PuzzLing Machines: A Challenge on Learning From Small Data

Gözde Gül Şahin[†], **Yova Kementchedjhieva**[α], **Phillip Rust**[†], **Iryna Gurevych**[†]
[†]Ubiquitous Knowledge Processing Lab (UKP),
Department of Computer Science, Technical University of Darmstadt
[α]Department of Computer Science, University of Copenhagen
[†]www.ukp.tu-darmstadt.de

Abstract

Deep neural models have repeatedly proved excellent at memorizing surface patterns from large datasets for various ML and NLP benchmarks. They struggle to achieve human-like thinking, however, because they lack the skill of iterative reasoning upon knowledge. To expose this problem in a new light, we introduce a challenge on learning from small data, *PuzzLing Machines*, which consists of *Rosetta Stone* puzzles from Linguistic Olympiads for high school students. These puzzles are carefully designed to contain only the *minimal* amount of parallel text necessary to deduce the form of unseen expressions. Solving them does not require external information (e.g., knowledge bases, visual signals) or linguistic expertise, but meta-linguistic awareness and deductive skills. Our challenge contains around 100 puzzles covering a wide range of linguistic phenomena from 81 languages. We show that both simple statistical algorithms and state-of-the-art deep neural models perform inadequately on this challenge, as expected. We hope that this benchmark, available at https://ukplab.github.io/PuzzLing-Machines/, inspires further efforts towards a new paradigm in NLP—one that is grounded in human-like reasoning and understanding.

1 Introduction

Kahneman (2011) discusses the two modes of human thinking which perfectly encapsulate the current (so called System1) and the desired state (System1+System2) of the deep learning field. System1 handles tasks that humans consider fast, intuitive and automatic, such as object detection and document classification. Recent deep learning (DL) models have shown great promise at this type of tasks—thanks to large training datasets. Yet, it is through slow, rational and sequential mechanisms that human-like abstract reasoning happens, to enable learning from just a few examples. This System2-style modeling is still in its early stages in DL, but is recognized as a much needed next step in the field (McClelland et al., 2019; Marcus, 2020; LeCun, 2020; Bengio, 2020). To foster research in this promising direction, we propose a unique challenge on "learning from small data": *PuzzLing Machines*, based on the Linguistic Olympiads—one of the 13 recognized International Science Olympiads targeted at high-school students.

Chikasaw	**English**
1. Ofi'at kowi'ã lhiyohli.	The dog chases the cat.
2. Kowi'at ofi'ã lhiyohli.	The cat chases the dog.
3. Ofi'at shoha.	The dog stinks.
4. Ihooat hattakã hollo.	The woman loves the man.
5. Lhiyohlili.	I chase her/him.
6. Salhiyohli.	She/he chases me.
7. Hilha.	She/he dances.
Now you can translate the following into Chickasaw:	
	The man loves the woman.
	The cat stinks.
	I love her/him.
Translate the following into English:	
Ihooat sahollo.	
Ofi'at hilha.	
Kowi'ã lhiyohlili.	

Table 1: The "Chickasaw" puzzle (Payne, 2005)

The *PuzzLing Machines* challenge is based on one of the most common puzzle types in the Linguistic Olympiads: the *Rosetta Stone* puzzles (Bozhanov and Derzhanski, 2013), a.k.a. translation puzzles. An example is given in Table 1.[1] Although these puzzles take the form of a traditional "machine translation" task, they are different in many ways: Rosetta Stone puzzles contain a minimal, carefully designed set of parallel expressions (words, phrases or sentences) in a for-

[1]Copyright University of Oregon, Department of Linguistics.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1241–1254
July 5 - 10, 2020.

eign and in a familiar language (e.g., Chickasaw-English). This minimal set is *just* enough to deduce the underlying translation model, which typically involves deriving mini-grammar rules, extracting a lexicon, and discovering morphological and phonological rules. The actual task then is to translate new expressions—generally in both directions—using the model deduced from the parallel data. The assignments are carefully designed so that the expressions cannot be generated through simple analogy, but rather through the application of the discovered rules. These properties distinguish the *PuzzLing Machines* challenge from the modern MT task, as it relies on deductive reasoning with linguistic concepts that are central to System2, rather than exploiting statistical properties from large datasets as in System1.

The lack of reasoning skills of statistical systems has recently gained a lot of attention. Various datasets that require a wide range of background knowledge and different types of reasoning abilities have been introduced, such as ARC (Clark et al., 2018), GQA (Hudson and Manning, 2019), GLUE benchmarks (Wang et al., 2018) and SWAG (Zellers et al., 2018). Our challenge distinguishes from previous benchmarks with some key properties. First, most of these reasoning tasks require external scientific or visual knowledge, which makes it hard to measure the actual reasoning performance. On the other hand, our challenge does not rely on any external, multimodal or expert-level information. Second, and more importantly, *PuzzLing* challenge consists of a minimal set of examples required for solution. That means, there exists no extra training data, ensuring that exploiting surface patterns would not be possible unlike in some of existing benchmarks (Gururangan et al., 2018).

In summary, this paper introduces a unique challenge, *PuzzLing Machines*, made up of ∼100 Rosetta Stone, a.k.a translation puzzles covering 81 languages from 39 different language families based on the Linguistic Olympiads. The challenge requires System2 skills—sequential reasoning and abstraction of linguistic concepts, discussed in detail in §2. We discuss the dataset and the linguistic phenomena in the resulting dataset supported with statistics and examples in §3. In §4, we present the results of intuitive baseline methods and strong MT baselines such as Transformers encoder-decoder (Vaswani et al., 2017) with integrated pretrained language models as applied to these puzzles. We show that, unsurprisingly, the puzzles cannot be easily or robustly solved by currently existing methods. We hope that this benchmark is going to evoke development of new deep MT/NLP models that operate in a human-like manner and reason upon linguistic knowledge, providing a new future research direction for NLP.

