Extracting Multi-valued Relations from Language Models

Sneha Singhania, Simon Razniewski, Gerhard Weikum
Max Planck Institute for Informatics
{ssinghan, srazniew, weikum}@mpi-inf.mpg.de

Abstract

The widespread usage of latent language representations via pre-trained language models (LMs) suggests that they are a promising source of structured knowledge. However, existing methods focus only on a single object per subject-relation pair, even though often multiple objects are correct. To overcome this limitation, we analyze these representations for their potential to yield materialized multi-object relational knowledge. We formulate the problem as a *rank-then-select* task. For *ranking* candidate objects, we evaluate existing prompting techniques and propose new ones incorporating domain knowledge. Among the *selection* methods, we find that choosing objects with a likelihood above a learned relation-specific threshold gives a 49.5% F1 score. Our results highlight the difficulty of employing LMs for the multi-valued slot-filling task, and pave the way for further research on extracting relational knowledge from latent language representations.

1 Introduction

Petroni et al. (2019) showcased the potential of relation-specific probes for extracting implicit knowledge from latent language representations. But the viability of materializing factual knowledge directly from LMs remain open problems (Razniewski et al., 2021; AlKhamissi et al., 2022).

Building upon the LAMA framework, where an LM predicts an object in the slot for given a cloze-style prompt such as “Dante was born in [MASK]”, several methods (Jiang et al., 2020b; Shin et al., 2020; Zhong et al., 2021; Qin and Eisner, 2021) design effective prompts for factual information extraction. Importantly, however, all these methods implicitly assume the existence of a single correct object per (subject, relation)-pair and evaluate on precision at rank 1. In reality, many relations have multiple correct values as illustrated in Figure 1.

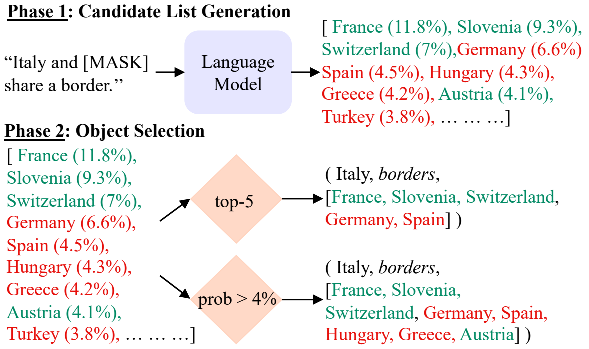
Phase 1: Candidate List Generation
“Italy and [MASK] share a border.”
Language Model
[France (11.8%), Slovenia (9.3%), Switzerland (7%),Germany (6.6%) Spain (4.5%), Hungary (4.3%), Greece (4.2%), Austria (4.1%), Turkey (3.8%), … … …]
Phase 2: Object Selection
[France (11.8%), Slovenia (9.3%), Switzerland (7%), Germany (6.6%), Spain (4.5%), Hungary (4.3%), Greece (4.2%), Austria (4.1%), Turkey (3.8%), … … …]
top-5
(Italy, *borders*, [France, Slovenia, Switzerland, Germany, Spain])
prob > 4%
(Italy, *borders*, [France, Slovenia, Switzerland, Germany, Spain, Hungary, Greece, Austria])

Figure 1: Probing LMs to extract objects for multi-valued relations.

In this paper, we focus on probing LMs directly for multi-valued slot-filling task. In a zero-shot setting, an LM is probed using prompts that include *a subject and a multi-valued relation* to generate a list of candidate objects. However, prior knowledge of the no. of correct objects is generally unknown. Since the LM’s probabilities alone do not provide a clear indication of the objects’ factual accuracy (Jiang et al., 2021; Holtzman et al., 2021), we apply various selection mechanisms on the generated list to choose the correct objects.

We probe LMs using existing prompting techniques and introduce new relation-specific prompts. The generated object lists are evaluated by their order, and our manual prompts result in higher-quality lists than the state-of-the-art automated methods. Our prompts outperform the best baseline, SoftPrompts, by ca. 5% points on the three most challenging relations (parts of chemical compounds, official languages of countries, instruments of musicians), while being competitive on the other four relations. While evaluating the output of selection mechanisms, our best approach achieves 54.1% precision, 50.8% recall, and 49.5% F1 score. These trends in scores demonstrate the difficulty of extracting complete knowledge from internal LMs representations.

Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023), pages 139–154
July 13, 2023 ©2023 Association for Computational Linguistics

2 Background

The idea of knowledge extraction using LMs representations was put forward by Radford et al. (2019); Petroni et al. (2019), and has since received much attention. Several approaches (Bouraoui et al., 2019; Goswami et al., 2020; Chen et al., 2022) use relational metadata and textual corpora to tune the LM using knowledge-enriched representations and efficiently extract knowledge. Jiang et al. (2020a); Kassner et al. (2021) focus on multilingual knowledge extraction, while Dhingra et al. (2022) looks at extracting temporal knowledge from LMs. Current methods usually sample from correct objects for a given subject-relation pair by employing the hits@k metric and do not enforce deliberate accept/reject decisions on the outputs.

3 Methodology

Candidate List Generation When probing an LM using a cloze-style prompt, the LM generates a probability distribution over the vocabulary tokens to fill in the masked position. For our task, we use prompts mentioning a subject-relation ⟨s, r⟩ pair and consider the resulting ranked list of tokens (w/ their corresponding probability scores) as candidates. In a zero-shot setting, the LM is probed using two types of prompts: (i) *discrete prompts* such as LPAQA (Jiang et al., 2020b) and AUTOPROMPT (Shin et al., 2020), and (ii) *continuous prompts* such as OPTIPROMPT (Zhong et al., 2021) and SoftPrompts (Qin and Eisner, 2021).

In addition, we propose a collection of carefully designed manual prompts to evaluate and compare the generated objects for multi-valued relations. We create 50 diverse relation-specific prompts by incorporating domain knowledge, relation type, and variations in sentence structure and grammar. Our prompts differ in verb form, tense, placement of the masked token (whether it is predicted in a prefix, suffix, or cloze-style), and whether or not there is a period and object type in the context.

Selection Mechanisms We experiment with the following parameterized mechanisms on the generated object lists to get a valid subset of triples.

Top-k: The most probable k objects are selected.

Prob-x: Objects with a probability greater than or equal to x are chosen.

Cumul-x: Retain all objects, in order of probability, whose summed probability is no larger than x. In difference with Prob-x, it would enable retaining candidates of similarly moderate probability.

Count Probe: We probe the LM again to get an object count prediction for a ⟨s, r⟩. A count probe could be as follows, *"Italy borders a total of* [MASK] *countries"*. From the list of generated tokens, the highest-ranked integer type token (either in alphabetical or numerical form) is used to subset the original object list.

Verification Probe: We probe the same LM again on each candidate object to factually verify the generated subject-relation-object ⟨s, r, o⟩ triple. A verification probe could be as follows, *"Italy and France share a border? Answer:* [MASK]*"*. We compare the relative probabilities of the _yes versus _no tokens in the masked position, using $(p_{\_yes} - p_{\_no} > \alpha)$, with α is a hyper-parameter, to determine the correctness of the original candidate object, France. All the original candidate objects satisfying the comparison condition are selected.

4 Experiment

Dataset The seven diverse multi-valued relations appearing in LAMA benchmark (Petroni et al., 2019) are chosen. For each relation, approx. 200 subjects and a complete list of objects from the popular Wikidata KB (Vrandečić and Krötzsch, 2014) are sampled. The subjects were picked based on popularity, measured using Wikidata ID and count of Twitter followers for person-type subjects[1].

Evaluation In the *ranking* phase, the quality of a candidate list is assessed using the maximally possible F1 score, defined as the highest possible F1 score achieved by applying the top-k selection mechanism with the optimal k value. The optimal k is found by iterating over all possible choices of k and calculating the respective F1 score on the subset of candidate objects and ground-truth objects. In the *selection* phase, the output triples obtained after applying a selection mechanism are evaluated using precision, recall, and F1 score.

