

Did the Cat Drink the Coffee? Challenging Transformers with Generalized Event Knowledge

Paolo Pedinotti

University of Pisa
paolo.pedinotti@phd.unipi.it

Giulia Rambelli

University of Pisa - Aix-Marseille University
giulia.rambelli@phd.unipi.it

Emmanuele Chersoni

The Hong Kong Polytechnic University
emmanuelechersoni@gmail.com

Enrico Santus

Bayer Pharmaceuticals
esantus@gmail.com

Alessandro Lenci

University of Pisa
alessandro.lenci@unipi.it

Philippe Blache

Aix-Marseille University
blache@lpl-aix.fr

Abstract

Prior research has explored the ability of computational models to predict a word semantic fit with a given predicate. While much work has been devoted to modeling the typicality relation between verbs and arguments in isolation, in this paper we take a broader perspective by assessing whether and to what extent computational approaches have access to the information about the typicality of entire events and situations described in language (*Generalized Event Knowledge*).

Given the recent success of Transformers Language Models (TLMs), we decided to test them on a benchmark for the *dynamic estimation of thematic fit*. The evaluation of these models was performed in comparison with SDM, a framework specifically designed to integrate events in sentence meaning representations, and we conducted a detailed error analysis to investigate which factors affect their behavior. Our results show that TLMs can reach performances that are comparable to those achieved by SDM. However, additional analysis consistently suggests that TLMs do not capture important aspects of event knowledge, and their predictions often depend on surface linguistic features, such as frequent words, collocations and syntactic patterns, thereby showing sub-optimal generalization abilities.

1 Introduction

People can discriminate between typical (e.g., *A cop arrested a thief*) and atypical events (e.g., *A thief arrested a cop*) and exploit this ability in on-line sentence processing to anticipate the upcoming linguistic input. Brains have been claimed to be “prediction machines” (Clark, 2013) and psycholinguistic research has shown that a crucial ingredient

of such predictive ability is the knowledge about events and their typical participants stored in human semantic memory, also referred to as **Generalized Event Knowledge** (GEK) by McRae and Matsuki (2009). To make an example, if we were asked to think about things that are played with a guitar, we would quickly and more or less unanimously think of words such as *song*, *piece* or *riff*.

Computational models of predicate-argument typicality, generally referred to as *thematic fit* in the psycholinguistic literature (McRae et al., 1998), extract typical arguments from parsed corpora. However, GEK is not just storing relations between words: The fact that this knowledge is generalized – that is, it is based on an abstract representation of what is typical – allows us to easily classify new argument combinations as typical or atypical. Furthermore, psycholinguistic studies (Bicknell et al., 2010; Matsuki et al., 2011) have shown that humans are able to combine and dynamically update their expectations during sentence processing: for example, their expectations given the sequence *The barber cut the _* differ from the ones given *The lumberjack cut the _*, since the integration of knowledge “cued” by the agent argument with the verb will lead to the activation of different event scenarios. In Distributional Semantics, sophisticated models of the GEK have been proposed to make predictions on upcoming arguments by integrating the cues coming from the verb and the previously-realized arguments in the sentence (Lenci, 2011; Chersoni et al., 2019). Since such knowledge is acquired from both first-hand and linguistic experience (McRae and Matsuki, 2009), an important assumption of this literature is that, at least for its “linguistic subset”, the GEK can be modeled with distributional information extracted

from corpora (Chersoni et al., 2017, 2021).

Language Models are trained to make predictions given a context, and thus, they can also be viewed as models of GEK. This approach is promising if one considers the success of recent Transformer-based Language Models (henceforth TLMS), which are trained on huge corpora and contain a massive number of parameters. Even if these models receive extensive training and have been shown to capture linguistic properties (Jawahar et al., 2019; Goldberg, 2019), it is not obvious whether they acquire the aspects of GEK that have been modeled explicitly in previous approaches. To the best of our knowledge, Transformers have never been tested on dynamic thematic fit modeling, nor their performance has been compared with traditional distributional models. Our current work is addressing this issue.

Contributions:

1. we propose a methodology to adapt TLMS to the dynamic estimation of thematic fit, using a dataset that contains several types of argument combinations differing for their typicality;
2. we present a comprehensive evaluation of various TLMS on this task, performed by comparing them to a strong distributional baseline;
3. we conduct further analysis aimed at identifying the potential limitations of TLMS as models of GEK.

Our results are relevant for researchers interested in assessing the linguistic abilities of TLMS, as well as those working on applications involving TLMS, such as text generation.

2 Related Work

In its classical form, the thematic fit estimation task consists in comparing a candidate argument or *filler* (e.g., *wine*) with the typical fillers of a given verb role (e.g., agent, patient, etc.), either in the form of exemplars previously attested in a corpus (Erk, 2007; Vandekerckhove et al., 2009; Erk et al., 2010) or in the form of a vector-based prototype (Baroni and Lenci, 2010; Sayeed and Demberg, 2014; Sayeed et al., 2015; Greenberg et al., 2015a,b; Sayeed et al., 2016; Santus et al., 2017; Chersoni et al., 2020). Additionally, recent studies explored the use of masked language modeling with BERT for scoring the candidate arguments (Metheniti et al.,

2020). Performance in the thematic fit task is typically measured with the correlation between the output scores of the model and human-elicited typicality judgments for verb-argument pairs (McRae et al., 1998; Ferretti et al., 2001; Padó, 2007; Zhang et al., 2019; Marton and Sayeed, 2021).