2 Meta-linguistics

Meta-linguistics is defined by Chomsky (1976) as "the knowledge of the characteristics and structures of language" as realised on the level of phonology, morphology, syntax and semantics. Any English speaker would likely have the linguistic capacity to produce the word *undo* when asked "What is the opposite of *do*?" Only a speaker with some level of meta-linguistic awareness, however, would further be able to reflect on the structure of the word they have produced: to identify *un-* as a unit that serves to negate words, to spot its similarity in function to other units like *dis-* and *de-*. He/she would also be aware that *un-* is not interchangeable with *dis-* and *de-*, since it attaches to the front of verbs and adjectives but not to nouns.

Meta-linguistic awareness is especially useful (and often improved) in the process of learning a new language, as it allows the learner to compare and contrast the structure and characteristics of the new language to those that he/she is already familiar with. It is desirable that systems for natural language processing possess meta-linguistic awareness, too, as that could hugely improve their cross-lingual generalizability, a problem that remains open after being approached from various engineering perspectives, often with little recourse to linguistics. However, measuring the meta-linguistic awareness of a system is not trivial. Existing probing techniques are mostly designed to measure how well neural models capture specific linguistic phenomena, e.g., whether a specific layer of an English language model can capture that *undo* is negative, instead of testing for meta-linguistic awareness. Our challenge takes a step further and tests whether the model can apply the underlying morphological processes, e.g. of verbal negation through prefixing. In addition, our challenge spans a wide-range of language families and covers a variety of linguistic phenomena (see §3.1), that qualifies it as a favorable testbed for measuring meta-linguistic awareness.

Let us demonstrate how meta-linguistic reasoning skills are used to solve the "Chickasaw puzzle" given in Table 1. The translation model is iteratively deduced as follows: (1) the word order in Chickasaw is Subject-Object-Verb (SOV), unlike the English SVO word order; (2) nouns take different suffixes when in a subject or object position (*at* and *ã*, respectively); (3) verbs take a suffix for 1st person singular pronomial subject or object (*li* and *sa*, respectively). Notice that, crucially, it is not possible to learn the function of the prefix *sa*, which corresponds to *me* in English, without deducing that *lhiyohli* corresponds to the verb *chases* and that third person agency in Chickasaw is not explicitly expressed. As demonstrated, inferring a translation model requires iterative reasoning on the level of words, morphemes and syntactic abstractions (classes), or, to put things differently, it requires meta-linguistic awareness.

3 The Dataset

The puzzles from Linguistic Olympiads cover many aspects of language such as phonetics, morphology, syntax and semantics. They are carefully designed by experts according to several key criteria: (1) The puzzles should be *self-contained* and *unambiguous*, meaning that no prior knowledge in the foreign language is requires, just the command of one's own native language and some level of meta-linguistic awareness and that a solution is guaranteed; (2) They should require no specialized external knowledge or formal linguistic knowledge, i.e. linguistic terms are either excluded from the instructions that accompany a puzzle or they are explicitly defined; (3) The foreign language used in a puzzle should be from a truly lesser known language family (e.g. Chickasaw, Lakhota, Khmer, Ngoni), such that there is no unfair advantage to participants whose native language is related.

We based our data collection efforts on a rich and publicly available database of language puzzles maintained by the organizers of NACLO.[2] This resource contains puzzles from IOL and a wide range of local competitions[3]. We only included puzzles written in English (or translated to English) to ensure a quality transcription and to enable error analysis. Expert solutions are available for most puzzles; we excluded the rest. In addition to the translation puzzle type shown in Table 1, we also collected 'matching' puzzles. These are two-step puzzles, in which the participants first align a shuffled set of sentences to obtain parallel data, and then translate a set of unseen sentences. We converted these puzzles to the translation puzzle format by referring to the solution files to align the training sentence pairs. Appendix A.1 describes how we selected the puzzles and how we transcribed them into a machine-readable format.

The final dataset contains 96 unique puzzles from 81 languages that span 39 different language families from all over the world, as well as two creoles and two artificial languages (see Appendix A.6 for the full list). Some of the large language families have multiple representatives, e.g. there are 13 Indo-European languages, seven Austronesian and six from the Niger-Congo family. But the majority of languages are single representatives of their respective family. This genealogical diversity leads to a great diversity in the linguistic phenomena attested in the data. Some puzzles are designed to explore a specific aspect of the unknown language in isolation, e.g. case markers on demonstrative pronouns in Hungarian (Pudeyev, 2009). In general, however, the correct solution of a puzzle involves processing on the level of syntax, morphology, phonology, and semantics all at once.

3.1 Linguistic Phenomena

The foreign languages used in linguistic puzzles are purposefully chosen to demonstrate some interesting linguistic phenomena, not found in English (or in the respective source language of the puzzle) (Bozhanov and Derzhanski, 2013), resulting in a challenging, non-trivial translation process between these diverse languages and English. In this section, we outline some key linguistic properties of the languages found in the dataset, but the list is by no means exhaustive.

