森羅2022最終報告会
分類タスク

東京農工大学
2023年1月18日
Ｄ３　秋山　賢二

報告内容

１．採用手法１　擬似ラベリング

２．採用手法２　データクレンジング

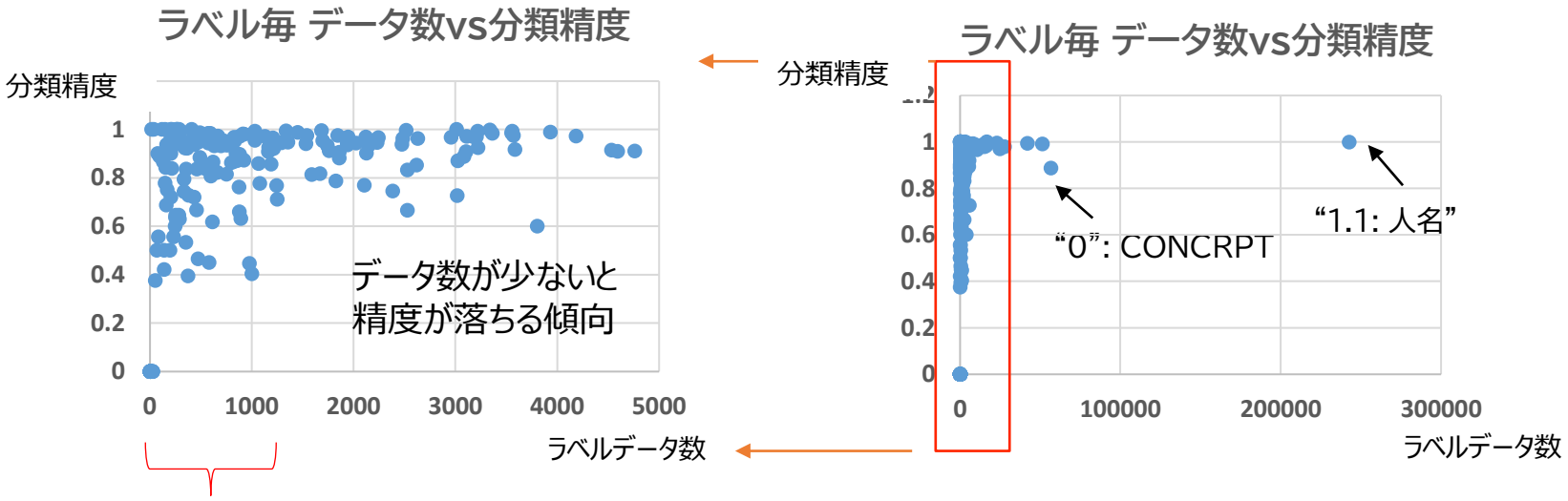
３．性能評価

４．クレンジングされたデータの確認

＜実験条件＞
MODEL_NAME = ”cl-tohoku/bert-base-japanese“
MAX_LENGTH = 128
optimizer: Adam
LEARNING_RATE = 5e-5
EPOCH = 5
BATCH_SIZE = 32
EVAL_BATCH_SIZE = 64

