複数データセットで情報を共有する固有表現抽出

大井拓　三輪誠
豊田工業大学
{sd23404,makoto-miwa}@toyota-ti.ac.jp

概要

本研究では，複数のデータセットにおけるラベル付けの違いを効果的にモデル化し，複数データセットで学習を行うために，条件付き変分オートエンコーダ（Conditional Variational Autoencoder; CVAE）をスパンベースの固有表現抽出（Named Entity Recognition; NER）モデルに統合する先行研究からエンコーダの変更と条件ベクトルの学習を行う．実験では，複数の生物医学データセットを用いた学習を行い，BioRED データセットでの評価で提案手法の有効性を示し，性能向上を確認した．

1　はじめに

固有表現抽出（Named Entity Recognition; NER）は自然言語処理の基本的なタスクであり，文書からの情報抽出における最初の手順として重要である．近年，ゼロショットまたは少数ショット学習による大規模言語モデルが注目されているが，高性能を達成するためには教師あり学習によるファインチューニングが依然として必要である [1, 2, 3]．教師あり学習を利用する NER モデルの性能は，通常，ラベル付きデータの量に依存するが，手動でのラベル付きデータ作成は高コストである．

ラベル付きデータの量を増加させる方法の一つとして，既存のラベル付きデータセットを統合する方法がある [4, 2]．しかし，同一の固有表現ラベルを対象としているデータセットであっても，その定義およびアノテーション基準が異なるため，複数のラベル付きデータセットを統一された学習データとして統合することは困難である．例えば，生物医学分野を対象としたデータセットである BioRED [5] と BC5CDR [6] は共に *Chemical* を対象の固有表現としてラベル付けを行っている．しかし，実際のラベル付けを見ると，“glucose” のようにどちらのデータセットでもラベル付けされている用語もあるが，“antipsychotics” のように BioRED ではラベル付けされているが BC5CDR ではラベル付けされていないような用語が存在する．これは BC5CDR のラベル付けの基準を記載したアノテーションガイドライン[1)]で “役割や類似の概念に基づいて名付けられた化学物質名（anti-HIV agents, anticonvulsants, ...）にはラベル付けしないでください．” と明記されていることに起因する．

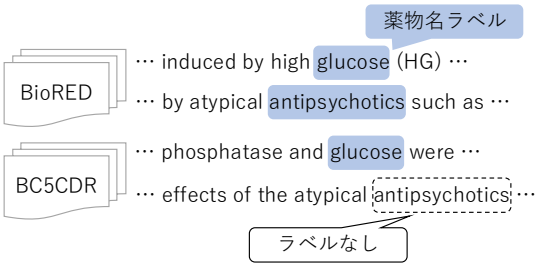

図 1　ラベル付けの違い

複数のデータセットを使用するための一般的なアプローチはマルチタスク学習である [7, 8, 9]．NER におけるマルチタスク学習では，エンコーダモデルはデータセット間で共有され，各データセットに対して別の分類層があるためデータセットの差異を意識せずに学習することができる．しかし，この手法は，独立した分類層により，異なるデータセットのラベル間の関係性を考慮することができない．この問題に対処するため，Luo ら [2] は，対象データセットとより適切に整合させるために追加のデータセットを処理することで，性能を向上した．しかし，この方法では，追加データセットのラベルを一つに制限し，アノテーションに対する編集を行っている．

そこで，先行研究 [10] において，学習時にラベルごとのベクトルを条件として与える CVAE (Conditional Variational Autoencoder) を組み込んだ NER モデルを提案した．この手法により複数データセットで学習を行う Luo らと同様の設定で性能向上を達成した．