Syntax: Three common configurations for the order between subject (S), verb (V) and object (O) in a sentence are exemplified in the dataset: SVO, SOV and VSO. In addition to these three, our dataset covers the rather rare OSV word order: see the example in Table 5 from the Australian language Dyirbal (Semenuks, 2012).

Morphology: We see examples of highly analytic languages (e.g. Yoruba from West Africa)

[2] http://tangra.cs.yale.edu/naclobase/

[3] NACLO (North America), OzCLO (Australia), UKLO (UK), Olimpíada Brasileira (Brazil), OLE (Spain), Panini (India), Russian LO, Russian Little Bear, Swedish LO, Polish LO, Estonian LO, Slovenian LO, Bulgarian LO, Netherlands LO and more.

	Language	Source sentence	Target sentence	Other accepted forms
1.	Chickasaw	Hilha.	(She/He) dances.	She dances. He dances.
2a.	Blackfoot	Nitoki'kaahpinnaan.	We.PL2- camped.	We camped.
2b.	Blackfoot	Oki'kaao'pa.	We.PL2 camped.	We camped.
3.	Wambaya	Bardbi ga bungmanya.	The old woman ran [away].	The old woman ran away.
4.	Euskara	Umea etorri da.	The child has (come/arrived).	The child has come. The child has arrived.

Table 2: Examples of special transcription notation.

Form	nyuk	duk	nuk	buk	guk	uk
After	vowel	n	r	m	ng	other

Table 3: Variants of a possessive suffix in Wembawemba and their phonological distribution.

as well as highly polysythetic ones (e.g. Inuktitut from Canada). Within the synthetic type, we see both agglutinative languages (e.g. Turkish) and inflectional ones (e.g. Polish). Some specific morphological properties explored in the puzzles are verbal inflection with its many categories concerning tense, aspect and mood, nominal declension and noun class systems. The aforementioned "Dyirbal" puzzle also exemplifies an interesting classification of nouns, wherein women and dangerous animals and objects are treated as one class, men and other animals constitute another class and a third class captures all remaining nouns. The choice of the articles *balan* and *bagu* in Table 5, for example, is guided by this classification.

Phonology: A wide range of phonological assimilation processes interplay with the morphological processes described above and obfuscate morpheme boundaries. These can concern voicing, nasality and vowel quality, among other features. As an example of morphological and phonological processes working together, consider the realization of pronomial possession in Australian language Wembawembda (Laughren, 2009). Unlike English, which expresses this feature with pronouns *his/her/its*, Wembawemba expresses it with a suffix on the noun it modifies, e.g. *wutyupuk* '(his/her/its) stomach'. The form of the suffix, however, depends on the ending of the noun it attaches to and can vary greatly as shown in Table 3.

Semantics: Semantics come into play when we consider the compositionality of language and figurative speech: the phrase "falepak hawei" in the

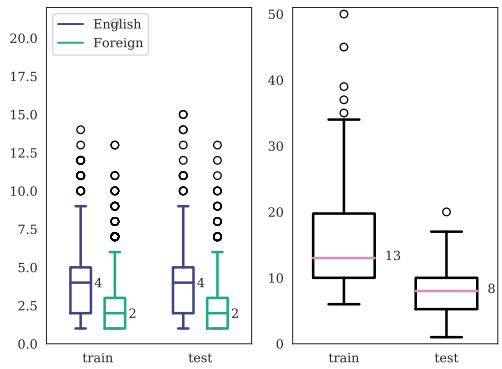

Figure 1: Box-plots for **Left:** Word# per language and split, **Right:** Sentence# per split.

Indonesian language Abui, for example, literally translates into "pistol's ear", but a more fitting translation would be "trigger" (Peguševs, 2017).

As a side note, it is important to note that while here we use extensive linguistic terminology to discuss the properties of the languages in our dataset, the high-school students who participate in Linguistic Olympiads need not and may not be familiar with any of the terminology. Their good performance depends on a well-developed metalinguistic awareness, not on formal linguistic training.

3.2 Dataset statistics

In total, 2311 parallel instances are transcribed—1559 training and 752 test. 63% of the test pairs are in the English → foreign direction, while the rest are in the foreign → English direction.

Statistics concerning the number of words per sentence[4] are shown on the left of Figure 1. The majority of both training and test pairs are fairly short, but length varies considerably. This is due to the fact that some puzzles in the dataset concern the

[4]We naively tokenize on space.

translation of individual words, some take scope over noun-modifier phrases and some, over entire sentences. English sentences are generally longer (median 4) than their translations (median 2). This is rather intuitive considering the synthetic nature of many of the foreign languages in the dataset, wherein a single long word in the foreign language may translate into 4-5 words on the English side, as in this translation from *tΛckotoyatih* in the Mexican language Zoque to the English *only for the tooth*.

Sentence statistics about the length of the train and test split per problem are shown on the right of Figure 1. Intuitively, train splits are bigger than test splits. However, the number of training instances varies greatly between the puzzles, which is related to a number of factors such as the difficulty and type of the task, as well as the linguistic properties of the foreign language.

3.3 Train versus Test Splits

One property of the data splits in linguistic puzzles, which diverges from the standard paradigm in machine learning, is that the *input* test data should not be considered "held out". On the contrary, in some cases, vocabulary items attested in the input of foreign→English test instances may be crucial to the translation of English→foreign test instances, and vice versa. So it is only the *targets* of test instances that should be truly held out. This specificity is not ubiquitous across the puzzles, but it should be accounted for by any approach to their solution, for example by building the system vocabulary over the union of the train and input test data.

