動的な環境における基盤化タスク設計の試み

宇田川拓真 [1]　　相澤彰子 [1,2]

[1] 東京大学　[2] 国立情報学研究所
{takuma_udagawa,aizawa}@nii.ac.jp

1 はじめに

自然言語対話における**基盤化**とは, 話者間で共通理解（**共通基盤**）を形成・修正・維持する一連のプロセスを指し, 人間の高度なコミュニケーションの重要な側面と考えられている [3]. 特に状況が絶えず変化する現実世界においては, 新たに共通基盤を形成するだけでなく, 状況の変化に応じて共通基盤を適切に更新し, 維持する能力が欠かせない.

しかし, 既存の対話タスクの多くは画像などの**静的**な情報を扱うものに限られており, **動的**な環境における共通基盤の形成・維持が十分に考慮されていない. 本研究では, 既存の基盤化タスク（OneCommon コーパス）[10] を時系列的に拡張し, 動的な環境における基盤化を評価する新たなタスク設計を行う.

このタスク設計に基づき, クラウドソーシングを用いて 5,617 対話を含む大規模なデータセットを新たに構築した. データセットの分析の結果, 提案タスクでは複雑な**時空間表現**を用いた基盤化が必要であること, また人間同士では過去の時点での共通基盤を利用してより正確かつ効率的に共通基盤の更新・維持ができていることが確認された。

最後に, 深層学習に基づく対話モデルを実装し, 提案タスクによる評価・分析を行った. その結果, 既存モデルは色・大きさ・位置などの空間的情報は（比較的）利用できているが, 動作・時制などの時間的情報はまだうまく扱えていないことが示唆された.

以上の考察を通じて提案タスクの有用性を検証し, 今後の課題について論ずる.

2 タスク設計

2.1 背景：協調的参照タスク

基盤化の能力を定量的・客観的に評価するために, OneCommon [10] では物体の共通認識を形成する**協調的参照タスク**を提案している. このタスクでは話者 A・B とエンティティの集合 $E = \{e_1, e_2, ..., e_m\}$ が存在し, 話者はそれぞれ E の観測 $obs_A(E)$・$obs_B(E)$ を持つとする. 話者同士は自然言語対話を通じて自由に情報を交換することができ, 最終的にそれぞれ観測中のエンティティを一つ選択する. 基盤化を通じて同じエンティティを選択できた場合のみタスクは成功とし, 異なるエンティティを選択した場合は失敗となる. 付録 A に実際の対話例を示す.

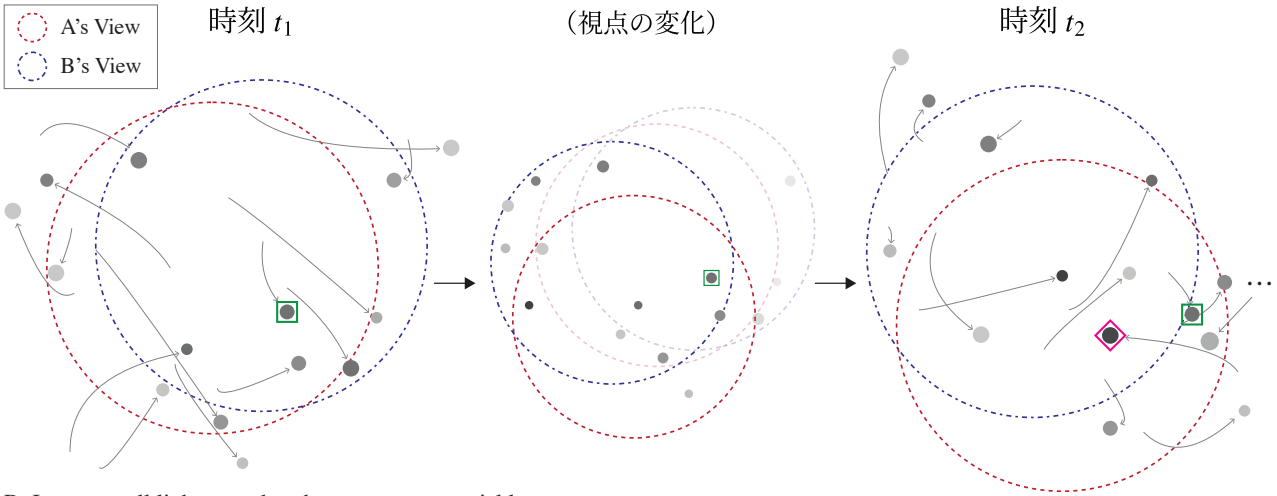
OneCommon ではさらに各エンティティ $e_i \in E$ の属性を**連続値**で表現することで複雑な曖昧性・不確実性の解消を要請している. また, 話者同士の観測が異なる**部分観測性**（$obs_A(E) \neq obs_B(E)$）を導入することで誤解・部分的な理解が生じやすくなっており, 理解の修正が要請されている.