1．採用手法１－　擬似ラベリング　（Pseudo Labeling）

半教師あり学習の擬似ラベリング手法を適用

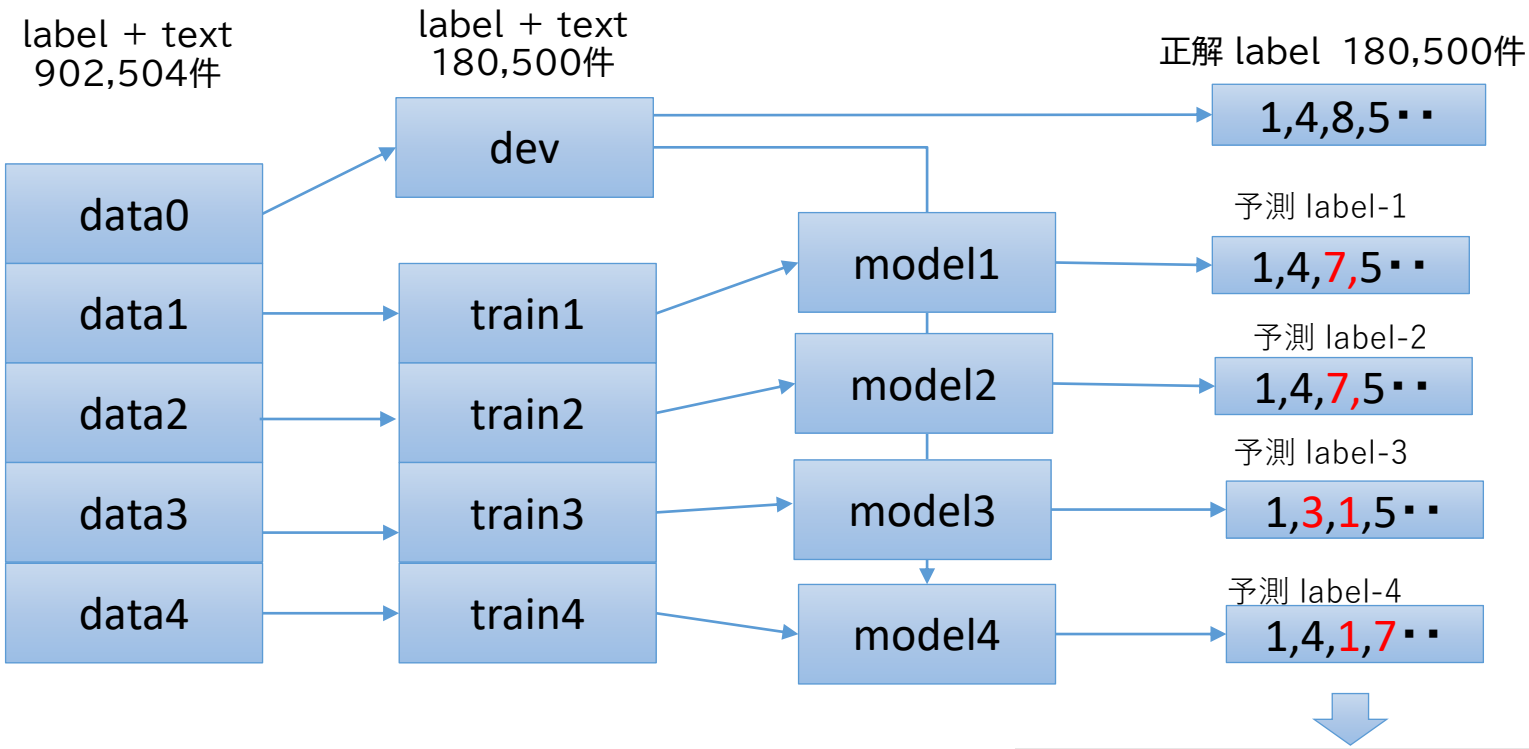


データ数がNに満たないラベルに対して
不足を補う擬似ラベルデータを付与

Nを変えて精度の変化を確認
- N=200　95.0249
- N=500　**96.0554**
- N=1000　95.8670

擬似ラベルデータ生成：
Wikipedia 200万件(森羅ラベル付きデータ以外）に対し、
Baseline手法でラベルを予測し、
logitsの高いものを擬似ラベルとして選択

2．採用手法２－ データクレンジング （Data Cleansing)