4 Baselines

We attemp to solve these puzzles with models of varying complexity, i.e. from random guessing to state-of-the-art neural machine translation systems.

Random Words (RW): Since the vocabularies of source and target languages are quite small, we test what random word picking can accomplish. We simply tokenize the training sentence pairs and then randomly choose a word from the target language's vocabulary for each token in the source sentence.[5]

FastAlign (FA): We use the translation alignment tool FastAlign (Dyer et al., 2013), to test whether the puzzles can be solved by early lexical translation models (Brown et al., 1993). Since FA produces alignments for each training pair, we post-process the output to create a translation dictionary separately for each direction. We then randomly choose from the translation entries for each token in source test sentence. [6]

Phrase Based Statistical Machine Translation (PBSMT) Since Koehn and Knowles (2017) report that PBSMT models outperform vanilla NMT models in case of small parallel training data, we use PBSMT as one of the baselines. For the foreign→English direction, we implement two models—one using no external mono-lingual English data and one otherwise.

4.1 Neural Machine Translation

We implement three different models based on Transformers (Vaswani et al., 2017) using the implementation of Ott et al. (2019). In the first scenario, we train an off-the-shelf Transformer encoder-decoder model for each direction, referred to as *Transformer*. Second, we use a strong pre-trained English language model, RoBERTa (Liu et al., 2019), to initialize the encoder of the NMT model for English to foreign translation. Finally, for foreign to English translation, we concatenate the translation features extracted from the last Transformer decoder layer, with the language modeling features extracted from RoBERTa (Liu et al., 2019), before mapping the vectors to the output vocabulary. These models are denoted as *Transformer+RoBERTa*.

5 Experiments

5.1 Experimental Settings

We first compile a subset from the puzzles that are diverse by means of languages and contain translation questions in both directions. During tuning, we use the test sentences on these puzzles to validate our models. Since our foreign languages are morphologically rich, we use BPE (Sennrich et al., 2016) to segment words into subwords. For the sentences in the foreign language, we learn the BPE from the training data, while for English sentences we use the already available GPT2-BPE dictionary to exploit English language prior. For convenience,

[5] We don't use frequency of the words, i.e., pick words that occur more often, since they are not that meaningful due to the tininess of the data.

[6] We add all aligned target phrases of the source token to the dictionary. Hence, when one target phrase is seen multiple times, it is more likely to be chosen during inference.

before we train the models, we lowercase the sentences, remove certain punctuations, remove pronoun tags and brackets, and augment training data with multiple reference translations.

PBSMT: We use Moses (Koehn et al., 2007) with default settings. We employ wikitext-103 corpus to train a 5-gram English LM for the model with access to external data. The other model only uses training sentences for the LM.

NMT: Following the suggestions for low-resource NMT systems by Sennrich and Zhang (2019), we use small and few layers and high dropout rates. Similarly we use the smallest available language model (RoBERTa Base) and freeze its parameters during training to reduce the number of trainable parameters. We tune the following hyper-parameters: BPE merge parameter, learning rate and number of epochs.

5.2 Evaluation Metrics

The submissions to Linguistic Olympiads are manually graded by experts. For a full mark, an exact solution has to be provided, as well as a correct and detailed discussion of the underlying processes that led to this solution, e.g., concerning findings about word-order, the function of individual morphemes, etc. Participants are also given partial marks in case of partial solutions or valid discussions. Since we don't have access to expert evaluation, we use readily available automatic machine translation measures. We also note grading of system interpretations or its solution steps as an interesting future research direction.

The first is the BLEU (Papineni et al., 2002) score since it is still the standard metric in MT. We use BLEU-2 to match the lower median of sentence lengths we observe across the English and the foreign data (see Fig 1). BLEU matches whole words rather than word pieces, which prevents us from assigning partial credit to subword matches, which could be especially relevant for foreign target languages with rich morphology. We therefore use three additional metrics that operate on the level of word pieces: CharacTER (Wang et al., 2016), ChrF (Popovic, 2016) and ChrF++ (Popovic, 2017). CharacTER is a measure derived from TER (Translation Edit Rate), where edit rate is calculated on character level, whereas shift rate is measured on the word level. It calculates the minimum number of character edits required to adjust a hypothesis, until the reference is matched, normalized by the length of the hypothesis sentence. For easier comparison, we report $1.0 - characTER$ scores. ChrF is a simple F-measure reflecting precision and recall of the matching character n-grams. ChrF++ adds word unigrams and bi-grams to the standard ChrF for a higher human correlation score. We experiment with different combinations of character n-grams ($n = 3, 5$ as suggested in Popovic (2016)) and word n-grams ($n = 0, 1, 2$ as suggested in Popovic (2017)).

Finally, we also measure the average exact match of the puzzles, which is calculated as 1 if the prediction and reference sentences match and 0 otherwise. As it is not feasible to report and compare results on all of these metrics (nine in total), we compute the pair-wise Pearson correlation coefficient between them, and average over all pairs to arrive at the following four metrics that show the least correlation with each other: BLEU-2, CharacTER, ChrF-3 and exact match. We note, however, that of these four, exact match is really the most meaningful metric. Since the sentences in the dataset are rather short and the puzzles are designed to be solvable and unambiguous, an exact match should be attainable. Moreover, as the puzzles in the dataset are of varying difficulty, the average exact match score can be seen as a continuous metric.