しかし, このタスクでは話者の観測およびエンティティ属性の時間変化は想定されておらず, 多様な状況変化に応じた共通基盤の更新・維持は要請されない. また, 共通基盤の維持能力の定量的・客観的な評価指標が存在しないという問題がある.

2.2 協調的参照タスクの時系列的拡張

これらの問題を解決するために, 本研究では各エンティティ $e_i \in E$ の属性が時間変化し, 各時刻 $t \in [t_0, \infty)$ において話者 A・B は E の観測 $obs_A(E,t)$・$obs_B(E,t)$ を持つとする. また, 話者同士は複数の時刻 $t_1, t_2, ... \in (t_0, \infty)$ において協調的参照を行い, 時刻 $t_k (k \in \mathbb{N})$ において失敗した時点でタスクは終了するものとする. これにより, 共通基盤を正確に維持する能力を協調的参照の成功時間長 $k-1$ により定量的・客観的に評価できるようになる.

本研究ではこの定式化に基づいて OneCommon を時系列的に拡張する. 具体的には, 各エンティティ $e_i \in E$ の二次元平面上の位置を各時刻間 $[t_{k-1}, t_k)$ で変化（移動）させる. 移動の軌跡は図 2 に示すように二次ベジェ曲線で記述し, 距離パラメーター $r_{i,1}$, $r_{i,2}$ および角度パラメーター $\Delta\theta_i$ は一様分布からサ



B: I see a small light grey dot, that moves very quickly.
It ends to the right of three larger, darker dots.

A: i think that one moves off my view. do you have two medium sized dark dots on top of one another at the start, with the upper one being a bit smaller and off to the left? they both move down and to the right

B: Yes, I think so. But the one that started off on the left moves very slowly? And another lighter grey one ends up almost in between them?

A: yes can you pick the slower moving one?

A and B select: ◉

A: do you still see the same one?

B: Nope, it barely squeaked out of view.
But a larger black ball comes towards the left into view. Do you see it?
It probably crosses underneath our old one.

A: yes

A and B select: ◆

図 1　提案タスクの対話例. エンティティは時刻 t_1・t_2 での位置, 軌跡は $[t_0, t_1)$・$[t_1, t_2)$ での移動を示している.

ンプリングする. また, $\theta_{i,0}$ はランダムに初期化し, 移動後に $\theta_{i,k} \leftarrow \theta_{i,k-1} + \Delta\theta_i$ に基づいて更新する. これによって多様かつ一貫性のある動的変化を導入できるだけでなく, テスト時には任意の状況・観測の変化を再現し, 検証することが可能になる.

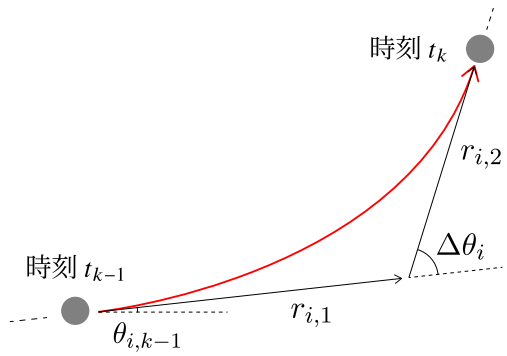


図 2　エンティティ e_i の $[t_{k-1}, t_k)$ 間の移動の軌跡.

