拡張固有表現に分類された31言語のWikipedia知識ベース

関根聡[1] 中山功太[1] 野本昌子[1] 安藤まや[2]
隅田飛鳥[1] 松田耕史[1]
[1]理化学研究所　革新知能統合研究センター [2]フリー
{satoshi.sekine, kouta.nakayama, masako.nomoto, asuka.sumida,koji.matsuda}@riken.jp
maya@cis.twcu.ac.jp

概要

説明可能なAI構築を目指し、Wikipediaを構造化することを目標とした「森羅プロジェクト」を推進している。2020、2021年には30言語のWikipediaを拡張固有表現に分類するタスクを実施し、2年間で14チームから24の参加システムを得た。その出力データを用いてアンサンブル学習を行い、大規模な固有表現の分類リソースを構築した。日本語と合わせて31言語で約3340万ページが分類されており、人手チェックも行った日本語の精度は98.5%、人手チェックのないその他の30言語の平均F値は86.9である。知識リソースは森羅ホームページで公開している。「協働による知識構築(Resource by Collaborative Contribution - RbCC)」と呼ぶ枠組みで、構造化知識を構築している「森羅プロジェクト」では、2022年も新たなタスクを実施する。

1 はじめに

Wikipediaは常時更新されている巨大な知識源であり、様々な言語処理技術への応用が期待されているが、構造化されていないため自然言語処理での利用が難しい。Wikipediaを分類し、構造化して知識グラフにすることで、幅広い活用が期待される。同様の目的で構造化された知識としては、DBpedia、YAGO、Wikidataなどがあるが、分類カテゴリーや構造フレームは主にクラウドによるボトムアップ方式で構築され、統一されたルールに基づいておらず、非常に多くのノイズが入っている。筆者らは、分類や構造フレームの設計はトップダウンに画一した形で行い、内容（コンテンツ）の構築をクラウドで行うことにより、言語処理にも利用しやすい知識が構築できると考えている。そこで、分類基準と構造フレームに幅広い固有表現カテゴリーを対象に定義されている「拡張固有表現」を利用し、そこにWikipediaに記載されている知識をはめ込むことにより、整理された構造化知識を構築することを目指している。これが「森羅プロジェクト」の目標である。プロジェクトでは、共有タスクを実施し、その複数の参加システムの出力に対してアンサンブル学習などを通じ精度の高い知識を自動的に構築する枠組みとして提唱している「協働による知識構築(Resource by Collaborative Contribution - RbCC)」を用いて知識構築を行っている。本論文では、そのプロジェクトのうち、2020、2021年に実施した、30言語のWikipedia分類タスクを通して構築したものに、日本語の分類データも含めた31言語のWikipediaを拡張固有表現に分類した知識リソースを紹介する。

2 拡張固有表現

拡張固有表現は、幅広い固有表現を対象とした固有表現オントロジーであり、今回利用したVersion 8.1では最大4階層で、末端カテゴリー数は219ある（Sekine et al. 2002, ENE-HP）。一般的な固有表現カテゴリー(Nadeau 2007)とされる組織名や地名の下位概念である企業名、政党名、市区町村名、湖沼名だけではなく、イベント名、製品名、自然物名などの新規のカテゴリーもあり、Wikipediaの項目を対象に拡充している。拡張固有表現のホームページに詳細の定義があるが、付録に記載した図2に全体像を示す。また、Wikipediaを対象にしているため、2つの特異なカテゴリーを設定している。一つは主に一般名詞などを対象にした"CONCEPT"というカテゴリーであり、もう一つは「転送ページ」「リストページ」「メタ情報ページ」などWikipedia特有のページを対象とした"IGNORED"である。拡張固有表現の多くのカテゴリーには構造化のために属性も定義され、構造化の際に利用する。

3 日本語分類データ

日本語 Wikipedia 分類データは 2019 年 1 月の Wikipedia ダンプを対象に作成した。被リンク数が 5 以下の 15 万ページ、シンプルな方法で除外した固有表現ではない 5 万ページを除いた 92 万ページから構成されている。3 万ページ程度の学習データを元に機械学習を用いた分類結果を人手ですべてチェックした（Suzuki et al. 2018）。サンプルデータを用いて分析した結果、精度は 98.5%程度であると推定されている。残りの 1.5%は、人が見ても曖昧なページなどであった。表 1 に頻度の高い上位 10 個のカテゴリーの情報を載せる。映画化された小説のページに両方の記載がある場合などには、1 つのページが複数のカテゴリーに属することとしており、その割合は 3%程度である。

