SHINRA2020-ML:30 言語の Wikipedia ページの分類

関根聡[1)] 野本昌子[1)] 中山功太[1)] 隅田飛鳥[1)] 松田耕史[1)2)] 安藤まや[3)]
1） 理化学研究所 AIP 2）東北大学 3）フリー
{satoshi.sekine, masako.nomoto, kouta.nakayama, asuka.sumida, koji.matsuda}@riken.jp , maya@kzd.biglobe.ne.jp

概要

Wikipedia に書かれている世界知識を計算機が扱えるような形に変換することを目的として、Wikipedia を構造化するプロジェクトを推進している本プロジェクトは、「協働による知識構築(Resource by Collaborative Contribution)」のスキームに基づき、評価型ワークショップを開催し、それに参加したシステムの結果を統合してより良い知識にまとめ上げ、それを公開していくことを目指す。SHINRA2020-ML タスクでは、30 言語の Wikipedia の項目を拡張固有表現に分類するタスクを実施し、分類データの構築を目指した。世界 7 カ国から 10 団体が参加し、17 のシステムによる結果が提出された。この論文では、本プロジェクトの概要を説明し、森羅／SHINRA2021 タスクの計画についても紹介する。

1 背景と目的

自然言語理解を実現するためには、言語的及び意味的な知識が必要なことは論を待たない。しかしながら、大規模な知識の作成は非常に膨大なコストがかかり、メンテナンスも非常に難しい問題である。名前を中心とした知識において、クラウドソーシングによって作成されている Wikipedia はコストの面でもメンテナンスの面でもそれ以前の百科事典の概念を一新した。しかし、この Wikipedia を自然言語処理のために活用しようと考えると障壁は高い。Wikipedia は人が読んで理解できるように書かれており、計算機が利用できるような形ではないためである。計算機の利用を念頭においた知識ベースには、CYC[11]、DBpedia[12]、YAGO[13]、Freebase[14]、Wikidata[15]などがあるが、それぞれに解決すべき課題がある。特に CYC ではカバレージの問題、他の知識ベースでは、首尾一貫した知識体系に基づいていない構造化の問題がある。この課題を解決するため、私たちは、名前のオントロジー「拡張固有表現」[3][10]に Wikipedia 記事を分類し、属性情報を抽出することで計算機が利用可能な Wikipedia の構造化を進めている[1][4][5][6][7]。本稿では、30 言語の Wikipedia ページの分類タスクである「SHINRA2020-ML」[2]について説明する。

2 協働による知識構築

Wikipedia の全データの構造化を人手で行うことはほぼ不可能に近い。特に、日々更新される Wikipedia を対象にしているため、将来の更新作業を考慮しても現実的ではない。しかし、分類や属性値抽出のような知識構築は様々な機械学習手法によってある程度の精度で自動化できることが分かっている。今回の分類タスクでも機械学習を活用するが、一つの機械学習システムだけで実現するのではなく、多くの違った種類のシステムが協力することによってより良いリソースを作成することを目標としている。現在の自然言語処理では「評価型ワークショップ」が多数行われている。この形式のワークショップは既存タスクにおける機械学習システムの最適化競争の側面があるが、これを逆に利用してリソースの作成を行おうと考えている。つまり、運営者側で訓練データとテストデータを用意し、多くのシステムに評価型ワークショップに参加していただく。この時にテストデータを参加者には知らせないことで、参加者には訓練データ以外の全項目を対象に結果を出すという仕組みを取り入れ、その結果は共有することを約束してもらう。この結果を利用しアンサンブル学習の手法を用いて、より信頼できるリソースを自動的に作る。また、信頼度の低いものを人手で確認訂正して次の学習時の訓練データにするアクティブ・ラーニングや、何度も訓練データの作成とシステムの実行を繰り返すブートストラッピング手法を取り入れることで、多くの参加者と協力しあって、精度の高いリソース作成を実現していくことを目標としている。本スキームは Resource by Collaborative Contribution (RbCC:協働による知識構築)と呼び、森羅プロジェクトの最も重要な骨格である。