また, エンティティの属性だけでなく話者の視点も随意に変化させることができる. 本研究では各時刻 t_k の直後に話者の視点をランダムに平行移動させることで, 観測の重複部分も変化させている.

図 1 にクラウドソーシングを用いて収集した提案タスクの対話例を示す. なお, 対話収集時は時刻の上限を t_5（最大の成功時間長を 5）としている.

2.3　関連研究との比較

提案タスクと既存の代表的な対話タスク（データセット）との比較を表 1 にまとめる.

まず, 基盤化に焦点を当てた対話タスクとして [4, 6, 10, 5, 13] などが挙げられる. これらはタスク成功率により共通基盤の形成能力を直接的に評価できるが, 全て静的な情報を扱うものに限られている.

基盤化を直接評価・分析するには不向きだが, 動的な情報を扱う対話タスクも近年現れている. 例えば AVSD[1] は動画に基づく対話を集めているが, 対話中に新たな情報が追加されるなどの**情報の更新**がないため, 共通基盤の更新は要請されない. 逆に Minecraft Dialog[8], SIMMC [7] などの仮想空間上の対話タスクでは（視点変化などを通じて）情報の更新は行われるが, 仮想空間自体は自発的に変化せず静的なものと考えられている. [9] は唯一動的な環境および情報の更新を扱っているが, 非タスク指向かつ不特定話者の対話であるため, 基盤化の評価・分析は非常に難しいと考えられる.

3　データセット

本研究では Amazon Mechanical Turk を利用し, 大規模かつ高品質な英語の対話データを収集した.

表 1　既存の代表的なデータセットとの比較.

データセット	環境（情報の種類）			情報の更新	情報源	基盤化の評価
	連続的	部分観測的	動的			
Twitch-FIFA [9]	✓	✗	✓	✓	仮想	N/A
AVSD [1]	✓	✓	✓	✗	実世界	間接的
Minecraft [8]	✗	✓	✗	✓	仮想	間接的
SIMMC [7]	✓	✗	✗	✓	仮想+実世界	間接的
GuessWhat?! [4]	✓	✗	✗	✗	実世界	形成
Photobook Dataset [6]	✓	✓	✗	✓	実世界	形成
OneCommon [10]	✓	✓	✗	✗	仮想	形成
本研究	✓	✓	✓	✓	**仮想**	**形成+維持**

3.1　定量的分析

表 2　データセットの統計値.

統計値	OneCommon	本研究
対話数	6,760	5,617
平均発話数	4.8	11.7
平均発話長	12.4	10.3
語彙数	3,621	3,895
語彙の重複	29.4%	
成功時間長	-	3.31
t_1 の成功率	76.8%	80.5%
t_2 以降の成功率	-	90.3%

データセット全体の統計値を表 2 に示す.

OneCommon と比較すると, 提案タスクでは協調的参照を複数回繰り返しているため対話中の発話数が多くなっており, より長く一貫した対話が必要となっている. 平均的な発話長は若干短くなっているが, これは t_2 以降に複雑な表現が *“same again?”* などのように省略されやすいからであり[1], 図 1 のように長く複雑な発話も多数見受けられた. また, OneCommon を拡張した設定であるため, 語彙的にはシンプルかつ重複が多いことが確認できた.[2]

時刻 t_1 での協調的参照の成功率も高く, 人間同士では正確に共通理解を形成できていることが分かる. また, 時刻 t_2 以降ではそれ以前の共通基盤を利用して, 成功率をさらに高めることができている.[3]

3.2　定性的分析

図 1 に例示されるように, 人間同士ではエンティティの動作（*“moves very quickly”*, *“come towards the left”*）や特定の時点での位置（*“started off on the left”*, *“ends to the right”*）などの**時空間表現**を用いて共通基盤を形成している. また, 時刻 t_2 以降では**それ以前の共通基盤**（*“still see the same one?”*, *“crosses underneath our old one”*）を利用していることが確認できる. さらに環境の連続性および部分観測性により *“a bit”*, *“almost”*, *“probably”* など**曖昧性・不確実性のニュアンス**が伴いやすいことも確認できる.