Setup We reuse the best prompts reported by each prompting baseline, which are tuned on much larger data. We probe BERT (Devlin et al., 2019) on each ⟨s, r⟩ and generate the 500 most probable candidate objects. The generated list is post-processed to remove stopwords and other type-irrelevant objects depending on the relation type, only to retain sensible candidate objects. Our dataset is split into train, dev, and test, with

[1] https://anonymous.4open.science/r/acl_dataset-00B6/

⟨subject, relation⟩	Ours	Mined	Para	AUTO	OPTI	Soft
compound, *has-parts*	**78.5**	38.1	29.5	51.4	68.1	71.6
country, *borders*	72.8	66.4	64.9	71.6	73.2	**75.6**
country, *official-lang*	**83.6**	81.9	75.2	71.8	75.9	79.9
person, *instrument*	62.5	**63.4**	61.3	61.7	52.7	57.0
person, *speaks-lang*	**72.8**	69.1	41.1	52.8	71.5	69.0
person, *occupation*	33.2	40.2	**44.9**	37.5	36.6	36.9
state, *borders*	25.7	23.8	24.3	24.4	25.9	**25.9**
Overall (avg.)	**61.3**	54.7	48.7	53.0	57.7	59.4

Table 1: Comparing generated object lists by probing BERT-large using macro-averaged max-F1(%) scores. Mined is LPAQA's mining-based prompts and Para is LPAQA's paraphrasing-based prompts. AUTO uses the open-sourced AUTOPROMPT templates. OPTI and Soft are OPTIPROMPT and SoftPrompts obtained by training the LM using author-released data and code.

100/∼50/50 subjects per relation, for tuning and estimating parameters in the selection mechanisms.

5 Results

Candidate List Generation We compare the candidate list generated by each prompting method in Table 1. Our prompts generate the best object lists in terms of macro-averaged max-F1 score. Unlike our prompts, discrete prompts have a lower performance, while continuous counterparts have a similar performance. Surprisingly, OPTIPROMPT obtained by initializing its continuous vectors using our prompts has a lower score.

To validate the effectiveness of our prompts and inspect if optimizing prompts on precision@1 is sufficient for extraction on multi-valued relations, we compare the best prompts in terms of precision@1 and max-F1 score. In Table 2, we see that prompts performing well on precision@1 are not necessarily the best on max-F1. Overall, prompts suitable for multi-valued extraction are more often of the prefix type, while prompts in the single-object case show more variance. In Appendix, Table 4 shows examples of generated lists, and Tables 9-15 lists all the prompt templates. The large gap between precision@1 and max-F1 indicates the difficulty in designing task-specific prompts and raises the need for developing more robust prompts.

Object Selection The candidate objects retained after applying a selection mechanism are compared against the ground-truth objects, and the results are shown in Table 3. The top-k method achieves the best overall F1 score, which is the macro-average of individual ⟨s, r⟩ tuple-specific F1 scores. The individual F1 scores and max-F1 (upper bound) have a large gap since the probabilities of predicted tokens are not calibrated enough to match the actual factuality of the ⟨s, r, o⟩ triple. Table 5 in Appendix gives the prompt templates and learned parameters of each selection mechanism.

6 Discussion

Although BERT was probed for 500 objects when generating object lists, only 119.7 objects were retained after post-processing. Objects with invalid types occur due to the zero-shot setting. Also, other eminent masked LMs, including BERT-base, RoBERTa-base, and RoBERTa-large, achieve 60.61%, 54.82%, and 58.90% max-F1 scores.

The max-F1 scores in Table 1 & 2 are far from 100%, i.e., LMs do not generate candidate lists that correctly rank all true objects above the false ones. In particular, max-F1 will not reach 100% when correct objects are ranked too low or absent. We found that 26.90% of valid objects in a candidate list were ranked below the optimal threshold and 27.75% of valid objects were not generated at all.

In Table 3, the top-k and prob-x achieve balanced precision and recall scores. The count-probe achieves a high recall since almost always a count greater than 10 is predicted, and in our dataset, the average count of ground-truth objects across all ⟨s, r⟩ is in [1,10] range. In the verification probe, the parameter α is near zero for most relations, and the probability of _yes is greater than _no, leading to a selection of all the candidate objects. Although Schick and Schütze (2021) and others show the effect of verbalizing labels to _yes and _no tokens in the few-shot setting on classification and inference tasks, optimally using them for factual knowledge extraction remains an open challenge.

Effect of Prompt Template In Table 3, each mechanism is evaluated on the candidate list generated using prompt templates shown in max-F1 column in Table 2. However, by choosing a different set of prompts, a higher overall F1 score of 51.3% with a lower 59.9% max-F1 can be achieved by using the prob-x method. Table 6 in Appendix gives more details. This change in F1 scores shows the hardness of designing robust prompts.

Effect of Relation Type The candidate lists generated for popular subjects tend to have higher precision and recall. For instance, F1 score for (state, *borders*) with top-k is the lowest due to the pres-

⟨subject, relation⟩	Our Prompts with best precision@1	hits@1	Our Prompts with best max-F1	max-F1
compound, *has-parts*	[X] contains [MASK] atom	78.50	[X] has [MASK], which is an atom.	78.52
country, *borders*	[X] and [MASK] share a border	84.86	[X] and [MASK] share a border.	72.82
country, *official-lang*	People of [X] mostly speak in [MASK].	93.37	[MASK] is the main language of [X].	83.57
person, *instrument*	Musician [X] plays [MASK].	67.50	[X] plays [MASK], which is an instrument	62.45
person, *speaks-lang*	In which language can [X] talk? Answer: [MASK].	92.50	[X] speaks in [MASK].	72.78
person, *occupation*	[X] is a well-known [MASK].	59.00	[X] is a well-known [MASK]	33.21
state, *borders*	[MASK], which is a [Y], borders [X].	37.50	[X] and [MASK] share a border	25.71

Table 2: Our best prompts among the 50 relation-specific prompts on precision@1 (%) and max-F1 (%). The [Y] slot takes the object-type information, e.g., in ⟨state, *borders*⟩, [Y] could be "state", "governate", "prefecture", etc.

⟨subject, relation⟩	top-k			prob-x			cumul-x			count-probe			verify-probe			max-F1
	p	r	f1	p	r	f1	p	r	f1	p	r	f1	p	r	f1	
compound, *has-parts*	62.5	78.1	68.0	60.3	76.4	65.4	37.8	74.7	37.8	54.2	78.7	61.8	16.1	70.7	22.9	78.5
country, *borders*	64.0	55.4	54.2	63.4	58.7	55.4	56.0	58.4	46.5	21.9	71.7	30.5	1.9	68.2	3.6	72.8
country, *official-lang*	96.0	74.1	80.1	94.0	75.8	80.5	52.8	71.3	43.2	28.9	82.3	40.5	27.0	32.4	4.9	83.6
person, *instrument*	46.0	42.3	38.8	51.7	40.8	39.1	51.8	41.4	33.8	18.2	60.9	25.5	7.1	24.4	4.8	62.5
person, *speaks-lang*	52.5	59.8	55.1	69.6	56.7	60.0	56.2	57.3	46.8	37.4	69.0	47.3	3.5	53.0	6.4	72.8
person, *occupation*	33.3	23.5	27.3	3.2	85.1	6.1	30.1	36.2	18.6	23.1	30.0	25.9	5.5	41.7	9.0	33.2
state, *borders*	24.4	22.6	22.9	63.1	21.1	24.9	21.9	18.3	13.7	10.0	24.3	13.9	2.4	26.2	4.3	25.7
Overall (averaged)	54.1	50.8	**49.5**	**57.9**	59.2	47.4	43.8	51.1	34.4	27.7	**59.5**	35.1	9.1	45.3	8.0	61.3

Table 3: Results on comparing triples using precision, recall, and F1 score when probing BERT-large and applying a selection mechanism. The bold-faced numbers are the highest achieved precision, recall, and F1 scores.

ence of long-tail subjects. Also for a large possible set of valid objects, e.g., in *occupation*, LM only generates common professions with a high probability and negatively impacts the F1 scores. This behavior, however, helps in *language* type relations. A similar difference in performance can be observed in Shin et al. (2020); Zhong et al. (2021).

