確信度を考慮した言語モデルの関係知識評価

吉川和 [1,2]　岡崎直観 [1]
[1] 東京工業大学　情報理工学院　[2] 富士通株式会社　富士通研究所
{hiyori.yoshikawa@nlp., okazaki@}c.titech.ac.jp

概要

本稿では，事前学習済み言語モデルが学習の過程で獲得した関係知識を評価する LAMA probe タスクにおいて，モデル出力に対する確信度を考慮した選択的予測の設定を導入し，その評価を行う．選択的予測の設定では，クエリに対する予測結果の確信度を算出し，予測結果を出力するか否かを決定する．これにより，予測を行った事例に対する精度に加え，誤った予測を出力するリスクをどの程度低減できるかを考慮した評価が可能となる．本稿では，言語モデルのパラメータと出力のみを用い，追加の訓練データを必要としない複数の確信度指標を提案し，LAMA probe タスクを選択的予測の設定で評価する．実験では，特定の確信度指標の組み合わせが複数のデータセットで有効であり，さらに予測に直接用いた場合に予測精度そのものを改善できることが示された．

1　はじめに

近年，大量のテキストで学習した言語モデルを様々な後段タスクに転用する研究が盛んである [1, 2]．このようなことが可能となる背景として，言語モデルが事前学習の過程で語彙や文法といった言語知識 [3, 4] だけでなく常識や世界知識 [5, 6] をテキストから獲得していることが示唆されている．しかし，これらの知識は言語モデルのパラメータに埋め込まれており，シンボリックな知識ベースに対するような明示的なアクセスや編集は困難である．

Petroni ら [7] は，言語モデルの保持する常識や事実といった関係知識の量を評価することを目的とし，ベンチマークタスク LAMA probe を提案した．LAMA probe では，問い合わせたい関係知識を自然文の穴埋めタスクに変換し，言語モデルが正しく穴埋めできた場合，言語モデルがその関係に関する「知識をもつ」と判断する．実験により，BERT 言語モデル [1] がテキストからの関係抽出に基づく手法と同等かそれ以上の精度を達成すると報告された．．一方で，このようにして言語モデルから取り出された知識の信頼性，逆に言うと誤った知識が出力されるリスクは LAMA probe の枠組みでは考慮されない．シンボリックな知識ベースにおいては，知識ベースに含まれる三つ組の信頼性を人手やシステムにより評価し，信頼性の高い三つ組のみを残すことで知識ベースの質が担保される [8, 9]．結果として，実用上は知識ベースに含まれる三つ組は信頼性の高いものとみなすことができる．しかしながら，知識への明示的なアクセスができない言語モデルにおいては，個々の関係知識の信頼性を評価し，編集・削除する方法が確立されていない．ゆえに，知識ベースや言語モデルから知識を取り出すとき，次のような違いが生じる．知識ベースでは，問い合わせに対して適切な関係知識が存在しない場合，空の結果を返す．一方， LAMA probe などでの言語モデルへの問い合わせでは，与えられた問い合わせに対して何らかの出力を行うことを前提としているため，学習で獲得できなかった知識に対して常に誤った出力を返す．誤った出力は正しい出力と一見して区別がつかないため，この違いは解答される知識の信頼性を重視する様々な応用において，重大な障壁となる．言語モデルの性能を向上させたとしても，学習コーパスに無い事実に関する問い合わせや，答えの存在しない問い合わせ [10] は存在しうるので，モデルが誤った知識を出力するリスクを無視できない．

こうした背景から，本研究では言語モデルが返す知識の誤りリスクを定量化するため，予測に対する振る舞いを考慮したシステムおよび評価方法を検討する．具体的には， LAMA probe タスクに**選択的予測 (selective prediction)** [11, 12] の設定を導入する．選択的予測は機械学習の枠組みの一つで，システムはモデルの予測結果に基づき予測を実際に出力するか控えるかを選択できる．本研究では特に，Geifman と El-Yaniv [12] により提案されたリスク保証のある設定を考える．この枠組みでは，予測に基