特に動作の表現に着目すると, 移動の軌跡が連続値で記述されているため（図 2）**語用論的推論を要する表現**が現れやすくなっている. 例えば図 3 に示すように *“straight down”* と言った時に正確な垂直を指していなかったり, 多様な移動の軌跡が *“right (and) then up”* と同一の表現で表されていたりする. また, 付録 B に示すように複数のエンティティの動作の関係性も同様に（語用論的に）表現されている. このように提案タスクでは必要な語彙はシンプルでありながら, 字義通りではなく語用論的推論を用いた高度な言語理解・生成が必要となっている.

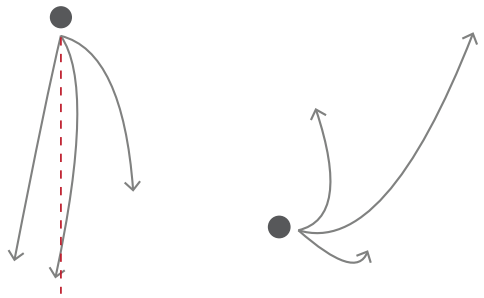


図 3　語用論的な動作の表現.

4　実験

最後に, ベースラインモデルを用いた実験を行う.

1）　例えば 5 トークン以下の短い発話の割合は提案タスクで 33.8%となっており, OneCommon の約二倍多くなっている.
2）　頻出語（10 回以上出現する語彙）に限ると, 重複は 53.0%とさらに高くなっている.
3）　一部のクラウドワーカーはより高い水準でタスクを解いており, 成功時間長・成功率の上限はより高くなると思われる.

表 3　実験結果.（データ分割含め）異なる乱数種を用いて各実験を 5 回繰り返した.

モデル	ターゲット選択（%）		成功率（%）		成功時間長
	t_1	t_2 以降	t_1	t_2 以降	
ベースライン	**76.4±1.7**	**77.1±0.3**	62.5±1.7	75.4±1.3	1.94±0.09
– 色	56.3±2.0	65.5±0.9	53.4±1.5	70.1±2.4	1.50±0.10
– 大きさ	58.4±1.3	66.7±0.5	57.1±0.9	69.3±2.0	1.58±0.07
– 位置	74.4±1.5	76.9±0.6	60.2±1.4	69.9±1.9	1.68±0.09
– 時間変化	75.1±2.3	76.9±0.3	**65.0±1.4**	**75.6±1.4**	**2.02±0.07**
人間	97.3±0.9	96.9±1.1*	80.5	90.3	3.31

4.1　評価

モデルの共通基盤の認識能力の評価には, 人間同士の対話・観測からどのエンティティが各時刻 t_k で選択されたかを予測する**ターゲット選択**タスクを用いる. 選択時の観測エンティティの数は 7 つに制約されているため, これは単純な分類問題として正解率で評価できる. 検証・テスト用データには成功時間長が 2 以上のランダムな対話を 500 ずつ用意する.

また, 各モデル同士で**提案タスク全体**を解かせることで共通基盤の形成・維持能力を評価する. 具体的には 2,000 の未知の環境で同じモデル同士を対話させ, 成功率・成功時間長によって評価を行う.

4.2　モデル

本研究では [11] を基にシンプルな深層学習対話モデルを実装する. 具体的には対話トークンを GRU[2] で埋め込み, 観測をエンティティレベルの埋め込み表現に変換する（付録 C）. 発話生成時にはエンティティの表現をアテンションスコアで重み付けし, 時刻 t_k でのエンティティの選択にはスコアが最大のものを選択する. また, 各モジュールは OneCommon コーパスを用いて事前学習を行なった.[4)]

このモデルをベースラインとして, どの素性が利用されているかを確認するために, エンティティの表現から色・大きさ・位置属性のアブレーションを行う. また, 観測の最終フレームのみを入力とすることで時間変化の情報のアブレーションも行う.

4.3　結果

実験結果を表 3 に示す. なお, ターゲット選択の人間性能は 3 名のアノテーターで確認を行なった.

