

BERT is Not a Knowledge Base (Yet): Factual Knowledge vs. Name-Based Reasoning in Unsupervised QA

Nina Poerner^{*†} and Ulli Waltinger[†] and Hinrich Schütze^{*}

^{*}Center for Information and Language Processing, LMU Munich, Germany

[†]Corporate Technology Machine Intelligence (MIC-DE), Siemens AG Munich, Germany
poerner@cis.uni-muenchen.de | inquiries@cislmu.org

Abstract

The BERT language model (LM) (Devlin et al., 2019) is surprisingly good at answering cloze-style questions about relational facts. Petroni et al. (2019) take this as evidence that BERT memorizes factual knowledge during pre-training. We take issue with this interpretation and argue that the performance of BERT is partly due to reasoning about (the surface form of) entity names, e.g., guessing that a person with an Italian-sounding name speaks Italian. More specifically, we show that BERT’s precision drops dramatically when we filter certain easy-to-guess facts. As a remedy, we propose E-BERT, an extension of BERT that replaces entity mentions with symbolic entity embeddings. E-BERT outperforms both BERT and ERNIE (Zhang et al., 2019) on hard-to-guess queries. We take this as evidence that E-BERT is richer in factual knowledge, and we show two ways of ensembling BERT and E-BERT.

1 Introduction

Imagine that you have a friend who claims to know a lot of trivia. During a quiz, you ask them about the native language of actor Jean Marais. They correctly answer French. For a moment you are impressed, until you realize that Jean is a typical French name. So you ask the same question about Daniel Ceccaldi (another French actor, but with an Italian-sounding name). This time your friend says “Italian, I guess.” If this were a Question Answering (QA) benchmark, your friend would have achieved a respectable accuracy of 50%. Yet, their performance does not indicate factual knowledge about the native languages of actors. Rather, it shows that they are able to reason about the likely origins of peoples’ names (see Table 1 for more examples).

Petroni et al. (2019) argue that the unsupervised BERT LM (Devlin et al., 2019) memorizes factual knowledge about entities and relations. They

	BERT _{base}	E-BERT _{base}
Jean Marais	French	French
Daniel Ceccaldi	Italian	French
Orane Demazis	Albanian	French
Kad Merad	Kurdish	French

Table 1: Native language of French actors according to BERT and E-BERT. Query: “The native language of [X] is [MASK].” (LAMA-T-REx:P103).

base this statement on the unsupervised QA benchmark LAMA (§2), where BERT rivals a knowledge base (KB) built by relation extraction. They suggest that BERT and similar LMs could become a “viable alternative to traditional knowledge bases extracted from text”. We argue that the impressive performance of BERT is partly due to reasoning about (the surface form of) entity names. In §2.1, we construct LAMA-UHN (UnHelpful Names), a more “factual” subset of LAMA-Google-RE and LAMA-T-REx, by filtering out queries that are easy to answer from entity names alone. We show that the performance of BERT decreases dramatically on LAMA-UHN.

In §3, we propose E-BERT, a simple mapping-based extension of BERT that replaces entity mentions with wikipedia2vec entity embeddings (Yamada et al., 2016). In §4, we show that E-BERT rivals BERT and the recently proposed entity-enhanced ERNIE model¹ (Zhang et al., 2019) on LAMA. E-BERT has a substantial lead over both baselines on LAMA-UHN; furthermore, ensembles of E-BERT and BERT outperform all baselines on original LAMA.

2 LAMA

The LAMA (LAnguage Model Analysis) benchmark (Petroni et al., 2019) is supposed to probe for “factual and commonsense knowledge” inherent in

¹Not to be confused with ERNIE 2.0 (Sun et al., 2019).

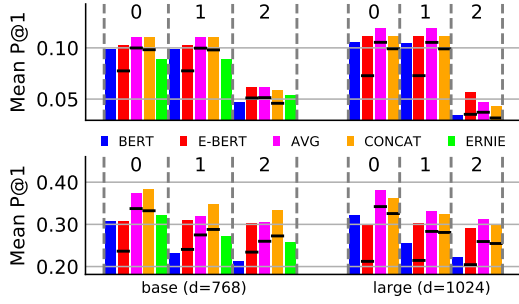


Figure 1: Mean P@1 on LAMA-Google-RE (top) and LAMA-T-REx (bottom). 0: unfiltered, 1: string match filter, 2: person name filter. Filters are applied sequentially. Black horizontal bars: Performance of wikipedia2vec without link graph loss.