Calibrate using Web Signals We used the query hit rate from Bing to calibrate and select objects from the candidate list. Bing receives each ⟨s, r, o⟩ triple in the form of a natural language query. The hit rate is used in two ways: (i) subset, objects with non-zero hit rate, (ii) rerank, objects with non-zero hit rate are calibrated to the highest probability. In contrast to prob-x F1 scores, we observe a high increase in precision and decrease in recall applying (I) with a lower overall F1 score of 34.7%, while the (II) method achieves higher recall and lower precision with a similar overall F1 score of 45.7%. Table 7 in Appendix shows all the scores.

Effect of LM Size We probed larger LMs like T5 (Raffel et al., 2020) and BART (Lewis et al., 2020), which can generate a list of tokens with likelihoods using Beam search decoding algorithm, similar to masked LMs. With top-k method, T5-large achieves 43.6% precision, 41.7% recall, and 40.3% F1 score. BART-large achieves an even lower 32.0% precision, 34.3% recall, and 30.8% F1 score. Table 8 in Appendix gives all the scores. These models tend to generate common objects and exhibit repetitive behavior. Also, the current trend is towards using autoregressive models like (chat-)GPT, like Alivanistos et al. (2022); Cohen et al. (2023) used to extract multi-valued relations. However, unlike in our method, with no control over the selection mechanism, the LM directly outputs one final list in both works. While internally autoregressive models also use token probabilities that could be used for our approach, once one generates a full list, previously generated list items conflate the probabilities of items.

7 Conclusion

In this work, we evaluate using LM's internal representation for materializing factual knowledge on multi-valued relations. We utilize existing prompt engineering techniques and propose new prompts tailored for multi-valued relations. The suggested selection mechanism approaches help to filter out valid triples. Our detailed analysis of the model's performance highlights the limitations of using zero-shot probing for the multi-valued slot-filling task. Future work could aim to improve overall precision and recall and measure the impact of using LMs to fill gaps in KBs.

References

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 *LM-KBC*.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. arXiv:2204.06031.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. 2019. Inducing relational knowledge from bert. In *AAAI Conference on Artificial Intelligence*.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, page 2778–2788, New York, NY, USA. Association for Computing Machinery.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the internal knowledge-base of language models. *Findings of EACL*.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Ankur Goswami, Akshata Bhat, Hadar Ohana, and Theodoros Rekatsinas. 2020. Unsupervised relation extraction from language models using constrained cloze completion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1263–1276, Online. Association for Computational Linguistics.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Simon Razniewski, Andrew Yates, Nora Kassner, and Gerhard Weikum. 2021. Language models as or for knowledge bases. In *Proceedings of the Workshop on Deep Learning for Knowledge Graphs (DL4KG 2021)*. CEUR Workshop Proceedings.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

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

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of ACM*, 57:78–85.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

A Appendix

⟨`subject`, `relation`⟩	Generated Object List	|GT|
(compound, *has-parts*)		
Calcium Carbonate	carbon (0.47), hydrogen (0.03), oxygen (0.02), calcium (0.01), silicon (0.01), nitrogen (0.002), sulfur (0.002)	3
Dopamine	hydrogen (0.09), nitrogen (0.05), carbon (0.05), oxygen (0.05), calcium (0.02), sodium (0.011), sulfur (0.009)	4
Sodium Chloride	carbon (0.15), hydrogen (0.09), oxygen (0.03), silicon (0.01), nitrogen (0.01), sulfur (0.006), sodium (0.006)	2
Thiocyanic Acid	hydrogen (0.1), carbon (0.1), oxygen (0.03), nitrogen (0.02), sulfur (0.01), silicon (0.003), sodium (0.002)	4
Water	oxygen (0.17), hydrogen (0.17), carbon (0.05), nitrogen (0.02), sodium (0.01), mercury (0.004), sulfur (0.003)	2
(country, *borders*)		
Germany	poland (0.14), austria (0.12), france (0.09), italy (0.06), belgium (0.05), russia (0.04), switzerland (0.04)	9
India	pakistan (0.34), bangladesh (0.19), myanmar (0.1), nepal (0.1), china (0.05), iran (0.02), bhutan (0.02)	8
Palau	japan (0.06), indonesia (0.05), taiwan (0.04), fiji (0.03), china (0.02), australia (0.02), philippines (0.01)	3
Malta	gibraltar (0.12), italy (0.11), cyprus (0.07), ireland (0.06), greece (0.05), tunisia (0.04), serbia (0.04)	1
Singapore	malaysia (0.7), thailand (0.1), indonesia (0.1), vietnam (0.02), myanmar (0.02), china (0.01), taiwan (0.01)	2
(country, *official-lang*)		
Algeria	french (0.47), arabic (0.4), spanish (0.04), english (0.03), algerian (0.007), italian (0.005), latin (0.003)	2
Bolivia	spanish (0.9), english (0.07), portuguese (0.01), french (0.006), arabic (0.003), italian (0.002), latin (0.001)	4
Ethiopia	somali (0.52), arabic (0.08), english (0.05), ethiopian (0.04), italian (0.02), spanish (0.01), french (0.01)	1
Singapore	english (0.7), malay (0.18), chinese (0.03), tamil (0.03), mandarin (0.02), indonesian (0.005), arabic (0.003)	4
South Africa	english (0.9), dutch (0.03), french (0.03), portuguese (0.02), spanish (0.01), german (0.006), arabic (0.004)	11
(person, *instrument*)		
A. R. Rahman	guitar (0.29), flute (0.18), piano (0.16), violin (0.08), saxophone (0.05), harmonica (0.04), clarinet (0.03)	16
Andy Hurley	guitar (0.36), piano (0.11), bass (0.06), violin (0.05), cello (0.04), accordion (0.03), drums (0.03)	1
Bruce Springsteen	guitar (0.54), piano (0.08), bass (0.05), drums (0.04), mandolin (0.03), harmonica (0.03), trumpet (0.03)	3
Owen Pallett	guitar (0.34), piano (0.15), violin (0.06), bass (0.05), cello (0.05), trumpet (0.03), drums (0.03)	4
Rino Sashihara	guitar (0.2815), flute (0.12), piano (0.12), violin (0.07), cello (0.04), accordion (0.04), clarinet (0.03)	1
(person, *speaks-lang*)		
Alessandra Ambrosio	italian (0.9), english (0.1), spanish (0.02), french (0.02), german (0.01), portuguese (0.01), latin (0.01)	3
Amy Jackson	english (0.6), spanish (0.1), french (0.1), japanese (0.03), german (0.03), russian (0.02), italian (0.02)	4
Gustavo Petro	spanish (0.7), english (0.2), italian (0.03), portuguese (0.03), french (0.02), german (0.01), catalan (0.006)	4
Gad Elmaleh	english (0.4), arabic (0.4), french (0.13), hebrew (0.03), spanish (0.01), persian (0.007), russian (0.007)	4
Petro Poroshenko	russian (0.4), ukrainian (0.3), english (0.13), polish (0.05), belarusian (0.03), bulgarian (0.006), german (0.006)	6
(person, *occupation*)		
Donald Trump	politician (0.0005), american (0.0005), speaker (0.0004), name (0.0003), personality (0.0003), person (0.0003)	17
Neil Gaiman	author (0.001), writer (0.001), novelist (0.0003), artist (0.0003), contributor (0.0002), character (0.0002)	11
Richard Dawkins	author (0.0024), biologist (0.0019), writer (0.0018), psychologist (0.001), philosopher (0.001), scientist (0.0008)	15
George R. R. Martin	author (0.0032), historian (0.0025), writer (0.0013), scholar (0.0012), biologist (0.0005), novelist (0.0004)	10
Yoko Ono	artist (0.001), singer (0.001), musician (0.0004), actress (0.0003), author (0.0002), writer (0.0002), painter (0.0002)	10
(river, *basins*)		
Aras River	russia (0.18), uzbekistan (0.08), azerbaijan (0.07), armenia (0.06), kazakhstan (0.04), iran (0.04), ukraine (0.03)	4
Draa River	somalia (0.07), ethiopia (0.07), afghanistan (0.02), turkey (0.02), egypt (0.02), algeria (0.02), morocco (0.02)	1
Mekong River	vietnam (0.2001), cambodia (0.19), thailand (0.05), laos (0.03), china (0.01), myanmar (0.003), cameroon (0.003)	6
Limpopo River	botswana (0.25), zambia (0.16), namibia (0.13), zimbabwe (0.08), mozambique (0.07), angola (0.02), africa (0.02)	4
Jordan River	jordan (0.33), israel (0.06), syria (0.05), iraq (0.02), iran (0.02), egypt (0.02), palestine (0.02), lebanon (0.01)	6
(state, *borders*)		
Alabama	mississippi (0.4), georgia (0.3), tennessee (0.1), louisiana (0.05), florida (0.04), arkansas (0.02), texas (0.02)	4
Castile and León	navarre (0.26), galicia (0.22), catalonia (0.14), aragon (0.04), castile (0.02), valencia (0.01), mexico (0.003)	10
La Rioja Province	mendoza (0.05), navarre (0.05), galicia (0.02), madrid (0.01), catalonia (0.007), piedmont (0.006)	5
Gelderland	utrecht (0.40), holland (0.06), hesse (0.02), hamburg (0.01), jersey (0.003), bremen (0.002), berlin (0.001)	7
Fukushima Prefecture	tokyo (0.23), hiroshima (0.14), okinawa (0.13), kyoto (0.1), osaka (0.03), nagoya (0.03), saga (0.01)	6

