



演題名：患者状態の包括的なアノテーション付きコーパスの構築

演者名：篠原恵美子、柴田大作、嶋本公德、河添悦昌

背景と目的

ICTやAI技術を用いた診療情報の利活用が期待されている。患者の症状・所見や診断、治療内容の選択理由などは重要な情報であるが、自由記載にのみ記録されているためそのままでは利用できない[1]。自由記載を自動で解析し情報を構造化するには自然言語処理技術が用いられ、その研究開発にはテキストとそこに含まれる情報(アノテーション)から構成されるコーパスが必要である。コーパス構築は一般に時間と労力のかかるものであるため、汎用的なアノテーションの付与されたコーパスを構築し様々な用途のアプリケーション開発に利用するのが効率的である。

我々は汎用的、すなわちカルテの自由記載の再利用されうるようなアノテーションスキーマの構築を目指している。ここでアノテーションスキーマとは、ある範囲の文字列に付与されるカテゴリであるエンティティタイプ、エンティティタイプが持つ属性、エンティティタイプ間に張られる関係のセットである。情報を網羅する医療テキストの内容は大きく症状・所見・診断といった患者の状態と投薬や手術などの治療に分かれると考え、本研究では前者の網羅性に焦点を当てたスキーマの作成を目的とする。

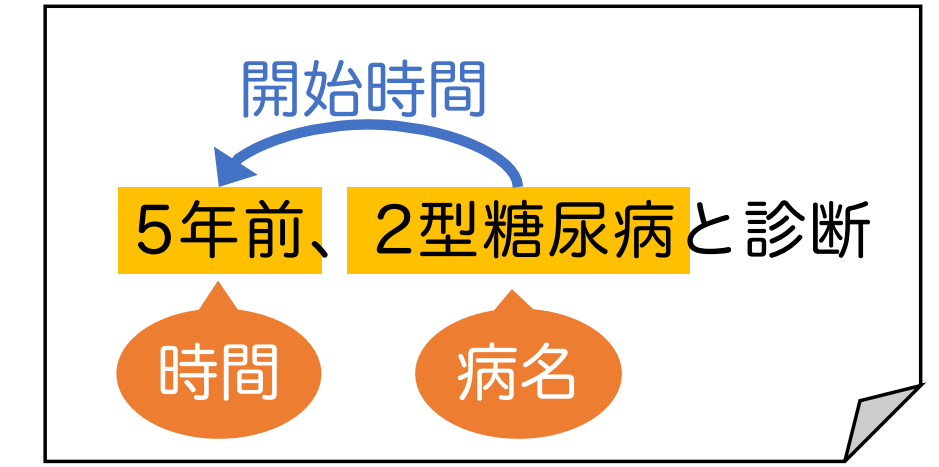


図1. アノテーション付きコーパスの例 「時間」「病名」がエンティティタイプ、「開始時間」が関係である。

方法

アノテーションスキーマの構築: 最初に少量の症例報告テキストを参考に必要と考えられた基準を仮のものとして作成し、その後アノテーション実施とその結果を受けての基準の修正を繰り返すことを行った。対象とするテキストには、臨床医学系雑誌に掲載された症例報告で、厚労省指定難病をタイトルに含むものの症例提示部分の本文テキストを用いた。症例報告を用いたのは内容や形式が退院時要約に類似しており、様々な診療記録の種別の中で退院時要約は必要十分な記載がされており、最初に取り組み対象として適していると考えたためであり、難病を条件としたのは診療科や疾患領域に限定が無く、幅広い症状・所見が記載されていると考えたためである。

アノテーションスキーマの評価: 構築したアノテーションスキーマの網羅性を評価するため、情報抽出タスクを扱ったi2b2コーパス[2-4]および医療自然言語処理器であるMedLEEの知識モデル[5]との比較を行った。

コーパスの公開: 自然言語処理分野では公開されたアノテーション付きコーパスを用いて研究が行われる場合が多く、コーパスの公開は研究や技術開発の促進につながると考えられる。本研究で収集した症例報告については公開には著作権処理が必要である。そこで、クリエイティブ・コモンズのもとで発行されたもの以外については、各発行団体に再配布の許諾を申請した。

