Thai Name Entity Recognition Benchmark on LST20 corpus

Thammakorn Kobkuachaiyapong

Literature review

Even though Thai NER is challenging due to its characteristic of the language, It still has some research papers on this topic. Firstly Thattinaphanich, Suphanut & Prom-On, Santitham. (2019)[1] use CRF as the system with two different segmentation words and syllables and use dictionary on specific word (i.e. “แห่ง”), unigram and bigram as features on BEST2009 corpus with three name entity tags(Person, Organization and Location). Their overall f1 score is 80.39% and 80.80 % respectively. Udomcharoenchaikit, Can & Vateekul, Peerapon & Boonkwan, Prachya. (2019)[2] their paper compare between CRF, LSTM, BLSTM, BLSTM-CRF and V-BLSTM-CRF on BEST2010 corpus without POS as a feature. Their overall f1 result is 54.9, 58.2, 62.1, 62.7 and 63.9 respectively. S. Thattinaphanich and S. Prom-on[3] use Bi-LSTM-CRF then apply word embedding from ULMFit on PythaiNLP dataset project which expands from BEST2009 with 13 different tags. Their f1 score is 86.9

This benchmarking will perform 4 models on LST20 corpus including Conditional Random Fields for traditional method, Long Short-Term Memory for neural network method and addition Bidirectional Long Short-Term Memory and Bidirectional Long Short-Term Memory with Conditional Random Fields. This benchmark will use macro f1 score as metrics due to large amount of tag “O” which will not affect the macro score while expected scores are 55, 58, 60, 62 based on Udomcharoenchaikit, Can & Vateekul, Peerapon & Boonkwan, Prachya. (2019)[2] result as shown in Table i.

Tabel I

	CRF	LSTM	Bi-LSTM	Bi-LSTM-CRF
Macro f1	55.0	58.00	60.00	62.00

Models

Every models use segmentation, label and split data from the LST20 dataset

Traditional model

Conditional Random Fields(CRF)

For features CRF use word, boolean isDigit, boolean isSpace, word size, boolean next word and last word isSpace, begin of sentence, end of sentence. The performance of CRF is represented with an overall f1 score = 43.60 while the Table II below shows precision, recall and f1 on respective tags. The traditional CRF used 29min 4s for training.

Table II

tag_name	precision	recall	f_score
B_BRN	36.84	14.89	21.21
E_BRN	25	12.5	16.67
I_BRN	33.33	20	25
B_DES	89.38	79.42	84.11
E_DES	67.11	51.52	58.29
I_DES	73.43	51.47	60.52
B_DTM	82.39	69	75.1
E_DTM	81.58	76.19	78.8
I_DTM	90.78	83.93	87.22
B_LOC	75.31	62.71	68.44
E_LOC	64.88	68.39	66.59
I_LOC	41.8	34.2	37.62
B_MEA	72	57.53	63.95
E_MEA	58.06	75.06	65.48
I_MEA	51.27	80	62.49
B_NUM	54.32	50.73	52.46
E_NUM	72.06	62.03	66.67
I_NUM	77.19	76.52	76.86
B_ORG	78.22	64.92	70.95
E_ORG	76.27	69.19	72.55
I_ORG	71.24	67.69	69.42
B_PER	90.8	81.55	85.93
E_PER	87.48	91.61	89.5
I_PER	86.51	90.77	88.59
B_TRM	90.48	74.22	81.55
E_TRM	40	16.67	23.53
I_TRM	25	13.33	17.39
B_TTL	95.4	97.82	96.6
E_TTL	0	0	0
I_TTL	0	0	0
Micro	0	77.56	72.27
Macro	0	47.59	43.6

Neural network model

For afterward neural networks model, this benchmark uses word as a feature via map distinct word to it index and use categorical cross entropy as loss function

Long Short-Term Memory(LSTM)

For LSTM model this benchmark uses macro f1 score as discussed above which is 56.78 and the Table III represents precision, recall and f1 score on respective tags while the table IV calculates and represent micro and macro f1 score with and without the “O” tag(note that on macro score with or without O tag does not have significant change as micro score has.). This model consumes 18min 39s for training.