LMS. In this paper, we focus on LAMA-Google-RE and LAMA-T-REx (Elsahar et al., 2018), which are aimed at factual knowledge. Contrary to most previous works on QA, LAMA tests LMs as-is, without supervised finetuning.

The LAMA probing task follows this schema: Given a KB triple of the form (S, R, O), the object is elicited with a relation-specific cloze-style question, e.g., (JeanMarais, native-language, French) becomes: “The native language of Jean Marais is [MASK].”² The LM predicts a distribution over a limited vocabulary to replace [MASK], which is evaluated against the known gold answer.

2.1 LAMA-UHN

It is often possible to guess properties of an entity from its name, with zero factual knowledge of the entity itself. This is because entities are often named according to implicit or explicit rules (e.g., the cultural norms involved in naming a child, copyright laws for industrial products, or simply a practical need for descriptive names). LAMA makes guessing even easier by its limited vocabulary, which may only contain a few candidates for a particular entity type.

We argue that a QA benchmark that does not control for entity names does not assess whether an LM is good at reasoning about names, good at memorizing facts, or both. In this Section, we describe the creation of LAMA-UHN (UnHelpfulNames), a subset of LAMA-Google-RE and LAMA-T-REx.

Filter 1: The *string match filter* deletes all KB triples where the correct answer (e.g., *Apple*) is a case-insensitive substring of the subject entity

name (e.g., *Apple Watch*). This simple heuristic deletes up to 81% of triples from individual relations (see Appendix for statistics and examples).

Filter 2: Of course, entity names can be revealing in ways that are more subtle. As illustrated by our French actor example, a person’s name can be a useful prior for guessing their native language and by extension, their nationality, place of birth, etc. Our *person name filter* uses cloze-style questions to elicit name associations inherent in BERT, and deletes KB triples that correlate with them. Consider our previous example (JeanMarais, native-language, French). We whitespace-tokenize the subject name into *Jean* and *Marais*. If BERT considers either name to be a common French name, then a correct answer is insufficient evidence for factual knowledge about the entity JeanMarais. On the other hand, if neither *Jean* nor *Marais* are considered French, but a correct answer is given nonetheless, then we consider this sufficient evidence for factual knowledge.

We query BERT for answers to “[X] is a common name in the following language: [MASK].” for both [X] = *Jean* and [X] = *Marais*. If the correct answer is among the top-3 for either query, we delete the triple. We apply this filter to Google-RE:place-of-birth, Google-RE:place-of-death, T-REx:P19 (place of birth), T-REx:P20 (place of death), T-REx:P27 (nationality), T-REx:P103 (native language) and T-REx:P1412 (language used). See Appendix for statistics. Depending on the relation, we replace “language” with “city” or “country” in the template.

Figure 1 (blue bars) shows that BERT is strongly affected by filtering, with a drop of 5%–10% mean P@1 from original LAMA to LAMA-UHN. This suggests that BERT does well on LAMA partly because it reasons about (the surface form of) entity names. Of course, name-based reasoning is a useful ability in its own right; however, conflating it with factual knowledge may be misleading.

3 E-BERT

BERT. BERT (Devlin et al., 2019) is a deep bidirectional transformer encoder (Vaswani et al., 2017) pretrained on unlabeled text. It segments text into subword tokens from a vocabulary \mathbb{L}_b . During training, some tokens are masked by a special [MASK] token. Tokens are embedded into real-valued vectors by an embedding function

²We omit [CLS] and [SEP] tokens for readability.

$\mathcal{E}_B : \mathbb{L}_b \rightarrow \mathbb{R}^{d_B}$.³ The embedded tokens are contextualized by the BERT encoder \mathcal{B} and the output of \mathcal{B} is fed into a function $\mathcal{M}_B : \mathbb{R}^{d_B} \rightarrow \mathbb{L}_b$ that predicts the identity of masked tokens. BERT can thus be used as an LM.