またアノテーションの質を高めるため、その時点のコーパスを教師データとして機械学習を行い、コーパスにアノテーションを自動で付与し修正候補としてアノテーション作業者に提示し、繰り返し修正作業を行った。

結果

333の指定難病のうち151に関する症例報告 362件が得られた。これらを対象にアノテーション実施とアノテーションスキーマの修正を繰り返した結果、アノテーションスキーマはエンティティタイプ46種、属性9種、関係36種から構成されるものとなった。患者状態は単一のエンティティタイプではなく複数のエンティティタイプの組み合わせの階層構造として表現される(図2)。

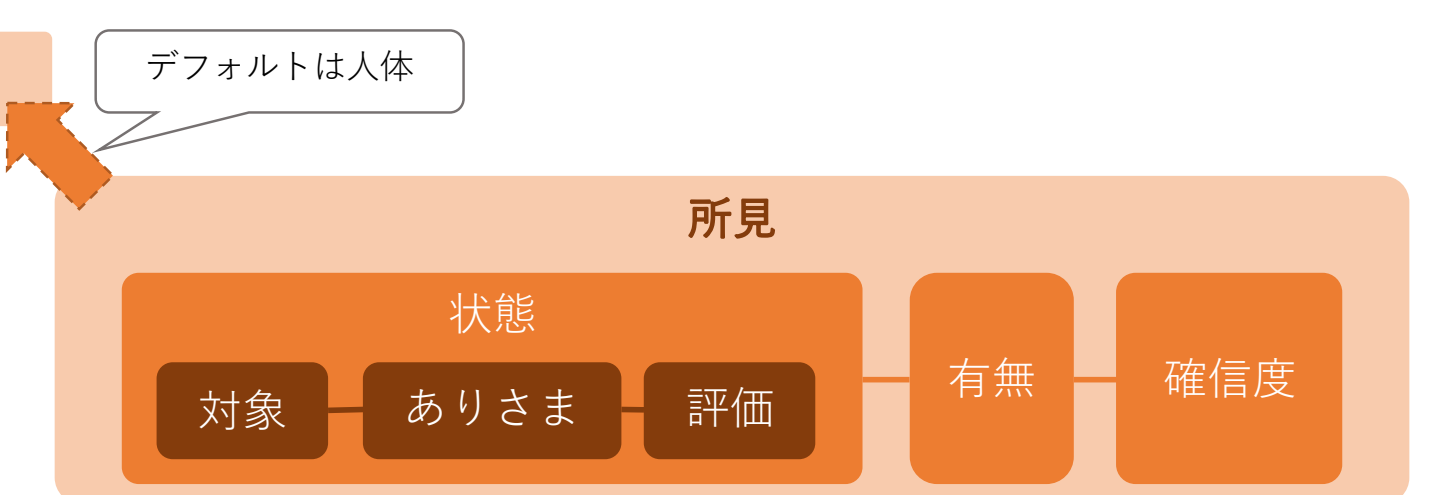


図2. アノテーションによる患者状態の表現

図3にこのスキーマに則ったアノテーションの例を示す。エンティティタイプは症状・所見を直接表すものだけでなく、人体部位や時間、臨床検査などさまざまである。先行研究では見られないものとして、状態の有無を表すものと「診断」など判断を伴うものを異なるエンティティタイプとして定義したことが挙げられる。これと所見間の因果関係・根拠関係を併用することで、従来は表現できなかった情報の表現ができるようになった。属性は状態や部位などのエンティティタイプが患者本人の情報であることを表すものや、判断を表すエンティティタイプがあり・無し・疑いなどのどれを表すかの分類などである。関係には観測手段関係、観測時間関係、因果関係、判断根拠などがある。

先行研究との比較を行った結果、構築したアノテーションスキーマはほとんどの情報をカバーしており(表1)、網羅性があることが示された。

179の症例報告について著作権処理が完了し、これらアノテーションとともに当講座のウェブサイト(<https://ai-health.m.u-tokyo.ac.jp/home/research/corpus>)で公開している。このコーパスについて機械学習によるエンティティ認識および関係抽出の精度を調査した。モデルとしてエンティティ認識はBERT-CRF、関係抽出はBERT-Attentionを用い、Micro-F1は0.920、0.852であった。

