HuLU: Hungarian Language Understanding Benchmark Kit

Noémi Ligeti-Nagy, Gergő Ferenczi, Enikő Héja, László János Laki, Noémi Vadász, Zijian Győző Yang, Tamás Váradi
HUN-REN Hungarian Research Centre for Linguistics
1068 Budapest, Benczúr utca 33.
{surname.firstname}@nytud.hun-ren.hu

Abstract

The paper introduces the Hungarian Language Understanding (HuLU) benchmark, a comprehensive assessment framework designed to evaluate the performance of neural language models on Hungarian language tasks. Inspired by the renowned GLUE and SuperGLUE benchmarks, HuLU aims to address the challenges specific to Hungarian language processing. The benchmark consists of various datasets, each representing different linguistic phenomena and task complexities. Moreover, the paper presents a web service developed for HuLU, offering a user-friendly interface for model evaluation. This platform not only ensures consistent assessment but also fosters transparency by maintaining a leaderboard showcasing model performances. Preliminary evaluations of various LMMs on HuLU datasets indicate that while Hungarian models show promise, there's room for improvement to match the proficiency of English-centric models in their native language.

Keywords: benchmarking, evaluation, less-resourced languages, Hungarian

1. Introduction

Over the past few years, the landscape of neural language models has seen dramatic advancements, most notably with the introduction of models like ChatGPT. These sophisticated models are not only reshaping the way we interact with technology but also challenging traditional benchmarking methodologies. Modern benchmarks have been pivotal for assessing models against diverse tasks, often being expansive corpus collections. The foundational standards set by the English GLUE and SuperGLUE benchmarks (Wang et al., 2018, 2020) have expanded to include benchmarks for languages like French (FLUE, Le et al., 2020), Spanish (GLUES, Cañete et al., 2020), and Russian (Shavrina et al., 2020). Moreover, with XGLUE's focus on multilingual models (Liang et al., 2020), the global drive towards more inclusive language processing is evident. Yet, as models like ChatGPT become more prevalent, the field may need to revisit and adapt its evaluation standards to remain relevant and effective.

Although with a slight delay, the pre-training of the most widely used LLM architectures on Hungarian corpora has begun (Nemeskey, 2021; Feldmann et al., 2021; Yang et al., 2023a,b, 2024a,b). In the future, we can expect even more models trained on Hungarian to emerge, and it will be essential to measure and compare their language comprehension.

While models like ChatGPT, equipped with instructive capabilities, have instigated a paradigm shift within the field of LLMs, the broader evaluation landscape still predominantly relies on traditional methods. As an example, in their recent 2023 paper, Laskar et al. (2023) introduced an approach they term 'Leaderboard-based Evaluation,' incorporating well-established benchmark datasets such as SuperGlue, leading to two significant observations. First, ChatGPT's performance falls short of the state-of-the-art single-task fine-tuned models when evaluated against the SuperGlue dataset. Second, despite ChatGPT's claimed multilingual capabilities, its performance in underrepresented languages remains notably deficient. Both of these observations strongly underscore the necessity for the creation of benchmark datasets for under-resourced languages, such as Hungarian. Moreover, Hungarian's unique linguistic features, including its discourse-configurational nature and highly agglutinative characteristics as described by É. Kiss (1995), may expose additional challenges related to LLMs. With this motivation in mind, we have embarked on the development of our database collection, which we refer to as the Hungarian Language Understanding Evaluation Benchmark Kit (HuLU).

A benchmark database for Hungarian has previously not yet been created. We chose the widely-used GLUE, considered a milestone and a defining benchmark of multi-task nature, and its successor, the SuperGLUE, as our starting points. However, we are also aware of the weaknesses and shortcomings of these databases (see for example Raji et al., 2021; Kiela et al., 2021, among others).

Therefore, our future goal is to further extend the scope of the benchmark datasets by complementing GLUE and SuperGLUE with additional resources. Our ultimate objective is to com-

LREC-COLING 2024, pages 8360–8371
20-25 May, 2024.

pile a consistent and comprehensive benchmark database for Hungarian.

In the case of GLUE, SuperGLUE, and even some benchmark databases compiled for other languages, researchers had the opportunity to select corpora from existing resources. However, when it comes to Hungarian, there are no dedicated corpora available that focus on specific tasks with proper annotation. Therefore, the corpus-building effort presented here serves a dual purpose: i) our goal is to produce several smaller, well-annotated, and task-specific corpora that address language comprehension challenges commonly encountered by language models, and ii) using these, we intend to compile a benchmark database that allows for the assessment and comparison of language model performance.

2. Corpora

As an initial step in introducing our Hungarian benchmark dataset kit, we present seven corpora. These have been selected from the 15 corpora featured in GLUE and SuperGLUE. The selected corpora can be classified into two categories: on one hand, there are databases that can be generated through translation, as the tasks they focus on are not tied to a specific language, and the dataset can be readily translated. On the other hand, there are datasets that resist translation due to either their language-specific attributes or the intricate nature of the texts within the corpus. In the following sections, we will present: i) our corpora created through machine translation of English corpora, coupled with translation verification and additional annotation efforts, and ii) our corpora derived from original Hungarian texts, crafted with the assistance of annotators.