表１．頻度の高いカテゴリ

カテゴリ	頻度	カテゴリ	頻度
人名	269,688	学校名	25,579
市区町村名	49,028	文学名	21,093
音楽名	46,889	映画名	19,381
番組名	33,747	鉄道駅名	18,296
企業名	30,120	競技会名	17,471

4 共有タスク

30 言語の 2019 年 1 月の Wikipedia ダンプを対象に、拡張固有表現に分類するタスクを実施した（関根ら 2021）。30 言語は、言語ごとの“active user”の数が多いものから選んだ。本タスクの教師データは、3 節で紹介した日本語の分類データと Wikipedia の言語間リンクから自動的に各言語に対して作成したものを利用した。したがって、完全なデータではなく、いわゆる“silver data”である。例えば、226 万ページあるドイツ語の場合には、日本語からの言語間リンクがあるページは 27 万ページあり、それを教師データとして、残りの 199 万ページの分類を行うことが本タスクとなる。（ただし、参加者にはアンサンブル学習での利用を考慮し、教師データ部分の分類も学習システムを使って実施することを求めている）参加チームは、2021 年は 10 チーム、2022 年は 3 チームであった。それぞれのシステムに関する論文は、森羅ホームページより閲覧できる（SHINRA-HP）。全ページのデータの提出を促すこと、将来のシステム評価との比較が公平にできることを目的として、評価の対象ページがどのページであるかは参加者には公表していない。評価結果を表 2 に示す。それぞれの言語での最高のシステムの F 値、アンサンブル学習システム F 値、最低 1 つのシステムに正解が含まれる割合を示している。

表２．タスクの評価結果

言語	最高システムのF値	アンサンブルシステムF値	正解が含まれる割合
Arabic	90.06	92.18	97.71
Bulgarian	86.94	88.32	92.77
Catalan	89.25	86.62	95.42
Czech	81.70	83.72	94.34
Danish	82.34	81.53	92.15
German	79.68	80.93	89.68
Greek	84.85	79.72	90.04
English	86.49	87.65	93.49
Spanish	85.54	86.51	94.69
Persian	89.63	90.87	94.63
Finish	85.95	86.36	95.47
French	83.77	87.20	92.83
Hebrew	81.80	81.74	90.74
Hindi	87.81	90.76	94.79
Hungarian	89.93	91.41	96.19
Indonesian	90.71	92.22	97.28
Italian	84.08	85.51	91.85
Korean	80.08	82.57	90.68
Dutch	85.88	85.88	91.51
Norwegian	85.89	86.10	93.53
Polish	84.44	85.06	94.00
Portuguese	88.96	89.83	96.11
Romanian	93.43	93.43	98.07
Russian	81.62	84.06	90.91
Swedish	84.28	86.28	91.85
Thai	84.72	85.48	95.31
Turkish	87.85	88.13	94.18
Ukrainian	85.14	84.53	91.20
Vietnamese	90.28	89.58	95.14
Chinese	88.00	88.72	95.10
平均	**86.04**	**86.76**	**96.75**

アンサンブル学習はシンプルな Voting により行なっている。ただし、1 参加者が 3 システムまで提出することができ、多くの場合にはそれらの出力は非常に似ているため、参加者ごとの票が 1 になるように Voting を補正している。表 2 の赤字は、アンサンブル学習システムが、最高性能のシステムを上回っていることを示す。多くの言語で上回っていることがわかる。また、上限はほとんどの言語で 90%を超え、平均は 97%と非常に高いことがわかる。上限に近いシステムを開発することは今後の課題である。

5 リソースのサンプルと統計情報

図 1 にサンプルデータを載せる。“pageid”と“title”で Wikipedia の該当ページを示し、ENEs でそのページの拡張固有表現カテゴリーを ID と名前で示している。言語ごとに 1 ファイルで構成され、31 言語分の 31 ファイルがある。