本研究では，この発展として，エンコーダモデル

1）https://ftp.ncbi.nlm.nih.gov/pub/lu/BC5CDR/bc5_CDR_data_guidelines.pdf

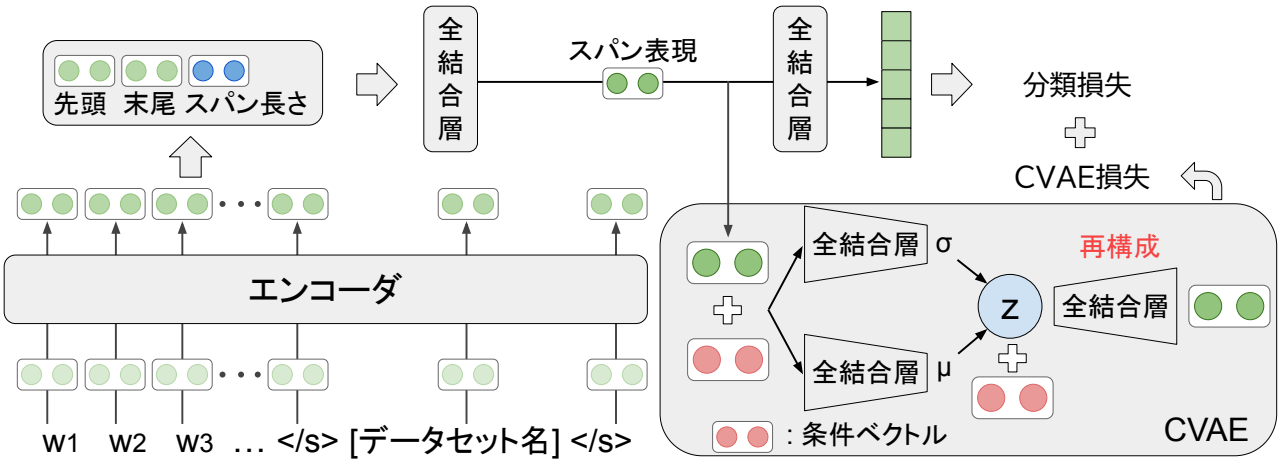

図 2 提案手法の全体図

の変更と条件として加えるベクトルの学習によって追加したデータセットでの性能向上を目指す．

2 関連研究

2.1 スパンに基づく NER

近年，文書内の特定のテキスト区間をまとめたスパン表現を直接分類するスパンベース NER が注目を集めている [1, 11]．これらのモデルは，トークン単位で固有表現を表現する系列ラベリングモデルとは異なり，その単純さと固有表現区間を直接モデル化する能力により注目されている．Zhong ら [1] は，スパン表現に最初と最後のトークンの埋め込みを結合し，スパンの長さの埋め込みを加えて使用している．得られたスパン表現は，全結合層を通した後，分類のために Softmax に入力される．その単純な手法にもかかわらず，このモデルは NER において高い性能を示した．

スパン表現は，スパンベースの NER モデルにおいて重要な役割を果たしている．Ouchi ら [12] は，スパン表現間の類似度を学習し，予測時に学習中のスパン表現との類似度を評価することで，解釈可能性を向上させた．Nguyen ら [13] は，VAE を用いてスパン表現の再構成と同義語生成を組み込んだスパンベース NER モデルを提案した．この手法ではスパン表現の再構成により，インスタンス固有の文脈などの情報がスパン表現に保持され，抽出性能が向上する．

2.2 複数データセットを使用した NER

Luo ら [2] は，BioRED データセットの性能向上を目指し，6 種類の固有表現（*Disease*，*Chemical*，*Gene*，*Species*，*Variant*，*Cell line*）を対象とした NER を複数のデータセットで学習した．追加のデータセットは，ラベル統合，削除，固有表現の範囲の調整により，各データセットを単一のラベルのみに制限し BioRED の基準に沿うように手動で編集された．この手法は既存のマルチタスク学習分類手法を上回る性能を示したが，コストがかかる手動での編集を必要とする．

3 提案手法

本研究では，複数データセットのラベル付けの違いを組み込むため，ラベルごとに作成したベクトルを条件として加える CVAE を導入した NER モデルを提案する．この条件ベクトルに対しても学習を行うことで複数データセットのラベル付けの違いを捉え，既存のラベル付きデータセットを効果的に活用した NER の学習を目指す．提案モデルの概要は図 2 に示す．