2.1. Translated Benchmark Corpora

2.1.1. Machine translation methodology

To facilitate machine translation, we constructed an English-Hungarian parallel corpus sourced from the OPUS corpus collection (Tiedemann, 2012). The sub-corpora utilized in this process comprises ParaCrawl, OpenSubtitles, Tatoeba, WikiMatrix, EUbookshop, the PHP manual, TED2020, KDEdoc, and KDE4. On the basis of this parallel corpus, we built a neural translation system employing a transformer encoder-decoder architecture using the Marian NMT framework (Junczys-Dowmunt et al., 2018). The trained model parameters include: 6 encoder layers and 6 decoder layers; 16 attention heads; word embeddings with a dimension of 1,024; an input length of 1,024; and a feed-forward network size of 4,096.

2.1.2. HuCoPA

CoPA (Choice of Plausible Alternatives, Roemmele et al., 2011) focuses on cause-and-effect relationships, comprising 1,000 questions. In each question, given a premise one has to select the more likely alternative that stands in a cause-and-effect relationship with the premise. Some questions require choosing the cause from the available options, while others involve selecting the effect. This dataset is an integral part of the SuperGLUE collection.

Machine translation was used to translate the 1,000 questions of CoPA. Subsequently, our annotators thoroughly reviewed and improved the machine-generated translations to ensure fluency. Following this, an annotator assigned the correct answer for each question. In cases where there was a discrepancy between the annotator's choice and the original label, we conducted a manual review of the specific instance. Interestingly, an incorrect label in the original CoPA was identified during this third step, specifically the question with ID 380 in the training set.[1] Consequently, the HuCoPA corpus was created, comprising 1,000 units.[2] To maintain the original distribution, 400 instances were allocated for training, 100 instances for validation and 500 instances for testing. Each sentence in the test set was subjected to evaluation by two different annotators to ensure the accuracy of the test set. The annotator agreement for the test set reached a high level of 0.95 in terms of Cohen's κ.

2.1.3. HuSST

The SST (Stanford Sentiment Treebank, Socher et al., 2013) is one of the most renowned English corpora containing sentiment annotations. For the compilation of the corpus, 10,662 sentences were collected from the Rotten Tomatoes website. The

[1] Premise: *The woman spotted her friend from across the room. What was the CAUSE of this?* 1st alternative: *The woman waved.* 2nd alternative: *The woman escaped.* The correct answer: 1. If we are looking for the cause in the premise and the 1st alternative is the correct answer, it is only possible if that particular *friend* is a woman. However, even then, it remains convoluted, and the sentence *The woman waved, therefore the woman spotted her friend...* doesn't make sense, because the reference of *the woman* differs in the two instances. It's more likely that the label is incorrect, as when looking for an effect instead of a cause, the sentence *The woman spotted her friend... therefore the woman waved* immediately makes sense. In this case, we replaced the original label; in a few other instances, we had to refine the translation to preserve the original cause-and-effect relationship.

[2] https://github.com/nytud/HuCoPA; the corpus is also available on Huggingface: https://huggingface.co/datasets/NYTK/HuCoPA

sentences were parsed using the Stanford Parser, and the resulting 215,154 phrases were individually annotated on a 25-point emotional scale[3]. In the version called SST-5, the 25-point scale was converted to a range from 0 to 4, and in SST-2, it was made binary. As part of GLUE, the authors include the SST-2, and notably, only the complete sentences, excluding phrase-level elements. From the GLUE database, one can thus download more than 70,000 sentences and their corresponding labels.

When we tackled the translation of the sentiment corpus, we departed from its original format in GLUE and opted for the so-called SST-5 dataset. As we acquired the SST-2 data from GLUE's subcorpora, we noticed a significant difference from the presentation in the GLUE paper. While the authors of the GLUE paper emphasized the exclusive use of complete sentences and framed the task as sentence classification,[4] we encountered a multitude of phrases within the files (some examples from the training set: *of saucy*, *in world cinema*, *a doa*). This variance explains the disparity in sentence count, with the SST presentation citing 10,662 sentences and the GLUE corpus containing 70,600 sentences.

It's worth noting that we identified 11,855 sentences that we translated into Hungarian using machine translation. Subsequently, we subjected these sentences to a series of verification steps, mirroring the process used for the HuCoPA corpus. Eventually, each Hungarian sentence was labeled by three annotators based on sentiment on a three-point scale. The sentiment labels were reviewed by a curator who provided the final labels for the sentences.[5] For 7,064 sentences (59.6%), there was complete agreement among the three annotators, while in 4,619 cases (38.96%) only two of the annotators agreed on the label. The final label in all cases was the curator's decision.[6] We do not use 172 sentences in the database, as they received three different labels from the three annotators. The dataset contains 11,683 sentences. The train, validation and test sets contain 9,347, 1,168 and 1,168 sentences, respectively.

[3]Annotators used a slider with 25 distinct positions, starting at a neutral point. For a practical illustration, please refer to Figure 3 in Socher et al. (2013), the paper that introduces the SST dataset.