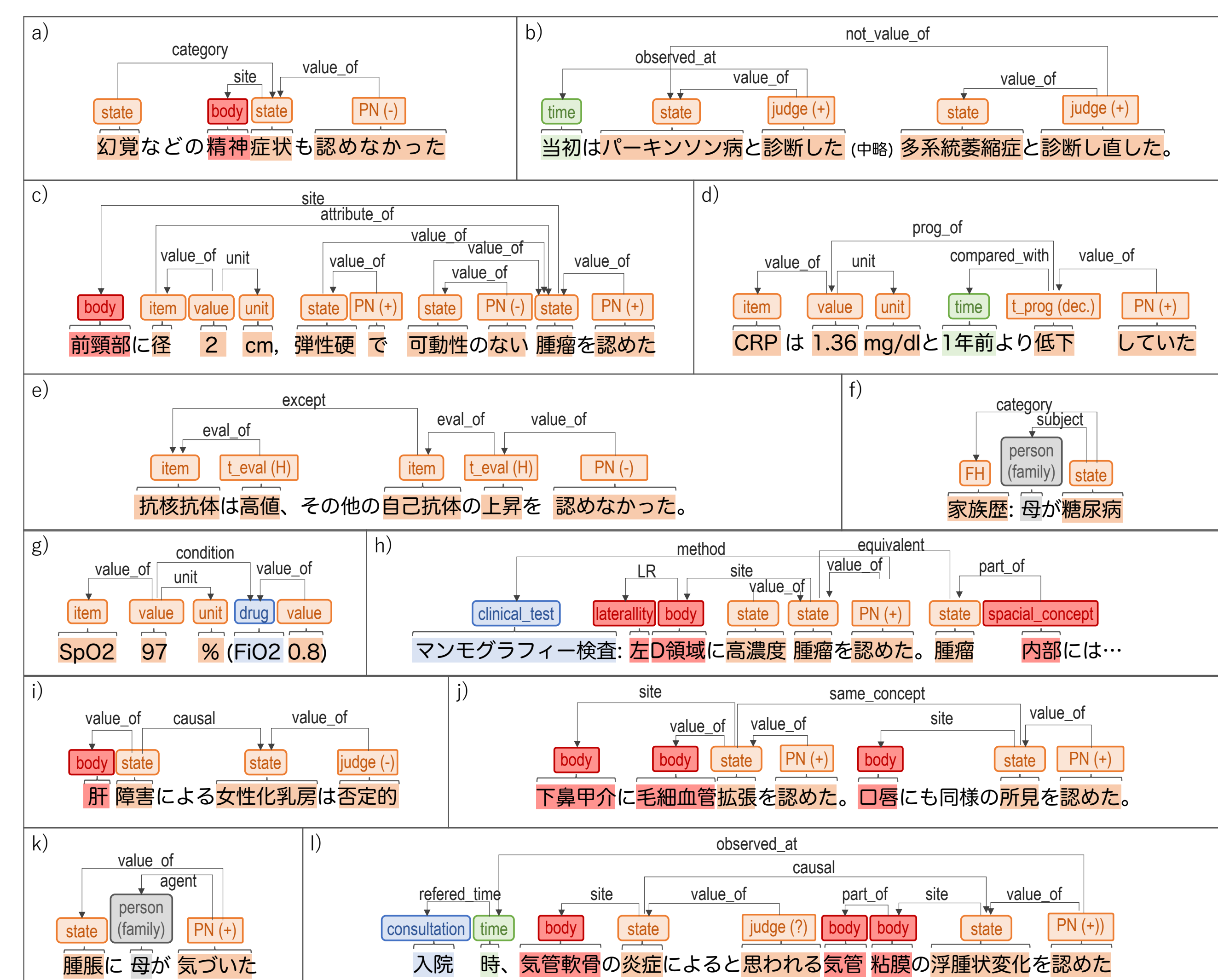


図3. 構築したスキーマに則ったアノテーションの例

表1. 先行研究(i2b2, MedLEE)との比較結果

i2b2 component	Evaluation	Corresponding components of our criteria	MedLEE component	Evaluation	Corresponding components of our criteria
Medications [2]	EXACT	ent:drug	Bodyloc Mod	EXACT	
Dosages [2]	EXACT	ent:value + ent:unit			
Modes [2]	EXACT	ent:treatment	Primary Loc	EXACT	ent:body ent:spacial_concept, rel:site rel:region_contain, →
Frequencies [2]	EXACT	ent:frequency	Spatial Mod	EXACT	ent:laterality, rel:LR, →
Durations [2]	EXACT	ent:time_span	Bodyloc Mod	EXACT	ent:body, rel:part_of, →
		rel:start_from	Mod		
		rel:end_at	Region Mod	EXACT	ent:spacial_concept, rel:part_of rel:referred_site
		rel:reason			rel:referred_m_site, →
Reasons [2]	EXACT	ent:state			ent:state, rel:value_of, →
List/narrative [2]	NONE	compared_with	Orientation Mod	EXACT	
Concept [3]	PARTIAL	ent:clinical_test	Quantity Mod	EXACT	ent:value, rel:value_of, →
Medical Problems Tests	EXACT	ent:item without rel:attribute_of	Change Mod	EXACT	
Treatments	PARTIAL	ent:treatment	Change Mod	EXACT	ent:quantity_progress ent:quality_progress, rel:prog_of, ←
		ent:drug	Concept	EXACT	ent:PN ent:judge, rel:value_of, ←
		ent:device	Mod		
Assertion [3]	EXACT	ent:PN with att:PN_value = Positive	Degree Mod	EXACT	ent:quality_evaluation, rel:eval_of, ←
Present	EXACT	ent:judge with att:certainty = Positive	Degree Mod	EXACT	ent:quality_evaluation, rel:eval_of, ←
Absent	EXACT	ent:PN with att:PN_value = Negative	Degree	EXACT	ent:quality_evaluation, rel:eval_of, ←
		ent:judge with att:certainty = Negative	Concept	EXACT	ent:quality_evaluation, rel:eval_of, ←
Possible	EXACT	rel:progress with att:prog = recover	Degree Mod	EXACT	ent:quality_evaluation, rel:eval_of, ←
Conditional	EXACT	ent:judge with att:certainty = Suspicious	Status Mod	EXACT	
Hypothetical	EXACT	rel:condition with att:patient = false	Status	EXACT	ent:quality_progress, rel:prog_of, ←