Wikipedia2vec. Wikipedia2vec (Yamada et al., 2016) embeds words and wikipedia pages⁴ (\approx entities) in a common space. It learns an embedding function for a vocabulary of words \mathbb{L}_w and a set of entities \mathbb{L}_e . We denote this function as $\mathcal{F} : \mathbb{L}_w \cup \mathbb{L}_e \rightarrow \mathbb{R}^{d_F}$. The wikipedia2vec loss has three components: (a) skipgram word2vec (Mikolov et al., 2013) operating on \mathbb{L}_w (b) a graph loss on the wikipedia link graph⁵ on \mathbb{L}_e (c) a version of word2vec where words are predicted from entity mentions. Loss (c) ensures that word and entity embeddings share a space. Figure 1 (black horizontal bars) shows that loss (b) is vital for our use case.

E-BERT. We want to transform the output space of \mathcal{F} in such a way that \mathcal{B} is fooled into accepting entity embeddings in lieu of its native subword embeddings. We approximate this goal by minimizing the squared distance of transformed wikipedia2vec word vectors and BERT subword vectors:

$$\operatorname{argmin}_{\mathcal{W}} \mathbb{E}_{x \in \mathbb{L}_b \cap \mathbb{L}_w} \|\mathcal{W}(\mathcal{F}(x)) - \mathcal{E}_B(x)\|_2^2$$

where \mathcal{W} is a linear projection obtained by least squares. Since \mathcal{F} embeds \mathbb{L}_w and \mathbb{L}_e into the same space, \mathcal{W} is applicable to members of \mathbb{L}_e , even though it was learned on members of \mathbb{L}_w .

Recall that BERT segments text into subwords, e.g., our previous example is tokenized as: *The native language of Jean Mara ##is is [MASK]*.

E-BERT replaces the subwords that correspond to the entity mention with the symbolic entity: *The native language of Jean_Marais is [MASK]*.

The entity (`true_type`) is embedded by $\mathcal{W} \circ \mathcal{F}$, while other tokens (*italics*) continue to be embedded by \mathcal{E}_B . The altered embedding sequence is fed into \mathcal{B} , where it is treated like any other embedding sequence. Neither \mathcal{B} nor \mathcal{M}_B are changed.

We ensemble BERT and E-BERT by (a) mean-pooling their outputs (AVG) or (b) concatenating the entity and its name with a slash symbol (CONCAT), e.g.: *Jean_Marais / Jean Mara ##is*.

³In our notation, \mathcal{E}_B does not include position embeddings or segment embeddings; we consider these to be part of \mathcal{B} .

⁴We map wikidata IDs to wikipedia page titles via query.wikidata.org/sparql.

⁵The (untyped) wikipedia link graph is not the same as the (typed) wikidata KB.

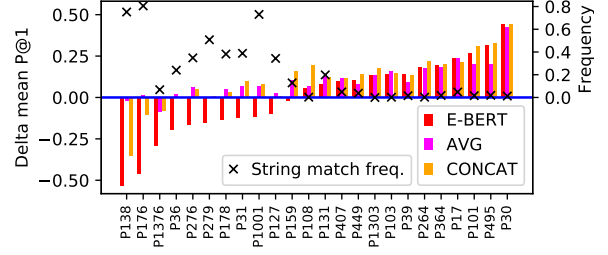


Figure 2: Delta in mean P@1 on unfiltered LAMA-TREx relations w.r.t. $\text{BERT}_{\text{base}}$ (blue baseline). Cross: frequency of triples where the object name is a substring of the subject name. We omit relations with absolute delta below 10% due to space constraints.

4 Experiments

Systems. We train cased wikipedia2vec on a recent wikipedia dump (2019-09-02), setting $d_F = d_B$. To learn \mathcal{W} , we intersect the wikipedia2vec word vocabulary with the cased BERT vocabulary.