In the simplest and most common version of this task, the typicality of verb argument-pairs is evaluated in isolation. Thematic fit is instead a *dynamic concept*: The expectations for an argument in a given verb role do not depend just on the verb, but also on the compositional combination with the other arguments in the sentence (Bicknell et al., 2010). To check the ability of computational models to account for the compositional update of argument expectations, Lenci (2011) framed the problem as a binary classification task: A system is presented a sentence pair, with one sentence expressing a typical real-world situation (*The journalist is checking the report*) and the other sentence expressing a plausible but less typical one (*The mechanic is checking the report*), and the task is to assign a higher thematic fit/typicality score to the former. Notice that the two sentences differ only for one argument, and that the “atypical” one might, however, be a common filler with respect to the verb target role (e.g., *report* is a typical patient for *check*, it is just less plausible in combination with *mechanic* as an agent).

Several models have tried to tackle the “dynamic” version of the thematic fit task, either based on classical distributional spaces (Chersoni et al., 2016, 2019) or on more sophisticated neural network architectures (Tilk et al., 2016; Hong et al., 2018). On the evaluation side, those works made use of the experimental materials of the study by Lenci (2011), which are, however, limited to agent-verb-patient triples. The recently-introduced DT-Fit dataset (Vassallo et al., 2018) is, in comparison, larger in size and provides more variety of fillers and roles (including instruments, locations and time). Other studies introduced larger datasets, but focused on more specific notions of event plausibility (e.g. the plausibility depending on the physical properties of the participants) (Wang et al., 2018; Porada et al., 2019; Ko et al., 2019).

3 Experimental Settings

3.1 Dataset

The DTFit (Vassallo et al., 2018) dataset has been specifically designed for the evaluation of dynamic

thematic fit.¹ The dataset contains pairs of tuples that differ only for one element, which can be either a typical or atypical filler of a given role in the event described by the tuple (cf. Table 1). The dataset includes tuples of different lengths, and the typicality of a given argument depends on its interaction with all the other elements. For each tuple, the authors collected typicality judgments by asking English native speakers how common was the event described. Scores range from 1 (very atypical) to 7 (very typical). The dataset mainly targets knowledge about professions, but also other typical everyday situations (e.g., what a dog typically eats, what a grandmother typically does).

The authors created several datasets, which differ with respect to the semantic role of the candidate filler. For our experiments, we selected the datasets created by the authors for the following relations: $\{\mathbf{Instrument}, \mathbf{Time}, \mathbf{Location}\}_{DTFit}$. Additionally, from the original dataset containing agent-verb-patient triples, we derived two datasets, that we named \mathbf{Agent}_{DTFit} and $\mathbf{Patient}_{DTFit}$. In \mathbf{Agent}_{DTFit} , the tuples forming a pair differ with respect to the typicality of the agent. In $\mathbf{Patient}_{DTFit}$, they differ for the typicality of the patient. We thus get a total of five datasets, each of which covers a different semantic relation. The latter two datasets have the same properties of the others, but they put stronger emphasis on the dynamic nature of thematic fit, as the atypical filler is still a typical complement of the verb alone. Conversely, the atypical candidate fillers in the other datasets are appropriate fillers of the role, but, in most cases, they do not relate to the other elements of the tuple. Therefore, \mathbf{Agent}_{DTFit} and $\mathbf{Patient}_{DTFit}$ are more challenging for computational models, as the typicality of a filler can only be determined through the composition of the verb with another argument. Accordingly, models have to update their predictions by accurately taking into account the whole context.

For each tuple in DTFit, the task for our models is to predict the upcoming argument on the basis of the previous ones. Models were evaluated in terms of Spearman correlation between the human ratings and the models’ scores. Moreover, we performed a second evaluation for \mathbf{Agent}_{DTFit} and $\mathbf{Patient}_{DTFit}$, consisting of measuring the accuracy of each system in assigning a higher thematic fit score

¹All the datasets used for the experiments described in this paper can be found at the link: https://github.com/giuliarambelli/transformers_thematic_fit.

Role	Tuple	Typical	Atypical
Agent	_ mix paint	painter	cook
Patient	tailor sew _	dress	wound
Instrument	cook clean fish _	knife	sponge
Time	cat chase bird _	hunting	marriage
Location	sailor mop deck _	boat	theatre

Table 1: Examples of tuples from DTFit.

to typical tuples. To the best of our knowledge, the only attempts to test computational models on this dataset have been done by the authors of the original paper and by Chersoni et al. (2019). In both works, distributional prototype models of thematic fit have been used.

3.2 Models

In our experiments, we compared the performance of TLMS with the Structured Distributional Model (SDM), which has been recently shown to be an efficient model for the dynamic estimation of thematic fit (Chersoni et al., 2019).

