Understanding and Estimating Pseudo-Log-Likelihood for Zero-Shot Fact Extraction with Masked Language Models

Riley Capshaw
riley.capshaw@liu.se
Linköping University
Linköping, Sweden

Eva Blomqvist
eva.blomqvist@liu.se
Linköping University
Linköping, Sweden

ABSTRACT

Knowledge Graph (KG) construction is a cumbersome task when performed manually. Instead, KGs are commonly extracted from existing knowledge sources, such as natural language text. With the emergence of large language models (LLMs), machine reading has taken a leap forward, but still several challenges remain. In particular, the task of extracting accurate statements from text is still an open research problem, and the work presented in this paper focuses on an important aspect of document-level fact extraction as a step toward solving that problem. To allow for flexibility in terms of input representations and emergence of new unseen terminology, we set our experiments in a zero-shot setting using masked language models (MLMs), rather than probing LLMs for facts seen in training. We first explore the correlation between the pseudo-log-likelihood (PLL) scores for various statements and their factuality. For statements derived from the DocRED data set, out-of-the-box MLMs will generally assign higher PLL scores to them if they are supported by some document. This correlation stayed consistent even when taking the influence of various high-level features into account. Since PLL cannot be calculated for non-token inputs like soft prompts, we additionally use these results to search for a suitable approximation to PLL with similar behavior. We examine four similarity measures for vectors and probability distributions, and find that of them, cosine similarity has the highest correlation to PLL. Finally, we outline how the knowledge gained from this explorative study can be used in future work on zero-shot document-level fact extraction for KG generation.

KEYWORDS

Knowledge Graphs, Masked Language Models, Machine Reading, Document-level Relation Extraction

ACM Reference Format:
Riley Capshaw and Eva Blomqvist. 2023. Understanding and Estimating Pseudo-Log-Likelihood for Zero-Shot Fact Extraction with Masked Language Models. In *Proceedings of The Twelfth International Joint Conference on Knowledge Graphs (IJCKG 2023).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

IJCKG 2023, December 08–09, 2023, Tokyo, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

A Knowledge Graph (KG) is a labeled, directed graph where the nodes are entities of interest and all edges represent the existence of a typed relation between two of those entities. A KG can also be viewed as a set of subject-predicate-object triples representing the edges of the graph. KGs are used for many tasks, such as information retrieval and storage, and can be incorporated into larger systems for tasks like recommendation and question answering. KGs are often structured according to an ontology to provide added semantics to the entities and relations, thus making up a knowledge base (KB) that can be queried and reasoned over. However, constructing such a KG is not an easy task. Commonly, the knowledge to be included in the KG already exists in an unstructured form as natural language text. In such cases, KG construction becomes a task of knowledge extraction from such sources, rather than gathering and constructing the KG from scratch. Automating this task is particularly important when constructing KGs for certain (narrow) knowledge domains, such as an enterprise KG or a KG for a very specific industry domain, where the use of other techniques, such as crowdsourcing, is not possible.

In this work, we explore the hypothesis that this knowledge extraction can be supported by masked language models (MLMs), a type of large language model (LLM) trained using a masked language objective [5]. The role of these MLMs would be to enable fact extraction from unstructured text documents. In order to minimize information leakage [6], we assert that an effective use of MLMs for this fact extraction cannot include specialized pre-training or fine-tuning. For instance, consider a scenario where a MLM extracts facts from news stories about election results in order to model them. To support fact checking those stories, the information in each story must be modeled correctly but separately. Given that multiple stories would discuss the same named entities and the same events, fine-tuning a MLM on any story risks leaking information about the entities it contains, yielding incorrect extractions for other stories. In this scenario, it should be clear that we are not concerned with the ground truth of a statement; rather, for the scope of this paper, we consider a statement to be true if it is supported by the document it was extracted from.

However, this fact extraction task as described here does not have a straightforward solution, especially since we wish to minimize information leakage by extracting facts directly from the underlying documents. In this paper, we attempt to craft a solution by looking at the problem from a language modeling point of view. Wang and Cho [20] show how the MLM BERT [5] can be used as a traditional language model (for generating text) through the use of pseudo-log-likelihood (PLL) scores. Salazar et al. [17] take this idea further and demonstrate the relationship between PLL scores

and linguistic acceptability [4] for short statements, better enabling the application of out-of-the-box MLMs to downstream tasks like translation. In order for MLMs to be applied to our problem in a robust way, we believe that three criteria must be met:

(1) There should be a strong correlation between scores and document support.
(2) The scoring mechanism should be able to handle out-of-vocabulary (OOV) inputs.
(3) Information leakage from the MLM's training data should not influence the scores.