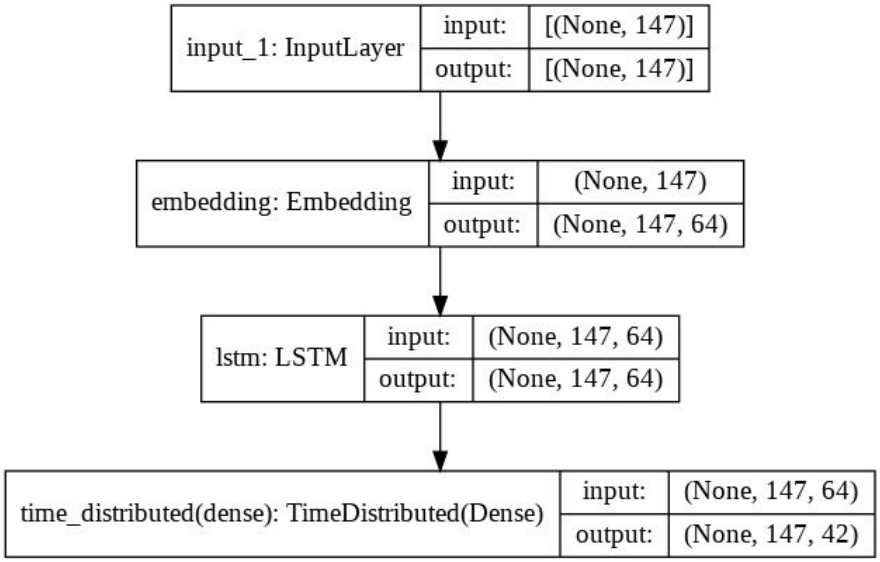

Table III: Score for LSTM on respective tags.

tag_name	precision	recall	f_score
B_BRN	38.88888889	16.27906977	22.95081967
I_BRN	-	0	-
E_BRN	-	0	-
B_DES	81.37687555	83.81818182	82.57948948
I_DES	85.13513514	37.27810651	51.85185185
E_DES	64.5320197	73.18435754	68.58638743
B_DTM	72.30769231	68.66883117	70.44129892
I_DTM	63.52331606	48.96166134	55.29995489
E_DTM	76.30057803	71.66123779	73.90817469
B_LOC	68.48691695	64.07663651	66.20841353
I_LOC	41	31.17870722	35.42116631
E_LOC	65.75562701	58.59598854	61.96969697
B_MEA	56.53198653	59.45467422	57.95650673
I_MEA	60.57692308	66.54929577	63.42281879
E_MEA	57.52330226	70.58823529	63.3895818
B_NUM	39.5480226	29.83802217	34.01360544
I_NUM	97.2972973	81.81818182	88.88888889
E_NUM	42.85714286	9.677419355	15.78947368
B_ORG	75.17580872	59.75405254	66.58361881
I_ORG	74.36882547	60.27580071	66.58476658
E_ORG	72.91414752	67.03724291	69.85230235
B_PER	86.76470588	85.39212386	86.07294317
I_PER	23.07692308	13.95348837	17.39130435
E_PER	53.85858586	88.68928809	67.01860231
B_TRM	65.95744681	27.67857143	38.99371069
I_TRM	6.25	20	9.523809524
E_TRM	100	9.090909091	16.66666667
B_TTL	90.39426523	95.24169184	92.75468922
I_TTL	-	0	-
E_TTL	-	0	-
O	96.40016107	97.36373578	96.87955253

Table IV: F1 score of LSTM

Metrics	precision	recall	f_score
Macro	65.04501519	56.78447002	60.63469297
MacroWithoutO	64.87324404	56.57947216	60.44317525
Micro	91.41879768	91.41879768	91.41879768
MicroWithoutO	89.56037536	89.23091762	89.39534295

Bidirectional Long Short-Term Memory(Bi LSTM)

As same as LSTM this model changes the unidirectional to bidirectional which affects the runtime and performance of the model as the micro f1 score is higher to 58.28 and runtime up to 23min 49s.