Table 4: Samples of generated object list for five unique subjects on multi-valued relations. The green highlighted valid objects, while the red ones are wrong. The |GT| column gives the total no. of ground-truth objects for the corresponding subject.

⟨subject, relation⟩	Metric	avg-cutoff	Our Prompts
compound, *has-parts*	top-k	4	[X] has [MASK], which is an atom
	prob-x	0.02	
	cumul-x	0.53	
	count-alpha	2.26	[X] consists of [MASK] elements.
	count-num	4.64	[X] consists of [MASK] elements
	verify-probe	$\alpha = 0.06$	[X] consists of [Y] atom. Is this correct? Answer: [MASK].
country, *borders*	top-k	3	[X] and [MASK] share a border.
	prob-x	0.05	
	cumul-x	0.79	
	count-alpha	2.54	[X] shares border with [MASK] countries.
	count-num	13.36	[X] shares border with [MASK] countries
	verify-probe	$\alpha = 0$	[X] and [Y] share a border. Is this correct? Answer: [MASK].
country, *official-lang*	top-k	1	[MASK] is the main language of [X].
	prob-x	0.22	
	cumul-x	0.91	
	count-alpha	3.42	[X] has [MASK] official languages.
	count-num	2.18	[X] has [MASK] official languages
	verify-probe	$\alpha = 0.11$	[Y] is the official language of [X]. Is this correct? Answer: [MASK].
person, *instrument*	top-k	2	[X] plays [MASK], which is an instrument
	prob-x	0.12	
	cumul-x	0.54	
	count-alpha	2.98	[X] plays [MASK] instruments.
	count-num	6.88	[X] plays [MASK] instruments
	verify-probe	$\alpha = 0.28$	[X] plays [Y]. Is this correct? Answer: [MASK].
person, *speaks-lang*	top-k	4	[X] speaks in [MASK].
	prob-x	0.05	
	cumul-x	0.87	
	count-alpha	4.16	[X] speaks in [MASK] languages.
	count-num	6	
	verify-probe	$\alpha = 0.24$	[X] can speak in [Y]. Is this correct? Answer: [MASK].
person, *occupation*	top-k	8	[X] is a well-known [MASK]
	prob-x	0	
	cumul-x	0.01	
	count-alpha	4.74	[X] had a total of [MASK] different professions.
	count-num	13.64	
	verify-probe	$\alpha = 0$	[X] is a well-known [Y]. Is this correct? Answer: [MASK].
state, *borders*	top-k	5	[X] and [MASK] share a border
	prob-x	0.04	
	cumul-x	0.75	
	count-alpha	3.16	[X] shares border with [MASK] states
	count-num	13.04	[X] shares border with a total of [MASK] states.
	verify-probe	$\alpha = 0$	[X] and [Y] share a border. Is this correct? Answer: [MASK].

Table 5: The prompt templates used for generating the object list. The avg-cutoff shows the learned parameters of each selection mechanism averaged across all subjects.

⟨`subject`, `relation`⟩	Our Prompts	avg-cutoff	Precision	Recall	F1 score	Max-F_1
compound, *has-parts*	`[X]` has `[MASK]`, which is an atom.	0.02	60.33	76.40	65.41	78.52
country, *borders*	`[X]` has borders with `[MASK]`.	0.07	74.73	55.41	58.45	71.41
country, *official-lang*	The official language of `[X]` is `[MASK]`.	0.15	94.33	79.97	83.38	83.54
person, *instrument*	`[X]` likes to play the `[MASK]`.	0.12	51.33	46.52	41.48	58.19
person, *speaks-lang*	`[X]` speaks in `[MASK]`.	0.05	69.60	56.72	59.99	72.78
person, *occupation*	`[X]` is a `[MASK]`.	0.01	22.79	31.26	25.42	29.30
state, *borders*	`[X]` and `[MASK]` share a border	0.04	63.10	21.12	24.91	25.71
Overall			62.32	52.49	51.29	59.9

Table 6: Higher F1 score achieved by using a different set of our proposed prompts and prob-x mechanism.

⟨`subject`, `relation`⟩	Our Prompts	Prob-x			Nonzero Hit-rate			ReRank		
		p	r	f1	p	r	f1	p	r	f1
compound, *has-parts*	`[X]` has `[MASK]`, which is an atom.	60.3	76.4	**65.4**	84.2	17.7	17.0	57.0	80.9	64.1
country, *borders*	`[X]` has borders with `[MASK]`.	74.7	55.4	58.5	82.3	53.5	56.0	58.9	69.4	**58.8**
country, *official-lang*	The official language of `[X]` is `[MASK]`.	94.3	80.0	**83.4**	91.7	68.3	71.1	69.3	81.6	69.3
person, *instrument*	`[X]` likes to play the `[MASK]`.	51.3	46.5	**41.5**	83.0	30.1	33.8	27.7	53.2	32.4
person, *speaks-lang*	`[X]` speaks in `[MASK]`.	69.6	56.7	**60.0**	83.0	19.7	22.1	48.5	66.3	51.2
person, *occupation*	`[X]` is a `[MASK]`.	23.0	31.3	**25.4**	19.7	23.7	20.3	19.8	24.1	20.5
state, *borders*	`[X]` and `[MASK]` share a border	63.1	21.1	**24.9**	93.0	18.3	22.3	69.0	21.9	23.4
Overall		57.9	59.2	**47.4**	76.7	33.1	34.7	50.1	56.8	45.7

Table 7: Results on calibrating object probabilities with Bing hit rates

⟨`subject`, `relation`⟩	Our Prompts	T5-large			BART-large		
		p	r	f1	p	r	f1
compound, *has-parts*	`[X]` has `[MASK]`, which is an atom.	67.3	64.9	61.5	58.7	56.2	56.2
country, *borders*	`[X]` and `[MASK]` share a border.	43.8	53.1	45.0	44.8	60.4	47.3
country, *official-lang*	`[MASK]` is the main language of `[X]`.	82.0	61.1	66.8	0	0	0
person, *instrument*	`[X]` plays `[MASK]`, which is an instrument	13.3	14.8	12.6	40.0	36.1	33.5
person, *speaks-lang*	`[X]` speaks in `[MASK]`.	47.0	40.4	43.0	22.1	28.7	24.5
person, *occupation*	`[X]` is a well-known `[MASK]`	27.2	36.4	30.8	36.5	37.1	34.4
state, *borders*	`[X]` and `[MASK]` share a border	24.6	21.5	22.3	21.5	21.4	19.8
Overall		43.6	41.7	40.3	32.0	34.3	30.8