[4]"We use the two-way (positive/negative) class split, and use only sentence-level labels.", (Wang et al., 2018, 3).

[5]Translation verification was done by 12 annotators, fluency improvement by 8. 11 took part in sentiment annotation, and 4 were involved in the curatorial task.

[6]https://github.com/nytud/HuSST, and https://huggingface.co/datasets/NYTK/HuSST.

2.1.4. HuRTE

A subset of the datasets from the Recognizing Textual Entailment (RTE) challenge was incorporated into GLUE: samples from RTE1 (Dagan et al., 2006), RTE2 (Bar-Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009) were collected, which originated from news texts and Wikipedia articles. In these, one has to determine whether a (sometimes multi-sentential) premise entails a single-sentence hypothesis or not. The task involves binary labeling, so for examples originally labeled with three classes, the *neutral* and *contradiction* labels were combined for the sake of consistency.

The portion of the RTE datasets selected for inclusion in GLUE was translated into Hungarian using the same machine translation system we used for the creation of the other corpora in HuLU. Subsequently, the 5,797 examples produced (representing roughly 18,000 sentences) were checked and corrected by annotators, aiming for fluency.[7]

As the correct labels were not provided for the test set in the original benchmark, we had to have their translations labeled. Thus, out of the 5,797 examples, 3,000 (the test set examples) were labeled by three annotators each. The agreement among them was 0.61 (Fleiss's κ). Only the examples where there was complete agreement among the annotators (2,123 examples) were included in the HuRTE test set.[8] With this step, we aim to ensure the purity of the test material, measuring the knowledge of language models on examples where we can judge the answer with greater certainty as correct or incorrect. Although we cannot discuss the methodology of evaluating language models within the scope of this study, for the complexity of the issue, the challenges of inferential tasks, and potential new evaluation methods, see e.g., Baan et al. (2022); Plank (2022).

After the translation and fluency check, the English examples available with their original labels were labeled by individual annotators. This revealed cases where the original inferential relationship between the sentence pairs was lost during the translation. Among these 2,797 examples, there's a conflict between the label from the original English database and the label obtained post-translation from a native Hungarian annotator in 357 cases from the initial training set and 35 from the validation set. We currently exclude these 392 examples from the HuRTE data. Thus, the HuRTE

[7]14 annotators conducted translation checks, and 7 performed fluency checks. 13 annotators labeled the test set examples, and 4 labeled the examples with original labels.

[8]Bentivogli et al. (2009) also used this method when compiling the RTE5 database: they retained only the examples where all three annotators agreed on the label.

corpus is split into training, validation, and test sets with 2,131, 242, and 2,123 examples respectively.[9], [10]

2.1.5. HuWNLI

The Winograd Schema Challenge (WSC, proposed by Levesque et al., 2012) requires the resolution of anaphora with the help of world knowledge and commonsense reasoning. In the original dataset of WNLI Winograd schemas has been transformed to an inference dataset to make them suitable to be training data for neural models. The original Winograd schemas were transformed to a classification task in which the model has to predict if the second sentence (the one with the substituted pronoun) is entailed by the first sentence (thus the labels are *entailment* and *not-entailment*, as in Example 1).

(1) a. What about the time you cut up tulip bulbs in the hamburgers because you thought they were onions?

b. You thought hamburgers were onions?

L. not-entailment

The schemas have already been translated into other languages, including Japanese, French, Portuguese, Chinese, Russian and Hebrew. For Hungarian, (Vadász and Ligeti-Nagy, 2022) first translated the English original, and then two annotators validated the output. They discarded certain schemas because were not able to translate them in a manner that retained the characteristics of the Winograd schemas. For example, *Lily spoke to Donna, breaking her (silence/concentration)*: the two English expressions cannot be translated into Hungarian in a way that only one word differs between the two sentences but still retains the possessive structure in both. In other cases, they adapted the original schema to Hungarian with slight modifications. The process and the result have been presented in Vadász and Ligeti-Nagy (2022).

The current version of HuLU includes the HuWNLI database, derived from Winograd schemas, as a corpus for testing coreference resolution, wherein anaphora resolution is formulated as an inference task. We created the NLI format that is part of HuLU by replacing the ambiguous pronoun in the schemas with every possible referent (the method is described in the study introducing GLUE, see Wang et al., 2018). We expanded the set of sentence pairs derived from the schemas by translating those sentence pairs which, together with the sentences from the Winograd schemas, make up the GLUE WNLI database. We release the corpus in three parts (with the number of elements for each set in parentheses): training set (562), validation set (59), and test set (134). The divisions follow the GLUE WNLI divisions but contain fewer examples since we had to discard several sentence pairs that couldn't be translated into Hungarian. The sentence pairs in the test set are all translated examples from the GLUE WNLI test set.

2.2. Locally Sourced Hungarian Benchmark Corpora

2.2.1. HuCOLA

The CoLA (Corpus of Linguistic Acceptability, Warstadt et al., 2018) contains 10,657 English sentences collected from linguistic literature. Included in GLUE, its purpose is to assess whether a model can determine the grammaticality of a sentence, making it a pivotal task. The binary labels indicate the acceptability of the sentence. The original labels given by the author were also compared with the judgments of human annotators.