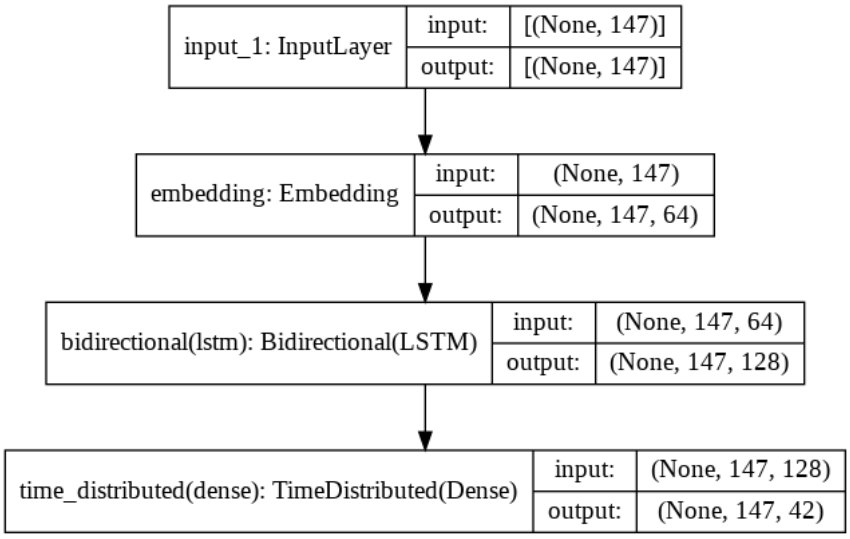

Table V: Score of Bi-LSTM on respective tags.

tag_name	precision	recall	f_score
B_BRN	40	18.60465116	25.3968254
I_BRN	-	0	-
E_BRN	-	0	-
B_DES	84.31718062	87	85.63758389
I_DES	71.42857143	47.33727811	56.93950178
E_DES	63.15789474	73.74301676	68.04123711
B_DTM	77.09923664	73.78246753	75.40439652
I_DTM	68.06387226	54.47284345	60.51464064
E_DTM	81.89550425	73.18132465	77.29357798
B_LOC	72.83746556	70.35657265	71.57552788
I_LOC	49.03846154	38.78326996	43.31210191
E_LOC	66.76646707	63.89684814	65.30014641
B_MEA	57.99936889	65.08498584	61.33822793
I_MEA	75.0929368	71.12676056	73.05605787
E_MEA	61.32450331	75.65359477	67.73957571
B_NUM	41.36645963	28.3887468	33.67037412
I_NUM	92.30769231	81.81818182	86.74698795
E_NUM	40	12.90322581	19.51219512
B_ORG	78.39486356	68.25041923	72.97176154
I_ORG	79.40871369	68.10498221	73.32375479
E_ORG	78.22028625	69.8721512	73.8109219
B_PER	89.17153661	87.31066981	88.23129252
I_PER	29.03225806	20.93023256	24.32432432
E_PER	55.46655987	92.14903526	69.25
B_TRM	89.47368421	30.35714286	45.33333333
I_TRM	25	20	22.22222222
E_TRM	50	9.090909091	15.38461538
B_TTL	94.69090909	98.33836858	96.48017784
I_TTL	-	0	-
E_TTL	0	0	-
O	97.13034801	97.64299723	97.38599796

Table VI: F1 score of LSTM

Metrics	precision	recall	f_score
Macro	65.0704477	58.28572204	61.49150117
MacroWithoutO	64.91791772	58.11610375	61.32899496
Micro	92.55444851	92.55444851	92.55444851
MicroWithoutO	90.8582245	90.68174032	90.76989663

Bidirectional Long Short-Term Memory with Conditional Random Fields(Bi LSTM - CRF)

This model adds the CRF layer to the Bi-LSTM model which results in an f1 score to 56.55 and uses 41min 8s for training.