Not associated with the patient	EXACT	rel:category to ent:FamilyHistory	Concept	EXACT	ent:PN ent:judge, rel:value_of, ←
relation [3]			Mod		
Treatment improves medical problem	EXACT	rel:after + ent:progress with att:prog = improve/recover	Procedure Mod	EXACT	ent:treatment ent:execute, rel:after rel:causal, ←
Treatment worsens medical problem	EXACT	rel:after + ent:progress with att:prog = worsen	Descriptor Mod	EXACT	
Treatment causes medical problem	EXACT	rel:after:causal + ent:progress with att:prog = start	Descriptor	EXACT	
Treatment is administered because of medical problem	EXACT	rel:reason	Descriptor	EXACT	ent:state, rel:value_of, ←
Treatment is not administered because of medical problem	EXACT	rel:reason + ent:execute with att:execute_type = didnot/stop	Descriptor	EXACT	ent:item, rel:attribute_of, ←
Test reveals medical problem	EXACT	rel:method	Descriptor	EXACT	ent:state, rel:value_of, ←
Test conducted to investigate medical problem	EXACT	rel:reason	Mod		ent:item, rel:attribute_of, ←
Medical problem	EXACT	ent:judge + rel:reason			
indicates medical problem	EXACT	ent:person			
Person [4]	EXACT				
Pronoun [4]	EXACT				
Co-reference (relation) [4]	EXACT				
Person, Pronoun					

[1] W.Q. Wei, et al., Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance, J. Am. Med. Inform. Assoc. 23 (2016) e20–e27.
[2] O. Uzuner, et al., Community annotation experiment for ground truth generation for the i2b2 medication challenge, J. Am. Med. Inform. Assoc. 17 (2010) 519–523.
[3] Ö. Uzuner, et al., 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. 18 (2011) 552–556.
[4] O. Uzuner, et al., Evaluating the state of the art in coreference resolution for electronic medical records, J. Am. Med. Inform. Assoc. 19 (2012) 786–791.
[5] C. Friedman, J.J. Cimino, S.B. Johnson, A schema for representing medical language applied to clinical radiology, J. Am. Med. Inform. Assoc. 1 (1994) 233–248.