To create the Hungarian CoLA corpus, we collected 9,944 examples from four major, comprehensive linguistic articles (Kiefer, 2015; Alberti and Laczkó, 2017a,b; É. Kiss and Hegedűs, 2021). The collection was guided by the following criteria:

- We extracted every example sentence from the articles, irrespective of the acceptability judgment given by the author.
- For examples of the type *Megnézzük (*a) Budapest hídjait.* ('We look at (*the) bridges of Budapest.'), we made two entries: *Megnézzük Budapest hídjait.* ('We look at the bridges of Budapest.') and **Megnézzük a Budapest hídjait.* ('*We look at the bridges of the Budapest.').
- If a sentence was deemed unacceptable because it couldn't convey a given meaning, we didn't collect it, e.g., **Megver Péter* which would mean 'Péter beats Péter' (Kiefer, 2015, 49.).
- Sentences containing nonsensical words were excluded.
- Sentences that were incorrect due to the position of the focus[11] were not collected.

[9] The surprisingly large proportion of the test set compared to the training material can be observed in several databases of both GLUE and SuperGLUE. For now, we determine the sizes of the training, validation, and test sets following these patterns.

[10] The HuRTE corpus is available on GitHub, https://github.com/nytud/HuRTE.

[11] A syntactic position in Hungarian marking a specific discourse function.

- We did not gather sentences that violate prescriptive rules (e.g., we do not start a sentence with *Hát*, which would be akin to starting an English sentence with 'Well,').

During the described collection process, both complete sentences and incomplete constructs, like phrases and clauses, were included. Since the targeted task is sentence classification, we expanded the incomplete sentence examples into full sentences.[12]

In the English corpus, after collection, the authors filtered the corpus to the 100,000 most common English words, replacing less frequent words. We did not apply such a filter to our corpus, as *subword* based tokenization makes this unnecessary for modern language models.

Every single sentence was labeled by four annotators[13]. Based on the guidelines, they had to determine whether the given sentence was acceptable and sounded like a proper Hungarian sentence.

During the collection, the sentences were also labeled based on the linguistic phenomena found within them.

While in the CoLA's English predecessor the sentence labels were those originally determined by the linguistic authors, we excluded those labels of our sentences from the analysis. This ensured that labels mistakenly recorded during collection or typographical errors did not affect data quality. In 69.2% of the sentences (6,883 sentences), all four annotators assigned the same label. In 22.2% (2,213 sentences), the sentences were labeled in a 3:1 ratio. Sentences annotated in a 2:2 ratio (8.5%, 848 sentences) were set aside and do not form part of the database. However, we make them available as they represent valuable linguistic research material.

The final label of the sentences in the case of 3:1 ratio annotation was determined based on the majority decision. Following the ratios found in GLUE, we release the data divided into training, validation, and test sets at a ratio of 80-10-10%.[14]

2.2.2. HuRC

We created the Hungarian HuRC corpus based on the English-language ReCoRD. Zhang et al. (2018) automatically compiled the ReCoRD: they extracted more than 120,000 examples from the CNN/Daily News[15] corpus. The daily news was divided into several parts (see Figure 1, left example): main text (*passage*), question – the proper name masked out in the last paragraph (*cloze-style query*), and reference answer (*reference answer*). The main text consists of the first few paragraphs of the article. In the last paragraph of the article, which serves as a kind of concluding passage, there must be a proper name that also appears in the main text. This proper name is the reference answer. In the actual reading comprehension task, this proper name is masked out, and the model must select the correct reference answer from a list.

For producing the Hungarian material, we based it on daily articles from Népszabadság Online[16], particularly the 396,886 articles that had a title, text, and summary (*lead*). If any component was missing from an article, we didn't use it. Then, we selected articles consisting of 3-6 paragraphs. A crucial criterion was that both the main text and the question (the last paragraph) contained proper names.

For proper name recognition, we trained our named entity recognition model with the help of huBERT (Nemeskey, 2021). For fine-tuning the NER model, we used the official training-validation-test datasets of the NYTK-NerKor corpus (Simon and Vadász, 2021) and the token-level classification library provided by Huggingface.[17] Our NER model achieved a 90.18 F-score on the test set.

In the final step, we looked for pairs of proper names that appeared in both the main text and the question. Several proper name pairs could occur in one article. In our example (see Figure 1, right example), besides *Presser Gábor*, *Tamás* also appears in both the question and the main text. In such cases, we included the same article multiple times in the database, with different proper name pairs. In total, 49,782 different types of articles (*type*) were selected, from which a total of 88,655 instances make up our dataset due to the multiple proper name pair phenomenon. The quantitative properties of our corpus produced with automatic methods are as follows: Number of articles: 88,655, different types of articles: 49,782, tokens: 27,703,631, types: 1,115,260, average length of text part (tokens): 249.42 (median: 229), average length of question (tokens): 63.07 (median: 56).

The resulting corrected dataset was verified in 100-unit batches by individual annotators. For the

[12] For the principles of sentence completion, see the annotation guidelines: https://github.com/nytud/HuCOLA.