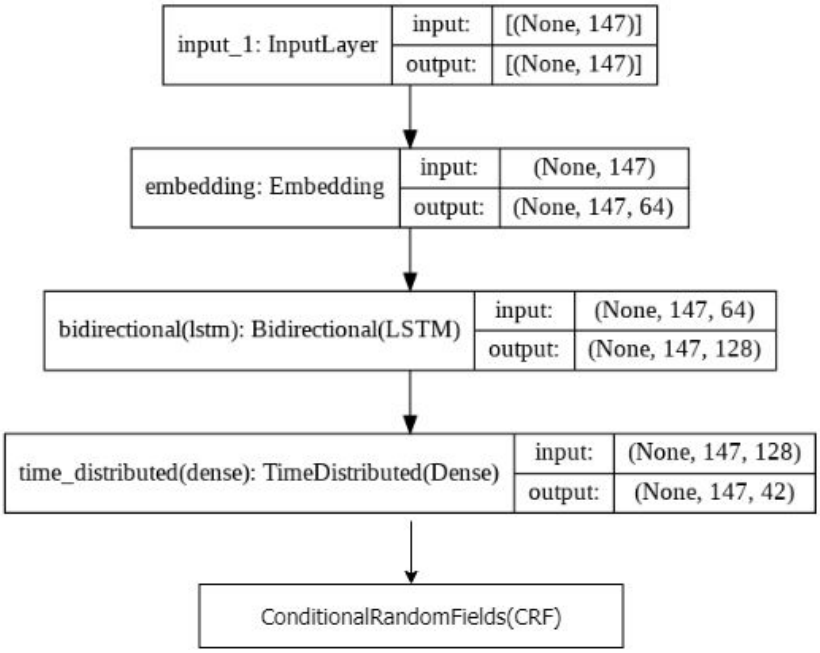

Table VII: Score of Bi-LSTM with CRF on respective tags

tag_name	precision	recall	f_score
B_BRN	38.88888889	16.27906977	22.95081967
I_BRN	-	0	-
E_BRN	-	0	-
B_DES	80.99585062	88.72727273	84.68546638
I_DES	31.42857143	6.50887574	10.78431373
E_DES	60.75268817	63.12849162	61.91780822
B_DTM	75.0617284	74.02597403	74.54025337
I_DTM	65.99576271	49.76038339	56.73952641
E_DTM	83.44198175	69.48968512	75.82938389
B_LOC	67.58756783	72.91112294	70.1484895
I_LOC	38.02816901	30.79847909	34.03361345
E_LOC	56.80628272	62.17765043	59.37072503
B_MEA	55.6760012	65.47450425	60.17900732
I_MEA	73.23420074	69.36619718	71.2477396
E_MEA	60.8401084	73.36601307	66.51851852
B_NUM	42.03431373	29.24126172	34.48969331
I_NUM	92.5	84.09090909	88.0952381
E_NUM	62.5	16.12903226	25.64102564
B_ORG	62.39958538	67.30016769	64.75729461
I_ORG	75.80645161	45.99644128	57.25359911
E_ORG	75.42778919	61.25625347	67.60736196
B_PER	82.08581272	88.21945473	85.0421804
I_PER	19.14893617	20.93023256	20
E_PER	75.05057316	74.05189621	74.54789015
B_TRM	73.33333333	29.46428571	42.03821656
I_TRM	50	20	28.57142857
E_TRM	20	9.090909091	12.5
B_TTL	95.14348786	97.65861027	96.38464406
I_TTL	-	0	-
E_TTL	-	0	-
O	96.57456286	97.38452223	96.97785138

Table VIII: F1 score of Bi LSTM with CRF

Macro	63.83694204	56.15362071	59.74928942
MacroWithoutO	63.66386031	55.95260934	59.55967628
Micro	91.55519248	91.55519248	91.55519248
MicroWithoutO	89.68668364	89.40985908	89.54805742