We demonstrate criterion 1 in Section 4.1 for PLL. Criterion 2 is based on the idea that if the MLMs should not be fine-tuned on the content of the documents, then the only way to further contextualize the predictions is through the input to the MLM. Thus, we believe that techniques like soft or continuous prompting [2] are necessary to achieve a reliable level of accuracy. However, these rely on OOV vector inputs which, as discussed further in Section 3.5, prevents PLL from being used in this setting. Therefore, we additionally search for suitable methods of approximating PLL which do work for soft prompts. Criterion 3 is left to future work, but is nonetheless important to mention. The experiments in this paper mostly rely on the information already learned by the selected MLMs to perform the scoring, and should be viewed in that light. Instead, we discuss in Section 5.2 one idea of how to make a MLM "blind" to named entities present in statements to remove some bias. For now, we focus on *understanding* and *estimating* PLL in its most common form for zero-shot applications.

With these challenges in mind, we perform an exploratory study in order to answer the following research questions:

RQ1: Can PLL scores be used to distinguish supported (factual w.r.t. some document) statements from unsupported statements?
RQ2: Given a statement which corresponds to a relation between two entities, can the PLL score for that statement indicate whether the types of those entities are consistent with the semantics of the relation?
RQ3: Does the fact that some entity mentions span multiple tokens have an effect on PLL scores?
RQ4: In situations where PLL cannot be calculated, how well can vector or probability similarity measures estimate it?

We consider the following points to be our main contributions with this work:

- An augmentation of the DocRED data set into a collection of positive and negative samples to capture document support (or a lack thereof).
- A better understanding of the relationship which exists between PLL scores for statements and their factuality defined as document support.
- An analysis of how certain statement characteristics impact PLL scores, such as adherence to the semantics of an underlying relation, or whether entity mentions in the statement spanned multiple tokens.
- Suggestions for alternatives to PLL with similar computational demands which can be used when token IDs are unavailable, such as when using soft-prompts.

We present these contributions with the following structure. First, we position our research questions within related work in Section 2. Then we discuss our approach for answering the research questions in Section 3, including a description of our augmentations to the DocRED data set. In section 4 we describe and evaluate our results. We discuss the limitations of these experiments as well as speculations about the implications of their results in Section 5, and offer some concluding remarks in Section 6.

2 RELATED WORK

The largest body of related work focuses on fact extraction by treating MLMs as knowledge bases in their own right [9, 16]. Some approaches use hand-crafted cloze-style (fill-in-the-blank) prompts to explore which token an MLM predicts is missing [16], while others automatically find the best prompt to maximize the score for some downstream task, such as knowledge base completion [1]. To simplify our exploration, we followed the former methodology and used manually-written prompts in our experiments (see Section 3.1).

For both prompting methodologies, the MLM performing the infilling can either be used out-of-the-box in a zero-shot fashion [8], or be fine-tuned to enhance results over a particular data set [7]. Given that MLMs capture various types of biases from their training data [10, 12, 15], we avoid any fine-tuning to remove the possibility that training on one document will change the predictions regarding statements about another document, such as in the scenario described previously around contradictory news articles.

For our prompts, we do not ask a MLM to fill in the two blanks. Instead, we automatically fill in both blanks with known entities and ask the MLM to score the resulting statement. However, MLMs are generally not trained to output actual likelihood values for a given text, unlike traditional unidirectional language models. To circumvent this limitation, both Shin [18] and Wang and Cho [20] show how to calculate pseudo-log-likelihood (PLL) scores from the output logits of a MLM as follows. Let M be a MLM and $\mathbf{V}^M$ be its vocabulary of size W. $\mathbf{V}^M$ maps an input token t to an index k such that $\mathbf{V}^M_t = k$. Let $\mathbf{S}$ be a sentence of length N and $\mathbf{S}_{\setminus i}$ be the same sentence but with the token t at index i replaced with a mask token. The output of $\mathrm{M}(\mathbf{S})$ is a matrix $\mathbf{M}^{\mathbf{S}}$ of size $N + 1 \times W$, where each row corresponds to the output logits for every $t \in \mathbf{S}$ and each column corresponds to some word in the vocabulary[1]. The pseudo-likelihood of t can then be seen as the likelihood of replacing the masked token in $\mathbf{S}_{\setminus i}$ with t if the replacement were randomly sampled from $\mathbf{V}^M$ weighted by $\mathbf{M}_i^{\mathbf{S}_{\setminus i}}$:

$$P_M(t \mid \mathbf{S}_{\setminus i}) = \mathbf{M}^{\mathbf{S}_{\setminus i}}_{i,k} \,. \tag{1}$$

PLL is then calculated by taking the average of the logarithm of the values for every $t \in \mathbf{S}$:

$$\mathrm{PLL}_M(\mathbf{S}) = \frac{1}{N} \sum_{i=1}^{N} \log\left(P_M(t \mid \mathbf{S}_{\setminus i})\right) \,. \tag{2}$$

Salazar et al. [17] demonstrate a strong correlation between PLL and the concept of linguistic acceptability [4]. They argue that this

[1] There are two important assumptions to note here. The first is that a `[CLS]` token is prepended to the sentence, which is not used in the score calculations. Second is that the MLM is in "masked language modelling" mode and outputting logits for every token in its vocabulary.

correlation enables MLMs to be useful across a wide variety of tasks without fine tuning, and exemplify this by using PLL to improve existing methods for scoring translations. Bias detection is another task for which PLL has been used effectively [12, 15]. We take inspiration from this and designed our experiments to answer RQ1 around the idea that a MLM is inherently biased by its training data and will likely find statements to be more acceptable if they are factual with regard to that training data.