[13] For this task, we chose annotators who did not have advanced linguistic knowledge, who weren't studying or did not graduate in linguistics. In total, 12 annotators worked on the corpus.

[14] https://github.com/nytud/HuCOLA, and the corpus is also available on Huggingface, the dataset card link: https://huggingface.co/datasets/NYTK/HuCOLA.

[15] https://github.com/abisee/cnn-dailymail

[16] http://nol.hu

[17] https://github.com/huggingface/transformers/tree/master/examples/pytorch/token-classification

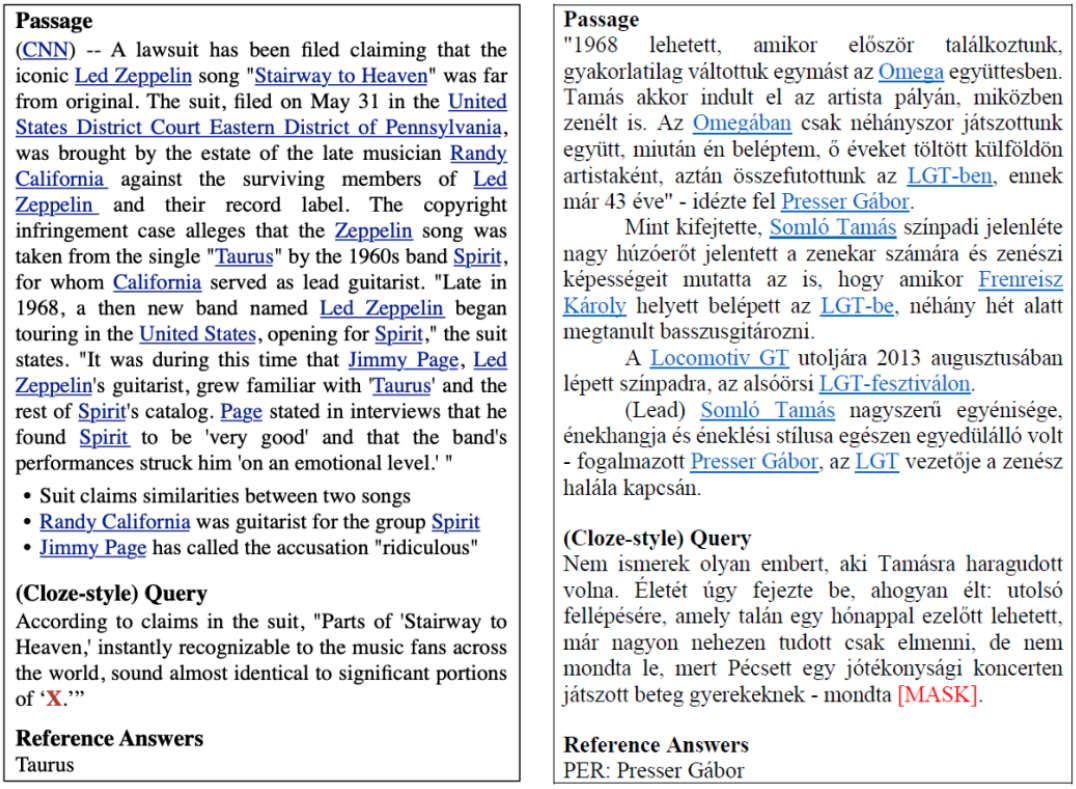

Figure 1: An example from ReCoRD (Zhang et al., 2018) and a HuRC example

annotation, we provided an annotating interface developed by us. The automatic masking had to be validated based on the following criteria: i) whether the named entity recognition and masking is correct (i.e., *Ferenc pápa* 'Pope Francis' was masked, not just *Ferenc* 'Francis', and *Gödöllőre* 'to Gödöllő' was masked as [MASK] and not [MASK]re '[MASK].to'), and ii) whether the masked named entity also appears in earlier parts of the article.[18] As a result of the verification, the database contains 80,587 automatically generated, manually validated text units.[19]

2.2.3. HuCommitmentBank

The CB (CommitmentBank, de Marneffe et al., 2019) consists of short text segments, each containing at least one sentence with a subordinate clause. Each subordinate clause is labeled according to the degree of commitment the writer of the text has towards the truth of the clause. In SuperGLUE, the task was transformed into a three-class inference task: the premise is the entire text segment, and the hypothesis is the embedded clause.

A key characteristic of the examples in the corpus is that every examined embedded sentence is syntactically under a logical entailment canceling operator. Question, modal, negation, antecedent of conditional are considered logical entailment canceling operators.

To create the Hungarian-language corpus, we searched for sentences using the Hungarian equivalents of 30 common matrix verbs/expressions from the English examples. The sources of the texts were, on the one hand, the spoken language sub-corpus of the Hungarian Gigaword Corpus (MNSZ2, Oravecz et al., 2014), on the other hand, texts from a few works of literary sub-corpus, and texts from online forum comments. The examples extracted in this way first had to be checked for the presence of entailment canceling operators. 4 annotators collected a total of 1,100 valid text segments. The collected texts were also validated among the annotators. The 1,100 examples were labeled by 5-5 annotators on a 7-point Likert scale (between -3 and 3, where 0 indicated that the speaker does not know whether the subordinate clause is true or false). Annotation was done using the LimeSurvey interface. A total of 9 native Hungarian annotators worked on the corpus. Their task was to read, understand, and judge 50 examples per hour. To facilitate their work, all 1,100 examples were also provided with a question, the answer to which assists in their decision. For a more detailed presentation of the creation of the Hungarian CB, see Hatvani (2022).