3.2.1 Structured Distributional Model

The **Structured Distributional Model** (SDM) proposed by Chersoni et al. (2019) combines word embeddings and formal semantics to specifically represent GEK and the dynamic construction of sentence meaning. Like traditional distributional models of thematic fit, it builds a prototype representation for a given role (e.g., the typical patient of *sing*) from its typical fillers, but its novelty is that the fillers are retrieved from an external resource called *Distributional Event Graph* (henceforth, *DEG*). *DEG* represents GEK as a graph automatically built from parsed corpora, where the nodes are words associated to a numeric vector, and the edges are labeled with syntactic relations and weighted using statistic association measures. Thus, given a lexical cue w , it is possible to identify the events in which w takes part and to retrieve words related to w on both the paradigmatic and the syntagmatic axis.

The formal structure at the basis of SDM consists of two semantic structures: the *linguistic condition* (*LC*), a context-independent tier of meaning that represents the lexical items in a sentence, and the *active context* (*AC*), which accumulates contextual information activated by lexical items. The crucial aspect of SDM is that it associates a vector representation to these formal structures: \vec{LC} is the sum of the embeddings of the lexical items of

a sentence; \vec{AC} , for each syntactic slot, is represented as the centroid vector built out of the role vectors r_1, \dots, r_n available in AC , corresponding to the syntactic associates of the lexical items that have been already processed.

In our implementation of SDM, the DEG was constructed by extracting syntactic relations from a concatenation of the ukWaC corpus (Baroni et al., 2009), a dump of Wikipedia 2018 and the British National Corpus (Leech, 1992). The final graph contains words with a minimum frequency of 300 and events with a minimum frequency of 30. We used as lexical embeddings the publicly-available FastText vectors extracted from Wikipedia.² For our experiments, we built a semantic representation for each tuple in the dataset, like in Chersoni et al. (2019). We used the information in LC and AC to assign a typicality score to each candidate filler of a role in the dataset. The scoring function for a given role filler is the following:

$$\frac{\cos(\vec{f}, \vec{LC}(sent)) + \cos(\vec{f}, \vec{AC}(sent))}{2} \quad (1)$$

where \vec{f} is the embedding of the candidate filler; $\vec{LC}(sent)$ is a vector obtained from the sum of the embeddings of the verb and of the argument other than f ; \vec{AC} stands for the updated expectation prototype for the role filled by f . In other words, we quantify the typicality of an argument given a tuple as the average of i.) the cosine similarity between the argument embedding and the additive combination of the other argument vectors (\vec{LC}), and ii.) the cosine similarity between the argument embedding and the prototype vector representing the active context (\vec{AC}). In the cases where \vec{AC} cannot be derived (because DEG does not store syntactic relations involving the context words), we take only the cosine between \vec{f} and $\vec{LC}(sent)$ as the final score.

3.2.2 Transformer-based Language Models

We experimented with four TLMS to test how different architectures, training objectives, and sizes of the training corpus affect performance.³

BERT (Devlin et al., 2019) consists of a series of stacked Transformer encoders. It was trained using both a masked language modeling objective (i.e.,

²<https://fasttext.cc/docs/en/english-vectors.html>

³For all experiments involving TLMS, we use pre-trained models available in the HuggingFace’s Python library Transformers (Wolf et al., 2019).

predicting a masked word from its left- and right-context), and a next sentence prediction objective (i.e., whether a sentence follows another sentence or not), on a combination of the BooksCorpus and English Wikipedia (13GB in total). The model uses WordPiece vocabulary. To test if the model size can affect BERT performance, we used both the `base` (Number of layers=12, Hidden size=768) and the `large` (L=24, H=1024) versions.

RoBERTa (Liu et al., 2019), which we used in the `large` version, is based on the same architecture as BERT, but it was trained on a much larger corpus (160GB) and without the next sentence prediction objective. In our experiments, we used the `large` version (L=24, H=1024).

In contrast with the bidirectional nature of BERT and RoBERTa, **GPT2** (Radford et al., 2019) is a uni-directional LM, which means that the training objective is to predict the next word, given all of the previous ones. It was trained on WebText, for a total of 8 million documents of data (40 GB). We employed the `medium` version of GPT2 (L=24, H=1024). We chose GPT2-medium since its dimensions are comparable to those of BERT and RoBERTa large. Moreover, both RoBERTa and GPT2 make use of a Byte-Pair Encoding tokenizer.

For our investigation, we designed the experiment as follows. First, we derived simple sentences from the tuples by adding definite articles to the words, [CLS] at the beginning of the input and a period to signal the end of the sentence (e.g., [CLS] The tailor sewed the dress.). Then, we masked the candidate filler (*dress* in the example) and we computed the probability distribution of the entire model’s vocabulary for that position. The model typicality score is the probability assigned to the candidate filler, when the candidate filler is included in the model’s vocabulary. In case a word to be scored is not included in the vocabularies of all the models that we used, we decided to disregard its tuple and the respective typical/atypical counterpart. For this reason, the final results only take in consideration a subset of the original datasets, which varies from model to model. Additionally, we computed a baseline for each Transformer model, where the model is prevented from attending to the other tokens in the sequence when making predictions.

	Coverage	SDM	BERT-base(line)	BERT-large	ROBERTA-large	GPT-2 medium
Agent _{DTFit}	105/134	0.58	0.46 (0.1)	0.53	0.64	-
Patient _{DTFit}	323/402	0.62	0.59 (0.06)	0.64	0.64	0.63
Instrument _{DTFit}	31/100	0.58	0.52 (0.08)	0.53	0.5	0.5
Time _{DTFit}	89/100	0.58	0.63 (0.06)	0.64	0.66	0.66
Location _{DTFit}	115/150	0.65	0.72 (0.06)	0.71	0.73	0.74

Table 2: Spearman Correlation for the DTFit datasets.