3 タスクの説明

SHINRA2020-ML タスクは、30 言語の Wikipedia ページを拡張固有表現のカテゴリーに分類するタスクである。このタスクで使用されている拡張固有表現と、教師データの作成方法について説明する。

3.1 拡張固有表現

「拡張固有表現」とは、[3][10]によって定義された固有表現に関する定義であり階層構造を持つ。図 1 に全カテゴリーを掲載する。「人名」、「地名」、「組織名」のみならず、「イベント名」、「地位職業名」、「芸術作品名」などを含む。また、例えば「地名」には「河川名」等の「地形名」や、「星座名」等の「天体名」を含む等、幅広い種類の下位カテゴリが含まれる。今回のタスクで使用されているバージョン 8.0 は最大 4 階層で、219 種類の「拡張固有表現」が定義されている。Version8 以前の「拡張固有表現定義書」は百科事典、新聞記事を対象とした質問応答システムや WordNet 等のオントロジーを参考に構築されてきたが、今回の更新ではより Wikipedia に即した定義となっている。Wikipedia は情報の更新頻度が高く、一般的に幅広く利用されているためである。拡張固有表現の旧定義書をもとに Wikipedia の項目を分類したところ、3 つの問題点が発見された。第 1 に、本来定義しきれない語が分類されるはずの「*__その他」というカテゴリに多くの項目が分類されたこと、第 2 に項目の分類の過程で、判断に迷うカテゴリが存在したこと、第 3 に Wikipedia 特有のメタページなど拡張固有表現に分類できないページが存在したことである。Wikipedia には、同じような表記の項目が複数ある場合の「曖昧さ回避」、項目に異表記がある場合等の「転送元」、「1930 年の日本公開映画」のように項目になるような言葉が並ぶ「一覧」といったページが存在する。これらは百科事典としての見出しではない項目(IGNORE)となる。詳細については[3]を参照されたい。

3.2 教師データと評価用データ

各言語の教師データは、すでに分類されている日本語の Wikipedia データ[8][9]と日本語からの言語間リンクによって作成された。例えば、日本語の Wikipedia からドイツ語の Wikipedia には 27 万余の言語間リンクがあり、それらの日本語 Wikipedia ページに分類されたカテゴリーを対応するドイツ語 Wikipedia ページにつける。これらを教師データーとしてドイツ語 Wikipedia の 200 万余のページを分類するのが今回のタスクである。日本語 Wikipedia の分類データは 92 万のページに対して、機械学習と人手によるチェックで構築したものである[3][10]。表 1 に 31 言語の統計データを示す。

言語	ページ数	日本語からのリンク数	割合
英語（en）	5,790,377	439.354	7.6
スペイン語（es）	1,500,013	257,835	17.2
フランス語（fr）	2,074,648	318,828	15.4
ドイツ語（de）	2,262,582	274,732	12.1
中国語（zh）	1,041,039	267,107	25.7
ロシア語（ru）	1,523,013	253,012	16.6
ポルトガル語（pt）	1,014,832	217,896	21.5
イタリア語（it）	1,496,975	270,295	18.1
アラビア語（ar）	661,205	73,054	11.0
日本語	1,136,222	-	-
インドネシア語（id）	451,336	115,643	25.6
トルコ語（tr）	321,937	111,592	34.7
オランダ語（nl）	1,955,483	199,983	10.2
ポーランド語（pl）	1,316,130	225,552	17.1
ペルシャ語（fa）	660,487	169,053	25.6
スェーデン語（sv）	3,759,167	180,948	4.8
ベトナム語（vi）	1,200,157	116,280	9.7
韓国語（ko）	439,577	190,807	43.7
ヘブライ語（he）	236,984	103,137	43.5
ルーマニア語（ro）	391,231	92,002	23.5
ノルウェイ語（no）	501,475	135,935	27.1
チェコ語（cs）	420,195	135,935	25.1
ウクライナ語（uk）	881,572	181,122	20.5
ヒンディー語（hi）	129,141	30,547	23.6
フィンランド語（fi）	450,537	144,750	32.1
ハンガリア語（hu）	443,060	120,295	27.2
デンマーク語（da）	242,523	91,811	35.6
タイ語（th）	129,294	59,791	46.2
カタルニア語（ca）	601,473	139,032	23.1
ギリシャ語（el）	157,566	60,513	38.4
ブルガリア語（bg）	248,913	89,017	35.7