[18] A total of 12 annotators worked on the corpus.

[19] https://github.com/nytud/HuRC and https://huggingface.co/datasets/NYTK/HuRC

The examples in the corpus are structured as follows: the *context* refers to the 1-2 sentences preceding the target sentence; the *target sentence* is the sentence containing the entailment canceling operator and the subordinate clause; for each example, we specifically highlight the *verb* and the *subordinate clause*; the *marker* indicates the type of entailment canceling operator present in the given example. There are a total of 10 types of markers observed in the corpus: modal, conditional, question, negation, modal negation, modal question, modal conditional, conditional question, rhetorical question, and negated question.
In SuperGLUE, only a subset of CB was included where there was at least 80% agreement between the annotators.[20] Thus, out of the original 1,200 examples, they selected a training set with 250 examples, a validation set with 57 examples, and a test set with 250 examples. The task was formulated as a three-class inference task.
Similar to SuperGLUE, this corpus appears in HuLU as an inference task, but the original 7-point label was replaced with a three-class categorization: the original labels of -1, 0, and 1 were condensed into the *neutral* category, -3 and -2 into the *contradiction* category, and 2 and 3 into the *entailment* category.
We measured the inter-annotator agreement (IAA) on the corpus in multiple ways (considering the Likert scale as an interval scale, for a debate on this see e.g., Wu and Leung, 2017). The Krippendorf's α was 60.5%. The standard deviation (SD) was 1.01. We only included examples in HuLU where $SD < 1$. Thus, we eventually created a training and test set, each with 250 examples, and a validation set with 103 examples from a total of 603 examples.[21] For the test set examples, $SD < 0.5$.

3. HuLU Web Service

After detailing the corpora, it's important to note that we also offer an additional resource for users, a web service: https://hulu.nytud.hu. It's designed to facilitate the quick and convenient evaluation of language models in a standardized manner and the publication of results. This ensures that anyone can effortlessly view the performance of individual models.
The web service, available in both Hungarian and English, mirrors the appearance and functionality of the GLUE and SuperGLUE interfaces. It's crafted to support users throughout the entire model development process. This encompasses everything from introductory explanations, granting access to publications, elucidating benchmark tasks, to making users familiar with practical steps. Directly from the site, users can download benchmark corpora (currently: HuCOLA, HuCommitmentBank, HuCoPA, HuRTE, HuSST, HuWNLI). These are partitioned into the conventional training, validation, and test sets. We uphold the confidentiality of the labels for the test material; they are only accessible to the evaluation module running on the HuLU server.
Subsequent to uploading a test set labeled in the designated format, the web service assesses the results based on the following metrics:

- HuCOLA, HuCoPA, HuRTE: Matthew's Correlation Coefficient (MCC). Despite GLUE/SuperGLUE computing absolute accuracy for the latter two, we opted for MCC due to the uneven distribution of labels in the test material.
- HuSST, HuWNLI: Absolute accuracy, paralleling the model of GLUE/SuperGLUE.
- HuCommitmentBank: Weighted F1 score, an optimal metric for managing classes of varying sizes in multi-class classification.

The results of the evaluation are communicated to the uploading user, but they do not automatically become public on the website. Approval from our staff is required for this.
The results that have already been made public are displayed in various ways: the performance of the models can be viewed chronologically on a graph and also in a table sorted by performance.
The HuLU web service allows for the evaluation of models on individual tasks at any time and in any order, whereas on the GLUE and SuperGLUE sites, models can only be evaluated across all tasks simultaneously.

3.1. Implementation of the web service

The website runs on servers in Docker containers, on the Portainer platform. Both the evaluation module and the web backend were implemented in Python, on top of the Django framework. For the appealing display of results, we used the Streamlit[22] system, which in itself is suitable for the publication of dynamic data. However, as in our case, it can also be embedded into a website.
The expected format for the test material to be uploaded for evaluation matches the format of the downloadable training and test datasets. Each must be uploaded in *json* format, containing the current entity identifier, as well as a result field with

[20] It is not clear how agreement was measured on the corpus.
[21] The HuCommitmentBank corpus is available on GitHub, https://github.com/nytud/HuCommitmentBank.
[22] https://streamlit.io

Database Name	Sentences / Instances	Source	Task type / Label proportion
HuCoPA	1,000	Translated	Bin. class. `cause:effect` 50%-50%
HuRTE	5,797	Translated	Bin. class. 51%-49%
HuSST	11,683	Translated	Three-class class. 32%-34%-32%, resp.
HuWNLI	755	Translated	Bin. class. 46%-54%
HuCOLA	9,944	Manually Created	Bin. class. 78%-22%
HuCommitmentBank	603	Manually Created	Three-class class. 45%-27%-27%, resp.
HuRC	88,655	Manually Created	Multiple choice