Table 8: Results on probing T5 and BART model with top-k selection mechanism

Prompt Template	optimal k	precision	recall	max-F1	p@1
[X] consists of [MASK].	81.08	42.68	68.58	45.14	31.66
[X] consists of [MASK]	251.54	22.73	26.18	20.08	11.05
[X] consists of [MASK] element.	48.12	49.8	66.64	49.28	37.56
[X] consists of [MASK] element	142.47	33.57	54.88	36.63	13.13
[X] consists of [MASK], which is an element.	33.44	67.84	72.2	64.11	68.00
[X] consists of [MASK], which is an element	39.69	60.18	70.56	57.03	54.27
The chemical compound [X] consists of [MASK].	23.11	50.65	73.09	52.23	42.00
The chemical compound [X] consists of [MASK]	196.56	26.73	27.74	23.52	13.00
The chemical compound [X] consists of [MASK] element.	26.91	47.05	64.6	43.38	33.00
The chemical compound [X] consists of [MASK] element	51.63	36.05	66.4	39.33	15.00
The chemical compound [X] consists of [MASK], which is an element.	21.43	63.34	69.99	59.83	56.50
The chemical compound [X] consists of [MASK], which is an element	28.12	56.23	69.19	51.97	48.50
[X] contains [MASK].	42.52	46.27	75.88	52.78	36.50
[X] contains [MASK] atom	20.13	72.82	78.49	72.45	78.50
[X] is composed of [MASK], which is an element.	23.99	66.23	74.15	64.01	65.50
[X] is composed of [MASK], which is an element	30.2	56.93	71.69	55.02	52.00
[X] is composed of [MASK] atom.	20.82	68.12	79.93	69.18	62.50
[X] is composed of [MASK] atom	21.11	70.26	75.46	67.22	71.00
[X] is composed of [MASK].	58.15	47.92	72.85	50.77	36.00
[X] is composed of [MASK]	172.04	28.28	35.87	26.18	14.75
[MASK] atom is present in [X].	32.04	35.01	79.21	45.23	4.50
[MASK] atom is present in [X]	35.25	33.03	76.98	42.03	5.50
[MASK] element is present in [X].	377.61	8.5	10.58	8.27	1.00
[MASK] element is present in [X]	312.44	12.47	15.98	12.47	2.51
[MASK] is present in [X].	79.54	27.81	66.19	35.53	7.54
[MASK] is present in [X]	111.5	27.42	61.21	33.08	8.25
[X] has [MASK], which is an element.	27.92	69.34	77.11	69.02	68.00
[X] has [MASK], which is an element	32.81	70.32	73.68	67.91	68.84
[X] has [MASK], which is an atom.	20.07	78.75	82.76	78.52	76.00
[X] has [MASK], which is an atom	20.02	74.81	81.77	75.95	74.50
[X] molecule is composed of [MASK], which is an element.	20.5	73.41	78.48	72.04	76.00
[X] molecule is composed of [MASK], which is an element	23.7	65.93	74.53	63.84	62.50
[X] molecule is composed of [MASK] atom.	20.62	70.35	80.6	71.32	69.00
[X] molecule is composed of [MASK] atom	20.77	71.7	77.44	69.94	73.00
[X] molecule is composed of [MASK].	21.23	63.81	80.4	67.24	53.00
[X] molecule is composed of [MASK]	149.31	37.28	34.97	30.58	24.00
[MASK] atom is present in [X] molecule.	23.69	41.1	79.8	50.42	10.50
[MASK] atom is present in [X] molecule	24.7	36.06	78.64	45.96	7.00
[MASK] element is present in [X] molecule.	144.71	19.53	47.54	25.17	1.50
[MASK] element is present in [X] molecule	190.69	20.99	38.21	22.93	4.50
[MASK] is present in [X] molecule.	71.54	31.73	68.47	37.97	10.00
[MASK] is present in [X] molecule	69.52	29.54	72.49	38.01	6.00
The [X] molecule consists of [MASK].	20.48	72.41	81.54	73.33	59.50
The [X] molecule consists of [MASK]	297.46	27.53	19.48	20.14	17.00
The [X] molecule consists of [MASK] element.	31.69	56.43	73.09	57.56	51.50
The [X] molecule consists of [MASK] element	54.96	39.27	64.37	43.4	11.50
The [X] molecule consists of [MASK], which is an element.	20.32	75.89	78.82	73.64	76.00
The [X] molecule consists of [MASK], which is an element	23.21	70.36	76.15	67.98	66.50
[X] molecule has [MASK], which is an element.	20.05	75.17	82.05	75.9	75.50
[X] molecule has [MASK], which is an element	20.1	75.67	81.32	75.68	77.00

Table 9: Our proposed prompts for (chemical compound, has parts) relation.

Prompt Template	optimal k	precision	recall	max-F1	p@1
[X] shares border with [MASK].	9.3	76.73	77.11	71.29	83.24
[X] shares border with [MASK]	10.92	65.44	70.6	62.02	70.81
[X] shares a border with [MASK].	9.49	78.63	76.59	72.1	83.78
[X] shares a border with [MASK]	11.62	61.24	69.24	58.22	65.95
[X] borders [MASK].	10.37	72.79	74.9	67.98	80.00
[X] borders [MASK]	25.04	25.82	60.28	28.67	12.97
[X] has borders with [MASK].	9.54	79.26	75.14	71.41	82.70
[X] has borders with [MASK]	19.76	49.49	62.74	43.75	48.65
[X] shares border with [MASK], which is a country.	10.01	77.89	77.04	71.49	82.16
[X] shares border with [MASK], which is a country	10.08	78.13	76.62	71.38	81.62
The neighbouring country of [X] is [MASK].	11.07	76.04	74.02	69.13	80.00
The neighbouring country of [X] is [MASK]	74.58	21.43	56.52	28.12	4.32
The neighbouring countries of [X] are [MASK].	46.29	35.97	61.2	39.71	21.08
The neighbouring countries of [X] are [MASK]	149.41	23.07	43.63	24.41	8.65
[X] shares a border with [MASK], which is a country.	10.46	77.38	77.05	71.45	82.16
[X] shares a border with [MASK], which is a country	10.27	78.38	76.07	71.32	83.78
[X] borders [MASK], which is a country.	10.08	78.15	75.92	71.54	82.16
[X] borders [MASK], which is a country	10.03	76.53	76.39	71.04	81.62
[MASK] is a neighbouring country of [X].	10.14	76.35	75.51	70.02	83.24
[MASK] is a neighbouring country of [X]	10.51	73.66	74.97	68.1	80.00
Which country shares border with [X]? Answer: [MASK].	8.74	58.42	71.62	57.49	53.51
Which country shares border with [X]? Answer: [MASK]	119.09	19.56	45.47	22.45	5.41
Which country is near [X]? Answer: [MASK].	8.12	55.89	70.71	56.9	45.95
Which country is near [X]? Answer: [MASK]	89.11	19.21	45.86	22.92	5.95
[MASK], which is a country, is near [X].	9.75	76.37	76.25	70.35	82.16
[MASK], which is a country, is near [X]	9.63	69.4	72.55	64.93	71.89
The country, [MASK], shares border with [X].	9.42	71.85	71.59	64.32	74.59
The country, [MASK], shares border with [X]	11.21	62.09	67.23	56.32	64.32
[MASK] is the bordering country of [X].	10.91	75.9	75.07	69.37	81.62
[MASK] is the bordering country of [X]	12.69	65.9	71.83	61.78	70.27
The country, [MASK], shares border with [X], which is a country.	8.52	71.85	74.25	66.52	75.68
The country, [MASK], shares border with [X], which is a country	8.69	71.83	73.57	66.28	76.76
[MASK], which is a country, shares a border with [X].	11.26	75.08	74.45	68.71	80.00
[MASK], which is a country, shares a border with [X]	11.98	68.41	72.85	64.18	74.05
[MASK], which is a country, borders [X].	11.23	75.63	75.7	69.52	80.54
[MASK], which is a country, borders [X]	11.9	72.8	74.06	67.23	75.68
[MASK], which is a country, has borders with [X].	13.42	72.52	74.88	67.53	77.84
[MASK], which is a country, has borders with [X]	12.77	67.38	71.64	62.24	71.89
The neighbouring country of [X] is [MASK], which is a country.	9.43	78.01	75.52	71.19	83.24
The neighbouring country of [X] is [MASK], which is a country	8.99	76.85	75.96	70.78	83.24
Which country shares border with [X]? The answer is [MASK].	234.62	27.4	28.8	24.76	21.62
Which country shares border with [X]? The answer is [MASK]	477.46	2.52	1.52	1.82	1.08
Which country shares border with [X]? Answer: [MASK], which is a country.	7.36	67.73	73.18	64.68	67.57
Which country shares border with [X]? Answer: [MASK], which is a country	7.92	61.61	73.56	61.44	61.62
[X] and [MASK] share a border.	9.19	80.08	76.7	72.82	83.78
[X] and [MASK] share a border	9.18	78.29	77.21	71.96	84.86
[X] and [MASK] are neighbouring countries.	10.86	73.27	73.9	67.27	77.30
[X] and [MASK] are neighbouring countries	11.32	72.26	73.86	66.7	77.30
[X] and [MASK] are neighbours.	11.08	77.04	75.67	70.76	83.24
[X] and [MASK] are neighbours	8.22	75.98	75.1	70.06	82.70

Table 10: Our proposed prompts for (country, shares borders) relation.