表１．31 言語の統計データ

３０言語に対してリーダーボード用データと評価用データを作成した。評価用データは RbCC の目的のため公開していない。現状のそれらのデータの作成方法についても、今後も継続する予定のタスクの平等性の観点から明らかにしない方針である。

4 タスクの実施と評価

4.1 スケジュールとリーダーボード

SHINRA2020-ML タスクは以下のスケジュールで NTCIR15 の下で実施した。

2020 年 1 月：データリリース
2020 年 4 月：HP および登録公開
2020 年 8 月 31 日：登録&結果締め切り
2020 年 9 月 16 日：評価結果変換
2020 年 12 月 8-11 日：結果報告会(NTCIR15)

また、今回はリーダーボードを作成し、多くの人に興味を持っていただける仕掛けを行なった。実際に 9 団体がリーダーボードに結果を提出し、この試みは成功であったと考えている。

4.2 参加団体

SHINRA2020-ML には 7 カ国から 10 団体が参加した。参加団体と参加した言語を表２に載せる。また、Appendix に一部の参加システムの説明を載せる。詳細は NTCIR15 に投稿された論文を参照されたい。[16][17][18][19][20][21][22]

グループ ID	国	参加した言語
CMVS	フィンランド	1 (ar)
FPTAI	ベトナム	30 (all)
HUKB	日本	30 (all)
PribL	ポルトガル	15 (ar, cs, de, en, es, fr, it, ko, nl, no, pl, pt, ru, tr, zh)
RH312	インド	6 (bg, fr, hi, id, th, tr)
TKUIM	台湾	30 (all)
Ousia	日本	9 (ar,de,es,fr,hi,it,pt,th,zh)
Uomfj	オーストラリア、日本	28 (except for el, sv)
Vlp	ベトナム	1 (vi)
LIAT	日本	30 (all)

表２．参加団体

4.3 評価結果

表３に参加システムの評価結果を載せる。(この表には提出データの一部しか提出しなかった参加者の結果と締め切り後に提出された結果の一部は載せていない）また、締め切り後に提出されたシステムについては”Late Submission”として表示されている。本タスクは１ページに複数のラベルが着くことが許され、参加システムも複数のラベルを提案することができる。システムの評価には精度と再現率から計算される F 値のマイクロ平均を用いた。

5 森羅／SHINRA2021 のタスク

SHINRA2020-ML は、当初の目的を達成し、過去の森羅の日本語タスク同様に RbCC の考え方の有効性が実証されたと考えている。森羅/SHINRA2021 では以下の３つのタスクを実施しようと考えている。

ML: 30 言語の分類

基本的に SHINRA2020-ML と同様の Wikiepdia 分類タスク。リーダーボードや評価データを拡充し、出力データ数を限ったタスクを追加する予定である。

CrowdML:クラウドソーシングによる分類精度向上

SHINRA2020-ML タスクの出力データを元に、クラウドソーシングを利用してより精度の高いデータを作成するタスク。クラウドソーシングにかかる費用は決まった上限まで主催者が負担し、その範囲内で効率的に精度の高いデータを作成することを評価目標とする予定である。

LinkJP:属性値をページにリンク

森羅 2020/2019/2018-JP で対象としたカテゴリーの属性値に対して、対応する Wikipedia のページをリンクさせるタスク。

詳細は論文執筆現在検討中であるが、３月上旬にはタスク仕様およびサンプルデータを完成させ、公開する予定である。全てのタスクは春にスタートし、９～10 月頃を締め切りにし、11、12 月の報告会の開催というスケジュールを計画している。また、これまで作成された全ての森羅データは一般公開されている[1]。ML タスクのデータは SHINRA-ML のホームページ[2]から入手されたい。