づき計算される確信度スコアにより，予測の出力可否を判断する．システムの評価では，決められた予測精度を保証しつつ，如何に多くの事例に対し予測を出力できるかを測定する．

本研究では，言語モデルの知識評価に選択的予測の設定を適用するため，言語モデルの出力のみを用いて計算可能な複数の確信度指標を提案し，実験を通じて最適な指標を比較検討する．実験の結果，最適な確信度指標はモデルやデータセットにある程度依存するものの，特定の確信度指標の組み合わせが一貫して有効であることが確認できた．さらに，確信度計算に有効な指標は予測の決定に直接用いた場合においても有効であることを確認した．

2 選択的予測

本節では機械学習一般における選択的予測 [11, 12] の枠組みについて説明する．選択的予測の設定では，入力に応じて予測結果を実際に出力するかを判断する **選択的分類器 (selective classifier)** を導入する．入力空間 $\mathcal{X}$ からラベル集合 $\mathcal{Y}$ への分類問題を考える．選択的分類器はもとの分類モデル $f : \mathcal{X} \to \mathcal{Y}$ および選択関数 (selection function) $g : \mathcal{X} \to \{0, 1\}$ の組 (f, g) として定義される．選択関数は，分類器が入力 $x \in \mathcal{X}$ に対する予測 $f(x) \in \mathcal{Y}$ を実際に出力するかを決定する:

$$(f, g)(x) := \begin{cases} f(x) & \text{if } g(x) = 1 \\ \textit{don't know} & \text{if } g(x) = 0 \end{cases}. \tag{1}$$

Geifman と El-Yaniv [12] は，確信度指標に基づく選択関数を用いたリスク保証ありの設定 (selection with guaranteed risk; SGR) を導入した．SGR では選択関数として，確信度に基づく以下の関数を考える:

$$g(x) = \begin{cases} 1 & \text{if } \phi(x) \geq \beta \\ 0 & \text{if } \phi(x) < \beta \end{cases}. \tag{2}$$

ここで $\phi(x) : \mathcal{X} \to \mathbb{R}$ は f の**確信度関数 (confidence score function)** で，その値が閾値 $\beta \in \mathbb{R}$ を超えたときに限り分類器は予測結果を出力する．この形式では，β の値を適切に設定することによってシステムに課す誤りリスクの許容範囲を調整できる．β が大きければ大きいほど，予測を出力する事例数が減少するが，誤った予測を行うリスクを低減できる．

リスク保証ありの設定においては，選択的分類器が誤った予測を行うリスク

$$\text{Risk} = \frac{1 - N_{\text{corr}}}{N_{\text{pred}}} \tag{3}$$

と実際に予測を行った事例の割合（カバレッジ）

$$\text{Coverage} = \frac{N_{\text{pred}}}{N} \tag{4}$$

との間にトレードオフ (risk-coverage tradeoff) が存在する．ここで，$N, N_{\text{pred}}, N_{\text{corr}}$ はそれぞれ全事例数，実際に予測を行った事例数，正しい予測を行った事例数である．選択的分類器の性能は選択関数 (式 (2)) の閾値 β を動かして得られるリスク-カバレッジ曲線の AUC (RC-AUC) を用いて評価する．RC-AUC の値が小さいほど，予測を行った事例に対してリスクが小さく，良い分類器とみなされる．

3 LAMA Probe タスクへの選択的予測の適用

3.1 LAMA Probe タスクとモデル出力

LAMA probe では，問い合わせたい関係知識をテンプレートにより自然文に変換して言語モデルに入力する．例えば "*Dante*" と `born-in` の関係にあるエンティティについて問い合わせる際には，"*Dante was born in* `[MASK]`." という文がモデルへの入力となる．ここで `[MASK]` は単語マスクを示す特別なトークンで，このトークンに対して予測された単語をクエリに対する回答と見なす.[1] テンプレートは関係の種類ごとに人手で作成されている．

