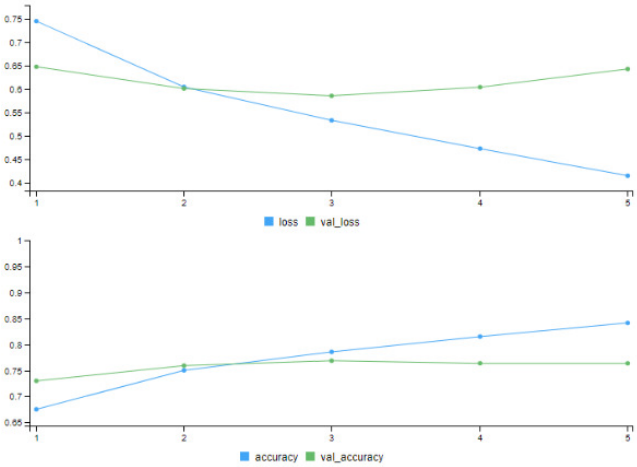

Fig. 2. Training one of the INFERSENT-based model in 5 epochs

For our proof-of-concept, we have chosen an INFERSENT architecture using GRU sentence encoder (i.e. encoders are GRU layers sharing the same parameters in premise and hypothesis part, i.e., Siamese architecture).

Sentences (premises/hypotheses) that are fed into the GRU layers are represented as sequences of word embeddings. Since we deal with Czech, we did not used GLOVE [4] as in the original INFERSENT model, but we exploited precomputed FASTTEXT embeddings from MUSE project[6]

Architecture and training details: the dimension of GRU layer was set to 512 as well as the dimension of the fully connected layer which follows the merging layer. The model was trained in 10 epochs using SGD optimizer, the implementation was written in R+Keras and rewritten in Python+Keras. Illustration of the training process is depicted in Figure II-E.

[5]https://nlp.stanford.edu/projects/snli/

[6]Available for download at https://github.com/facebookresearch/MUSE.

On the Czech TEST set we achieved 78.69 accuracy within the setting described above. This result may serve as a strong baseline for the Czech MT version of SNLI corpus.

III. TRANSFER LEARNING USE CASE – SEMANTIC TEXTUAL SIMILARITY IN CZECH

As an application of supervised sentence embeddings in Czech, i.e., for transfer learning, we chose a well known task of semantic textual similarity (in Czech).

A. Semantic Textual Similarity (STS) - Task Description

Semantic Textual Similarity (STS) can be defined by a metric over a set of documents with the idea is to finding the semantic similarity between them [12]. It was introduced for short texts (sentences) in [13]. Given two text snippets/sentences the task is to assign a numeric value from an interval $[m, n]$ for this pair, where the n value stands for identity, m corresponds with total unrelatedness of sentences considered.

STS is an intensively studied problem for years, the great development in this area was accelerated by SemEval challenges [14], [15] etc. In the framework of these challenges, this task was standardized into the following form: given a sentence pair, the task is to assign them a similarity score between 0 and 5, where 5 corresponds with (total) semantic equivalence and 0 corresponds with complete unrelatedness.

Each integer value refers to the following meanings [15]:

- § 5 – identical,
- § 4 – strongly related,
- § 3 – related,
- § 2 – somewhat related,
- § 1 – unrelated,
- § 0 – completely unrelated,

STS has many downstream applications including question answering systems, computer-aided translation (translation memory systems) etc. [16].

NLI and STS both deal with sentence pairs, however, there are substantial differences between these two tasks. Formally, STS is a regression task (in contrast to NLI, which is considered as a classification task). Another difference in the form is "symetry": entailment relation obviously depends on the "direction", whereas in STS the order of the two sentences does not matter.

As an evaluation metric for STS task, a Pearson correlation coefficient is traditionally used.

B. STS Corpus in Czech

Although STS in English is a well resourced problem, the same does not hold for STS in other languages, including Czech. At the time, there exists *only one* STS annotated dataset for Czech introduced in [17]. It contains 1,425 annotated pairs. It was developed upon the English sentence pairs from SemEval challenges (2013–2015) corpora. The sentence pairs were *manually* translated by four Czech native speakers ensuring the high quality of produced final corpus. The original labels were simply transferred (the assumption is the same as in the case of NLI corpora translation). The Czech STS contains two distinct domains/topics: news headlines and image captions. The corpus is split into two parts: TRAIN (925 instances) and TEST (500 instances) having no DEV subset. The structure of the corpus is summarized in Table III.