6 まとめ

Wikipedia の構造化データ「森羅」の作成を目指したプロジェクトを推進している。前章に記した通りこのプロジェクトは多くの方の協力なしには進まない。これまでの森羅のタスク協力いただいた皆様、特に評価に参加いただいた全ての団体にはここで感謝を述べたい。今後もより深い知識処理を実現するためにも、本プロジェクトに多くの協力をいただけるようにお願いしたい。

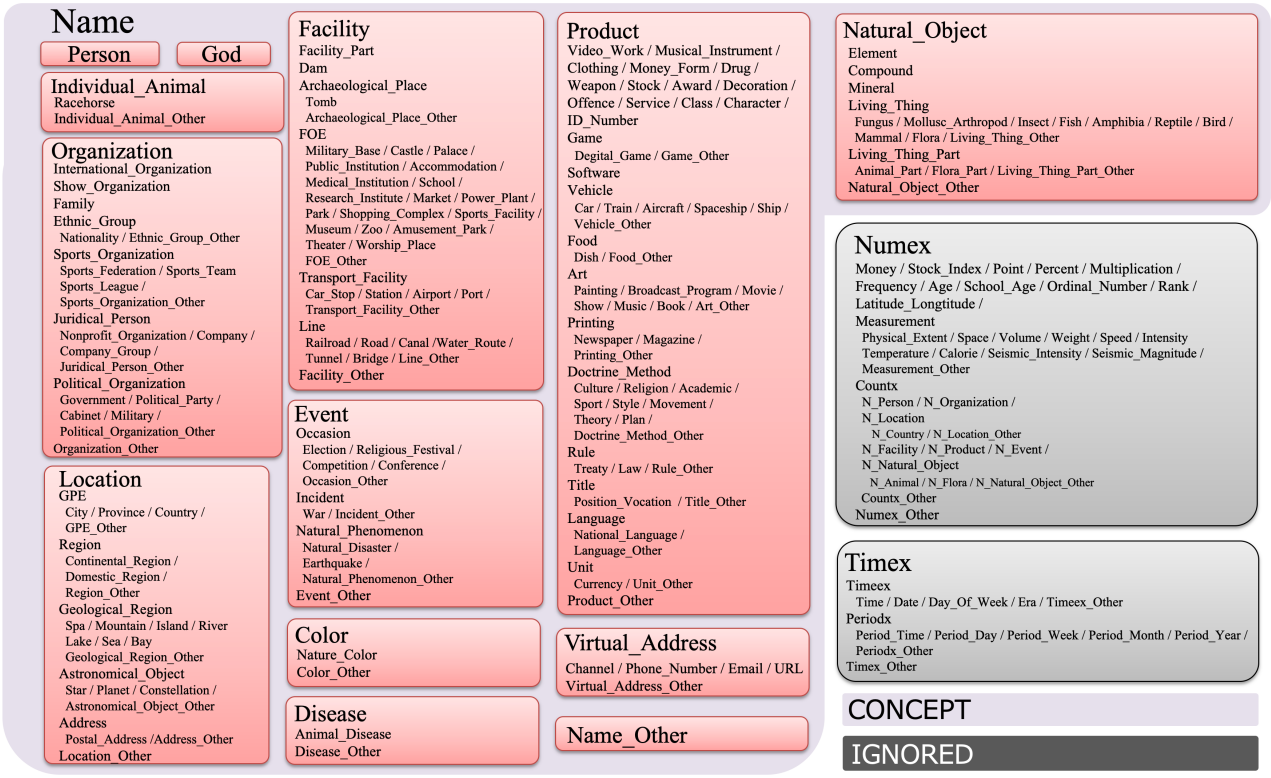

図１．拡張固有表現カテゴリー

Group ID		FPTAI	LIAT	PribL	PribL	RH312	ousia	uomfj	uomfj	uomfj	FPTAI	HUKB	HUKB	HUKB	LIAT
Method ID		BERT	ML-BERT	BERTGRU	BERTLINCONCAT	RnnGnnXlmr	RoBERTa+wiki2vec+wikidata	jointrep	jointrepPostprocess	jointrepUnionPostprocess	BERT	AB	ABC	AC	ML-BERT
Late Submission											Y	Y	Y	Y	Y
ar	Arabic	73.25	63.16	76.27	75.45	-	70.52	64.55	64.55	64.55	73.25	30.98	30.98	13.51	-
bg	Bulgarian	83.77	75.20	-	-	82.13	-	83.07	83.07	83.07	83.28	60.86	61.06	28.09	-
ca	Catalan, Valencian	52.55	76.28	-	-	-	-	79.82	79.82	79.82	81.10	42.34	42.54	16.26	-
cs	Czech	84.47	79.46	-	81.19	-	-	81.29	81.29	81.29	83.74	52.61	52.61	18.86	-
da	Danish	82.30	74.80	-	-	-	-	80.56	80.56	80.56	81.74	49.01	49.01	13.99	-
de	German	22.62	79.49	80.24	79.83	-	81.86	81.03	81.03	81.03	81.26	53.72	53.82	26.81	-
el	Greek, Modern (1453-)	84.40	72.43	-	-	-	-	-	-	-	84.10	7.51	7.51	7.51	-
en	English	82.23	78.56	81.27	80.12	-	-	82.73	82.57	82.68	81.96	45.11	45.11	11.92	-
es	Spanish, Castilian	80.60	77.73	80.30	80.72	-	80.94	81.39	81.39	81.39	80.60	49.21	49.11	19.50	-
fa	Persian	81.70	75.42	-	-	-	-	80.38	80.38	80.38	81.52	45.59	45.59	15.66	-
fi	Finnish	83.62	79.13	-	-	-	-	80.91	80.91	80.91	83.36	53.15	53.45	17.06	-
fr	French	21.59	76.88	77.93	78.52	80.31	81.01	78.21	78.21	78.21	80.68	43.84	43.74	11.23	-