Our primary baselines are $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$ ⁶ as evaluated in Petroni et al. (2019). We also test ERNIE (Zhang et al., 2019),⁷ a $\text{BERT}_{\text{base}}$ type model that uses wikidata TransE entity embeddings (Bordes et al., 2013) as additional input. ERNIE has two transformers, one for tokens and one for entities, which are fused by a trainable feed-forward module. To accommodate the new parameters, ERNIE is pre-trained with (a) standard BERT loss and (b) predicting Wikipedia entities.

Note that wikipedia2vec and TransE have low coverage on LAMA-Google-RE (wikipedia2vec: 54%, TransE: 71%). When an entity embedding is missing, we fall back onto original BERT. Coverage of LAMA-TREx is $> 98\%$ for both systems.

LAMA. In keeping with Petroni et al. (2019), we report P@k macro-averaged over relations. Macro-averaging ensures that every relation has the same impact on the metric before and after filtering.

Figure 1 shows that E-BERT performs comparable to BERT and ERNIE on unfiltered LAMA. However, E-BERT is less affected by filtering on LAMA-UHN, suggesting that its performance is more strongly due to factual knowledge. Recall that we lack entity embeddings for 46% of Google-RE subjects, i.e., E-BERT cannot improve over BERT on almost half of the Google-RE tuples.

Figure 2 plots deltas in mean P@1 on unfiltered LAMA-TREx relations relative to BERT, along with the frequency of tuples whose object entity

⁶github.com/huggingface/transformers

⁷github.com/thunlp/ERNIE

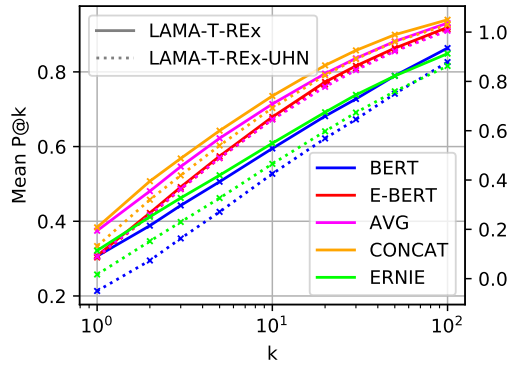


Figure 3: Mean P@k for different k on LAMA-T-REx (before filtering) and LAMA-T-REx-UHN. Architecture: BERT_{base}.

name is a substring of the subject entity name – i.e., the ratio of queries that would be deleted by the string match filter. We see that E-BERT losses relative to BERT (negative red bars) are mostly on relations with a high percentage of trivial substring answers. By contrast, E-BERT typically outperforms BERT on relations where such trivial answers are rare. The ensembles are able to mitigate the losses of E-BERT on almost all relations, while keeping most of its gains (purple and orange bars). This suggests that they successfully combine BERT’s ability to reason about entity names with E-BERT’s enhanced factual knowledge.

Figure 3 shows that the lead of E-BERT and the ensembles over BERT and ERNIE in terms of mean P@k is especially salient for bigger k.

FewRel. We also evaluate on the FewRel relation classification dataset (Han et al., 2018), using the setup and data split from Zhang et al. (2019) (see Appendix for details). Table 2 shows that E-BERT beats BERT, and the ensembles perform comparable to ERNIE despite not having a dedicated entity encoder.

5 Related work

Factual QA is typically tackled as a supervised problem (e.g., Das et al. (2017); Sun et al. (2018)). In contrast, LAMA (Petroni et al., 2019) tests for knowledge learned by LMs without supervision; similar experiments were performed by Radford et al. (2019). Their experiments do not differentiate between factual knowledge of LMs and their ability to reason about entity names.

The E-BERT embedding mapping strategy is inspired by cross-lingual embedding mapping on identical strings (Smith et al., 2017). A similar

	P	R	F1
BERT _{base}	85.87	85.74	85.71
E-BERT _{base}	87.29	87.33	87.18
AVG ensemble	88.28	88.34	88.19
CONCAT ensemble	88.58	88.54	88.46
BERT _{base} (Zhang et al., 2019)	85.05	85.11	84.89
ERNIE _{base} (Zhang et al., 2019)	88.49	88.44	88.32

Table 2: Macro Precision, Recall, F1 (%) on FewRel.

method was recently applied by Wang et al. (2019) to map cross-lingual FastText subword vectors (Bojanowski et al., 2017) into the multilingual BERT subword embedding space. Schick and Schütze (2019) mimic BERT subword embeddings for rare English words from their contexts and form.