More recently, Kauf and Ivanova [11] point out several shortcomings of MLM-based PLL scores, including biases toward multi-token words[2] and statement length, then present two alternative formulations for PLL. One yields scores more like those obtained from traditional (autoregressive) left-to-right language models, while the other performs whole-word masking when scoring multi-token words. While we do not use their alternatives in our study, we explore similar effects (see Section 4.3) with the differing goal of understanding which aspects influencing PLL can be controlled.

3 APPROACH

In order to answer the research questions, we need a way to empirically study measures such as PLL on a dataset containing already known facts. In this section we describe the outline of the work, including the dataset used. All data and code are available at https://github.com/LiUSemWeb/understanding-pll.

3.1 Data

The data used in the following experiments was derived from the development portion of the Document-Level Relation Extraction Data set (DocRED) [21]. DocRED is based on Wikidata [19] and includes 1000 documents and 96 unique relation types, and is intended for measuring the accuracy of relation-extraction systems in a challenging setting where some relations can only be concluded by reasoning over multiple sentences. We do not directly seek to solve the task presented by DocRED. Instead, we use it as a source of document-supported facts that can be easily represented as consistent short statements. We use these short statements to analyze the behavior of MLM scoring measures by contrasting the scores for statements which are supported by DocRED with the scores for those which are unsupported but mention the same set of entities.

To generate these statements, we wrote a short fill-in-the-blanks prompt for each relation. While a prompt for any given relation can be written in many different ways, each yielding different PLL scores, we felt that keeping the prompts short and concise was sufficient to capture the patterns we wanted to examine given the large number of negative samples we generate. For example, the relation P17 "country" was converted to "?x is located in the country of ?y." Then, for every unique pair of entity mentions in a given document, the two variables were populated to generate one statement. For document 127, this yields "Florida is in the country of United States" as a supported statement and "Barry University is in the country of Florida" as an unsupported statement, among many others. Note that five documents were excluded due to poorly formatted mentions, two relations were excluded due to accidentally identical prompts, and one relation was excluded for having a prompt which partially overlapped with another when populated. This yielded a final data set spanning 995 documents, 93 unique relation types, 11,577 supported statements, and 42,315,885 unsupported statements. For the graphs in Section 4, we uniformly sampled approximately 1% of the unsupported statements, but all conclusions drawn apply equally to the full data. Additionally, the number of supported statements varies slightly between language models, but the variation is too small to affect our conclusions. We believe this variation is due to slight differences in tokenizers, resulting in some statements being tokenized identically for some MLMs. Where appropriate, all figures include the number of statements per analyzed category.

[2]Note that in their work, out of vocabulary refers to any word which the tokenizer decomposes into multiple tokens, such as 'tokenizer' becoming 'token' and '##izer'. We instead refer to these as multi-token words and use out of vocabulary to refer to input vectors which do not map to any token.

3.2 Assessing Support with PLL

For these experiments, we chose to gather scores with the base and large variants of both BERT [5] and RoBERTa [13], the two most-commonly used MLMs in PLL-based literature. As discussed in Section 2, we used PLL to score every generated statement in the data. We then plotted the kernel density estimates of the scores for statements which have particular attributes. These allow us to visually and numerically analyze the score distributions and identify any significant differences between them. For significance, we used the two-sided Kolmogorov–Smirnov (K-S) test for goodness of fit, with the null hypothesis that the two distributions in question are identical. Rejecting the null hypothesis in this case is strong support that the two distributions are different and that the feature which defines the split (such as the presence of multi-token entities described in Section 3.3) has a meaningful effect on the scores. Where we make such a claim in Sections 4.1, 4.2 and 4.3, the tests yielded extremely small p-values, well below 0.0005, and so were not directly reported.

3.3 Multi-token Entities

In our prior experiments with BERT [3], statements which included long strings of non-English words were scored higher than initially expected due to how the BERT tokenizer works, with longer, uncommon words being broken down into sub-word tokens. Kauf and Ivanova [11] reported this effect as well in their experiments. While instances of these tokens are uncommon in the overall data set, they tend to be extremely common when appearing together as a sequence, which makes it easy for the masked language modeling learning objective to predict missing tokens if only one is masked at a time. This in turn means that pseudo-likelihood often scores a token based on the very narrow context of its immediate neighbors, rather than a statement as a whole. Take the unsupported statement "École nationale supérieure des Beaux - Arts was born in Paris." BERT assigned a pseudo-likelihood of 0.998 or higher to all nine tokens representing the mention "École nationale supérieure des Beaux - Arts", and an overall PLL score of -0.54 to the statement (a surprisingly high value, as illustrated later in Figure 1). To get a better idea of the overall effect this has on PLL scores, we extend the data by labeling every statement with whether it contains at least one multi-token entity (MTE). In this experiment, MTEs refer to any entity mention which contains multiple tokens, without regard for the occurrence of multi-token words.