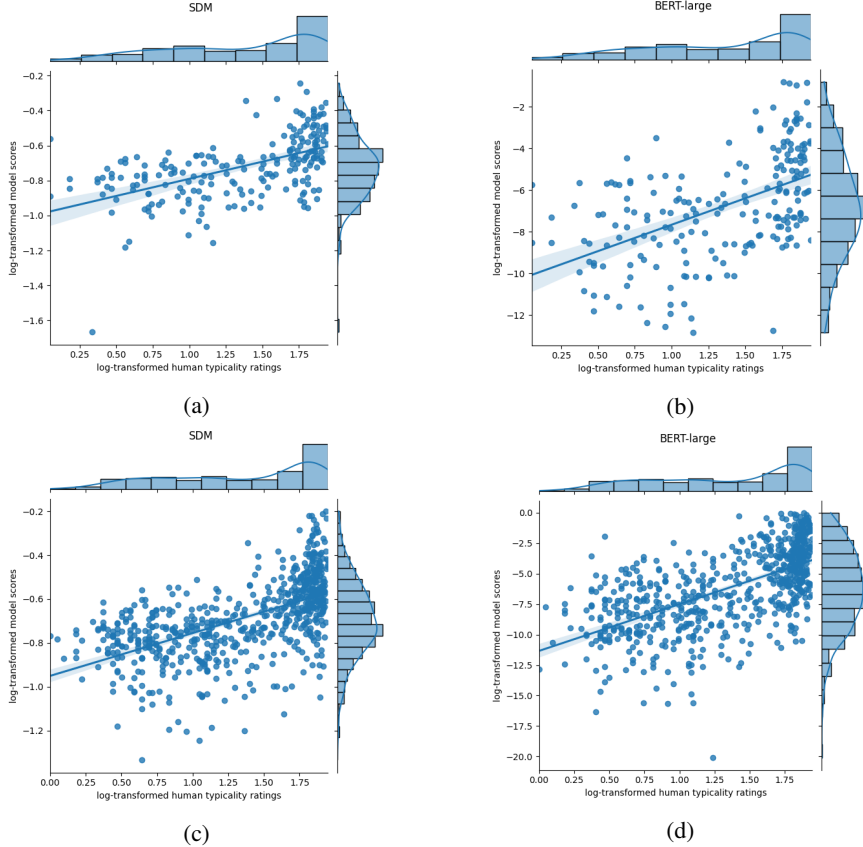


Figure 1: Correlation of elicited judgments and model-derived scores for **Agent**_{DTFit} (a-b) and **Patient**_{DTFit} (c-d) datasets.

4 Results and Analysis

In this section, we provide the results of the experiments on the DTFit datasets. Since the models cover different portions of the original tuples, we performed the evaluation over the common pairs.

Table 2 reports the correlation scores for all the five datasets.⁴ Values in brackets refer to the Spearman correlation obtained by the baseline. As the baseline scores are very similar across models, we reported the results only for BERT-base.

At a glance, we observe that both SDM and TLMS obtain quite strong correlations, going from 0.46 to a maximum of 0.74 across datasets and

⁴We do not computed GPT-2 scores for **Agent**_{DTFit}, as the model cannot make predictions based on context because the candidate filler occurs at the beginning of the sentence.

models. Specifically, we notice that TLMS tend to reach higher performances compared to the distributional approach. However, a marginally significant improvement of the correlations over SDM is obtained only for **Location**_{DTFit} ($p < 0.05$ for Locations, $p < 0.1$ for the other roles).⁵ This result is interesting, considering that SDM is trained on a really small corpus compared to TLMS (for instance, RoBERTa is trained on 160 GB of text). Another remark is that even if TLMS differ for architecture, training objective and data, BERT-large, RoBERTa and GPT-2 tend to achieve very similar performances, while correlation scores are lower for BERT-base.

As there is no significant difference between

⁵The p -value was computed with Fisher’s r -to- z transformation, one-tailed test.

SDM and TLMS results, we plotted an example of the relationship between the human ratings and the model-derived scores to provide a better picture of the models’ predictions. For visualization purposes, we applied a logarithmic transformation to the scores. For **Agent**_{DTFit}, we observe that SDM and BERT-large have a different trend. In the former (see Figure 1a), the majority of the points follow a roughly linear relationship, and there is a small variation around the regression line (with few outliers). On the contrary, BERT-large scores show more variance (Figure 1b). This trend is confirmed (even if it is less evident) for **Patient**_{DTFit}, where both SDM (Figure 1c) and BERT-large (Figure 1d) have a large amount of variance, and quite a few outliers. To verify these observations, we compared the sum of the BERT-large residuals with that of SDM (we first normalized the models’ scores with min-max scaling in order to make them comparable). For both subsets, the sum of residuals is higher for BERT-large than SDM, which is especially the case for **Agent**_{DTFit} (31.43 versus 17.85; 67.04 versus 63.47 for **Patient**_{DTFit}).