Comparison

Table IX: Performance comparison

tag_name	f_score_CRF	f_score_LSTM	f_score_BiLSTM	f_score_BiLSTM_CRF
B_BRN	21.21	22.95081967	25.3968254	22.95081967
I_BRN	25	-	-	-
E_BRN	16.67	-	-	-
B_DES	84.11	82.57948948	85.63758389	84.68546638
I_DES	60.52	51.85185185	56.93950178	10.78431373
E_DES	58.29	68.58638743	68.04123711	61.91780822
B_DTM	75.1	70.44129892	75.40439652	74.54025337
I_DTM	87.22	55.29995489	60.51464064	56.73952641
E_DTM	78.8	73.90817469	77.29357798	75.82938389
B_LOC	68.44	66.20841353	71.57552788	70.1484895
I_LOC	37.62	35.42116631	43.31210191	34.03361345
E_LOC	66.59	61.96969697	65.30014641	59.37072503
B_MEA	63.95	57.95650673	61.33822793	60.17900732
I_MEA	62.49	63.42281879	73.05605787	71.2477396
E_MEA	65.48	63.3895818	67.73957571	66.51851852
B_NUM	52.46	34.01360544	33.67037412	34.48969331
I_NUM	76.86	88.88888889	86.74698795	88.0952381
E_NUM	66.67	15.78947368	19.51219512	25.64102564
B_ORG	70.95	66.58361881	72.97176154	64.75729461
I_ORG	69.42	66.58476658	73.32375479	57.25359911
E_ORG	72.55	69.85230235	73.8109219	67.60736196
B_PER	85.93	86.07294317	88.23129252	85.0421804
I_PER	88.59	17.39130435	24.32432432	20
E_PER	89.5	67.01860231	69.25	74.54789015
B_TRM	81.55	38.99371069	45.33333333	42.03821656
I_TRM	17.39	9.523809524	22.22222222	28.57142857
E_TRM	23.53	16.66666667	15.38461538	12.5
B_TTL	96.6	92.75468922	96.48017784	96.38464406
I_TTL	0	-	-	-
E_TTL	0	-	-	-

Table X: F1 score comparison

	CRF	LSTM	Bi LSTM	Bi LSTM CRF
Macro	43.6	60.63469297	61.49150117	59.74928942

Table XI: Runtime comparison

Minute, Second	CRF	LSTM	Bi LSTM	Bi LSTM CRF
Runtime	29, 4	18, 39	23, 49	27, 17

Analysis

The result shows that Bidirectional performs better than unidirectional LSTM which refer to NER tasks that could use past and future context information.The result also reports that additional CRF layer on Bi-LSTM got an unexpected performance which slightly worse than both Uni-LSTM and Bi-LSTM without CRF layer. This may be caused by the huge amount of “O” tags which result in less good relation between neighbors or some incorrect hidden feature in the model. When combined with time needed for training Bi-LSTM-CRF model in Table XI, it infer that the model may not be a good candidate for this task. As the stated problem may be solved by performing with better distributed data or finding the better feature such as adding POS as a part of the features.

For the real world, one of the many use case is web keyword search by try to understand the entities in query then match with indexed word from document to gain better response for search result and as discussed above Bi LSTM without CRF may be the best one from four model that suit this case.

Reference

[1] Thattinaphanich, Suphanut & Prom-On, Santitham. (2019). Thai Named Entity Recognition Using Bi-LSTM-CRF with Word and Character Representation. 10.1109/INCIT.2019.8912091.

[2] Udomcharoenchaikit, Can & Vateekul, Peerapon & Boonkwan, Prachya. (2019). Thai Named-Entity Recognition Using Variational Long Short-Term Memory with Conditional Random Field: Selected Revised Papers from the Joint International Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2017). 10.1007/978-3-319-94703-7_8.

[3] S. Thattinaphanich and S. Prom-on, "Thai Named Entity Recognition Using Bi-LSTM-CRF with Word and Character Representation," 2019 4th International Conference on Information Technology (InCIT), Bangkok, Thailand, 2019, pp. 149-154, doi: 10.1109/INCIT.2019.8912091.

[4] Boonkwan, Prachya & Luantangsrisuk, Vorapon & Phaholphinyo, Sitthaa & Kriengket, Kanyanat & Leenoi, Dhanon & Phrombut, Charun & Boriboon, Monthika & Kosawat, Krit & Supnithi, Thepchai. (2020). The Annotation Guideline of LST20 Corpus.