3.4 Domain and Range restrictions

Another aspect of linguistic acceptability which we identified as possibly impacting PLL scores was the general concept of domain and range of a relation. For instance, while the syntax of the statement from the prior section is correct, its semantics are inconsistent with those of the underlying relation, since the subject should not be a school or institution (ORG according to DocRED). DocRED includes seven entity types, so we were able to mine general domain and range restrictions for each relation. Therefore, we examine the relationship between PLL scores and semantic correctness by additionally labeling every statement as being either "accepted" or "rejected" by these restrictions.

3.5 Estimating PLL

Some techniques, e.g. soft prompting, feed vectors into a MLM which do not correspond to any token from the vocabulary. Lv et al. [14] show that the use of soft prompts can enhance the performance of zero-shot entity recognition and relation extraction. However, PLL requires all inputs to the MLM to have token IDs, so it can no longer be calculated in these scenarios. In order to apply these techniques to a zero-shot fact-extraction scenario, we need a way to estimate PLL.

It is important to note that PLL can already be estimated for any input if a given MLM is fine-tuned to output approximate PLL scores directly [17]. Unfortunately, this relies on well-selected sets of sentences to calculate the ground truth from, which may not sample well from the set of "less correct" sentences, meaning that the method will be further biased by both the fine-tuning corpus and the negative sampling method. We seek to find a method for estimating these scores directly that does not have such a limitation, so any method which required fine-tuning any part of the model, even an external probing layer, was excluded.

An alternative formulation for pseudo-likelihood is as follows:

$$\mathbf{M}_i^{\mathrm{S}\setminus i} \cdot \mathbf{1}_k \,, \tag{3}$$

where $\mathbf{1}_k$ is a W-length one-hot vector with element k set to 1. In this sense, pseudo-likelihood can be seen as a similarity measure between those two vectors, and is maximal only when $\mathbf{M}_i^{\mathrm{S}\setminus i} = \mathbf{1}_k$ due to the use of softmax. In a setting which uses soft prompts, however, k may be undefined for some inputs. Instead, to approximate $\mathbf{1}_k$, we substitute it with $\mathbf{M}_i^{\mathrm{S}} = P_M(t \mid \mathrm{S})$. The use of the dot product in Equation 3 inspired our choice of cosine similarity as one of the evaluated measures. The other three measures were mean squared deviation, Jensen-Shannon divergence, and Hellinger distance, all of which measure some form of similarity between (discrete) probability distributions. These fit well if we interpret the MLM's output logits as probability distributions over the model's vocabulary. By using these measures, we are generally treating $\mathbf{M}_i^{\mathrm{S}}$ as an expected distribution and $\mathbf{M}_i^{\mathrm{S}\setminus i}$ as a prediction of it. Where appropriate, we normalize scores to be between 0 and 1, with 1 representing perfect similarity.

Cosine Similarity (CS) is a measure of the angle between two vectors, and is often used in natural language processing tasks involving comparing vector representations of concepts. Given two vectors $\mathbf{a}$ and $\mathbf{b}$, their cosine similarity is defined as

$$s_{\cos}(\mathbf{a}, \mathbf{b}) := \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}. \tag{4}$$

Mean squared deviation (MSD) is a measure of the quality of a predictor for a given distribution, and is often used in loss functions (as mean squared error). Given two discrete probability vectors $\mathbf{a}$ and $\mathbf{b}$, MSD can be used to define their similarity by assuming that $\mathbf{b}$ acts as a predictor of $\mathbf{a}$:

$$s_{\mathrm{msd}}(\mathbf{a}, \mathbf{b}) := 1 - \frac{1}{W} \sum_{i=0}^{W} (\mathbf{a}_i - \mathbf{b}_i)^2 \,. \tag{5}$$

Jensen-Shannon divergence (JSD) is a symmetric alternative to Kullback-Leibler (KL) divergence for measuring the similarity between discrete probability distributions. For the same probability vectors $\mathbf{a}$ and $\mathbf{b}$ as before, JSD calculates a midpoint mixture distribution $\mathbf{c}$:

$$\mathbf{c} = \frac{1}{2}(\mathbf{a} + \mathbf{b}) \,, \tag{6}$$

then uses that to find the average KL divergence from $\mathbf{c}$:

$$s_{\mathrm{jsd}}(\mathbf{a}, \mathbf{b}) := \frac{1}{2} \left(\mathrm{KL}(\mathbf{a} \| \mathbf{c}) + \mathrm{KL}(\mathbf{b} \| \mathbf{c})\right) \,. \tag{7}$$

KL divergence in this setting is calculated as:

$$\mathrm{KL}(\mathbf{a} \| \mathbf{b}) := \sum_{i=0}^{W} \mathbf{a}_i \log\left(\frac{\mathbf{a}_i}{\mathbf{b}_i}\right) \tag{8}$$

Hellinger distance (HD) is another measure of the similarity between probability distributions. For the same two vectors $\mathbf{a}$ and $\mathbf{b}$ as before, we calculate their similarity with HD as

$$s_{\mathrm{hd}}(\mathbf{a}, \mathbf{b}) := \frac{1}{\sqrt{2}} \sqrt{\sum_{i=0}^{W} \left(\sqrt{\mathbf{a}_i} - \sqrt{\mathbf{b}_i}\right)^2} \,. \tag{9}$$

For each per-token measure d, we calculate the total score for a sentence identically to PLL:

$$S_d(\mathrm{S}) = \frac{1}{N} \sum_{i=1}^{N} \log\left(s_d\left(\mathbf{M}_i^{\mathrm{S}}, \mathbf{M}_i^{\mathrm{S}\setminus i}\right)\right) \,. \tag{10}$$

4 EVALUATION AND RESULTS

In this section we present the results of the experiments for answering our research questions.