data0の3番目のデータは全て不正解
→ 不正解データとしてマーク

同様のやり方でdata1～data4の中の
不正解データをマーク

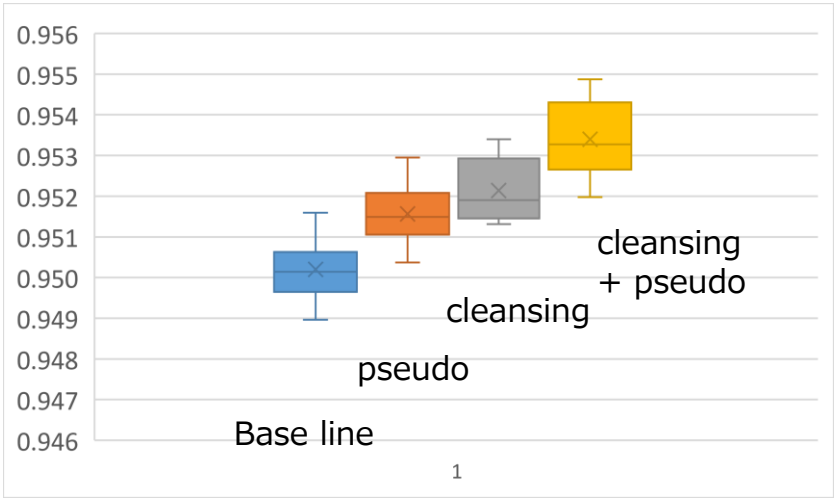
全データ(902,504件)から不正解データとしてマークした17097件(1.9%)を
削除したものを訓練データとし、評価データからは残して評価

関連研究　Zihan Wang et.al.（ACL2019）
CrossWeigh: Training Named Entity Tagger from Imperfect Annotations
訓練セットを分割してモデルを作成し、別の分割でテストをして間違えたラベルには重み付をする

3. 性能評価

全 902,500件 を10分割したデータで訓練と評価を実施
平均値とBase Lineのデータに対する平均値の違いをT検定で確認

Pattern	Base line	pseudo	cleansing	cleansing+ pseudo
train	data0	data0	data0	data0
dev_data0	-		-	-
dev_data1	94.971	95.120	95.224	95.487
dev_data2	95.044	95.168	95.190	95.262
dev_data3	95.011	95.135	95.340	95.327
dev_data4	94.960	95.149	95.166	95.309
dev_data5	95.073	95.091	95.132	95.269
dev_data6	95.159	95.295	95.329	95.466
dev_data7	95.053	95.165	95.158	95.348
dev_data8	95.014	95.248	95.257	95.395
dev_data9	94.896	95.037	95.131	95.198
平均	**95.020**	**95.156**	**95.214**	**95.340**
T検定 pvalue		0.000799	0.000034	0.000000



擬似ラベリング、データクレンジング、及び組合せ手法に関して性能向上を確認

■投稿データの精度

森羅BERTベースライン	94.2519%
森羅RoBERTaベースライン	95.2224 %
BERT Cleansing （ID291）	95.4678%
BERT Cleansing + Pseudo （ID293）	95.4678%