{"pageid": "22059861", "title": "Tarlach Rua Mac Dónaill", "ENEs": [{"ENE_id": "1.1", "ENE_name": "Person"}]}
{"pageid": "53177250", "title": "90th Scripps National Spelling Bee", "ENEs": [{"ENE_id": "1.9.1.3", "ENE_name": "Competition"}]}
{"pageid": "5724950", "title": "Hatley, Quebec (township)", "ENEs": [{"ENE_id": "1.5.1.1", "ENE_name": "City"}]}

図１．サンプルデータ（英語）

表 3 に、言語毎の分類結果の頻度を示す。名前(1)の下位カテゴリーは、その次の階層のカテゴリーでまとめている。例えば 1.4 の組織名には、国際組織名(1.4.1)、企業名(1.4.6.2)、政党名(1.4.7.2)などの組織名の下位にあるカテゴリーの頻度の総和が示されている。頻度の高いカテゴリーは言語共通であり、全体では“地名” (23.91%), “人名” (23.16%), “自然物名” (13.52%), “製品名” (13.65%), “施設名” (6.14%), “組織名”(5.58%)が頻度の高いカテゴリーである。

表 2 で示した通り、日本語を除いた 30 言語のリソースの F 値は 86.76 であり、完璧ではない。どのカテゴリーがどのカテゴリーに間違っているかを分析した。付録の表 4 に混同行列の形で、正解とアンサンブル学習システムの出力の頻度を示す（30 言語の合計を示している）。この階層レベルで見た時の全体の誤りの 85%は“CONCEPT”か“IGNORED”に関連している。特に、正解が“IGNORED”である時に、様々なカテゴリーが付いている場合と、正解が様々なカテゴリーである場合に“CONCEPT”のカテゴリーとなっている場合が目立つ。これらの誤り以外では、例えば、“組織名”と “地名”と“施設名”の 3 つのカテゴリーの混同がある。施設名は、公園、学校、空港、道路など、組織名や地名に関係するものが対象となっており、このような混同はある程度理解できる。システムは“CONCEPT”と“IGNORED”以外は理解できる範囲の結果を出力し、信頼できる知識になっていると考えている。“CONCEPT”と“IGNORED”の問題については今後の課題として取り組む考えである。

6 関連研究

知識オントロジーは自然言語処理にとって重要なリソースであると考えられてきた。例えば、1980 代の人工知能ブームでは日本では EDR(日本電子化辞書 1995)、米国では Cyc(Lenat 1995)という大きな国家プロジェクトで知識構築が行われたが、構築コスト、メンテナンス、カバレージなどの問題（Knowledge Acquisition Bottleneck）に直面し、当初の目的を達成したとは言い難い。

不特定多数の協働によって百科事典知識を構築する Wikipedia プロジェクトによって上記の問題に対するパラダイムは大きく変化した。ただし、この知識は人間が読むために作られたものであり、計算機利用が容易にできるものではない。そこで、計算機利用な知識構築を目指した取り組みが行われている。DBpedia は Wikipedia 内の infobox など構造化されている部分を利用し、構造化知識を構築している(Lehmann et al. 2015)。YAGO は Wikipedia と WordNet を結びつけることにより利用しやすいオントロジーを作ろうというプロジェクトである(Mahdisoltani et al. 2015)。Wikidata は Freebase などの経験を生かし、既に存在する構造化知識とクラウドを活用し、構造化知識を構築していこうというプロジェクトである(Vrande cic and Krotzsch 2014)。ただ、上記のどの知識も、カテゴリーや構造的な知識部分にノイズが多くある Wikipedia に頼っていたり、カテゴリーや構造フレームの部分もクラウドによるボトムアップな方法で構築しているため、多くのノイズが存在し、計算機による利用が難しいのが現状である。

7 協働による知識構築

本リソースを構築した森羅プロジェクトの大きな特徴の一つは「協働による知識構築(Resource by Collaborative Contribution - RbCC)」である(関根ら 2018)。共有タスクによる言語処理技術の向上が幅広く行われている。そのおかげで様々な技術が開発され精度向上が達成されてきた。しかし、多くの場合には、システムの最適化（Optimization）の競争に過ぎない形でのプロジェクトとなり、タスクが終了すると開発したシステムも放置され、その結果が何らかの貢献につながらないことが多い。この状況は次に挙げる３つの改良によって改善すると考えている。

