CLaC at SMM4H Task 1, 2, and 4

Parsa Bagherzadeh, Nadia Sheikh, Sabine Bergler

CLaC Labs
Department of Computer Science and Software Engineering
Concordia University, Montreal
{p_bagher, n_she, bergler} @ encs.concordia.ca

Abstract

CLaC Labs participated in Tasks 1, 2, and 4 using the same base architecture for all tasks with various parameter variations. This was our first exploration of this data and the SMM4H Tasks, thus a unified system was useful to compare the behavior of our architecture over the different datasets and how they interact with different linguistic features.

1 Base system

The base system is a feed-forward neural network with a recurrent neuron. We decided to explore that architecture for independent purposes and used the SMM4H tasks to compare performance on different datasets and task descriptions.

We considered three variations of this architecture:

Full: A recurrent neuron that outputs a 20 dimensional vector is followed by a 3 layer feed-forward neural net, all embedded in two decision neurons with soft-max activations. The feed-forward network has 50, 25 and 12 neurons in first, second and third layers respectively. Unless otherwise mentioned, the network has been trained for 100 epochs.

The recurrent neuron consists of an LSTM cell using $tanh$ activations [Hochreiter and Schmidhuber, 1997]. The activation functions for the feed-forward networks are also $tanh$.

NR: Only the recurrent neuron and the decision neurons are used, the feed-forward (N)etwork is (R)emoved.

Full+At: Attention is added to the full architecture. In contrast to Full, where the LSTM cell outputs a single vector, in Full+At, the recurrent neuron outputs the sequence of each time step.

We used the Keras package [Chollet and others, 2015] to implement the neural networks using TensorFlow as backend [Abadi *et al.*, 2015].

1.1 Input parameters

Tweets are normalized to a size of 25, padded with leading zeros or shortened from the end as required.

The input per tweet consists thus of 25 word vectors of size 100 compiled by the Word2Vec method [Mikolov *et al.*, 2013] over the training data. The Gensim package [Řehůřek and Sojka, 2010] is used for the training of word vectors. The minimum number of occurrences for a word to be considered in the vocabulary is 1 and the window size has been set to 5. Other parameters involved in word vector training were left to the default values of the Gensim package.

Tweet representations are then binned to a batch size of 5, unless otherwise indicated.

2 Text features and knowledge sources

We also experimented with a few linguistic text features and a gazetteer list to see whether they might influence the results.

2.1 Gazetteer

Inspired by Task 1, detection of drug mentions, we scraped the *name* field of *product* fields in DrugBank [Wishart *et al.*, 2017] to compile a gazetteer list for drugs. Due to time constraints, this resource was only minimally refined and contained many multi-word drug names such as *One A Day* and dosage specifications (*Aspirin 80mg*). The gazetteer information was appended to the word vector. Runs that use the gazetteer are identified as +*Gaz*.

Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task, pages 70–73
Brussels, Belgium, October 31, 2018. ©2018 Association for Computational Linguistics

Table 1: Training and validation results for Task 1

	WV trained on Task 1 data					WV trained on Task 1 +Task 2 data				
Architecture	Train Acc	Valid. Acc	Precision	Recall	F1	Train Acc	Valid. Acc	Precision	Recall	F1
Full	0.76	0.55	0.87	0.54	0.66	0.95	0.64	0.87	0.66	0.75
Full+Gaz	0.82	0.56	0.90	0.52	0.66	0.96	0.60	0.88	0.60	0.71
Full+POS	0.75	0.59	0.89	0.57	0.70	0.93	0.59	0.89	0.57	0.70
Full+Modality	0.76	0.55	0.87	0.53	0.66	0.94	0.60	0.87	0.61	0.72
Full+Gaz+POS	0.82	0.57	0.89	0.54	0.67	0.93	0.64	0.88	0.64	0.74
Full+Gaz+POS+Mod	0.81	0.50	0.90	0.42	0.58	0.96	0.62	0.88	0.64	0.73
NR	0.74	0.64	0.87	0.65	0.75	0.94	0.61	0.87	0.62	0.72
NR+Gaz	0.82	0.53	0.88	0.50	0.63	0.96	0.60	0.88	0.60	0.71
NR+POS	0.74	0.57	0.86	0.57	0.69	0.92	0.59	0.88	0.64	0.74
NR+Gaz+POS	0.81	0.59	0.87	0.59	0.70	0.94	0.63	0.87	0.63	0.73
Full+All	0.82	0.64	0.85	0.69	0.76	0.95	0.64	0.90	0.63	0.74
Full+All+At	0.85	0.65	0.86	0.70	0.77	0.95	0.65	0.91	0.68	0.78