Table 1: Summary statistics for the Hungarian benchmark dataset kit, including task type or label proportion for each dataset. Datasets are grouped by source and listed alphabetically within each group. The 'Task type / Label proportion' column specifies the classification challenge associated with each dataset: 'Bin. class.' indicates datasets used for binary classification tasks, 'Three-class class.' refers to datasets involving three-class classification challenges, and 'Multiple choice' describes tasks where multiple options are provided for each query.

the same name as the test material. The evaluation program checks the expected format for every upload and verifies that predictions have been received for each element. Consistent with the entire system, the evaluation system was also built modularly, enabling easy integration of new metrics and addition of new benchmark tests to the system.

4. Submitted results

Currently, 8 different models have been evaluated on various HuLU tasks from four different research efforts. Yang et al. (2023a) have trained various language models for the Hungarian language. A GPT-3 model (PULI GPT-3SX, with 6.7 billion parameters), a GPT-2 (PULI GPT-2) and a BERT-Large (PULI BERT-Large) model were pretrained and evaluated (see the first block in Table 2). They also compared their results with fine-tuned huBERT and XLM-RoBERTa base (Conneau et al., 2020) (XLM-R) models. Yang and Ligeti-Nagy (2023) have explored the effectiveness of prompt programming in the fine-tuning process of a Hungarian language model. In this research the prompting method were employed to enhance the fine-tuning performance of the huBERT (Nemeskey, 2021) model on several benchmark datasets of HuLU. As can be seen in the second section of Table 2, they achieved state-of-the-art results using this method. Yang et al. (2023b) have announced a trilingual (Hungarian-English-Chinese) GPT-3 large languge model (PULI GPTrio, with 7.67 billion paramteres), furthermore using this model and the Stanford Alpaca corpus (Taori et al., 2023), the first GPT-3 model designed to follow instructions was introduced for the Hungarian language (Instruct PULI GPTrio). In this research, the pre-trained PULI GPTrio was compared with PULI GPT-3SX in few-shot experiments. Additionally, in zero-shot experiments, the Instruct PULI GPTrio model was compered with the ChatGPT (Ouyang et al., 2022) and text-davinci-001 (Brown et al., 2020) models. In their evaluation tasks, HuCOLA, HuSST and HuRTE benchmarks were used. In their paper (Yang et al., 2023b), they used accuracy and balanced accuracy metrics for the evaluation. However, in the current table (see third and fourth block of Table 2), the MCC metric was used for HuCOLA and HuRTE. All the results that shown in Table 2 are submissions from the the HuLU Web Service (see Section 3).

The numbers in Table 2 are the initial results on the HuLU datasets. In fact, it represents a new research direction to achieve results similar to the English SOTA with Hungarian models, as each model and task requires different hyperparameter settings. For instance, a larger model may require a lower learning rate, but the batch size can also influence performance. HuSST and HuCOLA can be tackled with a simple sentence-level classification method. HuCoPA is a multiple-choice task. It's advisable to approach the HuRTE and HuWNLI corpora with a solution similar to entity-oriented sentiment analysis, where we aim to understand the relationship between the two examined sentences.

For different tasks, in addition to hyperparameters, it's worth experimenting with different input data structures, known as prompts (Shin et al., 2020), which also can affect performance.

	HuCOLA (MCC)	HuCoPA (MCC)	HuRTE (MCC)	HuSST (Accuracy)	HuWNLI (Accuracy)
huBERT	70.9	56.1	48.7	79.4	64.9
PULI GPT-2	49.9	-	41.5	72.8	61.9
PULI BERT-Large	**71.1**	41.4	51.7	**79.9**	65.7
XLM-R	55.9	3.2	33.3	66.1	63.4
huBERT prompt	-	**56.4**	53.4	-	**85.8**
PULI GPT-3SX few-shot	8.2	-	5.6	64.3	-
PULI GPTrio few-shot	6.6	-	9.1	61.6	-
Instruct PULI GPTrio zero-shot	4.7	-	1.7	64.3	-
ChatGPT zero-shot	27.7	-	70.2	71.8	-
text-davinci-001 zero-shot	31.6	-	**79.8**	52.8	-

Table 2: Performance of several Hungarian language models on various HuLU corpora. Columns represent individual corpora and their respective evaluation metrics, while rows represent the models. The results are from (Yang et al., 2023a; Yang and Ligeti-Nagy, 2023; Yang et al., 2023b; Brown et al., 2020)

5. Conclusion

Benchmark datasets and collections serve to adequately measure and compare the performance of neural language models, which consistently outperform earlier rule-based or traditional statistical models in various language technology tasks. In this article, we presented the datasets that comprise HuLU. These datasets serve to evaluate language models that have been exposed to Hungarian texts and thus have proficiency in Hungarian. We showcased the web service developed for HuLU, which allows easy evaluation of model results even on just a single database. The comparability of results is enhanced by a leaderboard. Our evaluations on the HuLU corpora indicate that there's still significant potential for refining Hungarian language models to achieve proficiency levels in Hungarian that are comparable to their performance in English.