6 Results and Analysis

We report the results for the best models in Fig. 2. The hyperparameter configuration and the development set results are given in Appendix A.4. The maximum exact match score among all results is only 3.4%; and the highest scores are consistently achieved by PBSMT models on both directions and dataset splits.

The overall results for foreign $\rightarrow$ English are generally higher than English $\rightarrow$ foreign. This may be due to (a) having longer sentences for English; (b) the scores (except from EM) being more suitable for English (even the character-based ones) or (c) the more challenging nature of translation into foreign languages, which needs another dedicated study.

English$\rightarrow$Foreign: Initializing the NMT encoder with RoBERTa has severely worsened the results, compared to standard Transformer model. We believe the main reason is the imbalance between encoder (huge encoder) and the decoder (tiny decoder), that makes training very challenging. The gap between the simplest baselines (RW,

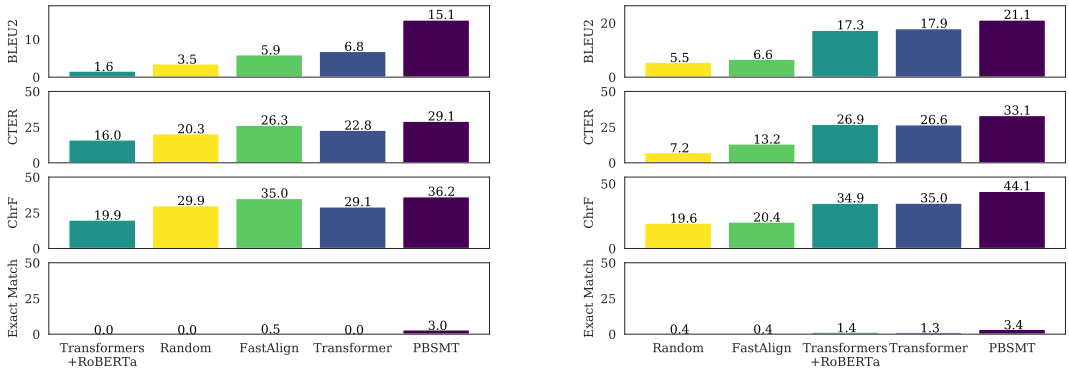

Figure 2: Main results (best viewed with color). **Left:** English→foreign **Right:** foreign→English.

FA) and more sophisticated models (Transformers, PBSMT) is also considerably small; FA even surpassing Transformers's CTER and ChrF performance. For most of the foreign languages, even when two words are semantically distant, there may still be significant morpheme overlap. These suggest that simple lexical alignment models (including random assignment) can achieve higher *partial* matching scores that hints at the unreliability of CTER and ChrF measures for the puzzles.

Foreign→English: We observe that the gap between the simple and more sophisticated baselines are higher in this direction by means of all measures, as we would expect. Using RoBERTa features in the decoder does not hurt the performance while providing a small increase in EM score compared to standard Transformers. It should be noted that the decoder is still tiny and LM features are only incorporated via a separate linear layer at a very late stage, which prevents the imbalance problem we saw in English → foreign.

We see similar results for the validation data with the exception that Transformer-based models achieve either higher or the same EM scores than PBSMT while surpassing PBSMT's BLEU-2 scores in foreign → English. It supports the findings of Sennrich and Zhang (2019), drawing attention to the importance of hyper-parameter tuning for low-resource NMT models.

6.1 Error Analysis

We perform manual error analysis on the predictions of our top two models for the Chickasaw puzzle presented in Table 1. The predicted translations are shown in Table 4. We also provide the predictions of the simple baselines in Appendix A.5 for comparison. Although the PBSMT model is best on average, we find that for this particular puzzle, the Transformer model did much better. PBSMT had very few hits overall: it correctly chose to include the lexical items *hattak* and *hollo* in (1), but the position and inflection of the former is incorrect. In (5) and (6) there are indications of correct lexicon induction, but the overall quality of the translations is very poor both in terms of accuracy and fluency. The Transformer model, on the other hand, predicts fluent translations in both directions. In the direction from English to Chickasaw, we see that the model correctly acquired the relevant morphological patterns: subjects take suffix *at*, objects take suffix *ã*, and, importantly, that first person agency is expressed through suffix *li*. The translations are still not perfect, though, due to lexical confusion: the words for *cat* and *dog* have been swapped in both (1) and (2), as well as the words for *love* and *chase* in (3). In the direction from Chickasaw to English, the Transformer's predictions remain fluent, but they hardly relate to the input. Contrary to the overall results, for this puzzle translation *to* English appears to be more challenging for the model.

7 Related Work

Recently, reasoning tasks and datasets that require natural language processing have been introduced, such as common-sense reasoning in the form of pronoun resolution e.g., WSC (Levesque, 2011), multiple-choice question answering e.g., SWAG (Zellers et al., 2018) and ARC (Clark et al., 2018); inference tasks in the form of binary or multi-label classification problems e.g., the GLUE benchmarks (Wang et al., 2018); and visual reasoning in the form of question answering (Zellers et al.,

	Chikasaw	English	PBSMT	Transformer
Now you can translate the following into Chickasaw:				
(1)	Hattakat ihooã hollo.	The man loves the woman.	the the woman hattakã hollo	ihooat hattakã hollo
(2)	Kowi'at shoha.	The cat stinks.	the lhiyohli stinks	ofi'at shoha
(3)	Holloli.	I love her/him.	i love him	lhiyohlili
Translate the following into English:				
(4)	Ihooat sahollo.	The woman loves me.	ihoothe sahollo	the woman loves the man
(5)	Ofi'at hilha.	The dog dances.	the(he/she) dances	the cat chases the dog
(6)	Kowi'ã lhiyohlili.	I chase the cat.	cat ch thei chase (him/her)	the dog stinks

Table 4: Predictions for the "Chickasaw" puzzle. Gold-standard target sentences are shown in yellow.