Finally, we also performed a *binary classification task* for **Agent**_{DTFit} and **Patient**_{DTFit}. In this case, we evaluated models on their ability to assign a higher score to the filler in the typical condition. As shown in Table 3 (left columns), the accuracy values are always high and the TLMS scores are comparable with the SDM ones.

5 Do Transformers Really Encode GEK?

The above results *prima facie* suggest that TLMS are able to model the dynamic interaction between the sentence elements to compute the typicality value of a candidate filler. However, analyzing the errors of the TLMS can be revealing of how they make their predictions.

Table 4 presents some of the **Patient**_{DTFit} pairs where BERT-base prefers the atypical filler. In all these cases, BERT simply seems to rely on frequent verb objects, without composing and integrating the verb expectations with information from other elements of the context (the agent in this case), which is a key aspect of human GEK and is reflected in the typicality judgments. However, we cannot make any claims about the event knowledge of TLMS from these examples alone, as only in some cases (such as *The cat drank the coffee*) the atypical tuples evoke events unlikely to take place in the real world (i.e., it may happen frequently that

	DTFit		Wang2018	
	Agent	Patient	Agent	Patient
SDM	.89	.91	.65	.66
BERT-base	.77	.85	.76	.63
BERT-large	.83	.89	.77	.65
ROBERTA-large	.89	.91	.76	.73
GPT-2 medium	-	.90	-	.64

Table 3: Accuracy in the binary classification task for DTFit (agent and patient roles) and Wang2018 datasets.

a chemist pours the juice, even if this is not a typical action for a chemist). To better understand if this can lead TLMS to make really implausible predictions, we carried out an additional experiment where we tested the models on a diagnostic dataset controlled for the frequency of the association between the verb and the filler. In this experiment, we also tried to address the question of whether TLMS rely more heavily on the local context when making predictions.

Furthermore, TLMS’ natural preference for what is more frequent could help them in the typicality task, as a typical event is often a frequent one. Their good performance could be due to the fact that they memorize frequent sequences during training. Therefore we tested TLMS on a different dataset, in which atypical but physically plausible events (e.g., *The cloth erased the cream*) are distinguished from atypical and implausible ones (e.g., *The cloth erased the house*). Frequency effects on performance should be alleviated in this setting, as both types of events in the dataset are atypical and, hence, rare. This task requires fine-grained knowledge of the properties of arguments, which is still an important component of GEK.

Additionally, different frequency variations in the training data could influence TLMS performance. Since the models’ knowledge of the world is mediated by language, it is likely that an argument filler may or may not be predicted depending on the frequency of the word chosen to refer to it. We investigated this issue by testing the models on another diagnostic dataset obtained by replacing typical fillers with low-frequency synonyms.

The last question we explored is whether TLMS can be influenced by the way statements of event typicality are syntactically expressed. So, we evaluated TLMS by feeding them with sentences encoding typical events with a transformed and more complex syntactic form than the one used in the DTFit experiments.

Tuple	Expected	Preferred
mason mix __	cement (H=6.65, M=-8.41)	soup (H=1.95, M=-5.54)
climber climb __	rock (H=6.8, M=-5.29)	staircase (H=5.6, M=-4.05)
blacksmith pour __	metal (H=6.5, M=-4.03)	wine (H=1.6, M=-1.6)
chemist pour __	compound (H=6.25, M=-8.4)	juice (H=2.75, M=-5.18)
cat drink __	milk (H=5.6, M=-2.89)	coffee (H=1.45, M=-3.65)

Table 4: Examples of errors (BERT-base, Patient_{DTFit}). H= Human score, M=Model’s log probability.

I. TLMS seem to prefer frequent collocations, but only when they are plausible.

Errors reported in Table 4 suggest the tendency of TLMS to predict frequent complements of the verbs, irrespective of whether they are coherent with the rest of the tuple. We questioned to what extent salient local word co-occurrences make the models “blind” to the rest of the context and thus compromise the plausibility of their predictions. To investigate this behavior, we generated a new diagnostic dataset. The dataset is a small (31 pairs) subset of Patient_{DTFit}, where the atypical filler in each pair was replaced with another noun that has a very strong association with the verb in the tuple. We computed the association between the verb and its direct objects using Local Mutual Information (LMI) (Evert, 2008). Since LMI is computed by multiplying the Pointwise Mutual Information and the frequency of the two words in a grammatical relation, it assigns higher values to combinations that are both common and informative. We chose the new atypical fillers among the words with the highest LMIs. We chose words that give rise to odd events when integrated with the rest of the context. To approximate the word distributions encountered in the training data, we extracted LMI values from a 2018 dump of English Wikipedia and we evaluated only the BERT model (base and large) on the new dataset, as Wikipedia is a considerable part of the training only for this model. Examples of the new test pairs are the following: *The terrorist released the **hostage/ album***, *The truck hit the **car/ ball***, *The soldier heard the **command/ case***.

To evaluate BERT performance, we computed the accuracy scores on the diagnostics dataset in the same way as in the main experiment (binary classification task). Results show that the models generally assign low probabilities to atypical fillers. They choose the atypical event in some cases (9 in BERT-base, 6 in large), but mainly when the contrast between the atypical event and our expectations is less evident (*The smuggler sold the **property*** is preferred to **weapon**, *The soldier throw the **ball*** is preferred to **bomb**).