Other contextualized models that incorporate entity embeddings are ERNIE (Zhang et al., 2019) (see §4) and KnowBert (Peters et al., 2019). KnowBert is contemporaneous to our work, and at the time of writing, the model was not available for comparison.

Both ERNIE and KnowBert add new parameters to the BERT architecture, which must be integrated by additional pretraining. By contrast, E-BERT works with the unchanged BERT model, and \mathcal{W} has an efficient closed-form solution. This means that we can update E-BERT to the newest wikipedia dump at little computational cost – the most expensive operation would be training wikipedia2vec, which takes a few hours on CPUs.

6 Conclusion

We have presented evidence that the surprising performance of BERT on the recently published LAMA QA benchmark is partly due to reasoning about entity names rather than factual knowledge. We have constructed more “factual” subsets of LAMA-Google-RE and LAMA-T-REx by filtering out easy-to-guess queries. The resulting benchmark, LAMA-UHN, is more difficult for BERT.

As a remedy, we proposed E-BERT, a simple extension of BERT that injects wikipedia2vec entity embeddings into BERT. E-BERT outperforms BERT and ERNIE on LAMA-UHN, which we take as evidence that E-BERT is richer in factual knowledge. Additionally, ensembling yields improvements over both BERT and E-BERT on unfiltered LAMA and on the FewRel relation classification dataset.

References

- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. [Enriching word vectors with subword information](#). *TACL*, 5:135–146.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. [Translating embeddings for modeling multi-relational data](#). In *NeurIPS*, pages 2787–2795, Lake Tahoe, USA.
- Rajarshi Das, Manzil Zaheer, Siva Reddy, and Andrew McCallum. 2017. [Question answering on knowledge bases and text using universal schema and memory networks](#). In *ACL*, pages 358–365, Vancouver, Canada.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *NAACL-HLT*, pages 4171–4186, Minneapolis, USA.
- Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frédérique Laforest, and Elena Simperl. 2018. [T-REx: A large scale alignment of natural language with knowledge base triples](#). In *LREC*, pages 3448–3452, Miyazaki, Japan.
- Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. [FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation](#). In *EMNLP*, pages 4803–4809, Brussels, Belgium.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Matthew E Peters, Mark Neumann, IV Logan, L Robert, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. [Knowledge enhanced contextual word representations](#). In *EMNLP-IJCNLP*, pages 43–54, Hong Kong, China.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. [Language models as knowledge bases?](#) In *EMNLP-IJCNLP*, pages 2463–2473, Hong Kong, China.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). *OpenAI Blog*, 1(8).
- Timo Schick and Hinrich Schütze. 2019. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. *arXiv preprint arXiv:1904.06707*.
- Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *ICLR*, Toulon, France.
- Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. [Open domain question answering using early fusion of knowledge bases and text](#). In *EMNLP*, pages 4231–4242, Brussels, Belgium.
- Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *NeurIPS*, pages 5998–6008, Long Beach, USA.
- Hai Wang, Dian Yu, Kai Sun, Janshu Chen, and Dong Yu. 2019. Improving pre-trained multilingual models with vocabulary expansion. *arXiv preprint arXiv:1909.12440*.
- Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. [Joint learning of the embedding of words and entities for named entity disambiguation](#). In *CoNLL*, Berlin, Germany.
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. [Ernie: Enhanced language representation with informative entities](#). In *ACL*, pages 1441–1451, Florence, Italy.

A FewRel training

We use the sentence classification setup from [Zhang et al. \(2019\)](#). We mark subjects and objects with the symbols # and \$, i.e., the inputs to BERT, E-BERT and the CONCAT ensemble look as follows:

BERT: *[CLS] \$ Tang ##ier \$ ' s # Ibn Bat ##to ##uta Airport # is the busiest airport in the region . [SEP]*

E-BERT: *[CLS] \$ Tangier \$ ' s # Tangier_Ibn_Battouta_Airport # is the busiest airport in the region . [SEP]*

CONCAT: *[CLS] \$ Tangier / Tang ##ier \$ ' s # Tangier_Ibn_Battouta_Airport / Ibn Bat ##to ##uta Airport # is the busiest airport in the region . [SEP]*

where entities (in `trueType`) are embedded by $\mathcal{W}_{\mathcal{OF}}$ and all other tokens (in *italics*) are embedded by $\mathcal{E}_{\mathcal{B}}$. Note that entity IDs are provided by FewRel. If we lack an entity embedding, we fall back onto the standard BERT segmentation.