まず, ターゲット選択ではベースラインの正解率は比較的高く, 既存モデルはある程度対話中の共通基盤を正しく認識できていると考えられる. また, アブレーションの結果からエンティティの色・大きさ・位置・時間変化全てが正解率に寄与していることが分かる. なお, 時刻 t_k（$k \geq 2$）での選択を予測する際には前時刻 t_{k-1} での選択の正解を入力としているため, t_2 以降の正解率は若干高くなっている.[5)]

全体タスクでは t_1 での成功率が最も低く, 動的な環境で共通基盤を新たに形成するのは難しいと考えられる. t_2 以降の成功率は比較的高くなっているが, 実際には前時刻での選択を繰り返してしまうケースが多く, 前時刻の選択エンティティが共有されなくなった場合の成功率に大きな向上は見られなかった.[6)] よって, 共通基盤をそのまま保持することはできるが変更することはまだ難しいと予想される. また, 時間変化の情報を利用しない場合に成功率は一貫して向上しており, 既存モデルは動的な情報を利用して基盤化を行えていないことが示唆された.

最後に, 全ての評価指標において人間の性能とは大きな乖離があり, 提案タスクが改善の余地のある難しいタスクであることが示された.

5　終わりに

本研究では動的な環境において物体の共通認識を維持する新たな対話タスクを設計し, 人間とベースラインモデルの基盤化の評価・分析を行なった.

将来的に対話システムの応用範囲が広がりより実世界に近い問題を扱うようになるほど, 状況の変化に応じて共通基盤を適切に更新・維持できることは重大な要件になると考えられる. 今後は参照表現 [11] や時空間表現 [12] のアノテーションなどを通じてより詳細な分析を可能にしつつ, 動画処理の手法も組み合わせてより汎用性の高い対話モデルの分析・改善を行う. これらの研究を通じて, 基盤化の観点からより高度な対話モデルを追究していきたい.

4)　これによって成功時間長で 0.6 程度の改善が確認された.

5)　人間のアノテーターには前時刻での正解は与えていない.

6)　むしろベースライン以外では成功率が下がっていた. なお, 人間同士ではこの状況でも 6%以上の成功率向上が見られた.

参考文献

[1]Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

[2]Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111. Association for Computational Linguistics.

[3]Herbert H Clark. 1996. *Using language*. Cambridge university press.

[4]Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.

[5]Rui Fang, Malcolm Doering, and Joyce Y. Chai. 2015. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 271–278, New York, NY, USA. ACM.

[6]Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

[7]Seungwhan Moon, Satwik Kottur, Paul Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2020. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[8]Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.

[9]Ramakanth Pasunuru and Mohit Bansal. 2018. Game-based video-context dialogue. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 125–136, Brussels, Belgium. Association for Computational Linguistics.

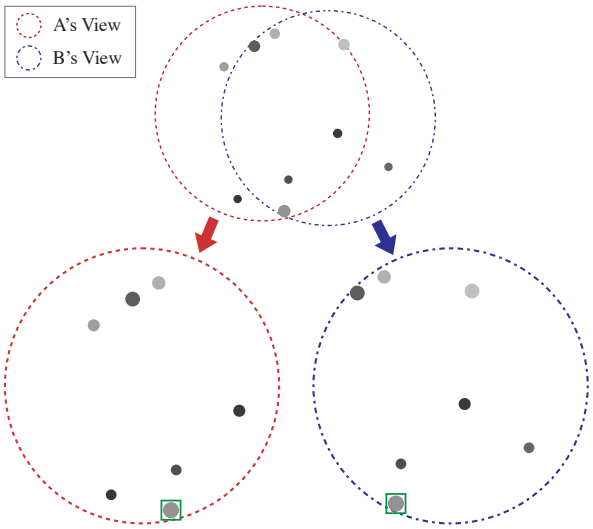
[10]Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127.

[11]Takuma Udagawa and Akiko Aizawa. 2020. An annotated corpus of reference resolution for interpreting common grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9081–9089.

[12]Takuma Udagawa, Takato Yamazaki, and Akiko Aizawa. 2020. A linguistic analysis of visually grounded dialogues based on spatial expressions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 750–765, Online. Association for Computational Linguistics.

[13]Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. PentoRef: A corpus of spoken references in task-oriented dialogues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 125–131, Portorož, Slovenia. European Language Resources Association (ELRA).