As already observed in the main experiment, BERT seems to be able to look beyond salient local associations and build representations of global events flexibly. However, this issue should be further explored for the other roles as well. For instance, given the sentence *The engineer completed the project in the __*, the models must consider more contextual elements to make the correct prediction.

On the other hand, even if SDM design aims at capturing this aspect of GEK, the manipulations we made in this dataset cause a drop in the model performance (14 pairs out of 31 are classified wrongly). This drop is probably due to aspects of the implementation such as data sparsity. Specifically, if there are no events in which the subject occurs with a direct object, the prototype of the patient is built only from the verb’s most associated patients, disregarding the fact they are implausible given the whole context.

II. TLMS know more about what is typical than what is possible.

The use of typicality datasets such as DTFit for the estimation of the models’ GEK has some limitations. TLMS’ ability to reproduce combinations encountered frequently during training could be the reason for high performances in the typicality task, since what is most typical often occurs most frequently. However, GEK is not just memory of exemplars, but it requires fine-grained knowledge of the properties of objects and it involves reasoning processes such as abstraction and comparison between objects and prototypical concepts.

To evaluate TLMS on a setting where frequency variations in the training corpus have a minor impact, we used the dataset realized by Wang et al. (2018) (henceforth, Wang2018). This dataset represents a benchmark for the task of semantic physical plausibility (Bagherinezhad et al., 2016), that is, distinguishing an atypical but physically plausible event such as *The student climbed the ship* from an atypical and physically implausible one such as *The student climbed the water*. The dataset contains agent-verb-patient (SVO) triples divided

into plausible and implausible. From the original dataset, which contains 1,540 plausible and 1,540 implausible triples, we derived two subsets containing pairs of triples differing either for the agent or for the patient role filler (obtaining 222 and 394 pairs respectively).

Table 3 reports the resulting accuracy values. In general, the models’ scores are lower than in the typicality task (min. 0.64, max. 0.77), and in some cases they are not much higher than random performance. Moreover, in many cases the models could be facilitated by the existence of an association between the plausible filler and the verb of the event, as in *The ant **built** the **wall*** and in *The chair **absorbed** the **water***. Nevertheless, the results demonstrate that the notion of plausibility is harder to model compared to typicality, and invite caution when making claims about TLMS world and event knowledge. In fact, the results suggest that even if it were true that TLMS develop some generalization skills from training, they still miss many predictions about possible events, which instead humans easily make on the basis of their common-sense knowledge.

This dataset is also difficult for SDM, which obtains scores lower than those of the TLMS (0.65 for Agent and 0.66 for Patient). Even if SDM should be better at reproducing generalization through the construction of prototypical fillers, the model’s distributional representations seem to fail to capture the specific properties that are relevant for the dataset items, namely physical properties of objects (liquid-solid, large-small, etc.). The lack of such properties constitutes a limitation of distributional models of word meaning based on text data only, which is why, in previous studies, world knowledge was explicitly injected into the models for the physical plausibility task (Wang et al., 2018).

III. TLMS do not extend fit judgments to low frequency synonyms. To test whether TLMS consider an entity more or less likely to take part in an event depending on the word used to refer to that entity, we evaluated them on a new diagnostic dataset of 39 pairs, generated from a subset of **Patient**_{D_TFit}. In this setting, the typical filler in each pair was replaced with a low-frequency word that is semantically related to the original one. To choose an appropriate substitute, we first extracted a set of synonyms according to two lexical resources (WordNet, Lexico.com). Then, we picked a word that 1) is less frequent than the original filler and

2) has a frequency lower than 300,000. For the same reasons described in the first additional experiment, we extracted statistics from a 2018 dump of English Wikipedia and evaluated only BERT on the new dataset. Examples of substitutions are the following: *The botanist examined the **plant** → **flora***, *The waiter cleared the **restaurant** → **tavern***, *The veterinarian examined the **dog** → **puppy***. It is interesting to observe that these variations pose serious difficulties to the models, as their accuracy on the diagnostics dataset is close or lower to the random level (BERT-base: 0.37, BERT-large: 0.53). For example, BERT considers *The terrorist released the **captive*** as less probable than *The terrorist released the **/book***, and the same occurs for *The mother prepared the **provisions/gun***, and *The carver built the **bust/house***.

These results cast doubts that current TLMS can constitute plausible models of event knowledge: they tend to reproduce the patterns that are frequently observed in the data, and their good performance is disrupted once these are replaced with semantically equivalent, but less frequent ones. This means that they lack the abstract semantic knowledge of human subjects, whose predictions are more flexible thanks to inference mechanisms such as generalization to concepts sharing semantic features. At least in principle, models aiming at building prototypes of ideal role fillers (such as the distributional models of thematic fit) are more cognitively realistic, since they are less dependent on specific words. However, they may still show sub-optimal performance in this diagnostic dataset as they are based on the quality of the distributional representations, which is lower for words that have low frequency in corpora. This is confirmed by the performance of SDM on the dataset (the accuracy is 0.51).

	transitive	cleft	wh-interrogative
Agent _{D_TFit}	0.64	-0.13	0.62
Patient _{D_TFit}	0.64	0.26	0.51
Instrument _{D_TFit}	0.5	0.10	0.6
Time _{D_TFit}	0.66	0.33	0.64
Location _{D_TFit}	0.73	0.67	0.73

Table 5: Spearman Correlation for D_TFit datasets using RoBERTa-large and input sentences with different word orders.