本研究における実験では Petroni ら [7] と同様に，双方向言語モデルを評価対象とする．モデルへの入力は，関係知識を問うクエリを単語マスクを含む自然文に変換したものである（例: "*Dante was born in* `[MASK]`."）．言語モデルは位置 t がマスクされた入力文 $x = W_{\setminus t} := (w_1, \ldots, w_{t-1}, \texttt{[MASK]}, w_{t+1}, \ldots, w_{|W|})$ を受け取り，位置 t の単語の予測確率 $P_{\text{LM}}(w_t | W_{\setminus t})$ を出力する．すなわち，言語モデルによる予測確率最大の単語 w' を予測単語とする:

$$f(x) = w' := \arg\max_{w_t} P_{\text{LM}}(w_t | W_{\setminus t}) \tag{5}$$

以下，マスク付き入力文 $W_{\setminus t}$ のマスク位置を予測単語 w' で置き換えた文を W' と表す．

3.2 確信度関数

LAMA probe はもともと言語モデルが学習で獲得した関係知識を評価するものであることから，確信度関数を構築するときには新しいデータによる言語

1) ここでは，言語モデルとして特に masked language model のような双方向言語モデルを想定している．また，簡単のため予測対象は一単語からなるエンティティに限定されている．

モデルの追加学習を必要としないことが望ましい．本節ではこの前提のもと，言語モデルの出力のみを利用するいくつかの確信度関数を提案する．

Token (T)　マスク位置における単語 w' の対数予測確率を用いる:

$$\phi_{\mathrm{T}}(x) = \log P_{\mathrm{LM}}(w'|W_{\backslash t}). \tag{6}$$

予測単語の推定（式 (5)）に用いた指標をそのまま確信度計算に用いることに相当する．

Sent (S)　文の容認性やファクトチェックの文脈では，文としての確からしさの指標として言語モデルによる文の生成確率が用いられている [13, 14]．ここでは，マスク付き言語モデルに基づく文の擬似尤度 (pseudo-log-likelihood; PLL) [15] を文長で正規化したものを採用する．マスク位置を予測単語 w' で置き換えた文 W' に対し，確信度は以下のように計算される:

$$\phi_{\mathrm{S}}(x) = \frac{1}{|W'|} \sum_{u=1}^{|W'|} \log P_{\mathrm{LM}}(w_u|W'_{\backslash u}). \tag{7}$$

Gap (G)　予測単語における予測確率が他の単語のものと比べて顕著に大きいときにモデルが確信をもって予測を行っているという仮定に基づく．マスク位置における予測確率が最大の単語 w' と二番目に大きい単語 w'' との対数予測確率の差を用いて，次にように確信度を計算する:

$$\phi_{\mathrm{G}}(x) = \log P_{\mathrm{LM}}(w'|W_{\backslash t}) - \log P_{\mathrm{LM}}(w''|W_{\backslash t}). \tag{8}$$

Reranking (R)　異なる指標で予測を評価した際にも，評価対象の予測が他の単語と比べて一貫して優位にあるとき，予測確信度が高いという仮定に基づく．まず，式 (5) にあるようなマスク位置における単語予測確率に基づき上位 K 件の予測候補 $\mathcal{W}$ を得る．次に，得られた予測候補を別の指標 ψ を用いて並べ替える．並べ替え後の w' の順位を $\mathrm{rank}_\psi(w')$ としたとき，確信度指標は以下の式で求められる:

$$\begin{aligned}\phi_{\mathrm{R}}(x) &= \log_2 \frac{K}{\mathrm{rank}_\psi(w')} \\ &= \log_2 K - \log_2 \mathrm{rank}_\psi(w').\end{aligned} \tag{9}$$

上式は本質的には並べ替え後の順位のみに基づく指標であり，言語モデルによるプライバシー情報の記憶リスクの評価にも用いられている [16]．実験では $K = 100$ とし，ψ としては Sent スコア $\phi_S(x)$ を用いる．

表 1　データセット毎の RC-AUC．確信度が X+Y のように表記されている場合，複数の確信度指標の和を用いている．GRE: Google-RE，TR: TREx，CN: ConceptNet，SQ: SQuAD，All: 全データセット．

モデル	確信度	GRE	TR	CN	SQ	All
BERT-base	T	.775	.478	.686	.755	.545
	S	.834	.594	.797	.815	.652
	G	.798	.470	.714	.794	.548
	R	.835	.597	.834	.798	.633
	T+S	.784	.512	.710	.766	.572
	T+G	.779	.468	.694	.769	.541
	T+R	**.772**	.465	**.685**	**.749**	**.535**
	S+G	.798	.482	.716	.786	.554
	S+R	.811	.557	.772	.795	.618
	G+R	.792	**.457**	.704	.777	.536
BERT-large	T	**.763**	.445	**.616**	**.669**	.506
	S	.815	.560	.738	.768	.614
	G	.801	.456	.650	.712	.525
	R	.826	.576	.792	.785	.609
	T+S	.771	.475	.645	.692	.532
	T+G	.779	.445	.627	.683	.510
	T+R	.764	**.437**	.617	.681	**.501**
	S+G	.788	.455	.645	.708	.520
	S+R	.796	.525	.712	.742	.583
	G+R	.795	.441	.638	.703	.511

4　実験

4.1　実験設定

本実験で用いる LAMA probe タスク [7] は言語モデルのもつ関係知識を評価するためのベンチマークタスクであり，Google-RE, TREx, ConceptNet, SQuAD の 4 つのサブタスクから構成される．事前学習済み言語モデルとしては BERT-base および BERT-large [1] を用いる．評価データ中のクエリの全てについて，Wikipedia 内に少なくとも一つの正解が存在する．BERT は Wikipedia で学習されているため，クエリに正解するための情報は学習データに含まれると考えられる．システムは，言語モデルがクエリに回答する知識を学習によって獲得できている場合には正答を返し，そうでない場合には予測を行わないことが求められる．

4.2　結果

4.2.1　モデル・データセットと確信度関数

表 1 に，各確信度指標を用いた場合の RC-AUC を示す．全データセットに対する結果を見ると，単独の確信度指標を用いた場合，Token (T) が最もよい性

表 2　予測に用いる指標を変えた場合の予測精度 (P@1) と，RC-AUC の最小値とそれを達成した確信度指標.

	予測	P@1	RC-AUC (best)	確信度 (best)
BERT-base	T	24.3	.535	$T+R_S$
	S	24.1	.538	$T+R_S$, T
	T+S	24.8	.534	$T+R_S$
	$T+R_S$	**25.3**	**.533**	$T+R_S$, T+G
	$S+R_T$	24.3	.535	$T+R_S$
BERT-large	T	26.1	.501	$T+R_S$
	S	26.1	.503	$T+R_S$, T, T+G
	T+S	26.9	.498	$T+R_S$
	$T+R_S$	**27.1**	**.497**	$T+R_S$, T
	$S+R_T$	26.3	.501	$T+R_S$

能を示した．２つの確信度指標を組み合わせた場合には，Token と Reranking を組み合わせた場合 (T+R) に最もよい性能となった．この傾向は BERT-base, BERT-large に共通であった．

データセット中の全ての事例に回答した場合の誤りリスクは BERT-base で 0.757，BERT-large で 0.739 である．これに対し，T+R を確信度指標として用いた場合，いずれのモデルについてもカバレッジを 0.3 程度に設定することで誤りリスクを 0.5 未満まで抑えることができた．具体的なリスク-カバレッジ曲線は付録 A 節に示す．

次に，データセット毎にそれぞれ RC-AUC を計算した結果を見ると，BERT-base では TREx を除いた 3 つのデータセットにおいて Token と Reranking の組み合わせ (T+R) が最高性能であるのに対し，BERT-large ではこれらのデータセットにおいて Token (T) を単独で用いた場合に最高性能であった．しかしながら，T+R を確信度指標として用いた場合の結果は，各データセットにおける最高性能と比較して大きく劣るものではなかった．このことから，T+R を確信度指標がいずれのデータセットにおいても比較的よい性能を示すが，最適な確信度指標はモデル・データセット毎に異なる場合があることが示唆される．TREx データセットについて関係の種類毎の比較を行った結果を付録 B 節に示す．

4.2.2　予測指標と確信度関数

3.2 節で提案した確信度指標の一部は，最適な予測を決めるための指標として式 (5) の代わりに用いることができる．そこで，各指標を予測に直接用いた場合に LAMA probe の精度 (P@1) を改善できるか，またそれぞれの予測指標に対しどの確信度指標を組み合わせることが最適かを調べた．予測指標として用いたのは Token (T)，Sent (S)，Reranking (R) およびそれらの組み合わせである．Reranking については，順位付けに用いる指標 ψ として Token (T) と Sent (S) の二種類を用い，R の添え字で区別する．S と R_S，T と R_T をそれぞれ単独で用いた場合の予測は定義から同一になることから，Reranking を単独で用いた場合の結果は省略する．また，Gap (G) は予測指標が決まって初めて定義される指標であるため対象外とする．なお，計算コストの都合上，T 以外の各予測指標に基づく予測結果の決定においては，まず T による指標（式 (5)）により候補を上位 100 件に絞り込んだのち，各予測指標を用いてリランキングすることで予測結果を近似した．

表 2 に結果を示す．BERT-base，BERT-large いずれにおいても，$T+R_S$ を予測に用いた場合に最高精度を達成した．さらに，いずれの指標を予測指標として用いた場合においても，$T+R_S$ を確信度指標として用いた場合に RC-AUC が最善となった．このことから，$T+R_S$ が予測指標としても確信度指標としても有効な指標であると考えられる．

5　まとめ

本稿では，言語モデルのもつ関係知識を評価する LAMA probe タスクに選択的予測の設定を導入し，モデル出力の精度に加え，モデル出力の正しさを何らかの確信度指標を用いて判別できるかを評価した．実験結果から，一般的に予測に用いられるマスク単語の予測確率値をそのまま用いるよりも，提案した Reranking スコアと組み合わせて用いることが予測・確信度計算双方において効果的であることが示唆された．また，適切な確信度指標は用いるモデルや評価対象のデータセットによって異なる場合があることを示した．今後は，対象タスクやモデルと最適な確信度指標の関係を詳細に調査し，多様なモデル・タスクにおけるモデル出力の信頼性を担保するための手法を検討したい．

謝辞

本研究は JSPS 科研費 19H01118 の助成を受けたものです．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chap-**

ter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

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

[3] Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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)**, pp. 2877–2887, Hong Kong, China, November 2019. Association for Computational Linguistics.

[4] Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. Linguistic profiling of a neural language model. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 745–756, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[5] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4762–4779, Florence, Italy, July 2019. Association for Computational Linguistics.

[6] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 5418–5426, Online, November 2020. Association for Computational Linguistics.

[7] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 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)**, pp. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[8] Chun How Tan, Eugene Agichtein, Panos Ipeirotis, and Evgeniy Gabrilovich. Trust, but verify: Predicting contribution quality for knowledge base construction and curation. In **Proceedings of the 7th ACM International Conference on Web Search and Data Mining**, WSDM '14, pp. 553–562, New York, NY, USA, 2014. Association for Computing Machinery.

[9] Shengbin Jia, Yang Xiang, Xiaojun Chen, Kun Wang, and Shijia. Triple trustworthiness measurement for knowledge graph. In **The World Wide Web Conference**, WWW '19, p. 2865âĂŞ2871, New York, NY, USA, 2019. Association for Computing Machinery.

[10] Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. Which linguist invented the lightbulb? presupposition verification for question-answering. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 3932–3945, Online, August 2021. Association for Computational Linguistics.

[11] Ran El-Yaniv and Yair Wiener. On the Foundations of Noise-free Selective Classification. **Journal of Machine Learning Research**, Vol. 11, No. 53, pp. 1605–1641, 2010.

[12] Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**, NIPS'17, pp. 4885–4894. Curran Associates Inc., 2017.

[13] Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. How furiously can colorless green ideas sleep? sentence acceptability in context. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 296–310, 2020.

[14] Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. Towards few-shot fact-checking via perplexity. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1971–1981, Online, June 2021. Association for Computational Linguistics.

[15] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2699–2712, Online, July 2020. Association for Computational Linguistics.