1. タスクを知識やリソースの構築などにつながる形で設計する
2. システムの出力結果は広く公開し、アンサンブル学習をはじめとした研究につなげる
3. タスクの出力結果をある種の学習データとして用いたシステムを開発できるように、再度タスクに参加できるような枠組みとする

「森羅プロジェクト」のタスクは、多言語の分類タスクに限らず 2018 年から日本語の属性値抽出タスク、リンク同定タスクを含め 6 回実施した。2022 年には、日本語で分類、属性値抽出、リンク同定の３つのタスクを同時に行い、構造化知識を構築するタスクを実施する予定である。これまで作成した全ての教師データと、構造化した 2019 年の Wikipedia 知識を教師データとして利用し、2021 年の Wikipedia を対象にする予定であり、これまで大きな課題であった知識メンテナンスの問題解決も目指している。興味ある方の参加を期待している。

7 まとめ

協働による知識構築（RbCC）の枠組みで構築した 31 言語の Wikipedia を拡張固有表現で分類した知識リソースについて報告した。日本語は 92 万ページを 98.5%の精度で、他の 30 言語は約 3250 万ページを 87 の F 値で分類されている。データは森羅ホームページで公開している。

表３.各言語のカテゴリーごとの頻度

ID	0	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	1.1	1.11	1.12	2	3	9
Category	Concept	Name Other	Person	God	Individual Animal	Organization	Location	Facility	Product	Virtual Address	Event	Natural Object	Disease	Color	Timex	Numex	IGNORED
ja	51039	5	269688	1278	3588	76345	89085	125232	250868	3645	32152	22312	2258	213	6087	347	13360
ar	52972	0	204207	513	3	26264	198716	21437	66559	2093	20277	38404	4937	97	7733	480	16849
bg	17533	0	73267	588	7	13016	53001	11304	32931	746	6704	31913	679	20	3651	51	4358
ca	51271	0	149258	946	7	30121	155361	41240	67885	1316	20958	64479	1652	84	4156	540	12471
cs	50169	0	116294	796	34	28476	74478	32128	68588	1515	18266	17454	1396	37	4982	1	15654
da	27238	1	73171	481	12	19399	33795	16822	45794	989	8853	7121	758	31	3764	786	3577
de	196031	0	768225	2837	279	192658	396296	192032	270062	4767	82408	64601	6401	84	10350	0	76738
el	74786	0	67372	1342	2	28297	48940	5861	53778	1339	13157	5978	1070	85	8462	1	25595
en	315531	0	1728797	4423	5822	432236	875918	472005	1005969	43520	279123	403739	11813	334	8829	811	204327
es	105543	2	376710	2265	89	90700	317901	92715	246103	4363	72964	153631	4378	187	8479	226	38106
fa	48311	0	138148	763	16	19230	241298	46761	100016	1354	9146	36456	2171	79	8104	711	8759
fi	42147	0	147556	784	48	34477	48877	21505	99635	1400	15659	23084	1599	31	4022	1	10071
fr	129098	0	595708	2393	573	135080	416779	149286	355881	6334	87986	119710	4722	254	11182	630	60099
he	31475	1	78479	484	15	16658	20233	10751	47355	859	8045	8761	1467	39	5501	0	7122
hi	14388	0	22242	406	2	5113	46309	6144	20433	848	3150	3508	643	47	5728	522	2693
hu	30551	1	110023	650	12	20570	137258	25902	60441	1266	18818	22436	688	18	4670	0	9952
id	26194	0	72630	669	22	21450	115722	20662	69492	1818	11991	98328	1050	32	3734	190	9556
it	98522	0	378736	1962	93	92536	313896	79661	325760	2873	109402	48390	4017	247	9276	7	32724
ko	52713	0	111535	761	26	32660	46817	54853	88858	2155	15243	14963	1325	70	6036	381	11935
nl	120574	0	217528	1349	30	68301	342682	88123	148271	2128	47810	874290	2611	68	5067	88	36880
no	43235	0	158039	696	16	39756	73512	45573	71868	1645	25076	23477	1084	33	3988	21	13680
pl	84346	0	354727	1805	55	79672	386271	93985	173646	2469	52842	51219	4422	49	4921	3	26274
pt	80353	1	224466	1520	52	65736	240881	43866	181481	4039	48010	90099	3473	94	5829	536	24859
ro	22821	0	57913	582	6	13518	201132	11420	33084	1020	5905	30486	657	20	3640	1	9209
ru	102057	1	463956	2005	57	98493	409426	76356	207009	2795	57119	53424	3288	96	7241	388	40425
sv	206567	0	231159	1301	288	53538	1729196	88261	119368	1459	23894	1281506	2160	74	4332	4	16006
th	14726	0	28041	378	2	11195	10381	8738	25335	905	6139	6944	833	32	3843	615	11570
tr	26188	0	83506	859	112	18642	83704	12515	59418	1755	15271	7721	1149	88	5139	237	9344
uk	67899	0	178244	1381	8	39675	348825	45919	101043	1759	21852	47983	1734	67	5596	2	20769
vi	46134	0	55154	430	20	9533	245542	10573	38259	882	8636	773560	966	56	4495	5	6613
zh	61095	0	217583	1633	267	56013	301227	103105	133350	3371	28580	100343	1985	114	7484	249	25499
Total	2230412	12	7534779	36647	11296	1813345	7702232	1951630	4435190	104056	1146856	4425977	75401	2666	172750	7238	779575
%	6.84%	0.00%	23.16%	0.11%	0.03%	5.58%	23.91%	6.14%	13.65%	0.32%	3.51%	13.52%	0.23%	0.01%	0.56%	0.02%	2.40%