Prompt Template	optimal k	precision	recall	max-F1	p@1
The official language of [X] is [MASK].	18.18	92.58	80.43	83.54	91.84
The official language of [X] is [MASK]	18.22	90.73	79.66	81.66	88.27
[MASK] is the official language of [X].	18.16	93.09	80.04	83.4	92.35
[MASK] is the official language of [X]	18.2	92.15	80.17	83.1	91.33
[X] has [MASK] as its official language.	18.3	87.26	79.62	79.51	83.16
[X] has [MASK] as its official language	18.43	83.35	79.92	77.36	75.00
The official languages of [X] are [MASK].	18.19	92.08	80.3	83.02	90.31
The official languages of [X] are [MASK]	25.24	29.18	77.04	33.08	16.84
The main language spoken in [X] is [MASK].	18.22	92.43	79.69	83.41	92.35
The main language spoken in [X] is [MASK]	18.42	86.31	79.28	79.09	82.14
People of [X] mostly speak in [MASK], which is a language.	18.2	92.17	80.03	83.26	91.84
People of [X] mostly speak in [MASK], which is a language	18.18	92.31	79.86	83.26	92.86
People of [X] mostly speak in [MASK].	18.13	93.37	79.66	83.41	93.37
People of [X] mostly speak in [MASK]	18.26	88.45	79.75	80.4	84.69
[MASK] is the main spoken language of [X].	18.14	93.22	79.66	83.43	92.35
[MASK] is the main spoken language of [X]	18.16	92.49	79.66	83.05	91.84
[MASK] is spoken in [X].	26.31	66.55	78.05	63.59	60.20
[MASK] is spoken in [X]	26.92	59.87	77.15	58.95	50.51
Language spoken in [X] is [MASK].	18.27	91.67	79.95	82.86	91.33
Language spoken in [X] is [MASK]	19.35	75.56	78.09	70.91	67.86
Languages spoken in [X] are [MASK].	18.95	75.13	78.81	71.35	66.33
Languages spoken in [X] are [MASK]	49.03	12.19	75.08	18.92	0.00
What are the main languages spoken in [X]? Answer: [MASK].	18.44	81.97	80.09	76.73	71.94
What are the main languages spoken in [X]? Answer: [MASK]	119	13.16	61.35	18.35	2.55
What are the official languages of [X]? Answer: [MASK].	18.4	83.09	80.47	77.59	72.45
What are the official languages of [X]? Answer: [MASK]	61.16	14.15	69.61	20.96	1.02
What is the official language of [X]? Answer: [MASK].	18.35	84.53	80.17	78.37	73.98
What is the official language of [X]? Answer: [MASK]	68.94	15.07	69.35	21.67	3.06
Which language is officially spoken in [X]? Answer: [MASK].	18.44	82.86	80.47	77.3	74.49
Which language is officially spoken in [X]? Answer: [MASK]	80.3	14.68	66.10	19.17	4.08
[MASK] is the main language of [X].	18.15	93.22	80.04	83.57	92.35
[MASK] is the main language of [X]	18.16	92.71	79.92	83.32	91.84
In [X], people speak in [MASK].	18.26	91.93	79.86	82.93	92.35
In [X], people speak in [MASK]	19.64	78.75	78.74	72.13	71.94
In [X], people speak in [MASK], which is an official language.	18.22	92.36	79.78	83.18	92.86
In [X], people speak in [MASK], which is an official language	18.23	92.39	79.69	83.18	92.35
In [X], people speak in [MASK], which is a language.	18.21	91.8	79.78	82.9	91.84
In [X], people speak in [MASK], which is a language	18.19	91.89	79.35	82.78	90.82
In [X], people mainly speak in [MASK].	18.14	92.73	79.66	83.15	92.35
In [X], people mainly speak in [MASK]	18.49	87.07	79.16	78.43	83.67
In [X], people mainly speak in [MASK], which is an official language.	18.17	92.46	79.92	82.96	93.37
In [X], people mainly speak in [MASK], which is an official language	18.21	92.35	79.52	83.05	92.35
In [X], people mainly speak in [MASK], which is a language.	18.21	91.78	79.78	82.91	90.82
In [X], people mainly speak in [MASK], which is a language	18.17	92.19	79.35	82.98	91.33
The national language of [X] is [MASK].	18.18	92.62	80.17	83.38	91.33
The national language of [X] is [MASK]	18.37	89.85	78.85	80.39	86.73
[X] is a country and [MASK] is the official language.	18.23	90.15	80.00	81.59	87.76
[X] is a country and [MASK] is the official language	18.28	88.03	80.17	80.49	83.67
[MASK] is the national language of [X].	18.2	92.24	80.43	83.23	91.84
[MASK] is the national language of [X]	18.2	92.24	80.17	83.2	91.84

Table 11: Our proposed prompts for (country, has official language) relation.

Prompt Template	optimal k	precision	recall	max-F1	p@1
[X] plays [MASK].	83.77	52.59	60.22	49.03	49.00
[X] plays [MASK]	56.31	31.27	66.57	32.52	20.50
[X] plays [MASK] instrument.	17.48	36.88	76.01	42.07	24.00
[X] plays [MASK] instrument	19.72	35.9	73.79	39.48	20.00
[X] plays [MASK] musical instrument.	59.91	16.09	62.72	19.48	5.00
[X] plays [MASK] musical instrument	71.2	11.67	61.89	15.84	2.50
[X] plays [MASK], which is an instrument.	14.12	65.16	74.57	61.24	61.50
[X] plays [MASK], which is an instrument	14.1	67.33	74.74	62.45	66.00
Musician [X] plays [MASK].	15.1	67.36	74.24	61.23	67.50
Musician [X] plays [MASK]	54.72	21.35	66.87	26.93	8.00
Musician [X] plays [MASK] instrument.	30.13	17.17	72.3	23.44	2.50
Musician [X] plays [MASK] instrument	35.04	17.36	68.15	22.46	2.50
Musician [X] plays [MASK], which is an instrument.	14.12	66.99	73.75	61.72	65.50
Musician [X] plays [MASK], which is an instrument	13.81	68.78	72.2	61.11	65.00
[X] played the [MASK].	14.77	63.76	74.19	59.37	63.00
[X] played the [MASK]	18.53	48.28	72.43	48.46	44.50
[X] played the [MASK], which is an instrument.	14.39	59.86	73.18	57.25	58.00
[X] played the [MASK], which is an instrument	14.21	63.68	70.63	56.69	57.50
[MASK], which is an instrument, was played by [X].	15.02	58.45	74.3	55.99	49.50
[MASK], which is an instrument, was played by [X]	14.68	61.6	73.21	56.85	52.00
[X] likes to play the [MASK].	14.29	66.32	70.17	58.19	62.50
[X] likes to play the [MASK]	24.56	42.42	65.56	41.5	26.00
[X] performed on (her|his) [MASK], which is an instrument.	14.08	63.14	72.04	57.15	64.00
[X] performed on (her|his) [MASK], which is an instrument	14.06	65.43	71.42	58.36	67.00
[X] likes to play the [MASK], which is an instrument.	13.83	66.84	69.79	58.34	61.00
[X] likes to play the [MASK], which is an instrument	13.68	68.08	69.25	58.29	63.50
[X] knows to play the [MASK].	13.97	65.89	69.83	58.75	62.00
[X] knows to play [MASK]	486.81	0.40	1.20	0.32	0.00
[X] knows to play the [MASK] instrument.	20.21	26.09	72.71	32.17	10.50
[X] knows to play [MASK] instrument	17.64	47.08	72.52	47.46	35.50
[X] knows to play the [MASK], which is an instrument.	13.96	64.67	70.09	57.21	58.00
[X] knows to play the [MASK], which is an instrument	13.93	64.96	69	56.24	57.50
[X] can play [MASK].	15.03	59.04	74.3	56.56	57.00
[X] can play [MASK]	292.6	9.42	26.31	11.35	2.00
[X] can play [MASK], which is an instrument.	14.16	67.2	73.89	61.7	66.00
[X] can play [MASK] instrument.	19.54	34.77	72.25	37.44	23.00
[X] is noted for playing [MASK] instrument.	32.89	14.19	70.87	19.81	3.00
[X] is noted for playing [MASK] instrument	35.96	16.07	71.73	21.05	1.50
[X] is noted for playing [MASK], which is an instrument.	14.14	66.44	74.48	61.75	64.50
[X] is noted for playing [MASK], which is an instrument	14.18	65.82	73.17	60.11	65.50
[X] is noted for playing [MASK].	17.33	63.09	73.6	58.68	64.50
[X] is noted for playing [MASK]	38.19	18.92	69.68	23.92	5.50
[X] is noted for playing [MASK], which is an instrument.	14.14	66.44	74.48	61.75	64.50
[X] is noted for playing [MASK], which is an instrument	14.18	65.82	73.17	60.11	65.50
[X] practised [MASK], which is an instrument.	14.31	66.18	70.75	58.3	61.50
[X] practised [MASK], which is an instrument	14.38	66.34	71.79	58.92	64.50
[X] taught [MASK], which is an instrument.	14.27	67.04	72.28	60.67	65.00
[X] taught [MASK], which is an instrument	14.1	67.6	72.64	61.16	66.50
[X] performed on (his|her) [MASK], which is an instrument.	14.44	51.89	73.1	51.84	40.50
[X] performed on (his|her) [MASK], which is an instrument	14.01	63.91	71.72	57.67	67.00