2019) e.g., GQA (Hudson and Manning, 2019). In these tasks, the required level of semantics is mostly limited to single sentences rather than a collection; almost all tasks target English; data is derived from running text and is mostly close-domain. In addition, some require external knowledge bases or high-level knowledge on physical models or experiments as in ARC classified by Boratko et al. (2018), which leaves room for accumulating errors from external parts and complicates the analysis of individual parts like reasoning.

Another body of early work on symbolic AI provides a different set of tools to model reasoning such as rule-engines, rule-induction algorithms, logic programs and case-based reasoning models (Kolodner, 1992). However, it is not trivial to represent and model our task in these frameworks, since they mostly require defining primitives, expressions, discrete features and cases. Furthermore, the strength of statistical/neural models has been repeatedly shown to surpass rule-based models. Our goal is to encourage researchers to incorporate reasoning into statistical models, rather than replacing them with symbolic models.

8 Conclusion and Future Work

The field of NLP has developed deep neural models that can exploit large amounts of data to achieve high scores on downstream tasks. Still, the field lacks models that can perform human-like reasoning and generalization. To mitigate this gap, we draw inspiration from the *Linguistic Olympiads* that challenge the meta-linguistic and reasoning abilities of high-school students. We create a new benchmark dataset from available Linguistic Puzzles that spans over 81 languages from 39 language families, which is released at `https://ukplab.github.io/PuzzLing-Machines/`. We implement and evaluate simple baselines such as alignment, and state-of-the-art machine translation models with integrated a pretrained English language model. We show that none of the models can perform well on the puzzles, suggesting that we are still far from having systems with meta-linguistic awareness and reasoning capabilities.

Acknowledgements

This work was supported by the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1 and grant GU 798/17-1). We would like to thank Liane Vogel, Marc Simon Uecker and Siddharth Singh Parihar for their great help during the project. We are grateful to Dragomir Radev for his feedback and continuous help with encoding problems encountered during puzzle transcription. We thank Adam Lopez and Ilia Kuznetsov for providing feedback on early drafts of the paper. We thank the area chairs and the senior area chair, whose comments helped us improve the paper.

References

Yoshua Bengio. 2020. Deep Learning for AI. Invited talk at AAAI.

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock. 2018. A systematic classification of knowledge, reasoning, and context within the ARC dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, Melbourne, Australia, July 19, 2018*, pages 60–70.

Bozhidar Bozhanov and Ivan Derzhanski. 2013. Rosetta stone linguistic problems. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 1–8.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathemat-

ics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Noam A Chomsky. 1976. *Reflections on language*. New York: Pantheon.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 107–112.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2019. Glottolog 4.1. Max Planck Institute for the Science of Human History.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709.

Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Janet L. Kolodner. 1992. An introduction to case-based reasoning. *Artif. Intell. Rev.*, 6(1):3–34.

Mary Laughren. 2009. Wembawemba expressing possession. http://cxielamiko.narod.ru/zadachi/ozclo-09-state.pdf, OzCLO.

Yann LeCun. 2020. Self-Supervised Learning. Invited talk at AAAI.

Hector J. Levesque. 2011. The winograd schema challenge. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Gary Marcus. 2020. The next decade in AI: four steps towards robust artificial intelligence. *CoRR*, abs/2002.06177.

James L. McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2019. Extending machine language models toward human-level language understanding. *CoRR*, abs/1912.05877.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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

Tom Payne. 2005. Chickasaw. http://lingclub.mycpanel.princeton.edu/challenge/chickasaw.php, Linguistics Society of America.

Aleksejs Peguševs. 2017. Abui. https://ioling.org/booklets/iol-2017-indiv-prob.en.pdf, IOL.

Maja Popovic. 2016. chrf deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 499–504.

Maja Popovic. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 612–618.

Victor Pudeyev. 2009. Hungarian. http://tangra.cs.yale.edu/naclobase/get.cgi?pid=198&version=1, NACLO.

Arthur Semenuks. 2012. Dyirbal. https://ioling.org/booklets/iol-2012-indiv-prob.en.pdf, International Linguistic Olympiad.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 211–221.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 353–355.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 505–510.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

A Appendices

A.1 Transcription of Puzzles

The puzzles are generally provided as pdf files. Many languages in the dataset use the Latin script, optionally with some diacritics. Some which use a non-Latin script (or have no writing system at all), are transcribed with IPA or transliterated into the Latin script. Only one language, Georgian, uses a non-Latin script, namely the Mkhedruli script. As there are various types of puzzles presented at the Olympiads, we identified the ones relevant to our task through automatic filtering for the keywords "translation" or "matching", and manually verified the results.

To represent linguistic puzzles in a unified, machine-readable format, we defined a JSON format shown in Appendix A.2. The relevant data was manually extracted from the PDF files and mapped to this format in a semi-automated fashion. We faced encoding issues with many of the puzzles. For some of these, the database owner kindly provided us with the source files of the pdf documents, which enabled us to generate UTF-8 encoding of the data; others we fixed manually. Some puzzles, which use pictorial scripts or are otherwise UTF-8 incompatible, were discarded.