Zhong らの手法 [1] に従い，分類対象とするある区間（x_1，…，x_n）をまとめるスパン表現を作成する．スパン表現は，エンコーダから得られる先頭と末尾のトークンの表現，およびスパン長さの埋め込みを結合したものを用いる．この際に使用するエンコーダを BERT から T5 に変更する．異なるデータセットからの入力をモデルが認識できるよう，データセット名を入力テキストの特殊トークンとして追加する．スパン表現は，2 つの全結合層と Softmax

を用いて分類される．損失関数としては交差エントロピー損失（L_{CE}）を採用する．

CVAE による損失関数については，スパン表現 $\boldsymbol{h}_{span}$ とラベルごとに作成した条件ベクトルを結合したベクトルの圧縮・再構成を行い，損失を計算する．各ラベルに対応して作成する条件ベクトルは正解ラベルを基にデータセットが異なるラベルの共有と非共有を表すようなベクトルを用いる．詳細は 4.2 節で説明する．また条件ベクトルは学習パラメータとして損失関数による更新を行うことで，実際のデータにおけるラベル付けの違いを吸収する．再構成した表現ベクトルと元のスパン表現から CVAE の損失関数を計算する．元のスパン表現と再構成したスパン表現の平均二乗誤差と予測した分布が正規分布 $N(0,1)$ から離れないように制約をかける KL ダイバージェンスを足し合わせて使用する．

最終的に，2 つの損失の重み付き和を，学習のための全体の損失 L として使用する．

$$L = L_{CE} + \alpha L_{CVAE}.$$

推論時は，正解ラベルを必要とする CVAE は使用せず，学習済みの分類器のみを用いる．

4 実験設定

実験で使用したデータセットとモデルについて説明する．

4.1 データセット

Luo ら [2] に従い，10 個の生物医学分野の文書にラベル付けを行ったデータセットを使用した．学習中の評価には BioRED の開発データを使用し，他のデータセットの開発データは学習データとして使用した．Luo らの手法 [2] では追加データセットの固有表現を BioRED の基準に従うように編集したが，本研究ではそのような編集を行わず，元のデータセットをそのまま使用した．

4.2 条件ベクトル

CVAE への条件として与える条件ベクトルは，2 つの one-hot ベクトルを連結して作成する．1 つ目のベクトルは，表 6 に記載されている全てのラベル（各データセットの負例ラベルを含む）を表現する 47 次元の one-hot ベクトルである．2 つ目のベクトルは，BioRED のラベルに基づく共有ラベル（6 種類＋負例）を表現する 7 次元の one-hot ベクトルである．追加データセットのラベルを BioRED のラベルの 1 つに対応付けた Luo ら [2] のマッピングを使用する（例えば，NLMgene の Gene ラベルと FamilyName ラベルを BioRED の Gene ラベルと同等に扱う）．この対応関係の詳細は表 6 に示す．

4.3 モデル

本稿では T5（Text-To-Text Transfer Transformer）-3B[14] のエンコーダ部分を使用した．以下の設定について比較を行った．また本稿の実験と合わせて先行研究 [10] の設定は入力の文末にデータセット名を加えたモデルと比較を行う．

- **Single** 単一の対象データセットで学習した固有表現抽出モデル．
- **Multi** 純粋なマルチタスク学習設定で複数のデータセットを用いて学習した固有表現抽出モデル．ここでは，基本となるエンコーダ層は共有しながら，各データセットに対して個別の分類層を用意する．
- **+CVAE** 4.2 節の条件ベクトルを組み込んだ CVAE を用いた固有表現抽出モデル．
- **+CVAE (Fixed)** 初期化後に条件ベクトルを固定した CVAE を用いた固有表現抽出モデル．
- **+CVAE (Unshared)** 4.2 節の 1 つ目のワンホットベクトル（各データセットのラベル）のみで条件ベクトルを初期化した CVAE を用いた固有表現抽出モデル．