6. Bibliographical References

Gábor Alberti and Tibor Laczkó, editors. 2017a. *Syntax of Hungarian. Nouns and Noun Phrases, Volume 1*. Comprehensive Grammar Resources. Amsterdam University Press.

Gábor Alberti and Tibor Laczkó, editors. 2017b. *Syntax of Hungarian. Nouns and Noun Phrases, Volume 2*. Comprehensive Grammar Resources. Amsterdam University Press.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernández. 2022. Stop Measuring Calibration When Humans Disagree. *CoRR*, abs/2210.16133.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

K. É. Kiss. 1995. *Discourse Configurational Languages*. Discourse Configurational Languages. Oxford University Press.

Katalin É. Kiss and Veronika Hegedűs, editors. 2021. *Syntax of Hungarian. Postpositions and Postpositional Phrases*. Comprehensive Grammar Resources. Amsterdam University Press.

Ádám Feldmann, Róbert Hajdu, Balázs Indig, Bálint Sass, Márton Makrai, Iván Mittelholcz, Dávid Halász, Zijian Győző Yang, and Tamás Váradi. 2021. HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben [HILBERT: Training a Hungarian BERT-Large Model in a Cloud Environment]. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 29–36, Szeged, Magyarország. Szegedi Tudományegyetem, Informatikai Intézet.

Péter Hatvani. 2022. A Corpus to Investigate Projection Methods: The Hungarian Commitment Bank. Master's thesis, Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar, Angol-Amerikai Intézet, Elméleti Nyelvészet Tanszék.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Ferenc Kiefer, editor. 2015. *Strukturális magyar nyelvtan 1. Mondattan. [Structural grammar of Hungarian 1. Syntax]*. Akadémiai Kiadó, Budapest.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking Benchmarking in NLP.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Dávid Márk Nemeskey. 2021. Introducing huBERT. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 3–14, Szeged, Magyarország. Szegedi Tudományegyetem, Informatikai Intézet.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the Everything in the Whole Wide World Benchmark.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Noémi Vadász and Noémi Ligeti-Nagy. 2022. Winograd schemata and other datasets for anaphora resolution in Hungarian. *Acta Linguistica Academica*, 69(4).

Huiping Wu and Shing-On Leung. 2017. Can Likert Scales be Treated as Interval Scales?—A Simulation Study. *Journal of Social Service Research*, 43(4):527–532.

Zijian Győző Yang, Réka Dodé, Gergő Ferenczi, Enikő Héja, Ádám Kőrös, László János Laki, Noémi Ligeti-Nagy, Kinga Jelencsik-Mátyus, Noémi Vadász, and Tamás Váradi. 2023a. Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre [The Big ones are Coming! BERT-Large, GPT-2 and GPT-3 Language Models for Hungarian]. In *XIX.*

Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Hungary. Szegedi Tudományegyetem.

Zijian Győző Yang, Réka Dodé, Enikő Héja, László János Laki, Noémi Ligeti-Nagy, Gábor Madarász, and Tamás Váradi. 2024a. ParancsPULI: Az utasításkövető PULI-modell [ParancsPULI: The instruction-following PULI model]. In *XX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 61–72, Szeged, Magyarország. Szegedi Tudományegyetem.

Zijian Győző Yang, László János Laki, Tamás Váradi, and Gábor Prószéky. 2023b. Mono- and multilingual GPT-3 models for Hungarian. In *Text, Speech, and Dialogue*, Lecture Notes in Computer Science, pages 94–104, Plzeň, Czech Republic. Springer Nature Switzerland.

Zijian Győző Yang and Noémi Ligeti-Nagy. 2023. Improve performance of fine-tuning language models with prompting. *Infocommunications Journal, Special Issue on Applied Informatics*, pages 62–68.

Zijian Győző Yang, Szilárd Szlávik, and Noémi Ligeti-Nagy. 2024b. Magyar nyelvű utasításkövető korpusz építése Stanford Alpaca promptok fordításával és lokalizálásával. In *XX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 243–255, Szeged, Magyarország. Szegedi Tudományegyetem.

7. Language Resource References

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The Second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment, Venice, Italy*, pages 1–9.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The Fifth PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the TAC Workshop*.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The Third PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, page 1–9.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A New Benchmark Dataset for Cross-lingual Pretraining, Understanding and Generation. *arXiv*, abs/2004.01401.

Csaba Oravecz, Tamás Váradi, and Bálint Sass. 2014. The Hungarian Gigaword Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Melissa Roemmele, Cosmin Bejan, and Andrew Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. *AAAI Spring Symposium - Technical Report*.

Tatiana Shavrina, Alena Fenogenova, Anton Emelyanov, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov,

Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. *arXiv preprint arXiv:2010.15925*.

Eszter Simon and Noémi Vadász. 2021. Introducing NYTK-NerKor, A Gold Standard Hungarian Named Entity Annotated Corpus. In *Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings*, volume 12848 of *Lecture Notes in Computer Science*, pages 222–234. Springer.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural Network Acceptability Judgments. *arXiv preprint arXiv:1805.12471*.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension.