Hybrid-SET: 意味的類似性とセットカバレッジを考慮した few-shot 例選出手法

朱晨成 [1]　谷口友紀 [2]　大熊智子 [2]　嶋田和孝 [1]
[1] 九州工業大学大学院　[2] 旭化成株式会社
zhu.chencheng822@mail.kyutech.jp
{taniguchi.tcr,okuma.td}@om.asahi-kasei.co.jp　shimada@ai.kyutech.ac.jp

概要

材料・化学分野のテキストに対して，固有表現抽出を試みる場合，学習に使えるデータが十分でないことやアノテーションに高度な専門知識が必要であるという問題点がある．一方で，大規模言語モデル（LLM）は少量事例（few-shot 例）に基づき，分類や推論が可能である．LLM は提供する事例によって回答の精度が大きく変化することから，適切な例を選択するための方法を設計することが重要である．本研究では，文の意味的類似性とセットカバレッジを考慮した few-shot 例選出手法 Hybrid-SET を提案する．実験結果から Hybrid-SET は既存の選択手法より優れていることを示す．

1　はじめに

自然言語処理技術は材料分野において大きな期待が寄せられている．材料・化学文書から情報や合成プロセスを抽出することで，知識ベースの拡充や開発の支援に貢献できる [1]．BERT に代表される事前学習済み言語モデルは，様々な情報抽出タスクにおいて高い性能が示されている．MatBERT [2] や MatSciBERT [3] のような材料科学に特化された言語モデルもある．しかし，それらのモデルのファインチューニングには大量の高品質な教師データが必要となる．さらに，材料・化学文書へのアノテーションには専門的な知識が必要であるため，教師データの作成にコストがかかる．

近年，大規模言語モデル（LLM）を用いた自動アノテーションが注目されており，人手に迫る精度でアノテーションを行えることが報告されている [4, 5]．従来の事前学習モデルと異なり，LLM は少量事例（few-shot 例）を通して，パラメータ更新を行わずに，様々なタスクに適応できる．一方，多くの研究が，提供事例によって LLM の回答精度が大きく変化することを指摘している [6, 7, 8]．したがって，LLM のポテンシャルを最大限に引き出すためには，few-shot 例の選出方法が重要な役割を担う．

一般に使われているアプローチとして，few-shot 候補例のプールからテスト例と高い関連性を持つ事例を検索する方法がある．何らかの関連性を測定し，候補例に対してスコアリングを行い，上位の事例を選択する．Liu ら [6] は文の意味的類似度を関連性の尺度とし，kNN を用いて訓練データから類似例を検索した．しかし，この手法を情報抽出タスクに適応する場合，次のような問題が考えられる．まず，文の意味的表現は文全体の情報を捉えることができるが，全てのトークンを均等に扱うため，トークンに注目するべきである固有表現抽出 (NER) タスクには適していない可能性がある [9]．次に，各事例の選出を独立に行うため，few-shot 例間の関係を考慮することができない．その結果，同じような情報が含まれた事例が選ばれる可能性があり，網羅性に問題がある．

本研究では，この 2 つの問題点を補うため，Gupta ら [10] が提案した SET-BSR を導入する．SET-BSR は，BERTScore [11] を関連度測定に用い，文全体の類似度ではなく，文脈に基づいたトークン間の類似度を重視する．さらに，網羅性（カバレッジ）の観点から，Greedy アルゴリズムによって，few-shot 例セット全体のテスト例に対するカバレッジを最適化している．

SET-BSR は，トークンとカバレッジの両方に着目した手法であり，NER タスクに適している．一方で，文の意味的類似度が Few-shot 例選出に有効に機能する場面も当然ながら存在する．そこで，本研究では，この 2 つの手法を組み合わせた Hybrid-SET を提案する．図 1 に提案手法である Hybrid-SET の

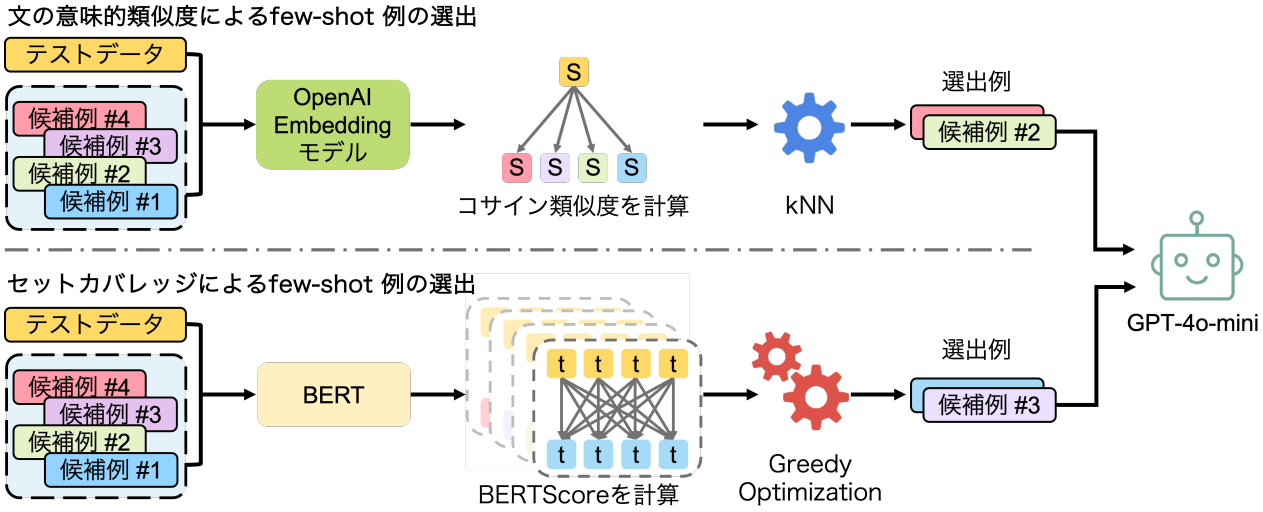

図 1　提案手法の概略.

概略を示す．この手法により，テストデータとの文レベルおよびトークンレベルの両方の関連性を考慮し，網羅性の高い few-show 例の選出が可能になる．

実験では，材料化学分野のデータセットを用いて，手法の有効性を評価する．Hybrid-SET が BSR などの単一モデルなどと比較して，最も優れていることを示す．

2　関連研究

従来の事前学習モデルはタスク固有の教師データを用いてファインチューニングを行う必要がある．それに対して，LLM は限られた「入力-出力」ペアの事例（few-shot 例）のみで様々なタスクを遂行することができる [12]．一方，提供される few-shot 例の内容によって LLM の回答精度が大きく影響受ける [6, 7, 8]．このような背景から，最適な few-shot 例を選出するための様々な手法が提案されている．その中，一般に使われているのは few-shot 候補例のプールからテスト例と最も関連性を持つ事例を検索する手法である．関連性の尺度として，文の意味的類似性を用いる研究は多く存在するが [6, 13]，エンティティを重視する NER タスクにとっては必ずしも適していないと指摘されている [9, 14]．Gupta ら [10] は BERTScore[11] を関連度の指標として利用する事例選出を行った．また，BERTScore を拡張し，few-shot 例セット全体のカバレッジを測る手法 SET-BSR を提案した．様々なタスクで評価した結果，SET-BSR は複雑な推論タスクに効果的であることを示した．

いくつかの研究により，LLM が NER タスクに適していないことが指摘されている．LLM は自然なテキストを生成することを事前学習の目的としており，LLM は要約や質問応答など，その生成処理とタスクが一致している場合には有効である．一方で，NER のように，複雑なスパンを抽出し，そのスパンのタグなどを分類・推定する必要があるタスクには様々な課題点を抱えている [15]．材料・化学分野のような専門性が高い NER タスクにおいては，さらに深刻な問題となる．Gutiérrez ら [16] は GPT-3 を用いて生物医学の情報抽出タスクを行った．その結果，文脈によるキャリブレーションや類似例を検索する手法を利用しても，ファインチューニングベースのモデルより大きく性能が劣ることがわかった．Zhu ら [17] は類似例によるミスリードを緩和するため，代表性サンプリングと検索ベースアプローチを組み合わせた few-shot 例選出法を提案した．しかし，代表性サンプリングは選出モデルの性能によって，選出結果が大きく変わるという問題点がある．

3　提案手法: Hybrid-SET

図 1 に示したように，本手法は文の意味的類似性に基づく事例選出（図上段）とトークンに着目したセットカバレッジに基づく事例選出（図下段）の 2 つから構成される．

3.1　文の意味的類似性による事例選出

文の意味的類似性による事例選出（Sentence Representation Similarity: SRS）では，まず，OpenAI の embedding モデルを用いて，ラベル付き候補例とテスト例をベクトルに変換する．次に，各テスト例

x_{test} に対して，文の埋め込み空間での距離に基づき，ラベル付き候補例プール $D = \{d_i\}_{i=1}^{N_i}$ から最も近い k_s 個の近傍 $d_1, d_2, \cdots, d_{k_d}$ を獲得する．定義された類似度尺度 sim（コサイン類似度など）が与えられると，近傍は $sim(d_i, s) \leq sim(d_j, s), i < j$ の順序でランク付けされる．

3.2 セットカバレッジによる事例選出

関連度の計算 本研究では，Gupta ら [10] の手法に倣い，BERTScore を用いてラベル付き候補例とテスト例の関連性を測る．BERTScore は言語モデルによって生成されたテキスト（例えば要約，翻訳など）と参照テキストとの類似度を測定するための評価尺度である．ここでは，few-shot 例を選出する際に BERTScore を利用する．候補例 $ca = \{c_1, c_2 \cdots c_n\}$ とテスト例 $x_{test} = \{x_1, x_2 \cdots x_n\}$ (c, x をトークンの文脈埋め込み表現とする) が与えられたとき，BERTScore（BSR[1)]）は式 1 のように定義される．

$$BSR(x_{test}, ca) = \sum_{x_i \in x_{test}} \frac{1}{|x_{test}|} \cdot \max_{c_j \in ca} Cosine(x_i, c_j) \tag{1}$$

本研究では，少量のラベル付きデータでファインチューニングされた BERT モデルを用いて，トークンの文脈埋め込み表現を獲得する．

セットカバレッジ（SET-BSR）の計算 カバレッジはテスト例の各トークンが，選ばれた few-shot 例によってどれだけ広く網羅されているかを表す．各事例のガバレッジは式 1 で計算される．

ここで，各々の few-shot 候補例のカバレッジを計算し，ランク付けし，独立して選出する場合の問題点を考える．まず，1 つの事例で全てのトークンをカバーすることは難しい．また，選出された few-shot 例の間の相補関係を考慮しないため，事例間でのトークンが重複する可能性も高くなる．このような場合，Few-shot 例のセット全体が持つ情報量が少なくなる．そこで，few-shot 例セット C のカバレッジを計算するため，式 2 を利用する．また，Greedy アルゴリズム 1 を用いてセットカバレッジを最適化する．

$$setC(x_{test}, C) = \sum_{x_i \in x_{test}} \frac{1}{|x_{test}|} \cdot \max_{ca \in C} \max_{c_j \in ca} \text{Cosine}(x_i, c_j) \tag{2}$$

1) 関連度の計算は BERTScore-Recall を用いる．

Algorithm 1 セットカバレッジの最適化

Require: Few-shot candidate pool $D = \{d_i\}_{i=1}^{N_i}$, test input x_{test}, the number of few-shot example k, coverage scoring function $setC$

$F \leftarrow \emptyset$ ▷ Selected Demonstrations
$F_{curr} \leftarrow \emptyset$ ▷ Current Set Cover
curr_cov $\leftarrow -\infty$
while $|F| < k$ **do**
  f^*, next_cov $\leftarrow \underset{f \in D-F}{\text{argmax}}$ setC$(x_{test}, F_{curr} \cup f)$
  if next_cov > curr_cov **then** ▷ Pick f^*
    curr_cov $\leftarrow$ next_cov
    $F \leftarrow F \cup f^*$
    $F_{curr} \leftarrow F_{curr} \cup f^*$
  else ▷ Start new cover
    $F_{curr} \leftarrow \emptyset$, curr_cov $\leftarrow -\infty$
  end if
end while
return F

4 実験

4.1 実験設定

本研究では，Materials Science Procedual Text Corpus (MSPT) [18] を用いて実験を行う．MSPT は，230 個の材料の合成プロセスに関する英語の学術論文を含むデータセットであり，材料分野に関する知識を持つ専門家 3 名で人手アノテーションを行っている．本実験では材料（MAT），操作（OPE），物性（PRO）の 3 つのエンティティを対象とする．

従来の教師あり学習と比較するため，少量のラベル付きデータを用いて BERT のファイチンチューニングを行い，NER モデルを構築した．また，このファインチューニング済み BERT はセットカバレッジによる事例選出において，テスト例と few-shot 候補例のトークン embedding を獲得するためにも利用される．事例を LLM に提供しない zero-shot，few-shot 候補例プールからランダムで k 個の事例を選択する手法（Random）もベースラインとする．

訓練データを十分に用意できない Low-resource 環境を想定し，MSPT の訓練データの 10% (170〜200 文) を BERT のファインチューニングに使う．few-shot 例も同じデータプールから選ぶ．結果の頑健性を確保するため，同じ割合でランダムにデータを抽出し，3 回実験をした平均値を結果とする．

表 1 ベースライン手法および異なる few-shot 例選出手法による自動アノテーション精度.

Entity (Support)		ベースライン			few-shot 事例選出手法			
		BERT	zero-shot	Random	SRS	BSR	SET-BSR	Hybrid-SET
MAT	(360)	0.507	0.508	0.641	0.657	0.657	**0.663**	0.660
OPE	(254)	0.719	0.619	0.765	0.788	0.793	**0.802**	0.801
PRO	(103)	0.071	0.155	0.275	0.264	0.269	0.284	**0.319**
Overall	(717)	0.547	0.495	0.623	0.633	0.637	0.645	**0.649**

BERT モデルのファインチューニングについて，HuggingFace Transformers library[2] を用いる．使用したハイパーパラメータは付録 A に記載する．なお，我々の別の研究において，GPT が作った検証データは人手アノテーションと同等に機能することが示されている [19]．したがって，本実験の検証データは，GPT-4o-mini を用いて，擬似ラベルを生成して利用した．本論文では最終的なアノテーション作業をする LLM として GPT を用いる．GPT の実装について，モデルは gpt-4o-mini [3]，Temperature は 0 に設定する．文の意味的類似性による事例選出 (SRS) で用いる embedding モデルについては OpenAI の text-embedding-3-small[4] を使用する．

各選出手法において，選ぶ few-shot 例の数が $k = 8$ を設定する．提案手法 Hybrid-SET では，SRS に 2 個，SET-BSR に 6 個をそれぞれ割り当てる．SET-BSR で選ばれた事例と SRS で選ばれた事例が被った場合，SET-BSR でランク付けた例から置換例を探す．評価指標として各エンティティには F1 スコア使用し，全体にはミクロ平均 F1 スコアを使用する．

4.2 実験結果

表 1 にベースライン手法および異なる few-shot 例選出手法による自動アノテーションの精度を示す．

まず，zero-shot を除き，全ての GPT ベースの手法はチューニング済み BERT よりも高い精度を得た．これは，ラベル付きデータが限られている場合，GPT-4o-mini がより適した解決策であることを意味している．また，全ての few-shot 事例選出手法がランダムサンプリングより優れた結果を得ている．

SRS と BSR の比較について説明する．図 2 に SRS と BSR によって選ばれた few-shot 例を示す．SRS で選ばれた事例は，SRS が文レベルでの類似性を扱っていることにも起因し，テスト例と構文的に類似した文構造を持っている傾向がある（例えば：[*noun*] was filtered と washed with [*noun*] ）．一方，BSR で選ばれた事例はトークン列の一致を重視する傾向がある（例えば “washed with deionized water”）．それにより，BSR は詳細な実験条件や手順などをカバーすることができる．実験結果から SRS と BSR の精度は全体的に大差がないとわかる．化学分野の NER タスクにおける事例選出では，文全体の意味的類似性とトークン毎の文脈類似性，この両方が重要な役割を果たしていることを示している．一方，事例を独立にスコアリングするこの 2 つの手法は SET-BSR より劣る結果になった．[5] これにより，Few-shot 例に対して，カバレッジを考慮して選出する方が，より有効であることがわかった．さらに，提案手法の Hybrid-SET は Overall において一番良い精度を得られた．特に，他の事例選出手法で精度が低かったエンティティである PRO の値が高かった．PRO は MAT や OPE よりも専門性が高く，また事例数も少ないというエンティティである．提案手法は 2 つの手法を相補的に利用することで，このような難しい問題・環境に対応できていると考えられる．

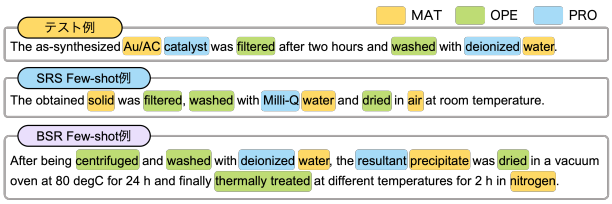

図 2 SRS と BSR によって選出された Few-shot 例.

5 おわりに

本研究では，GPT による化学文書への自動アノテーションにおいて，文の意味的類似性と，セット全体でのカバレッジを考慮した few-shot 例選出手法 Hybrid-SET を提案した．実験結果から，Hybrid-SET は専門性の高いエンティティに対してより効果的であることがわかった．

今回セットカバレッジの計算に文脈を考慮した BERTScore を用いた．今後は BM25 など，より表層的なカバレッジの計算方法も検討する．

2) https://huggingface.co/google-bert/bert-base-uncased
3) https://platform.openai.com/docs/models/gpt-4o-mini
4) https://platform.openai.com/docs/guides/embeddings
5) 選ばれた few-shot 例の比較は付録 C に記載する．

参考文献

[1] 有馬隆広, 大熊智子, 出羽達也. 新規用途探索を目的とした技術文書からの材料情報抽出. 言語処理学会第 29 回年次大会 発表論文集, pp. 512–515, 2023.

[2] Amalie Trewartha, Nicholas Walker, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. Quantifying the advantage of domain-specific pretraining on named entity recognition tasks in materials science. **Patterns**, Vol. 3, No. 4, p. 100488, 2022.

[3] Yu Song, Santiago Miret, and Bang Liu. MatSci-NLP: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3621–3639, 2023.

[4] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. Is GPT-3 a good data annotator? In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 11173–11195, 2023.

[5] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. **Proceedings of the National Academy of Sciences**, Vol. 120, No. 30, p. e2305016120, 2023.

[6] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In **Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures**, pp. 100–114, Dublin, Ireland and Online, 2022.

[7] Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 9134–9148, Abu Dhabi, United Arab Emirates, 2022.

[8] Hongfu Liu and Ye Wang. Towards informative few-shot prompt with maximum information gain for in-context learning. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 15825–15838, Singapore, 2023.

[9] Chenxiao Wu, Wenjun Ke, Peng Wang, Zhizhao Luo, Guozheng Li, and Wanyi Chen. Consistner: Towards instructive ner demonstrations for llms with the consistency of ontology and context. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 38, No. 17, pp. 19234–19242, 2024.

[10] Shivanshu Gupta, Matt Gardner, and Sameer Singh. Coverage-based example selection for in-context learning. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 13924–13950, Singapore, December 2023. Association for Computational Linguistics.

[11] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**, 2020.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda et al Askell. Language models are few-shot learners. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901, 2020.

[13] Rishabh Adiga, Lakshmi Subramanian, and Varun Chandrasekaran. Designing informative metrics for few-shot example selection. In **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 10127–10135. Association for Computational Linguistics, 2024.

[14] Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. GPT-RE: In-context learning for relation extraction using large language models. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 3534–3547, 2023.

[15] Jiawei Chen, Yaojie Lu, Hongyu Lin, Jie Lou, Wei Jia, Dai Dai, Hua Wu, Boxi Cao, Xianpei Han, and Le Sun. Learning in-context learning for named entity recognition. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 13661–13675, 2023.

[16] Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. Thinking about GPT-3 in-context learning for biomedical IE? think again. In **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 4497–4512, 2022.

[17] Chencheng Zhu, Kazutaka Shimada, Tomoki Taniguchi, and Tomoko Ohkuma. Staykate: Hybrid in-context example selection combining representativeness sampling and retrieval-based approach – a case study on science domains. **CoRR**, Vol. abs/2412.20043, pp. 1–11, 2024.

[18] Sheshera Mysore, Zachary Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In **Proceedings of the 13th Linguistic Annotation Workshop**, pp. 56–64, 2019.

[19] 朱晨成, 谷口友紀, 大熊智子, 嶋田和孝. 生成 ai による化学文書への自動アノテーションとその評価. 言語処理学会 第 30 回年次大会 発表論文集, pp. 2914–2919, 2024.

A　ハイパーパラメータ

BERT のファインチューニングについて，最適化アルゴリズムに AdamW，学習率は 2e-5 とし，損失関数には CrossEntropy を用いる．Epoch 数は 20，過学習抑制のために EarlyStopping を用いる．

B　プロンプトの全体像

プロンプトの全体像を図 3 に示す．

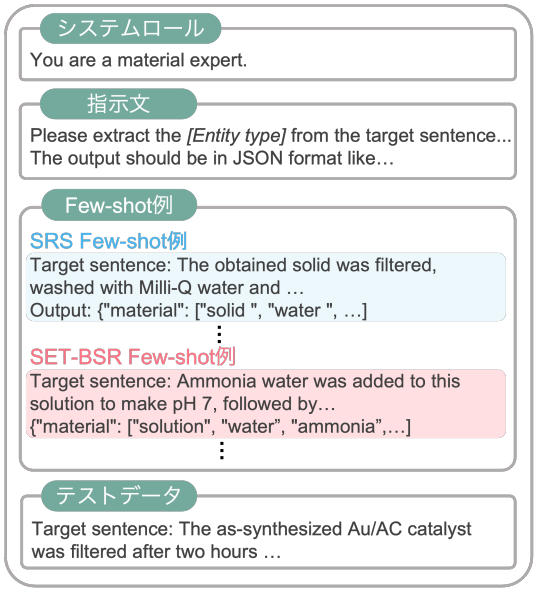

図 3　プロンプトの全体像．

C　異なる事例選出手法によって選ばれた few-shot 例

図 4 に異なる事例選出手法で選らばれた few-shot 例（2-shot まで）を示す．4.2 節で説明したように，SRS はテスト例と構文的に類似した文構造を持つ事例を検索する傾向がある．BSR はテスト例とトークン列がなるべく一致する事例を検索する傾向がある．しかし，どちらの手法も，事例を独立に選出するのは同じである．したがって，この 2 つの手法で選ばれた few-shot 例のセットは同じ情報を持っている事例が含まれる傾向がある．例えば，SRS で選ばれた 2 例は “[*noun*] was dried” という情報が繰り返されている．この問題点はトークンでマッチする BSR でも同様に生じている．例えば，BSR の事例でも “washed (…) with deionized water” が重複している．

一方，SET-BSR は few-shot 例の相補関係を考慮するため，同じような表現を含んでいる事例を選択する可能性が低くなる．例えば，SET-BSR の 2 番目

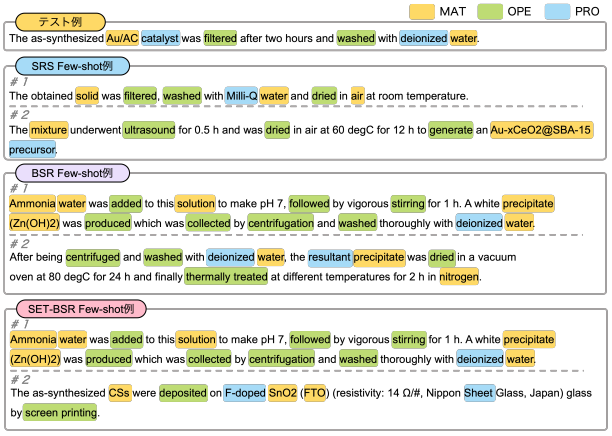

図 4　異なる事例選出手法によって選ばれた few-shot 例．

の事例では，1 番目の事例には含まれていないが，テスト例には含まれている “The as-synthesized” を含んでいる．このように，SRS や BRS で選出される few-shot 例のセットよりも，SET-BSR では，テスト例に対して網羅性が高いセットが得られる．