During the transcription we came across various formats of linguistic annotation in the puzzles. This kind of information was not consistently provided across puzzles, but we included it where available, as it can be both helpful and crucial to a correct solution. In the next paragraphs, we provide details on the different types of annotated information and the standardized format we used to encode those.

Gender distinction in pronouns: When the foreign language does not mark gender on pronouns (or omits pronouns altogether), singular pronouns in the English translations are provided consistently as *(he/she)* and *(him/her)*, or *(he/she/it)* and *(his/her/its)*, as in Ex. 1 in Table 2. During evaluation, instances of this notation are accepted, as well as instances of the individual alternatives.

Number marking on pronouns: When the foreign language marks two levels of plurality for the second person pronoun *you*, they are marked accordingly as *you.SG* and *you.PL*. Some languages make a distinction between plural forms concerning two objects and plural forms concerning three or more objects. In this case, we mark pronouns (not just *you*, but also *we* and *they*) with the notation *.PL2* and *.PL3*, respectively. Some languages also make a distinction between an inclusive *we* 'you and me' and an exclusive *we* 'me and someone else'. We reserve *we.PL2* for the inclusive sense, and mark the exclusive sense with *we.PL2-*. See examples 2a and 2b in Table 2. The notation presented here holds for both personal pronouns, e.g. *you*, and possessive pronouns, e.g. *your*. During evaluation, we disregard this notation on the side of the target language.

Zero-to-one matching: Words that are semantically implied or required by English grammar, but not directly expressed on the side of the foreign language are shown in square brackets in some of the puzzles, as in Table 2-Ex. 3. This bracketing exists only to aid the learning of a translation model. During evaluation, we remove these brackets from the target test sentences.

Notice that number marking and special notation for zero-to-one matching is not ubiquitous across the puzzles. We included it only when it was provided in the original puzzle.

Multiple reference translations: Occasionally, several possible translations are listed in a puzzle for a given word, phrase or sentence–see Table 2-Ex. 4. We represent these options inside parenthesis separated with a slash (/), e.g., (alternative1/.../alternative N). Since the alternatives are of different granularity, nested bracketing may sometimes occur. During evaluation, we calculate the scores between the prediction and all possible references, and take the maximum.

Additional information Roughly half of the puzzles contain remarks on individual characters and diacritics in the inventory of the foreign language, e.g. "In the following orthography a colon (:) marks a long vowel, and the ʔ symbol marks a glottal stop (like the sound in the middle of uh-oh)". In many cases, the instructions state that these are pronunciation notes, i.e. they are made available only to allow the participants to vocalize the words they see on the page. On some occasions, however, they might introduce a character that is not present in the training data, but is relevant to the translation of the test sentences, e.g. the voiceless counterpart of a voiced consonant in a language with voice assimilation. As this kind of information cannot be mapped to the parallel data format, we include it in a separate field in the JSON files, directly as it

Source	balan	waymin	bambun	baNgu	jugaNgu	jamiman.
Gloss	OBJ	mother-in-law	healthy	SUBJ	sugar-SUBJ	fat-MAKE.
Target	Sugar makes the healthy mother-in-law fat.					

Table 5: Example sentence in Dyibral.

appeared in the puzzles. [7]

With the aforementioned guidelines, each puzzle was transcribed by one transcriber and verified by at least one other transcriber. For the test pairs, the direction of translation is stored as well, since a possible and singular solution is only guaranteed in the direction as given in the puzzle.

A.2 JSON File Format

Each puzzle is represented with a JSON file containing the following fields: SOURCE_LANG, TARGET_LANG, META, TRAIN and TEST. Each field is explained in Table 6.

A.3 Development Results

The results on the validation set are given in Fig. 3.

A.4 Hyperparameter Settings

The best hyperparameters found for each NMT model is given as following. *FA:* word to word alignments; PBSMT for English→Foreign: word alignment with external English LM; PBSMT for Foreign→English: BPE with 30 merge operations. For both Transformers-based models in Foreign→English direction, we used BPE with 10 merge operations, learning rate of 0.001 and 500 epochs; while for the standard Transformer in English→Foreign direction, BPE with 30 merge operations have been used. For all models except from Transformers with RoBERTa encoder, both the encoder and decoder had 1 layers, and all hidden dimesions were set to 128, dropout was set to 0.3, and the models were trained with Adam optimizer. For Transformer with RoBERTA LM Encoder for English→Foreign, we have used 0.0001 learning rate with reduction on plateau, batches of size 2, dropout of 0.1, 1 layer, 64 embedding units, 128 hidden units, and BPE with 5 merge operations.

[7] We believe that even if all instances of such remarks are ignored, the puzzles should remain mostly solvable, but we note that without this information, the ceiling for human performance would not be quite 100 percent.

A.5 Chickasaw Additional Predictions

In Table 7, the predictions of RW and FA are shown for comparison.

A.6 List of Languages and Families

The full list for the languages and the families they belong to, as classified in WALS (Dryer and Haspelmath, 2013) and, where WALS lacks an entry, Glottolog (Hammarström et al., 2019), are given in Table 8.