[16] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In **Proceedings of the 28th USENIX Conference on Security Symposium**, SEC'19, pp. 267–284, USA, 2019. USENIX Association.

A　リスク-カバレッジ曲線

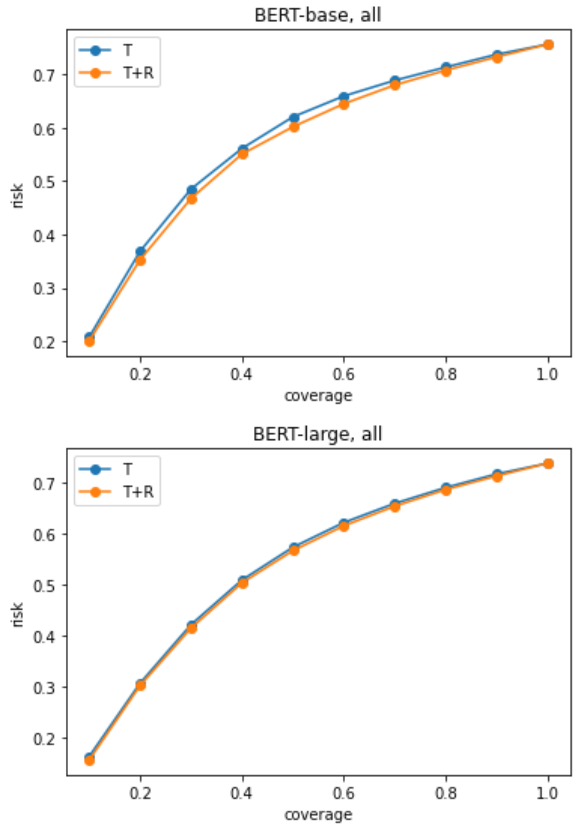

図 1　LAMA probe データセット全体に T および T+R を用いた場合のリスク-カバレッジ曲線.

図 1 に予測指標を T, 確信度指標を T または T+R とした場合のリスク-カバレッジ曲線を示す．図中の最も右，カバレッジが 1.0 の場合のプロットが全ての事例に対し予測を行った場合のリスクに相当する．いずれのモデルにおいても全ての事例に対し予測を行った場合の誤りリスクは 0.7 を超えているのに対し，いずれかの確信度指標を用いて選択的予測を行った場合，カバレッジを 0.3 程度に抑えることでリスクを 0.5 未満に抑えることができる．

B　関係タイプによる比較

実験に用いたデータセットのうち，TREx データセット中では一対一関係 (1-1)，多対一関係 (N-1)，多対多関係 (N-M) をもつ複数の関係タイプに関するクエリが混在している．関係タイプごとに RC-AUC を計算した結果を表 3 に示す．用いる確信度指標による性能差は，多対一や多対多の関係においてより顕著であることが見て取れる．関係の種類ごとに最適な確信度指標は異なるが，BERT-base では G+R, BERT-large では T+R が一貫して最適な確信度指標に匹敵する性能を示している．

表 3　TREx データセットにおける，関係タイプ毎の RC-AUC．表記方法は表 1 に準ずる．

モデル	確信度	1-1	N-1	N-M	All
BERT-base	T	.118	.434	.611	.478
	S	.163	.549	.776	.594
	G	.133	.422	**.604**	.470
	R	.218	.568	.756	.597
	T+S	.119	.465	.674	.512
	T+G	.125	.420	.605	.468
	T+R	**.115**	.416	.606	.465
	S+G	.126	.429	.636	.482
	S+R	.151	.505	.746	.557
	G+R	.127	**.404**	**.604**	**.457**
BERT-large	T	.085	.409	**.575**	.445
	S	.119	.520	.740	.560
	G	.092	.412	.597	.456
	R	.149	.534	.747	.576
	T+S	.085	.440	.619	.475
	T+G	**.084**	.404	.582	.445
	T+R	.085	.395	.576	**.437**
	S+G	.086	.409	.605	.455
	S+R	.108	.478	.713	.525
	G+R	.087	**.393**	.589	.441