TABLE III
STRUCTURE OF THE CZECH STS CORPUS [17]

Dataset	Split	No. of Pairs
SemEval 2014–15 Images CZ	TRAIN	550
SemEval 2013–15 Headlines CZ	TRAIN	375
SemEval 2014–15 Images CZ	TEST	300
SemEval 2013–15 Headlines CZ	TEST	200

TABLE IV
EXAMPLE OF CZECH STS CORPUS ITEMS

Sentence 1: *Dva černí psi si hrají na trávě.* (original: *Two black dogs are playing on the grass.*)
Sentence 2: *Dva černí psi si hrají na travnaté planině.* (original: *Two black dogs are playing in a grassy plain.*)
Label: 4.60
Sentence 1: *Skupina čtyř dětí tancujících na dvorku.* (original: *A group of four children dancing in a backyard.*)
Sentence 2: *Skupina dětí se protahuje na barevných podložkách.* (original: *A group of children do stretches on colored mats.*)
Label: 1.60
Sentence 1: *Žena drží dítě, zatímco muž se kouká na jiného muže držícího dětské hodinky.* (original: *A woman holds a baby while a man looks at it as another man holding a child watches.*)
Sentence 2: *Žena stojí v obchodě s rukama venku, zatímco jiná žena drží kameru.* (original: *A woman stands with her arms out in a store while another woman holds a camera.*)
Label: 0.40
Sentence 1: *Žena držící noviny.* (original: *A woman holding a newspaper.*)
Sentence 2: *Muž na kolečkových bruslích na kovové tyči.* (original: *A man rollerblading on a metal bar.*)
Label: 0.00

The corpus is publicly available for download[7].

Again, in order to provide a better overview of the corpus, we also provide several examples taken from the "Image" part of the corpus, see Table IV.

The authors in [17] proposed several approaches to STS in Czech (over their corpus), based on (strong) text preprocessing (stemming/ lemmatization) and feature engineering (n-grams, TF-IDF scores etc.) as well as bag-of-words (BOW) approaches with FASTTEXT embeddings. The authors achieved a Pearson correlation coefficient $\mathbf{0.7887}$ on the TEST set/Image part, using linear regression over feature vectors.

C. Architecture for STS that Uses INFERSENT Encoders

Analogously to NLI, STS (regression) task has inputs in the form of sentence pairs, hence we can exploit similar architectures as in Figure II-E assuming that sentence encodings are already prepared. The difference is obviously in the last layer, since we do not elaborate on classification, but regression. We squeezed the output interval from $[0, 5]$ to "more natural" $[0, 1]$ (without loss of generality, since Pearson correlation coefficient is invariant to linear transformations). The final output is provided by a *sigmoid layer.*

[7]https://github.com/Svobikl/sts-czech

TABLE V
RESULTS ON THE CZECH STS CORPUS

Merging fnc	Train set	TEST	IMG
f_1	TRAIN-FULL	0.7086	0.8046
f_1	TRAIN-IMG	0.6902	0.8170
f_2	TRAIN-FULL	**0.7488**	0.8412
f_2	TRAIN-IMG	0.7409	**0.8511**
f_3	TRAIN-FULL	0.7123	0.8096
f_3	TRAIN-IMG	0.6906	0.8198
f_4	TRAIN-FULL	0.4879	0.5698
f_4	TRAIN-IMG	0.3690	0.4944
f_5	TRAIN-FULL	0.7447	0.8358
f_5	TRAIN-IMG	0.6857	0.8410

We investigated the following settings regarding merging sentence embeddings (corresponding to architecture from Figure II-E):

» $f_1(u, v) = (u, v, |u - v|, u * v)$
» $f_2(u, v) = (|u - v|, u * v)$
» $f_3(u, v) = (u, v, |u - v|$
» $f_4(u, v) = (u, v, u * v)$
» $f_5(u, v) = (|u - v|)$

We performed experiments with architectures described in the previous subsection. Moreover, we also used different subsets of TRAIN and TEST splits of Czech STS – we elaborated on the following scenarios:

» training on the entire TRAIN split, abbr. TRAIN-FULL,
» training only on the "Image part" of the TRAIN split, abbr. TRAIN-IMG.

D. Results

Table V summarizes our results. Training was done in 24 epochs using Adam optimizer [18]. The fully connected layer following the merging layer had 28 units (set using grid search) using *elu* activation.

The evaluation uses Pearson correlation coefficient of predictions and gold labels.

IV. DISCUSSION

We have achieved results comparable to those obtained by feature-based approaches presented in [17]. In case of "Image part" of Czech STS corpus (for both training and test), we strongly outperformed results presented in [17] (0.8511 vs. 0.7887). The reason most likely arises from the fact that sentence encoders were trained on the same domain.