IV. TLMS can be influenced by the surface structure of sentences Finally, we analyzed to what extent TLMS’ ability to predict the fit of a

word in a given role arises from the observation of recurrent word order patterns during pre-training (e.g., the fact that an actor’s award-winning event is canonically expressed with active sentences, in which *award* follows the words *actor* and *won*), rather than being based on a deep understanding of the semantic relations between the words.

To explore this issue, we modified DTFFit tuples to create two different versions of the dataset, each with examples of a syntactic construction different from the English canonical word order. Specifically, we experimented with **cleft** (*It was the award that the actor won, It was on the ring that the boxer delivered the punch*) and **wh-interrogative** sentences (*Which award did the actor win?, On which ring did the boxer deliver the punch?*).

We evaluated this new set of sentences using RoBERTa-large (cf. Table 5). We observe that the model is not particularly affected by the interrogative structure. Conversely, the model suffers from the cleft construction for all semantic roles except for Location ($\rho=0.67$). If we ask the model to generate the most likely words to appear in that position, we observe that word predictions in the new construction are more general and less dependent on the GEK associated with the other words in the sentence, proving that TLMS are affected by the surface syntactic shape of the linguistic input, since the cleft construction is less frequent and presents a less canonical word order. For instance, given the sentence *It was with the [MASK] that the guard opened the door*, RoBERTa generates the following possible fillers: *gun* ($P=0.044$), *crowd* ($P=0.020$), *sword* ($P=0.016$), and then *key* ($P=0.016$), while in the active sentence *key* is correctly predicted as the most probable word ($P=0.22$). In this specific case, it seems that the model only looks at the word nearby (*guard*) to make a prediction, disregarding the entire context. Generally, the agent role shows the worst results, obtaining -0.13 . Note that SDM is not affected by these variations by design, since its predictions are based on semantic roles derived from the syntactic analysis of the sentence, which is explicitly provided to the model.

6 Conclusions

In this paper, we tested Transformer-based Language Models on tasks related to Generalized Event Knowledge. In the main experiment, we evaluated their ability to model event typicality, that is, discern typical from atypical events, on a dataset

designed for this task, DTFFit. Results show that TLMS scores positively correlate with human judgments. However, they do not significantly outperform the distributional prototype-based model (SDM) that we selected for comparison. This confirms the ability of SDM to dynamically update the semantic representation of a sentence, which was recently shown for the challenging task of logical metonymy interpretation (Rambelli et al., 2020).

However, we decided to go beyond the simple evaluation against human judgments. We carried out several additional small-scale experiments with the specific aim to understand which factors could affect the predictions of TLMS. The results suggest that models are often too dependent on what they observe during training and lack some key aspects of human event knowledge. In particular, we observed that, in some cases, they are unable to compose all elements of the input to make predictions, and they tend to rely more on salient local associations between words. However, further analysis is needed. Secondly, their performance drop on the physical plausibility task, which requires the ability to infer physical properties necessary for an object to participate in a given event. Lastly, their probabilities are dependent on the specific words that have to be predicted rather than on their meaning, and on the canonical word order in which these words tend to occur. Noticeably, even a distributional model of event knowledge (SDM) showed similar limitations, generally likely to be due to data sparsity and inherent limitations of distributional representations obtained from text data.

To conclude, we believe that the experiments we reported are the first step towards a deep investigation of “how general” is the Generalized Event Knowledge in computational models. Future work might include the creation of a larger version of our diagnostic datasets, in order to make available to NLP researchers a more robust benchmark for tasks related to Generalized Event Knowledge.

Acknowledgements

This work, carried out within the Institut Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government, managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). We thank the anonymous reviewers for their insightful feedback.