2.2 Linguistic features

We used a CLaC pipeline in the GATE environment to extract linguistic features for each tweet. Third party processing resources in our pipeline include the ANNIE Twitter Tokenizer [Cunningham *et al.*, 2002], the Hashtag Tokenizer [Maynard and Greenwood, 2014] and the Stanford Part-Of-Speech Tagger with a model trained on tweets [Toutanova *et al.*, 2003].

Following sentence splitting, tweets were tokenized and Twitter specific tokens (@name and URLS) were removed from the token set. The remaining tokens were assigned one of 36 part-of-speech tags, resulting in a feature value range of integers from 1 to 36.

Following [Doandes, 2003], the part-of-speech tags were used to identify verb clusters. Voice, tense and aspect were assigned to each verb cluster, and the main verb in each verb cluster was identified. These features were also added to the respective word vectors of the main verbs.

We selected only indicative tenses for our binary *tense* feature.

Tokens were also checked against two ad hoc gazetteer lists of explicit negation triggers and modality terms and the binary features *neg* and *mod* were added to the respective word vectors.

Thus we created 4 linguistic features (*tense, voice, POS, and modality*) in addition to the gazetteer feature, that can be appended to word vectors for those words onto which the features project.

3 Task 1

Task 1 was a basic binary categorization task, identifying tweets where a drug was mentioned in its medical sense (the detailed description of the tasks and data can be found in the overview paper [Weissenbacher *et al.*, 2018]). The training data consisted of over 9000 tweets, balanced in both categories.

Table 1 shows the results from some of the runs we compared in order to evaluate the effectiveness of our features. We selected a validation set of 1000 tweets from the training data and trained on the remaining tweets. We compared the training accuracy and the validation accuracy to get some indication of the degree of overtraining. We observe that the difference for training accuracy and validation accuracy is surprisingly small for such a small dataset. Moreover, the differences between our different feature bundles is also rather small. The gazetteer list led to a marked improvement for training accuracy, but not necessarily validation accuracy. Paradoxically, the two best validation accuracy performances came from NR and Full+All (with Full+All+At adding a percentage point). That means that on the validation data, the contribution of the neural net plus gazetteer plus all linguistic feature (plus attention) was matched by simply removing the neural net (NR).

We achieved a greater performance increase in training accuracy across all our configurations when training on Task 2 training data as well as on Task 1 training data. This improvement carries over to validation accuracy and F1 measure, but inconsistently. However, the overall results of different configurations showed less variation when also training on task 2 training data. We speculate that this stabilization may stem from some disruptive effect of data from another task (but that can be expected to contain drug mentions) which might counterbalance overfitting. Our competition runs were all trained on both, Task 1 and Task 2 training data.

It was clear from the beginning that our architecture is severely mismatched to the simple categorization task. The very small difference that our different experiments generated show that the variations do not truly access different tweets. The extremely high training accuracy indicates to us a high degree of overfitting, with the danger of making the entire system somewhat brittle. Table 2 shows that our best competition run on Task 1 was with the Full architecture, the addition of the gazetteer list and two linguistic features reduced the performance. But the near equal performance of Full and NR+Gaz+POS[1] confirms the findings of the validation data, namely that the performance contribution of the network can be matched by the gazetteer list plus some linguistic features. Interestingly, our official test results top the results we obtained on our validation set, which shows that the performance in this case was stable. The performance difference between the best and the last system was 0.1399.