he	Hebrew	83.79	79.11	-	-	-	-	81.09	81.09	81.09	84.21	59.95	60.05	15.78	-
hi	Hindi	76.43	16.49	-	-	71.70	69.75	66.67	66.67	66.67	75.65	39.70	39.51	22.02	-
hu	Hungarian	85.46	78.93	-	-	-	-	85.02	85.02	85.02	84.78	69.15	69.44	26.09	-
id	Indonesian	81.93	72.45	-	-	77.56	-	78.51	78.51	78.51	81.65	44.07	44.47	16.28	-
it	Italian	26.55	81.36	81.92	81.89	-	81.21	82.02	82.02	82.02	82.81	45.55	45.55	12.06	-
ko	Korean	83.67	80.38	81.51	81.04	-	-	82.51	82.51	82.51	83.77	63.68	63.98	13.95	-
nl	Dutch, Flemish	83.29	79.86	80.95	81.26	-	-	81.64	81.64	81.64	83.17	42.36	42.45	17.12	-
no	Norwegian	80.53	76.50	-	78.39	-	-	78.79	78.79	78.79	80.17	34.58	34.58	11.33	-
pl	Polish	84.53	80.60	82.73	83.46	-	-	84.52	84.52	84.52	84.07	62.72	63.51	32.55	-
pt	Portuguese	83.23	78.49	82.36	81.88	-	81.40	80.87	80.87	80.87	82.70	42.32	42.62	16.10	-
ro	Romanian, Moldavian, Moldovan	84.60	76.17	-	-	-	-	80.83	80.83	80.83	84.60	57.60	57.70	28.50	-
ru	Russian	84.08	79.09	82.60	83.07	-	-	82.90	82.90	82.90	83.44	42.04	42.24	11.30	-
sv	Swedish	83.18	71.63	-	-	-	-	-	-	-	83.44	50.32	50.62	21.98	79.58
th	Thai	81.26	49.58	-	-	76.77	76.36	65.02	65.02	65.02	81.16	39.98	40.38	24.05	-
tr	Turkish	86.50	77.19	84.36	83.23	83.28	-	84.85	84.85	84.85	86.03	61.88	62.48	16.73	-
uk	Ukrainian	83.12	78.71	-	-	-	-	81.61	81.61	81.61	82.61	60.29	60.19	22.51	-
vi	Vietnamese	80.34	75.24	-	-	-	-	77.06	77.06	77.06	80.42	60.38	60.48	22.14	-
zh	Chinese	81.25	77.97	78.38	79.37	-	79.76	78.58	78.58	78.58	80.60	21.22	21.42	17.57	-