5 結果

Luo ら [2] のモデル，先行研究 [10] との比較を表 1 に，先行研究と Multi+CVAE 設定における追加したデータセットのテストデータでの評価を表 2 に示す．結果からエンコーダの変更と条件ベクトルの微調整によって BioRED のテストデータに対する性能が向上し，追加のデータセットに対しても向上する傾向が見られた．条件ベクトルの条件を変更した場合のアブレーション実験の結果を表 3 に示す．Fixed 設定と Unshared 設定の両方とも提案モデルと比較して性能が劣っており，共有ラベルの使用と学習中の条件ベクトルの調整の有効性が示された．

また抽出結果を分析すると，図 1 に示した “antipsychotics” は Multi では追加データセットでの学習の影響により BioRED の開発データでも薬物名として誤って抽出していたが Multi+CVAE では抽出

表 1　BioRED のテストデータによる評価結果．評価指標として F 値［%］を用いる．各タイプにおける最高スコアは太字で示す．

モデル	All	*Disease*	*Chemical*	*Gene*	*Variant*	*Species*	*Cell line*
Luo ら [2][2)]	91.26	88.07	90.98	92.40	88.51	97.50	90.53
先行研究 [10]	94.18	**93.48**	92.29	**96.39**	**95.18**	97.69	76.54
Single	89.89	86.52	87.99	91.19	90.56	97.73	86.67
Multi	93.99	92.82	92.16	95.89	90.07	98.00	82.76
Single + CVAE	89.87	85.81	89.45	90.72	89.18	97.62	91.84
Multi + CVAE	**94.36**	92.84	**92.89**	95.77	94.29	**98.23**	**78.57**

しなかった．

表 2　追加データセットのテストデータに対する評価（F 値 [%]）

データセット	先行研究 [10]	Multi + CVAE
BC5CDR	91.16	**91.59**
BioID	**91.89**	90.01
GNormPlus	78.15	**80.68**
Linnaeus	84.80	**93.84**
NCBIdisease	80.13	**84.11**
NLMchem	- [3)]	**82.12**
NLMgene	**85.57**	84.76
SPECIES800	76.02	**77.64**
tmVar3	91.22	**91.87**

さらに，*Variant* ラベルと tmVar3 の *DNAMutation*, *DNAAllele* の学習データに対して，両方の学習データに含まれる事例を色分けしてスパン表現の可視化を行った結果を図 3 に示す．Multi 設定での結果に比べて提案手法ではこれらの事例が *DNAAllele*（緑の点）とともに tmVar3 の別のラベル（DNAMutation）と分かれるように表現できているとわかる．

6　おわりに

本研究では，学習データの量を増やすために，複数データセットのラベル付け基準の違いを緩和することを目指した．この目的を達成するため，CVAE に基づく損失関数を固有表現抽出モデルに組み込む手法を提案した．

実験結果から，条件ベクトルの学習による調整を行うことで先行研究に対して対象データセット，追加データセットでともに性能向上した．

今後の課題として，共有ベクトルの定義が必要ないエンコーダベースの条件ベクトルの検討や，生物医学分野以外の他のデータセットによる評価が挙げられる．

表 3　アブレーション実験

モデル	F 値 [%]
Multi+CVAE	**94.36**
Multi+CVAE (Fixed)	93.99
Multi+CVAE (Unshared)	94.12