4.1 Identifying Supported Statements

Figure 1 compares kernel density estimates for PLL scores based on whether a statement is supported (considered true) or not. Four different MLMs were used (BERT base and large, RoBERTa base and large), none of which were fine-tuned for any task. All MLMs showed a clear trend where supported statements have higher PLL values, which seems to confirm for RQ1 that a high PLL score acts as an indication for whether a statement is supported.

Figure 2 shows the average pseudo-likelihood value per statement, again separated by support. Again, there is a clear distinction between supported and unsupported statements. Of note here is the behavior by RoBERTa base and large, with a much larger number of tokens in unsupported statements scoring near-zero, which is not as noticeable for BERT base or large.

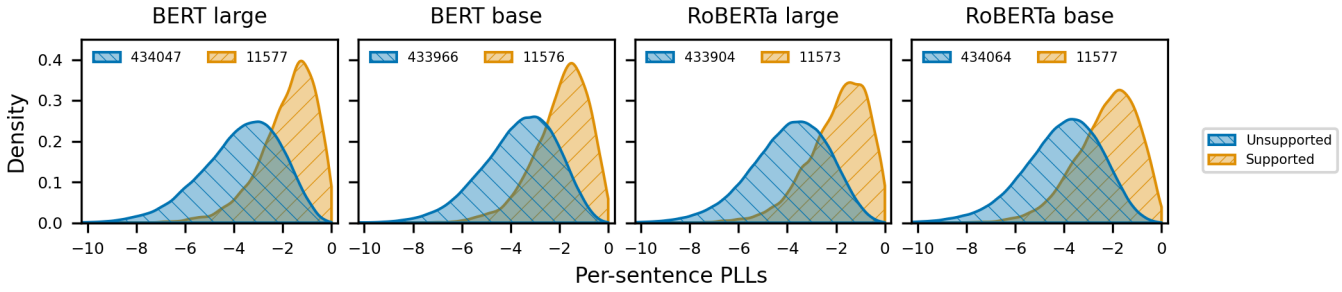

Figure 1: PLL scores for supported (orange) and unsupported (blue) statements. Supported statements score higher fairly consistently, but the population size for unsupported statements makes these density plots somewhat misleading.

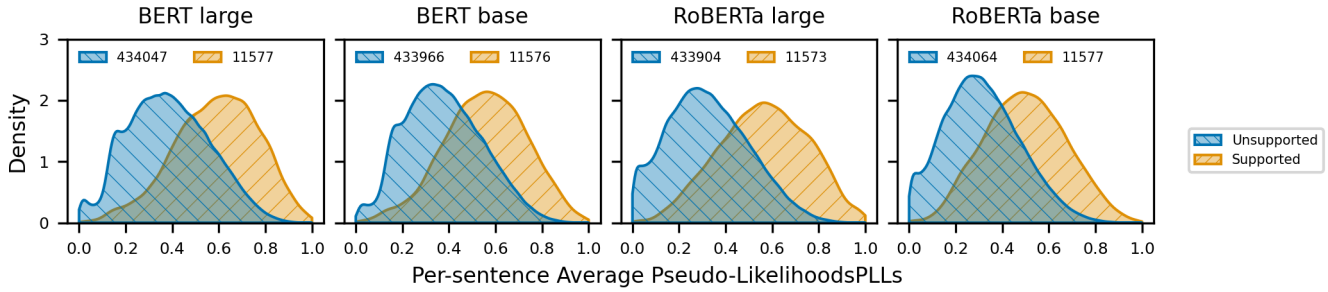

Figure 2: Average pseudo-likelihood (PLL) scores for supported (orange) and unsupported (blue) statements. Supported statements again score higher fairly consistently.

Despite these graphs, we know that PLL alone is not sufficient to distinguish supported and unsupported statements [3]. This is primarily due to the sheer volume of unsupported statements; for RoBERTa large, there are 46,376 unsupported statements with scores above -2, but only 6844 supported ones. We see a similar trend with the average PLL scores, where a cutoff of 0.6 yields 30,618 unsupported statements but 5010 supported ones. For both, the unsupported statements were counted after the 1% sampling, so the actual number is roughly 100 times larger. This does not mean that PLL is useless. A low PLL is still a very strong indicator of a lack of support, and a more effective method for generating candidate statements could reduce the number of false statements such that a reliable cutoff point could be found. Likely, PLL simply needs to be used in conjunction with other evidence.