謝辞

本研究はJSPS科研費 JP20269633の助成を受けたものです。

参考文献

SHINRA-HP. Shinra project homepage: https://shinra-project.info.

SHINRA2020-ML-HP. Shinra 2020-ml homepage: http://shinra-project.info/shinra2020ml/.

ENE-HP. Extended named entity homepage: https://ene-project.info.

関根聡, 野本昌子, 中山功太, 隅田飛鳥, 松田耕史, 安藤まや (2021). SHINRA2020-ML:30 言語のWikipedia ページの分類, 言語処理学会第27回年次大会発表論文集

関根聡, 小林暁雄, 安藤まや (2019). Wikipedia構造化プロジェクト「森羅2018」,言語処理学会第25回年次大会発表論文集, pp.69-72, 2019.

日本電子化辞書研究所 (1995). EDR電子化辞書シンポジウム

Tushar Abhishek, Ayush Agarwal, Anubhav Sharma, Vasudeva Varma, and Manish Gupta (2020). Rehoboam at the ntcir-15 shinra2020-ml task. In The 15th NTCIR Conference Evaluation of Information Access Technologies (NTCIR-15).

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Soren Auer, and Christian Bizer (2015). Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web, 6(2):167-195.

Douglas Lenat (1995). Cyc: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11):33-38.

pages 94-100. AAAI Press.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian M Suchanek (2015). Yago3: A knowledge base from multilingual wikipedias. CIDR.

David Nadeau, Satoshi Sekine (2007). A survey of Named Entity Recognition and Classification". Lingvisticae Investigationes 30 (1), 3-26.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata (2002). Extended named entity hierarchy. In the Third International Conference on Language Resources and Evaluation (LREC'02).

Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoki Okazaki, and Kentaro Inui. (2018) A joint neural model for fine-grained named entity classification of Wikipedia articles. IEICE Transactions on Information and Systems, E101.D(1):73-81.

Denny Vrande cic and Markus Krotzsch (2014). Wikidata: A free collaborative knowledgebase. Commun. ACM, 57(10):78-85.