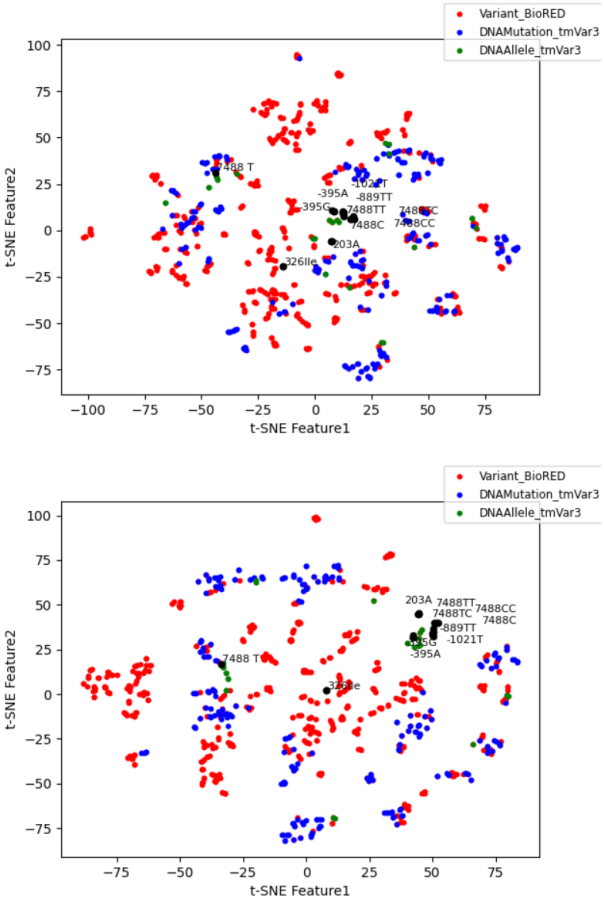

図 3　*Variant* に対するスパン表現の可視化．上図が Multi モデルのスパン表現，下図が Multi+CVAE のスパン表現．

2）　論文中の値
3）　テストデータ中に BERT の入力長を超える文があるため評価していない．

参考文献

[1] Zexuan Zhong and Danqi Chen. A Frustratingly Easy Approach for Entity and Relation Extraction. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 50–61, Online, June 2021. Association for Computational Linguistics.

[2] Ling Luo, Chih-Hsuan Wei, Po-Ting Lai, Robert Leaman, Qingyu Chen, and Zhiyong Lu. AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning. **Bioinformatics**, Vol. 39, No. 5, p. btad310, 05 2023.

[3] Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. Instructuie: Multi-task instruction tuning for unified information extraction, 2023.

[4] Rezarta Islamaj, Robert Leaman, Sun Kim, Dongseop Kwon, Chih-Hsuan Wei, Donald C Comeau, Yifan Peng, David Cissel, Cathleen Coss, Carol Fisher, et al. Nlm-chem, a new resource for chemical entity recognition in pubmed full text literature. **Scientific data**, Vol. 8, No. 1, p. 91, 2021.

[5] Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. Biored: a rich biomedical relation extraction dataset. **Briefings in Bioinformatics**, Vol. 23, No. 5, p. bbac282, 2022.

[6] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. **Database**, Vol. 2016, , 05 2016. baw068.

[7] Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. Cross-type biomedical named entity recognition with deep multi-task learning. **Bioinformatics**, Vol. 35, No. 10, pp. 1745–1752, 10 2018.

[8] Mei Zuo and Yang Zhang. Dataset-aware multi-task learning approaches for biomedical named entity recognition. **Bioinformatics**, Vol. 36, No. 15, pp. 4331–4338, 05 2020.

[9] Nicholas E. Rodriguez, Mai Nguyen, and Bridget T. McInnes. Effects of data and entity ablation on multitask learning models for biomedical entity recognition. **Journal of Biomedical Informatics**, Vol. 130, p. 104062, 2022.

[10] 大井拓, 三輪誠, 佐々木裕. Cvae による複数データセットからの固有表現抽出. 言語処理学会 第 30 回年次大会 発表論文集, 2024.

[11] Mohammad Golam Sohrab and Makoto Miwa. Deep Exhaustive Model for Nested Named Entity Recognition. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2843–2849, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[12] Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, Sho Yokoi, Tatsuki Kuribayashi, Ryuto Konno, and Kentaro Inui. Instance-based learning of span representations: A case study through named entity recognition. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6452–6459, Online, July 2020. Association for Computational Linguistics.