4.2 Domain and Range Violations

Figure 3 shows the PLL for statements further divided by whether they are consistent (accepted) or inconsistent (rejected) with the entity-type restrictions of their underlying relation. There is a clear trend toward higher values the more "correct" a statement gets. Specifically, the scores for rejected statements are generally lower than those for accepted statements for both supported and unsupported statements. There is also a clear separation of the two categories of unsupported statements. We believe that this implies that entity-type restrictions are one important type of semantic feature captured by the MLM's notion of linguistic acceptability, an aspect that will allow us to better treat more domain-specific relations as well.

4.3 Multi-token Entities

PLL scores for tokens that compose a multi-token word, as is the case in many entities, tend to be fairly high due to their treatment in the masked language model learning objective. Figure 4 confirms this effect by further partitioning the statements into those which have multi-token entities (MTEs) and those which do not. To remove the effects seen in Section 4.2, only the statements marked "accepted" were used in this analysis.

We can see from the figures that statements with MTEs tend to be higher than those without MTEs. Interestingly, supported statements without MTEs and unsupported statements with MTEs seem to score roughly similarly for both BERT models (K-S test reported $p = 0.012$ for base, $p = 0.217$ for large), which implies strongly that this is a category for which BERT has trouble distinguishing supported from unsupported statements. For both RoBERTA models, these two categories were distinct, ($p < 0.0005$ for base, $p = 0.003$ for large), but with the unsupported category scoring higher. Overall, these graphs show that statements which do not have MTEs are likely to be scored lower than those which do, regardless of support.

4.4 Estimating PLL for Out-of-Vocabulary Inputs

The final goal of our exploration was to identify a suitable metric which can act as an estimator of PLL for when it cannot be calculated. As described in Section 3.5, to estimate PLL with a new metric, we compare the output logits for a token t_i at position i in a given statement with the output logits of that same statement but

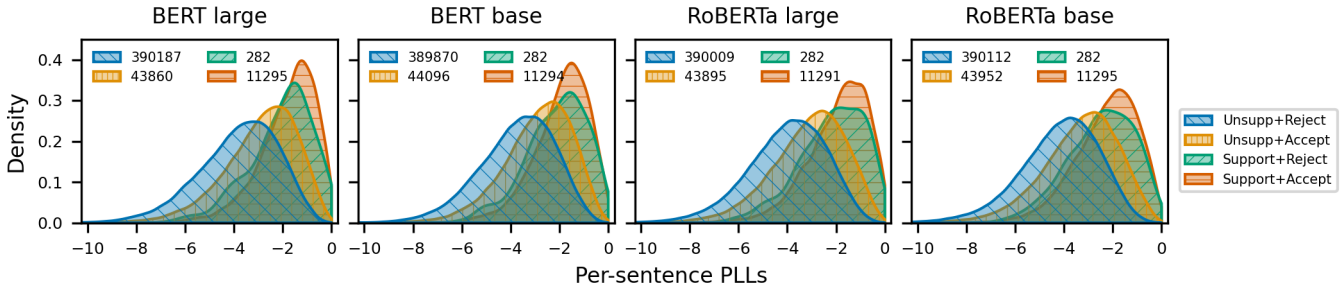

Figure 3: PLL scores for supported and unsupported statements, further divided by whether the entities used in the statements conform (Accepted/Rejected) to the domain and range of the relation they represent.

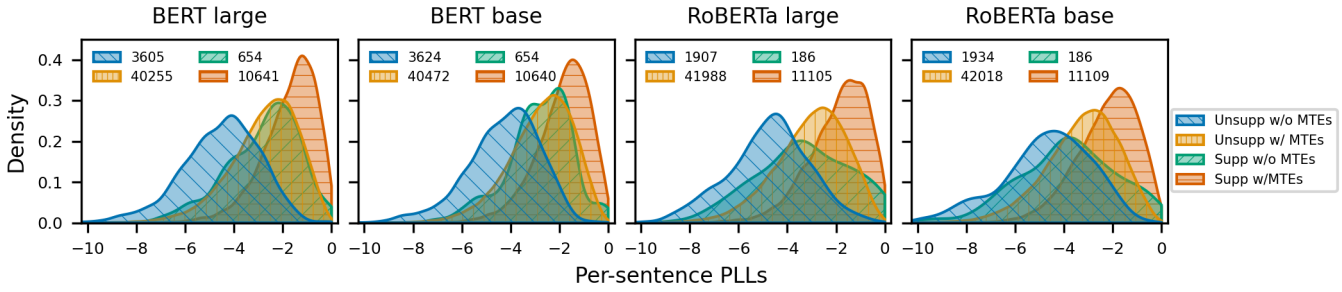

Figure 4: PLL scores for true and false statements, further divided by whether a statement contains a multi-token entity (MTE). Note that only "Accepted" statements are used in this analysis.

with t_i replaced with the mask token ([MASK]). If we consider the PLL scores to be the ground truth and the metric-based scores to be an estimate, then we can use the Spearman rank-order correlation coefficient to determine the quality of the estimation. In brief, a high Spearman coefficient indicates a strong correlation and implies that the metric will rank statements similarly to PLL.