表３．システムの評価結果

参考文献

1. SHINRA project homepage: https://shinra-project.info
2. SHINRA 2020-ML homepage: http://shinra-project.info/shinra2020ml/
3. Extended Named Entity homepage: https://ene-project.info

4. Satoshi Sekine, Akio Kobayashi, and Kouta Nakayama. 2019. SHINRA: Structuring Wikipedia by Collaborative Contribution. In *Proceedings of the 1st conference on the Automatic Knowledge Base Construction AKBC-2019*.

5. 関根聡，小林暁雄，安藤まや，馬場雪乃，乾健太郎. Wikipedia 構造化データ「森羅」構築に向けて. 言語処理学会第 24 回年次大会(2018)

6. 小林暁雄，中山功太，安藤まや，関根聡. Wikipedia 構造化プロジェクト「森羅 2019」. 言語処理学会第 26 回年次大会(2020)

7. 小林暁雄，関根聡，安藤まや. Wikipedia 構造化プロジェクト「森羅 2018」言語処理学会第 25 回年次大会(2019)

8. 関根聡,安藤まや,小林暁雄,松田耕史,Duc Nguyen,鈴木正敏,乾健太郎 「拡張固有表表現+Wikipedia」データ（2015 年 11 月版 Wikipedia 分類作業完成版）.言語処理学会第 24 回年次大会(2018)

9. 鈴木 正敏，松田 耕史，関根 聡，岡崎 直観，乾健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与. 言語処理学会第 22 回年次大会 (2016)

10. Satoshi Sekine. 2008. Extended Named Entity Ontology with Attribute Information. In *Proceedings of the Sixth Internationa Conference on Language Resource and Evaluation (LREC08).*

11. Douglas B. Lenat. CYC: a large-scale investment in knowledge infrastructure. ACM 38, pp. 32–38.

12. Lehmann, J., Isele, R., Jakob, M., Jentzch, M., Kontokostas, D., Mendes, P.N., Hellman, S., Morsey M., Kleef, P., Auer, S. and Bizer, C. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal, 6(2) :167—195

13. Farzaneh Mahdisoltani, Joanna Biega, Fabian M. Suchanek. YAGO3: A Knowledge Base from Multilingual Wikipedias. Proceedings of the Conference on Innovative Data Systems Research (CIDR 2015).

14. Bollacker, K., Evans, C., Paritosh, P., Sturge, T. and Taylor, J. Freebase: a collaboratively created graph database for structuring human knowledge. Proc. International conference on Management of data (SIGMOD '08). ACM, pp.1247-1250.

15. Vrandečić, D. and Krötzsch, M. Wikidata: a free collaborative knowledgebase. Commun. ACM57, pp. 78-85.

16. The Viet Bui and Phuong Le-Hong. Cross-lingual Extended Named Entity Classification of Wikipedia Articles. In Proceedings of NTCIR-15.

17. Rúben Cardoso, Afonso Mendes and Andre Lamurias. Priberam Labs at the NTCIR-15 SHINRA2020-ML: Classification Task. In Proceedings of NTCIR-15.