Field	Definition	Example
SOURCE_LANG	Name of the source language	Foreign language e.g., Kiswahili, Pali
TARGET_LANG	Name of the target language	English
META	Additional information about the foreign language provided in the original puzzle (as free text)	”The sound represented as ã is a ’nasalized’ vowel. It is pronounced like the ’a’ in ’father’, but with the air passing through the nose, as in the French word ’ban’.”
TRAIN	Parallel training sentences given as a list of lists	[[“Bonjour”, “Good morning”], [“chat”, “cat”]], where the source and the target language is French and English respectively.
TEST	Parallel test sentences with direction information	[[“Bonjour”, “Good morning”, >], [“chat”, “cat”, <]]. “>” implies that the translation is required from source to target language, vice versa for “<”

Table 6: JSON file format used in the linguistic puzzles shared task

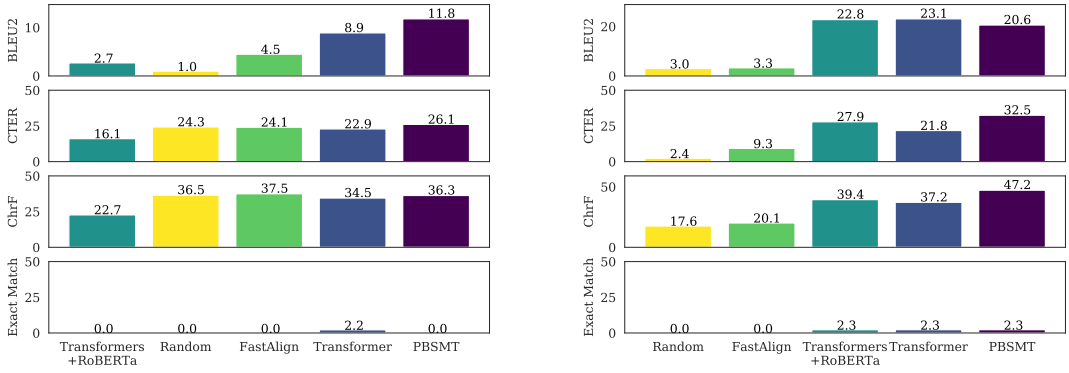

Figure 3: Development set results. **Left:** English→foreign **Right:** foreign→English

	Chikasaw	English	RW	FA
(1)	Hattakat ihooã hollo.	The man loves the woman.	Ihooat lhiyohli hollo salhiyohli ofi’at.	The hollo loves the woman.
(2)	Kowi’at shoha.	The cat stinks.	Lhiyohlili lhiyohlili kowi’ã.	The lhiyohli shoha.
(3)	Holloli.	I love her/him.	Ofi’ã hilha lhiyohlili.	I love lhiyohlili.
(4)	Ihooat sahollo.	The woman loves me.	Dog loves	Ihooat sahollo
(5)	Ofi’at hilha.	The dog dances.	I the	ofi’at he dances
(6)	Kowi’ã lhiyohlili.	I chase the cat.	stinks cat	Kowi’ã I chase (him/her).

Table 7: Predictions of the simple baseline models for the “Chickasaw” puzzle. Gold-standard target sentences are shown in yellow.

Language	Family	Language	Family
Abkhaz	Northwest Caucasian	Luiseño	Uto-Aztecan
Abma	Austronesian	Madak	Austronesian
Abui	Timor-Alor-Pantar	Malay	Austronesian
Afrihili	Artificial	Maori	Austronesian
Amele	Trans-New Guinea	Mayangna	Misumalpan
Ancient Greek	Indo-European	Miwoc	Penutian
Bambara	Mande	Muna	Austronesian
Basque	Basque	Nahuatl	Uto-Aztecan
Beja	Afro-Asiatic	Ndebele	Niger-Congo
Benabena	Trans-New Guinea	Nen	Trans-New Guinea
Blackfoot	Algic	Nepali	Indo-European
Bulgarian	Indo-European	Nhanda	Pama-Nyungan
Central Cagayan Agta	Austronesian	Norwegian	Indo-European
Chamalal	Nakh-Daghestanian	Nung	Tai-Kadai
Chickasaw	Muskogean	Old English	Indo-European
Choctaw	Muskogean	Pali	Indo-European
Cupeño	Uto-Aztecan	Papiamento	creole
Danish	Indo-European	Persian	Indo-European
Dyirbal	Pama-Nyungan	Polish	Indo-European
Esperanto	Artificial	Proto-Algoquian	Algic
Fula	Niger-Congo	Quechua	Quechuan
Georgian	Kartvelian	Somali	Afro-Asiatic
Guaraní	Tupian	Swahili	Niger-Congo
Haitian Creole	Creole	Tadaksahak	Songhay
Hmong	Hmong-Mien	Tanna	Austronesian
Hungarian	Uralic	Teop	Austronesian
Icelandic	Indo-European	Tok Pisin	creole
Ilokano	Austronesian	Tshiluba	Niger-Congo
Inuktitut	Eskimo-Aleut	Turkish	Altaic
Irish	Indo-European	Udihe	Altaic
Jaqaru	Aymaran	Waanyi	Garrwan
Kabardian	Northwest Caucasian	Wambaya	Mirndi
Kayapo	Macro-Ge	Warlpiri	Pama-Nyungan
Kimbundu	Niger-Congo	Welsh	Indo-European
Kunuz Nubian	Eastern Sudanic	Wembawemba	Pama-Nyungan
Kurdish	Indo-European	Witsuwit'en	Dené–Yeniseian
Lakhota	Siouan	Yidiny	Pama-Nyungan
Lalana Chinantec	Oto-Manguean	Yolmo	Sino-Tibetan
Latvian	Indo-European	Yonggom	Nuclear Trans New Guinea
Lopit	Nilo-Saharan	Yoruba	Niger-Congo
		Zoque	Mixe-Zoque

Table 8: Full list of languages and their families.