Table 1 shows the Spearman rank-order correlation coefficients for the various metrics with regards to the PLLs given by RoBERTa large. Given two statements **S** and **T**, we interpret a high correlation between PLL and a metric M as saying that if $PLL(\mathbf{S}) < PLL(\mathbf{T})$, then it is likely that $M(\mathbf{S}) < M(\mathbf{T})$, thus preserving ranking while also being linearly correlated. We consider a score of 0.8 to be a high correlation. The measure with the highest correlation for supported statements was HD, but only by a small amount. CS correlated better with unsupported statements, which is likely why its correlation to all statements is high as well, and is on par with HD for supported statements.

It is important to note that deviations from PLL might actually be favorable. For instance, if a measure ends up being more predictive of support than PLL, the deviation of the two may be large. However, here we are only considering whether a metric is a good estimate of PLL, flaws and all, so we base our judgments on the assumption that a higher correlation is preferred.

Table 2 further examines the four measures by considering their Spearman coefficients for subsets of the data as broken down in the earlier experiments. Cosine similarity again has the highest correlation, except for the two categories where statements were both supported and accepted. This seems to follow a general trend, where the correlations are highest for supported statements and lowest for unsupported statements across all measures.

Metric	All Statements	Supported	Unsupported
CS	**0.796**	0.921	**0.789**
JSD	0.742	0.914	0.731
MSD	0.675	0.793	0.664
HD	0.756	**0.925**	0.746

Table 1: Spearman rank-order correlation coefficients between aggregate scores under the metric schemes and PLL scores, all from logits output via RoBERTa large. We consider scores above 0.8 to indicate a strong correlation. Three of the four metrics seem to have a moderately high correlation to PLL in all categories, with cosine similarity (CS) having the highest overall correlation except for true statements. All p-values are below 0.0005.

5 DISCUSSION AND FUTURE WORK

In this section we discuss the limitations of our current work, and provide an outlook towards the larger aim of our work.

5.1 Limitations

In this work, we only considered the case where the context (the world in which the statements are factual) was the background

Metric	Supp + Acc	Supp + Rej	Unsupp + Acc	Unsupp + Rej	Supp W/MTEs	Supp w/o MTEs	Unsupp W/MTEs	Unsupp w/o MTEs
CS	0.921	**0.924**	**0.847**	**0.779**	0.921	**0.865**	**0.850**	**0.620**
JSD	0.914	0.892	0.815	0.717	0.914	0.862	0.819	0.615
MSD	0.793	0.802	0.713	0.651	0.789	0.834	0.709	0.551
HD	**0.925**	0.911	0.827	0.733	**0.925**	0.864	0.830	0.604

Table 2: Spearman rank-order correlations between aggregate scores for each measure and PLL scores, all from logits output via RoBERTa large The categories are the same as in Figures 3 and 4. We consider scores above 0.8 to indicate a strong correlation. Cosine similarity (CS) again has the highest overall metric score in most categories, but Hellinger Distance (HD) shows slightly higher correlations for the "most correct" statements (supported statements which conform to relational restrictions). All p-values are below 0.0005.

knowledge of the MLMs being tested, and made the assumption that the information in the testing documents is also captured by those models, due to being based on Wikidata. Further, we only examined primarily commonsense relations. Our results may not generalize well to statements that capture more complicated or domain-specific relations not present in the MLM's training data. Further, the data in this work was primarily monolingual, with only a portion of the entities originating from languages other than English. While the transformer architecture of MLMs should not inherently be biased toward languages with specific characteristics (e.g. text direction or syntactic structure), availability of training data will be a limiting factor. As such, caution should be exercised when extrapolating the results of this work to MLMs trained on other languages, especially those which are resource-poor.

5.2 Outlook and Future Work

We discussed in Section 1 the need to avoid the bias inherent in MLMs for effective fact extraction, while all experiments in this work took advantage of that bias. However, still, we feel that our work so far is a first step toward this goal. The next step is to develop a method for calculating scores for statements where the MLM does not have knowledge of the entities. Based on the experiments in this paper, we believe that the following are valid speculations: The experiments with relation restrictions confirm that the semantics of the entities in questions influences the scores. The experiments regarding MTEs hint that such a method will need to represent entities with multiple tokens, to avoid the tendency of MLMs to assign high pseudo-likelihoods to pronouns when only one masked token is present. We further speculate that in order to remove background knowledge about an entity, its representation will likely need to be replaced with OOV tokens somehow contextualized exclusively on the document from which they come, such as during the generation of a soft prompt. Such OOV representations then necessitate the use of one of the PLL approximations shown earlier. We intend to explore all of these routes in future work.

More general future work should include the application of our approach to data sets beyond DocRED, such as for domain-specific fact extraction, in particular with MLMs trained on domain-specific documents. In the end, our goal is to build a KG extraction and querying framework, using a MLM as the front end to a virtual KG that is extracted on-the fly based on the user's queries.