性能向上が見られず？

4. クレンジングされたデータの確認

Labelの正誤判定　　分類器の予測

pid	Label	判定	Pred0	Pred1	Pred2	Pred3	Text
685268	0	○	1.5.0	1.5.0	1.5.0	1.5.0	**七辻。**七辻（ななつじ）は、7本の道路が交わる交差点。七差路・七叉路（ななさろ）とも呼ばれる。 日本国内では、東京都大田区の七辻が有名。この交差点は、大田区の南蒲
1488622	0	×	1.5.3.2	1.5.3.2	1.5.3.2	1.5.3.2	**七北田丘陵。**七北田丘陵（ななきたきゅうりょう）は、宮城県中南部において、奥羽山脈から東に延びる舌状台地であり、松島丘陵（狭義）と平行に、その南側で東西に横たわる
600438	0	×	1.6.3.1	1.6.3.1	1.6.3.1	1.6.3.1	**カリバタ英雄墓地。**カリバタ英雄墓地（カリバタえいゆうぼち）は、インドネシアの南ジャカルタ地区にある国立追悼施設である。現地の名称はTaman Makam Pah
377693	0	×	1.6.4.18	1.6.4.18	1.6.4.18	1.6.0	**安国寺利生塔。**安国寺（あんこくじ）と利生塔（りしょうとう）は、南北朝時代に足利尊氏、直義兄弟が、北海道、沖縄を除く日本各地に設けた寺院と仏塔。 臨済宗の夢窓疎石
847329	0	○	1.6.4.18	1.6.4.18	1.6.4.18	1.6.4.18	**子守神社。**子守神社（こもりじんじゃ、こまもりじんじゃ）は、日本各地に存在する神社である。名前の通り、子供の守護神としての「子守神」を祀る神社であり、氏神とされた
31865	0	×	1.6.4.18	1.6.4.18	1.6.4.18	1.6.4.18	**東光寺。**東光寺（とうこうじ）は、仏教寺院の名称。「東光」は東方浄瑠璃世界（中国語版）（浄琉璃世界、琉璃世界、琉璃光世界、東方浄土）を意味し、薬師如来を本尊とする
2589516	0	×	1.6.4.6	1.6		6	**チェンマイラーム病院。**チェンマイラーム病院（英語:Chiangmai Ram Hospital、タイ語：โรงพยาบาลเชียงใหม่ ราม）は、タイ北
119108	0	×	1.6.4.7	1.6.4.7	1.6.4.7	1.6.4.7	**天真楼。**天真楼（てんしんろう）は、杉田玄白の医学・蘭学塾。天真楼塾ともいう。 天真楼が開かれた時期は未詳。 杉田玄白が天真楼という塾を開いていたことは確かである
1820992	0	○	1.6.4.7	1.6.4.7	1.6.4.7	1.6.4.7	**大学校**（1869年）。大学校（だいがっこう）は、明治2年7月（1869年8月）、明治新政府により東京に設立された官立教育機関群、もしくは教育行政官庁。この記事
836876	0	×	1.7.0	1.7.0	1.7.0	1.7.0	**愚者。**愚者（ぐしゃ、英：The Fool, 仏：Le Mat）は、タロットの大アルカナに属するカードの1枚。英語ではThe Jesterと呼ばれることもある。
1542171	0	×	1.7.0	1.7.0	1.7.16.0	1.7.0	**セボン。**セボンとは、アース製薬が製造・販売する水洗トイレ用洗浄芳香剤である。 これまで水洗トイレ用洗浄芳香剤は小林製薬のブルーレットくらいであったが、ブルーレッ
3031049	0	×	1.7.0	1.7.0	1.7.19.5	1.7.19.8	**メフォ手形。**メフォ手形（メフォてがた、ドイツ語: Mefo-Wechsel）とは、ナチス・ドイツ時代のドイツにおいて、軍事費調達のために創出された割引手形。19
2750744	1.5.0	×	1.6.5.5	1.6.5.5	1.6.5.5	1.6.5.5	**三宅ジャンクション。**三宅ジャンクション（みやけじゃんくしょん）は、大阪府松原市にある阪神高速道路6号大和川線と14号松原線とのハーフジャンクションである。松原線
369158	1.5.0	○	1.5.1.1	1.5.1.1	1.5.1.1	1.5.1.1	**綱島。**日本 > 神奈川県 > 横浜市 > 港北区 > 綱島 綱島（つなしま）は、神奈川県横浜市港北区の地名。現行行政地名は綱島上町、綱島台、綱島西一丁目から綱島
21687	2.1.3	×	0	0	0	0	曜日。曜日（ようび）とは、七曜（7つの天体）が守護するとされる日のことをいい、曜日が循環する7日の組の事を週と呼ぶ。日本語では現在でも各曜日を日曜日、月曜日、火
11902	2.1.3	○	2.1.2	1.7.13.2	2.1.2	2.1.2	**火曜日。**火曜日（かようび）は、月曜日と水曜日の間にある週の1日。 日本語や朝鮮語、また、ロマンス諸語の名称は、七曜の1つである火星 (Mars) の日にちなむ。
11905	2.1.3	○	2.1.2	1.7.13.2	2.1.2	2.1.2	**金曜日。**金曜日（きんようび）は、木曜日と土曜日の間にある週の1日。 日本語や朝鮮語の名称は、七曜の1つである 金星 (Venus) の日にちなむ。五行思想のは、

0：CONCEPT　一般名称

1.5　地域名

1.6　施設名

1.7　プロダクト名

2.1.3　曜日表現

2.1.2　日付表現

個人判定では482件中　正：239　誤：243　（およそ半数は違えたラベルが付け）
⇒誤ったラベルを削除することで精度が上がった可能性あり

END