Table 12: Our proposed prompts for (person, plays an instrument) relation.

Prompt Template	optimal k	precision	recall	max-F1	p@1
[X] speaks in [MASK].	2.61	82.13	72.35	72.78	83.50
[X] speaks in [MASK]	5.73	54.6	61.08	48.62	47.50
[X] can speak in [MASK].	2.56	81.63	69.37	70.76	82.50
[X] can speak in [MASK]	7.26	63.35	66.07	57.14	60.50
[X] communicates in [MASK].	2.94	77.04	71.28	69.93	74.50
[X] communicates in [MASK]	6.71	48.41	64.69	47.07	36.50
[X] spoke in [MASK].	2.59	82.62	70.24	71.32	86.00
[X] spoke in [MASK]	7.28	38.36	59.67	39.74	20.00
[X] communicated in [MASK].	3.09	77.38	71.76	69.41	79.50
[X] communicated in [MASK]	13.03	32.09	58.54	35.86	11.00
[X] knows the [MASK] language.	3.18	76.9	68.85	67.49	78.50
[X] knows the [MASK] language	4.04	70.16	70.24	63.91	75.50
[X] learnt [MASK], which is a language.	2.68	79.65	70.03	70.23	81.00
[X] learnt [MASK], which is a language	3.18	74.11	72.93	68.82	79.00
[X] knows [MASK], which is a language.	2.58	82.19	68.29	69.92	88.50
[X] knows [MASK], which is a language	2.54	81.04	68.78	70.12	87.00
Languages spoken by [X] are [MASK].	2.64	78.99	68.52	68.6	86.50
Languages spoken by [X] are [MASK]	19.64	24.57	57.63	29.39	5.50
In which language can [X] speak? Answer: [MASK].	2.45	82.8	65.31	68.18	91.50
In which language can [X] speak? Answer: [MASK]	110.26	19.29	39.12	23.86	1.00
In which language can [X] talk? Answer: [MASK].	2.53	82.43	64.67	67.2	92.50
In which language can [X] talk? Answer: [MASK]	157.13	18.64	37.15	22.08	3.00
In which language can [X] communicates? Answer: [MASK]	142.3	19.28	36.47	22.15	3.00
In which language can [X] communicates? Answer: [MASK]	142.3	19.28	36.47	22.15	3.00
[X] knows to speak in [MASK].	2.6	82.13	69.34	70.88	86.50
[X] knows to speak in [MASK]	18.97	50.01	61.15	45.91	42.50
[X] speaks in [MASK], which is a language.	2.85	78.28	70.92	70.02	82.00
[X] speaks in [MASK], which is a language	3.06	74.5	70.92	68.25	78.00
[X] can speak in [MASK], which is a language.	2.61	81.86	70.11	71.22	85.50
[X] can speak in [MASK], which is a language	2.88	77.43	71.18	69.82	84.50
In [MASK], [X] spoke.	3.12	72.21	62.23	61.27	74.50
In [MASK], [X] spoke	9.81	39.6	57.87	40.93	26.50
In [MASK], which is a language, [X] spoke.	2.83	77.59	69.16	68.24	80.50
In [MASK], which is a language, [X] spoke	2.91	77.5	69.2	68.11	78.50
[X] learned to speak [MASK] fluently.	2.89	78.81	71.55	70.46	80.00
[X] learned to speak [MASK] fluently	2.91	80.46	72.49	71.38	83.50
[X] learned to speak [MASK], which is a language.	2.73	80.59	70.28	70.4	82.50
[X] learned to speak [MASK], which is a language	3.26	75.33	71.25	68.07	80.00
[X] communicates in [MASK], which is a language.	2.44	83.7	67.28	69.82	85.50
[X] communicates in [MASK], which is a language	2.47	82.43	68.62	70.25	80.50
[X] spoke in [MASK], which is a language.	3.17	74.57	69.58	66.67	79.00
[X] spoke in [MASK], which is a language	3.92	67.89	71.21	63.95	73.00
[X] knows to speak in [MASK], which is a language.	2.48	81.9	68.69	70.36	85.00
[X] knows to speak in [MASK], which is a language	2.97	77.49	70.76	69.88	81.00
[X] learned to speak [MASK].	2.99	76.54	70.91	68.95	80.00
[X] learned to speak [MASK]	76.98	18.98	42.24	22.06	3.50
[X] learnt [MASK] language.	4.46	60.35	66.61	56.26	55.00
[X] learnt [MASK] language	4.62	62.27	68.51	58.51	63.00
[X] addressed in [MASK], which is a language.	2.52	81.3	68.71	70.15	82.00
[X] addressed in [MASK], which is a language	2.69	79.85	70.65	70.75	82.00

Table 13: Our proposed prompts for (person, speaks a language) relation.