6 CONCLUSIONS

In this paper we have studied how well the MLM scoring method PLL correlates with the support of statements in input texts, in particular for certain common MLMs. We found for RQ1 that this correlation is high, although the large number of possible unsupported statements that can be generated means that it can not act as a discriminator on its own. We further found for RQ2 that there was a noticeable lowering of PLL scores for statements with entities which were of an incorrect type given the semantics of the relation they represented. Since the relations in DocRED are mostly commonsense, we assume that a violation of the semantics of these relations is likely to be perceived as a lack of fluency, drawing a parallel back to linguistic acceptability. For RQ3 we showed that statements without MTEs generally scored lower, regardless of whether they were supported. We believe that this is related to the heavy use of pronouns in English, where masked single-token nouns are often scored low because MLMs assign high pseudo-probabilities to pronouns (e.g. it or she).

Finally, for RQ4 we studied how well common similarity metrics can estimate PLL for cases where it cannot be calculated, which will be a situation important to our envisioned, more generic KG extraction setting. We showed that cosine similarity has the highest overall correlation to PLL among the measures we tested, and by the use of K-S test additionally showed that it generally preserved rankings, making it a suitable substitute. Hellinger distance was a close second, but showed lower correlations for unsupported statements. While we focused exclusively on estimating PLL, future work should assess whether lower correlations have a positive impact on the actual predictive performance.

Overall we find that PLL sufficiently well represents the support of a fact in a document, and there are ways to cope with the identified aspects of domain-specific sentences. We also identify a clear next step as the need to account for MLM biases in scoring, such that we can properly handle fact extraction from documents without concern for the information about an entity in one document leaking into the extractions from another.

ACKNOWLEDGMENTS

This work was funded by the Swedish National Graduate School in Computer Science (CUGS). Portions of this work were carried out using the AIOps/Stellar facilities funded by the Excellence Center at Linköping–Lund in Information Technology (ELLIIT).

REFERENCES

[1] Dimitrios Alivanistos, Selene Báez Santamaría, Michael Cochez, Jan Christoph Kalo, Emile van Krieken, and Thiviyan Thanapalasingam. 2022. Prompting as Probing: Using Language Models for Knowledge Base Construction. In *2022 Semantic Web Challenge on Knowledge Base Construction from Pre-Trained Language Models, LM-KBC 2022*. CEUR-WS. org, 11–34.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Riley Capshaw and Eva Blomqvist. 2023. Towards Tailored Knowledge Base Modeling using Masked Language Models. In *Proceedings of TEXT2KG, Co-located with ESWC 2023, CEUR-WS.*

[4] Noam Chomsky. 1957. *Syntactic structures*. Mouton de Gruyter.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] Aparna Elangovan, Estrid He, and Cornelia Verspoor. 2021. Memorization vs. Generalization: Quantifying Data Leakage in NLP Performance Evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL 2021)*. Association for Computational Linguistics, 1325–1335.

[7] Leandra Fichtel, Jan-Christoph Kalo, and Wolf-Tilo Balke. 2021. Prompt tuning or fine-tuning-investigating relational knowledge in pre-trained language models. In *3rd Conference on Automated Knowledge Base Construction*.

[8] Adi Haviv, Jonathan Berant, and Amir Globerson. 2021. BERTese: Learning to Speak to BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 3618–3623.

[9] Benjamin Heinzerling and Kentaro Inui. 2021. Language Models as Knowledge Bases: On Entity Representations, Storage Capacity, and Paraphrased Queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1772–1791. https://doi.org/10.18653/v1/2021.eacl-main.153

[10] Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask–evaluating social biases in masked language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11954–11962.

[11] Carina Kauf and Anna Ivanova. 2023. A Better Way to Do Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 925–935. https://aclanthology.org/2023.acl-short.80.pdf

[12] Bum Chul Kwon and Nandana Mihindukulasooriya. 2022. An Empirical Study on Pseudo-log-likelihood Bias Measures for Masked Language Models Using Paraphrased Sentences. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*. Association for Computational Linguistics, Seattle, U.S.A., 74–79. https://doi.org/10.18653/v1/2022.trustnlp-1.7

[13] 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* (2019).

[14] Bo Lv, Xin Liu, Shaojie Dai, Nayu Liu, Fan Yang, Ping Luo, and Yue Yu. 2023. DSP: Discriminative Soft Prompts for Zero-Shot Entity and Relation Extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 5491–5505. https://doi.org/10.18653/v1/2023.findings-acl.339

[15] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1953–1967.

[16] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2463–2473.

[17] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2699–2712.

[18] Joonbo Shin, Yoonhyung Lee, and Kyomin Jung. 2019. Effective Sentence Scoring Method Using BERT for Speech Recognition. In *Proceedings of The Eleventh Asian Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 101)*, Wee Sun Lee and Taiji Suzuki (Eds.). PMLR, 1081–1093. https://proceedings.mlr.press/v101/shin19a.html

[19] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[20] Alex Wang and Kyunghyun Cho. 2019. BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. 30–36.

[21] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 764–777.