A　OneCommon コーパス



A: I see three in a line going up and to the right.
The middle one is the largest and darkest
B: I don't see that. I have one large, medium gray dot
that's under a small, darker gray dot
A: Is the larger dot slightly to the left
B: yes, slightly, let's choose the larger one
A and B select:

図 4　OneCommon コーパスの対話例

図 4 に示すように, OneCommon ではエンティティおよび話者 A・B は二次元平面上に位置しており, A・B はそれぞれ自分の周囲の一定の半径以内のエンティティのみ観測できる. この状況で, 自然言語対話による基盤化を通じて同一の・共有されているエンティティを一つ選択しなければならない.

B　複数エンティティの動作

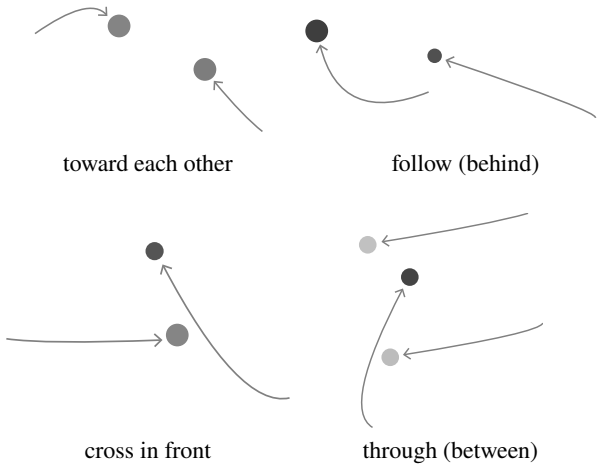


図 5　複数エンティティの動作の表現.

エンティティの動作はランダムに生成しているため, 話者はそれらのインタラクション（相互作用）を特徴的な時空間関係に見立てて表現していると考えられる. よって, 図 5 の他にも *pass by*, *move along with* のように多様かつ語用論的な複数エンティティの動作表現が多数見受けられた.

C　エンティティの埋め込み表現

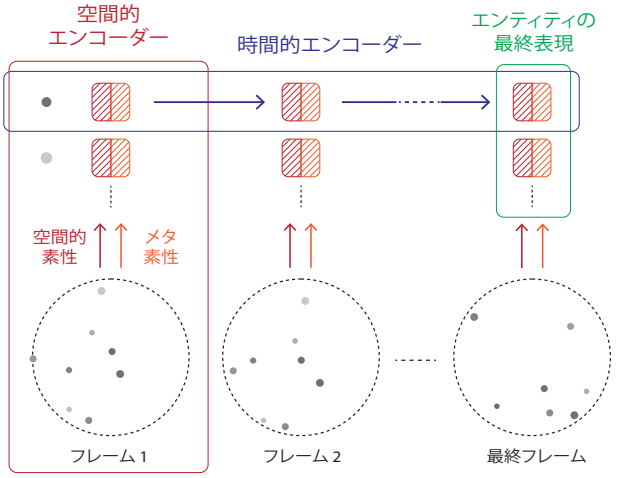


図 6　エンティティの埋め込み表現の生成手法.

時刻間 $[t_{k-1}, t_k)$ の観測アニメーションからエンティティの埋め込み表現を得る手法を図 6 に示す.

まず, 各時刻での観測フレームから空間的エンコーダーを用いて**空間的素性**と**メタ素性**の二つを生成する. 空間的素性は [11] と同様にその時刻でのエンティティの属性（色・大きさ・位置）等を多層パーセプトロン（MLP）で埋め込む. なお, フレーム中に存在しない場合はデフォルトの値 $(0, 0)$ を観測上の位置とする. メタ素性はエンティティのメタ情報（フレーム中に存在するか, 前時刻 t_{k-1} に存在したか, 前時刻 t_{k-1} で選択したか等）を二値トークンとして表現し, MLP を用いて埋め込み表現に変換する. 各時刻（観測フレーム）でのエンティティの表現は空間的素性とメタ素性の和とする.

続いて, 各観測フレームにおけるエンティティ表現を時間的エンコーダー（GRU）で埋め込み, 最終時刻での状態をエンティティの最終表現とする.