To predict the relation, we feed the contextualized embedding of the [CLS] token into a linear classifier. During training we finetune all network parameters except for the embeddings. For hyperparameter tuning, we use the ranges from [Zhang et al. \(2019\)](#) except for the number of epochs, which we fix at 10. The AVG ensemble averages over BERT’s and E-BERT’s output distributions. Experiments were run on two GeForce GTX 1080 Ti GPUs with data-parallel training.

B A note on casing

The cased BERT vocabulary is a superset of the LAMA vocabulary. This ensures that BERT can in principle answer all LAMA queries correctly. The uncased ERNIE vocabulary does not have this property. For ERNIE, we therefore lowercase all queries and restrict the model output to the intersection of its vocabulary with the lowercased LAMA vocabulary. As a result, ERNIE selects an answer from $\sim 18\text{K}$ candidates (instead of the standard $\sim 21\text{K}$), which should work in its favor. We verify that all lowercased object names from LAMA-T-REx and LAMA-Google-RE appear in ERNIE’s vocabulary, i.e., ERNIE is in principle able to answer all lowercased queries correctly.

	relation	% del	example of a deleted query
string match filter	P176:manufacturer	81%	[Fiat Multipla] is produced by [Fiat].
	P138:named after	75%	[Christmas Island] is named after [Christmas].
	P1001:applies to jurisdiction	73%	[Australian Senate] is a legal term in [Australia].
	P279:subclass of	51%	[lenticular galaxy] is a subclass of [galaxy].
	P31:instance of	39%	[Tantalon Castle] is a [castle].
	P178:developer	38%	[IBM AIX] is developed by [IBM].
	P276:location	35%	[Cologne Cathedral] is located in [Cologne].
	P127:owned by	34%	[Atari Interactive] is owned by [Atari].
	P361:part of	32%	[South Asia] is part of [Asia].
	P36:capital	24%	The capital of [Aberdeenshire] is [Aberdeen].
	P131:located in territory	20%	[California State Route 9] is located in [California].
	P527:has part	18%	[apple strudel] consists of [apple].
	P159:headquarters location	13 %	The headquarter of [Paris Saint-Germain F.C.] is in [Paris].
person name filter	P1412:language used	63%	[Fulvio Tomizza] used to communicate in [Italian]. (1,1) [Rajit Kapur] used to communicate in [Hindi]. (1,1)
	P103:native language	58%	The native language of [Tommy Nilsson] is [Swedish]. (-,1) The native language of [Andrey Malakhov] is [Russian]. (1,1)
	P27:nationality	56%	[Harumi Inoue] is a [Japan] citizen. (1,-) [Yves Mirande] is a [France] citizen. (1,-)
	P20:place of death	31%	[Avraham Harman] died in [Jerusalem]. (1,-) [Pierre Cartellier] died in [Paris]. (2,1)
	P19:place of birth	23%	[Christel Bodenstein] was born in [Munich]. (3,3) [Masako Natsume] was born in [Tokyo]. (2,1)
	Google-RE:place-of-birth	17%	[Marcel Bertrand] was born in [Paris]. (-,3) [Jan Jacob Kieft] was born in [Amsterdam]. (2,-,-)
	Google-RE:place-of-death	14%	[Bernardo López Piquer] died in [Madrid]. (-,1,-) [Nikolay Alexandrovich Milyutin] died in [Moscow]. (1,1,1)

Table 3: Statistics and examples of deleted queries. String match filter (top): most stongly filtered relations only. Person name filter (bottom): Numbers in brackets indicate which part(s) of the person name triggered the filter, and at what rank. For instance, (-,1) means that the correct answer was ranked first for the person’s last name, but was not among the top-3 for their first name. Architecture: BERT_{base}.