A 付録

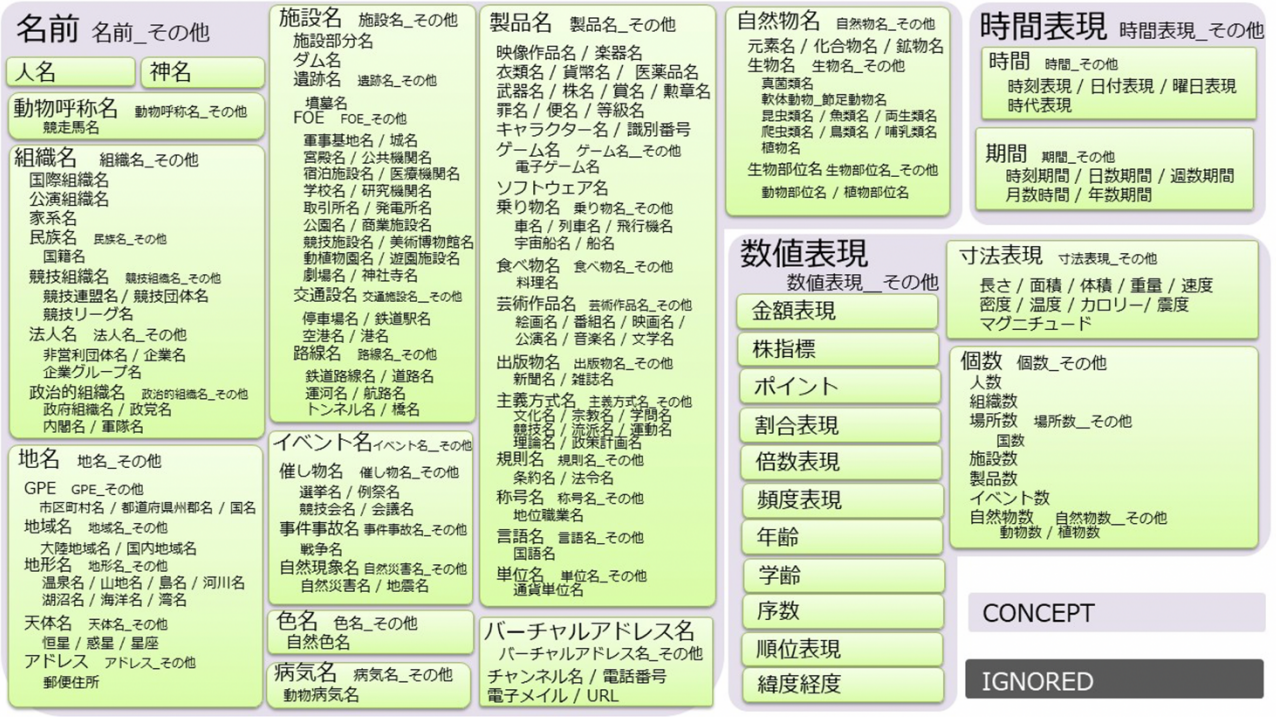

図２．拡張固有表現階層

表４．正解とアンサンブルシステム出力の混同行列
（縦軸が正解カテゴリー、横軸がアンサンブルシステムの出力カテゴリー）

ID	0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	1.1	1.11	2	3	9
Category	Concept	Person	God	Individual Animal	Organization	Location	Facility	Product	Virtual Address	Event	Natural Object	Disease	Timex	Numex	IGNORED
0:Concept	587	-	-	-	1	2	-	24	-	2	65	1	-	-	1
1.1:Person	9	3675	-	-	4	-	-	5	-	-	-	-	-	-	3
1.2:God	-	-	14	-	-	-	-	-	-	-	-	-	-	-	-
1.3:Individual_Animal	-	1	-	-	-	-	-	1	-	-	1	-	-	-	-
1.4:Organization	15	6	-	-	760	20	12	5	-	3	-	-	-	-	3
1.5:Location	11	-	-	-	4	3371	20	1	-	-	1	-	-	-	2
1.6:Facility	12	3	-	-	11	26	906	6	-	1	1	-	-	-	3
1.7:Product	103	12	1	-	13	1	10	1824	-	9	3	-	1	-	4
1.8:Virtual_Address	1	-	-	-	5	-	-	1	28	-	-	-	-	-	-
1.9:Event	5	1	-	-	7	-	1	11	-	416	-	-	2	-	8
1.10:Natural_Object	2	-	-	-	-	-	-	1	-	-	91	-	-	-	-
1.11:Disease	-	-	-	-	-	-	-	-	-	-	-	23	-	-	-
2:Timex	-	-	-	-	-	-	-	1	-	3	-	-	46	-	-
3:Nume	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1
9:IGNORED	287	171	2	-	73	181	33	138	1	45	9	-	18	2	398