[13] Nhung T. H. Nguyen, Makoto Miwa, and Sophia Ananiadou. Span-based named entity recognition by generating and compressing information. In **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 1984–1996, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[14] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[15] Cecilia Arighi, Lynette Hirschman, Thomas Lemberger, et al. Bio-id track overview. In **BioCreative VI Workshop**, pp. 28–31, Bethesda, MD, USA, 2017. BioCreative.

[16] Chih-Hsuan Wei, Hung-Yu Kao, Zhiyong Lu, et al. Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. **BioMed research international**, Vol. 2015, , 2015.

[17] Martin Gerner, Goran Nenadic, and Casey M Bergman. Linnaeus: a species name identification system for biomedical literature. **BMC bioinformatics**, Vol. 11, No. 1, pp. 1–17, 2010.

[18] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. **Journal of Biomedical Informatics**, Vol. 47, pp. 1–10, 2014.

[19] Rezarta Islamaj, Chih-Hsuan Wei, David Cissel, Nicholas Miliaras, Olga Printseva, Oleg Rodionov, Keiko Sekiya, Janice Ward, and Zhiyong Lu. Nlm-gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. **Journal of biomedical informatics**, Vol. 118, p. 103779, 2021.

[20] Evangelos Pafilis, Sune P Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. The species and organisms resources for fast and accurate identification of taxonomic names in text. **PloS one**, Vol. 8, No. 6, p. e65390, 2013.

[21] Chih-Hsuan Wei, Alexis Allot, Kevin Riehle, Aleksandar Milosavljevic, and Zhiyong Lu. tmvar 3.0: an improved variant concept recognition and normalization tool. **Bioinformatics**, Vol. 38, No. 18, pp. 4449–4451, 2022.

表 4 BioRED データセット [5] の学習データ，開発データ，テストデータにおけるラベル別の詳細な統計．

	学習	開発	テスト
Document	400	100	100
All	13,351	3,533	3,535
Disease	3,646	982	917
Chemical	2,853	822	754
Gene	4,430	1,087	1,180
Variant	890	250	241
Species	1,429	370	393
Cell line	103	22	50

表 5 ハイパーパラメータの設定

ハイパーパラメータ	値
学習率	7e-4
α	1e-4
バッチサイズ	64
最大スパン長	10
スパン長さ埋め込みの次元数	150
CVAE の隠れ層の次元数	150
LoRA 適用層	T5 の全層
LoRA ランク	32

A　データセットの詳細

BioRED データセットの統計を表 4 に，その他のデータセットの対象とする固有表現ラベルの一覧と各ラベルの対応付けした BioRED ラベルを表 6 に示す．

B　ハイパーパラメータ

本実験で使用したハイパーパラメータを表 5 に示す．

表 6 追加で使用した学習データのラベルごとのデータ数と共有するように定めた BioRED のラベル．

データセット	固有表現ラベル	対応ラベル
BC5CDR [6]	*Disease*	*Disease*
	Chemical	*Chemical*
BioID [15]	*Cell line*	*Cell line*
GNormPlus [16]	*FamilyName*	*Gene*
	Gene	*Gene*
	DomainMotif	-
Linnaeus [17]	*Species*	*Species*
NCBIdisease [18]	*DiseaseClass*	*Disease*
	SpecificDisease	*Disease*
	CompositeMention	*Disease*
	Modifier	*Disease*
NLMChem [4]	*Chemical*	*Chemical*
	NonStandardRef	-
	OTHER	-
NLMGene [19]	*Gene*	*Gene*
	FamilyName	*Gene*
	Cell	-
	DomainMotif	-
	ChromosomeLocation	-
SPECIES800 [20]	*Species*	*Species*
tmVar3 [21]	*Gene*	*Gene*
	Species	*Species*
	Disease	*Disease*
	DNAMutation	*Variant*
	ProteinMutation	*Variant*
	OtherMutation	*Variant*
	Cell line	*Cell line*
	AcidChange	*Variant*
	SNP	*Variant*
	DNAAllele	*Variant*
	ProteinAllele	*Variant*