References

- Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2016. Are Elephants Bigger than Butterflies? Reasoning about Sizes of Objects. In *Proceedings of AAAI*.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.
- Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):673–721.
- Klinton Bicknell, Jeffrey L Elman, Mary Hare, Ken McRae, and Marta Kutas. 2010. Effects of Event Knowledge in Processing Verbal Arguments. *Journal of Memory and Language*, 63(4):489–505.
- Emmanuele Chersoni, Philippe Blache, and Alessandro Lenci. 2016. Towards a Distributional Model of Semantic Complexity. In *Proceedings of the COLING Workshop on Computational Linguistics for Linguistic Complexity*.
- Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Logical Metonymy in a Distributional Model of Sentence Comprehension. In *Proceedings of *SEM*.
- Emmanuele Chersoni, Ludovica Pannitto, Enrico Santus, Alessandro Lenci, and Chu-Ren Huang. 2020. Are Word Embeddings Really a Bad Fit for the Estimation of Thematic Fit? In *Proceedings of LREC*.
- Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2021. Not All Arguments Are Processed Equally: A Distributional Model of Argument Complexity. *Language Resources and Evaluation*, pages 1–28.
- Emmanuele Chersoni, Enrico Santus, Ludovica Pannitto, Alessandro Lenci, Philippe Blache, and C-R Huang. 2019. A Structured Distributional Model of Sentence Meaning and Processing. *Natural Language Engineering*, 25(4):483–502.
- Andy Clark. 2013. Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *Behavioral and Brain Sciences*, 36(3):181–204.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.
- Katrin Erk. 2007. A Simple, Similarity-Based Model for Selectional Preferences. In *Proceedings of ACL*.
- Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics*, 36(4):723–763.
- Stefan Evert. 2008. Corpora and Collocations. *Corpus Linguistics. An International Handbook*, 2:1212–1248.
- Todd R Ferretti, Ken McRae, and Andrea Hatherell. 2001. Integrating Verbs, Situation Schemas, and Thematic Role Concepts. *Journal of Memory and Language*, 44(4):516–547.
- Yoav Goldberg. 2019. Assessing BERT’s Syntactic Abilities. *arXiv preprint arXiv:1901.05287*.
- Clayton Greenberg, Vera Demberg, and Asad Sayeed. 2015a. Verb Polysemy and Frequency Effects in Thematic Fit Modeling. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.
- Clayton Greenberg, Asad B Sayeed, and Vera Demberg. 2015b. Improving Unsupervised Vector-Space Thematic Fit Evaluation via Role-Filler Prototype Clustering. In *Proceedings of HLT-NAACL*.
- Xudong Hong, Asad Sayeed, and Vera Demberg. 2018. Learning Distributed Event Representations with a Multi-task Approach. In *Proceedings of *SEM*.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn About the Structure of Language? In *Proceedings of ACL*.
- Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Linguistically-informed Specificity and Semantic Plausibility for Dialogue Generation. In *Proceedings of NAACL*.
- Geoffrey Neil Leech. 1992. 100 Million Words of English: The British National Corpus (BNC). *Language Research*, 28.
- Alessandro Lenci. 2011. Composing and Updating Verb Argument Expectations: A Distributional Semantic Model. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*.
- 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. *arXiv preprint arXiv:1907.11692*.
- Yuval Marton and Asad Sayeed. 2021. Thematic Fit Bits: Annotation Quality and Quantity for Event Participant Representation. *arXiv preprint arXiv:2105.06097*.
- Kazunaga Matsuki, Tracy Chow, Mary Hare, Jeffrey L Elman, Christoph Scheepers, and Ken McRae. 2011. Event-Based Plausibility Immediately Influences On-Line Language Comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(4):913.

- Ken McRae and Kazunaga Matsuki. 2009. People Use their Knowledge of Common Events to Understand Language, and Do So as Quickly as Possible. *Language and Linguistics Compass*, 3(6):1417–1429.
- Ken McRae, Michael Spivey-Knowlton, and Michael Tanenhaus. 1998. Modeling the Influence of Thematic Fit (and other Constraints) in On-Line Sentence Comprehension. *Journal of Memory and Language*, 38(3):283–312.
- Eleni Metheniti, Tim Van de Cruys, and Nabil Hathout. 2020. How Relevant Are Selectional Preferences for Transformer-based Language Models? In *Proceedings of COLING*.
- Ulrike Padó. 2007. *The Integration of Syntax and Semantic Plausibility in a Wide-Coverage Model of Human Sentence Processing*. Ph.D. thesis.
- Ian Porada, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. Can a Gorilla Ride a Camel? Learning Semantic Plausibility from Text. In *Proceedings of the EMNLP Workshop on Commonsense Inference in Natural Language Processing*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.
- Giulia Rambelli, Emmanuele Chersoni, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2020. Comparing Probabilistic, Distributional and Transformer-Based Models on Logical Metonymy Interpretation. In *Proceedings of ACL-IJCNLP*.
- Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Measuring Thematic Fit with Distributional Feature Overlap. In *Proceedings of EMNLP*.
- Asad Sayeed and Vera Demberg. 2014. Combining Unsupervised Syntactic and Semantic Models of Thematic Fit. In *Proceedings of CLIC.it*.
- Asad Sayeed, Vera Demberg, and Pavel Shkadzko. 2015. An Exploration of Semantic Features in an Unsupervised Thematic Fit Evaluation Framework. *Italian Journal of Computational Linguistics*, 1(1).
- Asad Sayeed, Clayton Greenberg, and Vera Demberg. 2016. Thematic Fit Evaluation: An Aspect of Selectional Preferences. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*.
- Ottokar Tilk, Vera Demberg, Asad Sayeed, Dietrich Klakow, and Stefan Thater. 2016. Event Participant Modelling with Neural Networks. In *Proceedings of EMNLP*.
- Bram Vandekerckhove, Dominiek Sandra, and Walter Daelemans. 2009. A Robust and Extensible Exemplar-based Model of Thematic Fit. In *Proceedings of EACL*.
- Paolo Vassallo, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2018. Event Knowledge in Sentence Processing: A New Dataset for the Evaluation of Argument Typicality. In *Proceedings of the LREC Workshop on Linguistic and Neuro-Cognitive Resources*.
- Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling Semantic Plausibility by Injecting World Knowledge. In *Proceedings of NAACL*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.
- Hongming Zhang, Hantian Ding, and Yangqiu Song. 2019. SP-10K: A Large-scale Evaluation Set for Selectional Preference Acquisition. In *Proceedings of ACL*.