18. Tushar Abhishek, Ayush Agarwal, Anubhav Sharma, Vasudeva Varma and Manish Gupta. Rehoboam at the NTCIR-15 SHINRA2020-ML Task. In Proceedings of NTCIR-15.

19. Hiyori Yoshikawa, Chunpeng Ma, Aili Shen, Qian Sun, Chenbang Huang, Guillaume Pelat, Akiva Miura, Daniel Beck, Timothy Baldwin and Tomoya Iwakura. UOM-FJ at the NTCIR-15 SHINRA2020-ML Task. In Proceedings of NTCIR-15.

20. Kouta Nakayama and Satoshi Sekine. LIAT Team's Wikipedia Classifier at NTCIR-15 SHINRA2020-ML: Classification Task. In Proceedings of NTCIR-15.

21. Masaharu Yoshioka and Yoshiaki Koitabashi. HUKB at SHINRA2020-ML task. In Proceedings of NTCIR-15.

22. Sosuke Nishikawa and Ikuya Yamada. Studio Ousia at the NTCIR-15 SHINRA2020-ML Task. In Proceedings of NTCIR-15.

A 付録

参加システムの説明

Group :Run	# of Best Scores	Basic Method	Method Description	Use of Datasets (1)	(2)	(3)	(4)	(5)	Use of External Info
(F-1)	19	BERT	BERT Linking	Y			Y		multilingual-bert-base-cased
(F-2)	6	BERT	BERT Linking	Y			Y		multilingual-bert-base-cased
(L-1)	0	BERT	ML BERT + Transformer Encoder	Y	Y				ML BERT
(L-2)	0	BERT	ML BERT + Transformer Encoder	Y	Y				ML BERT
(P-1)	1	BERT, GRU	M-BERT's pooled output for the first 512 tokens of the wiki pages is provided to a GRU layer that sequentially predicts the correct label for each hierarchical level	Y	Y			Y	BERT base multilingual cased
(P-2)	0	BERT	M-BERT embeddings pooled into a single representation through concatenation and mean operations, followed by a classifier consisting of a linear layer.	Y	Y			Y	BERT base multilingual cased
(U-1)	2	BERT, Inception -v3	Combine document representation trained with different source of information	Y	Y			Y	pre-trained BERT model, pre-trained embeddings of Wikidata graph, Wikipedia pages in HTML format, Inception-v3 model
(U-2)	1	BERT, Inception -v3	Combine document representation trained with different source of information, followed by post-processing	Y	Y			Y	pre-trained BERT model, pre-trained embeddings of Wikidata graph, Wikipedia pages in HTML format, Inception-v3 model
(U-3)	1	BERT, Inception -v3	Combine document representation trained with different source of information, followed by post-processing. combine outputs of multiple models	Y	Y			Y	pre-trained BERT model, pre-trained embeddings of Wikidata graph, Wikipedia pages in HTML format, Inception-v3 model
(O-1)	2	XLM-R	XLMRoBERTa (large) + entity の wikipedia2vec+ entity の wikidata graph embedding で fine-tune	Y	Y	Y			wikipedia2vec, wikidata pretrained graph embedding.
(R-1)	0	XLM-R	Used xlmr as embedding with is learned jointly by RNN and GNN	Y	Y		Y	Y	
(H-1)	0	Wikipedia category	Using Wikipedia categories information of the page for classification.	Y	Y	Y	Y	Y	Recent Wikipedia dump to extract language links and category hierarchy information.
(H-2)	0	Wikipedia category	Using Wikipedia categories information of the page for classification. w/o language links exp.	Y	Y	Y		Y	Recent Wikipedia dump to extract language links and category hierarchy information.
(H-3)	0	Wikipedia category	Using Wikipedia categories information of the page for classification. w/o bigrams exp.	Y	Y	Y	Y	Y	Recent Wikipedia dump for extracting category relationships.

Datasets:

(1) Training Data (2) Wikipedia dump (3) Wikipedia classification (4) Language links (5) ENE definition