Prompt Template	optimal k	precision	recall	max-F1	p@1
[X] is a [MASK] by profession.	30.87	18.52	35.37	20.78	8.00
[X] is a [MASK] by profession	28.39	20.08	34.17	21.23	13.00
[X] is a [MASK].	19.23	35.01	35.98	29.3	40.50
[X] is a [MASK]	20.02	34.14	33.8	26.34	41.50
[X] is a [MASK], which is a profession.	23.15	30.17	36.72	26.81	31.00
[X] is a [MASK], which is a profession	17.87	35.11	31.63	26.43	40.00
[X]'s profession is [MASK].	61.53	13.22	44.71	17.2	3.00
[X]'s profession is [MASK]	111.9	6.14	22.31	6.72	1.50
[X] worked as a [MASK].	25.75	21.29	33.97	22.68	20.50
[X] worked as a [MASK]	16.41	30.66	27.78	24.41	25.50
[X] was a [MASK] for a living.	20.94	37.44	27.59	23.97	48.00
[X] is a [MASK] for a living.	15.66	38.35	30.19	27.14	53.00
[X] is a well-known [MASK].	14.58	39.76	33.36	30.8	59.00
[X] is a well-known [MASK]	11.64	45.35	32.37	33.21	54.50
[X] worked as a [MASK], which is a profession.	23.93	25.12	35.08	24.33	22.50
[X] worked as a [MASK], which is a profession	22.28	25.84	34.95	24.62	25.00
[X] worked as a [MASK] for a living.	29.44	25.35	32.45	21.22	27.00
[X] worked as a [MASK] for a living	27.66	23.66	32.75	21.08	26.00
[X] works as a [MASK].	32.09	17.79	38.41	21.05	19.00
[X] works as a [MASK]	21.62	25.43	31.52	23.02	20.50
What is the profession of [X]? Answer: [MASK].	158.97	2.68	20.45	4.11	0.00
What is the profession of [X]? Answer: [MASK]	216.16	5.51	6.41	3.68	1.50
What did [X] do for a living? Answer: [MASK].	225.71	1.8	8.83	2.33	0.00
What did [X] do for a living? Answer: [MASK]	326.04	3.28	3.27	2.23	1.00
[X] worked as a professional [MASK].	20.84	26.52	36.39	26.9	24.00
[X] worked as a professional [MASK]	39.51	14.9	27.85	15.78	1.00
[X] works as a professional [MASK].	22.99	25.73	38.29	26.45	23.00
[X] works as a professional [MASK]	35.24	16.39	30.64	17.54	2.00
[X] is a professional [MASK].	33.5	19.12	42.3	23.4	2.50
[X] is a professional [MASK]	30.51	15.1	31.41	18	3.50
[X] was a [MASK].	35.97	24.36	28.63	20.46	13.00
[X] was a [MASK]	29.84	31.56	26.78	21.66	33.00
[X] served as a [MASK].	34.84	19.08	28.5	17.56	3.00
[X] served as a [MASK]	30.13	20.31	25.71	17.14	9.00
[X] served as a [MASK], which is a profession.	26.62	25.43	29.37	21.94	13.00
[X] served as a [MASK], which is a profession	35.59	16.22	24.68	16.07	0.00
[X] became a [MASK].	24.79	22.97	30.52	21.71	20.00
[X] became a [MASK]	21.07	28.28	24.12	19.7	26.00
[X] became a [MASK], which is a profession.	24.82	23.87	35.55	24.4	24.50
[X] became a [MASK], which is a profession	23.56	23.95	30.41	21.78	19.00
[X] is an [MASK].	6.75	63.21	14.61	19.86	54.50
[X] is an [MASK]	14.35	26.24	13.94	15.54	11.00
[X] was an [MASK].	6.71	61.71	13.59	18.83	50.00
[X] was an [MASK]	20.49	31.84	13.04	14.65	18.00
[X] was a professional [MASK].	30.19	17.73	36.75	21.38	4.50
[X] was a professional [MASK]	51.69	12.68	29.00	14.65	0.50
[X] joined as a [MASK], which is a profession.	20.01	33.47	31.79	26.26	36.50
[X] joined as a [MASK], which is a profession	17.54	27.09	29.91	25.06	9.50
[X] joined as a [MASK].	21.72	23.88	31.06	24.16	8.00
[X] joined as a [MASK]	26.56	19.02	30.65	20.96	0.00

Table 14: Our proposed prompts for (person, has an occupation) relation.

Prompt Template	optimal k	precision	recall	max-F1	p@1
[X] shares border with [MASK].	285.62	35.78	21.85	25.18	32.50
[X] shares border with [MASK]	291.63	25.94	19.15	19.14	26.50
[X] shares a border with [MASK].	285.64	36.08	22.02	25.39	32.50
[X] shares a border with [MASK]	296.5	22.06	18.2	17.32	20.00
[X] borders [MASK].	285.81	35.48	21.51	24.63	34.00
[X] borders [MASK]	316.65	9.03	16.07	9.48	2.50
[X] has borders with [MASK].	286.25	32.6	21.67	24.1	30.00
[X] has borders with [MASK]	306.04	10.59	16.58	10.83	3.50
[X] shares border with [MASK], which is a [Y].	285.38	37.32	21.48	25.21	35.50
[X] shares border with [MASK], which is a [Y]	285.53	36.7	21.8	25.16	34.50
The neighbouring [Y] of [X] is [MASK].	290.67	32.64	20.87	23.52	32.50
The neighbouring [Y] of [X] is [MASK]	371.29	6.00	8.67	6.08	0.50
The [X] [Y] shares border with [MASK].	285.53	35.92	21.6	25.06	32.00
The [X] [Y] shares border with [MASK]	293.25	14.16	17.2	12.97	6.50
[X] shares a border with [MASK], which is a [Y].	285.4	37.27	21.6	25.3	35.00
[X] shares a border with [MASK], which is a [Y]	285.56	36.81	21.77	25.21	35.50
[X] borders [MASK], which is a [Y].	285.51	36.36	22.08	25.1	35.00
[X] borders [MASK], which is a [Y]	285.56	36.13	22.08	25.15	34.00
[MASK] is a neighbouring [Y] of [X].	285.45	35.96	21.79	25.04	34.50
[MASK] is a neighbouring [Y] of [X]	285.51	34.73	21.43	24.52	31.50
Which [Y] shares border with [X]? Answer: [MASK].	293.96	23.81	20.94	20.63	25.00
Which [Y] shares border with [X]? Answer: [MASK]	429.81	6.28	5.16	4.78	3.00
Which [Y] is near [X]? Answer: [MASK].	286.75	23.85	20.66	20.44	20.00
Which [Y] is near [X]? Answer: [MASK]	432.43	5.58	6.21	4.67	1.50
[MASK], which is a [Y], is near [X].	285.42	37.09	21.85	25.39	35.50
[MASK], which is a [Y], is near [X]	286.12	29.84	21.02	22.02	29.50
The [Y], [MASK], shares border with [X].	285.78	32.22	21.64	24.1	31.50
The [Y], [MASK], shares border with [X]	286.05	28.00	21.31	22.26	26.50
[MASK] is the bordering [Y] of [X].	285.41	37.06	21.94	25.43	35.50
[MASK] is the bordering [Y] of [X]	285.54	35.88	21.9	24.69	36.50
The [Y], [MASK], shares border with [X], which is a [Y].	285.61	34.00	21.25	24.23	34.00
The [Y], [MASK], shares border with [X], which is a [Y]	285.64	33.48	21.88	24.29	34.50
[MASK], which is a [Y], shares a border with [X].	285.37	37.96	21.86	25.49	37.00
[MASK], which is a [Y], shares a border with [X]	285.41	37.29	21.5	25.1	36.50
[MASK], which is a [Y], borders [X].	285.35	38.46	21.8	25.65	37.50
[MASK], which is a [Y], borders [X]	285.46	36.25	21.46	24.68	36.00
[MASK], which is a [Y], has borders with [X].	285.37	37.9	21.78	25.4	37.00
[MASK], which is a [Y], has borders with [X]	285.41	36.57	21.07	24.57	36.00
The neighbouring [Y] of [X] is [MASK], which is a [Y].	285.46	36.14	21.78	24.75	35.50
The neighbouring [Y] of [X] is [MASK], which is a [Y]	285.68	33.13	21.73	23.75	32.50
Which [Y] shares border with [X]? The answer is [MASK].	407.63	8.70	6.99	6.63	5.50
Which [Y] shares border with [X]? The answer is [MASK]	499.0	0.00	0.00	0.00	0.00
Which [Y] shares border with [X]? Answer: [MASK], which is a [Y].	285.71	32.77	21.72	23.74	30.50
Which [Y] shares border with [X]? Answer: [MASK], which is a [Y]	291.11	26.46	20.07	20.93	24.00
[X] and [MASK] share a border.	285.49	37.49	22.3	25.69	37.50
[X] and [MASK] share a border	285.49	37.54	22.3	25.71	37.00
The [X] [Y] shares border with [MASK], which is a [Y].	285.43	37.26	21.72	25.36	36.00
The [X] [Y] shares border with [MASK], which is a [Y]	285.51	37.11	21.55	25.22	35.50
[X] and [MASK] are neighbours.	288.2	33.55	21.78	23.96	33.00
[X] and [MASK] are neighbours	290.68	33.77	21.68	23.92	33.50

Table 15: Our proposed prompts for (state, shares border) relation.