From the results we have also seen that including the separate sentence embeddings that form an input pair does not lead to improvements. The architecture which yields the best results on Image subset used only "merged representations/embeddings" (concatenated vectors $|u - v|$ and $u * v$, where u, v are corresponding embeddings of sentences of the STS task). We can observe that in case of "Image part" of the TEST set, all architectures omitting separate u, v (i.e., using merging functions f_2 and f_5) and hence using only "fusions" of u, v provide better results than all other architectures (using merging with separate sentence embeddings).

Relatively poor results achieved on "Headlines part" of the TEST set (causing lower results on the whole set compared to the "Image part" only) are probably caused with a large amount of out-of-vocabulary words (the vocabulary used in tokenization was derived from Czech MT version of SNLI, i.e., from "image domain" that perfectly fits to "Image part" of Czech STS corpus, but, in contrast, it is not so suitable for news headlines which often contain proper names – surnames, locations' names etc. not covered by the dictionary). Future experiments and datasets augmentations are needed – mainly in the sense of adding labeled data to Czech MT version of SNLI corpus. One of possible (and feasible) approaches is probably a machine translation of MultiNLI corpus that contains more genre-diverse sentence pairs, however, the "methodology" may stay unchanged.

V. CONCLUSION AND FURTHER WORK

In this paper we have introduced a Czech MT version of SNLI corpus and state an INFERSENT (GRU) baseline of the corpus, together with obtaining sentence encoders in Czech. These encoders were directly used in transfer learning approach to semantic textual similarity task in Czech. We achieved notable results on particular "Image captions" dataset (0.8511 in terms of Pearson correlation coefficient).

This work primarily demonstrates the feasibility of this general approach to sentence embeddings available for all target languages, where suitable English-target language MT system / translation API exists. Thanks to simplicity of this process, it can be easily implemented even in cases when only limited computational resources are available.

Further Work

Our presented results indicate that supervised sentence embeddings obtained from NLI task is a promising way of investigations. There are several research questions arising from this initial work, mainly:

1) How does the INFERSENT particular architecture used affects the result in Czech comparing to English?
2) Are there any statistically significant differences in accuracy achieved with same architectures on different languages?

Another direction of further research is extrinsic evaluation of sentence embeddings obtained on different transfer tasks (including tasks like sentiment analysis, CST relations classification [19] etc.) in different languages.

A related issue to this direction of research is investigating the impact of quality of machine translation on the quality of final sentence embeddings obtained.

Sentence embeddings are generally an emerging topic. In contrast to English, where this topic is intensively and deeply studied, the research for other languages is in the beginning. However, there some attempts including Slavic BERT [20]. A solid comparison of our proposed INFERSENT based approach and BERT approach for Czech is also an open issue.

Remark: This position paper contains several results from the author's PhD thesis – submitted after the the FedCSIS deadline, currently under the review.

REFERENCES

[1] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[2] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Jun. 2018, pp. 1112–1122.

[3] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[4] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[5] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[6] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[7] R. Sifa, M. Pielka, R. Ramamurthy, A. Ladi, L. Hillebrand, and C. Bauckhage, "Towards contradiction detection in german: a translation-driven approach," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2019, pp. 2497–2505.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[9] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluation the role of bleu in machine translation research," in *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[12] G. Majumder, P. Pakray, A. Gelbukh, and D. Pinto, "Semantic textual similarity methods, tools, and applications: A survey," *Computación y Sistemas*, vol. 20, no. 4, pp. 647–665, 2016.

[13] Y. Li, D. McLean, Z. A. Bandar, J. D. O'shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE transactions on knowledge and data engineering*, vol. 18, no. 8, pp. 1138–1150, 2006.

[14] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "Semeval-2012 task 6: A pilot on semantic textual similarity," in ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 2012, pp. 385–393.

[15] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, "* sem 2013 shared task: Semantic textual similarity," in *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, 2013, pp. 32–43.

[16] R. Gupta, H. Bechara, and C. Orasan, "Intelligent translation memory matching and retrieval metric exploiting linguistic technology," *Proc. of Translating and the Computer*, vol. 36, pp. 86–89, 2014.

[17] L. Svoboda and T. Brychcín, "Czech dataset for semantic textual similarity," in *International Conference on Text, Speech, and Dialogue*. Springer, 2018, pp. 213–221.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] P. Kędzia, M. Piasecki, and A. Janz, "Graph-based approach to recognizing cst relations in polish texts," in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 2017, pp. 363–371.

[20] M. Arkhipov, M. Trofimova, Y. Kuratov, and A. Sorokin, "Tuning multilingual transformers for language-specific named entity recognition," in *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, 2019, pp. 89–93.