Table 2: Official Task 1 results for CLaC Difference between best and last system score for this task was 0.1399

	P	**R**	**F**
NR+Gaz+POS	0.75	0.80	0.77
Full	0.79	0.77	0.78
Full+Gaz+Mod+POS	0.76	0.76	0.76
Competition Mean	0.89	0.87	0.88

4 Task 2

Task 2 had a semantic component that Task 1 lacked: it concerned distinguishing actual medication intake from possible medication intake and mere mention of a medication in a 3-way decision. We augmented the basic architecture with a third decision neuron for this task.

The training data size for Task 2 was 14482 tweets that were highly imbalanced. Again, a validation set of 1000 tweets was randomly selected from the training data.

Table 3 shows that the richer task definition led to a greater variance in team performance: the difference between the first and last placed team's best runs is .341 micro-averaged F measure. Unlike for Task 1, our performance was not commensurate with our validation performance: in validation runs Full+All+At was also the best run with a validation accuracy of 0.85. Note, that our performance is determined in part by the lowest recall. These results suggest to us that firstly, a custom tailored architecture that better addresses the task can make a greater difference and that our architecture showed more signs of overfitting than in Task 1.

Table 3: Official Task 2 micro-averaged results

Team	P	R	F
UChicagoCompLx	0.654	0.783	0.713
Light_task2	0.520	0.491	0.505
Tub-Oslo-task2-predictions	0.478	0.458	0.468
IRISA_team_task2	0.434	0.501	0.465
IIT_KGP	0.408	0.407	0.408
UZH	0.371	0.437	0.401
CLaC Full+All+At	0.402	0.366	0.383
Techno	0.327	0.432	0.372

5 Task 4

Task 4 was the most semantics oriented task we attempted. The binary task was to identify tweets that clearly indicate that someone received, or intended to receive, a flu vaccine.

Of the 8000 tweets mentioned in the task description, only 4502 tweets could be downloaded for our training data. Despite the very small size of the training data and the potentially deeper semantic distinction, our system performed the closest to the competition mean. Note that the general drug name gazetteer list was not useful for this task.

Table 4: Official Task 4 results for CLaC.

	P	**R**	**F1**
Full+All	0.70	0.89	0.78
NR	0.76	0.46	0.57
Full+Voice+Tense	0.75	0.65	0.69
Competition Mean	0.82	0.85	0.84

CLaC's best run was by Full+All. It is interesting that what appeared to us as the semantically most difficult task has our best performance (measured in distance to the competition mean) due to a recall of .89. We speculate that there may be certain linguistic patterns that our features were able to detect that made this task more amenable to our architecture (in comparison) based on the fact that Full+All outperforms NR and Full+Voice+Tense significantly.

[1] less obvious due to rounding in Table 2

6 Conclusion

CLaC decided late to participate in SMM4H with a uniform architecture to test across several tasks that was not inspired by them. Our conclusion is that the architecture and in particular the input binning and normalizing techniques have to be carefully reviewed, as they risk ignoring key terms in the input. The linguistic features showed some effect, as did the addition and removal of the network. Repeatedly, trial runs showed that removing the network could be offset by adding linguistic features to the recurrent neuron. The detailed interplay of these parameters has to be further studied.

However, we conclude that using the same architecture across several tasks (that are related, but differ significantly) is an interesting exercise and allowed us to gain additional insight. Despite its potential for gross overfitting, the architecture has shown promise. The linguistic features also proved effective, and most importantly, the two components interplay effectively as demonstrated in the fact that in two tasks Full+All was our best performing run.

While each of the three tasks is interesting in itself and clearly has relevance to society at large, we find the juxtaposition of the three tasks very interesting for the ML/NLP researcher.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

François Chollet et al. Keras. https://keras.io, 2015.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, 2002.

M. Doandes. Profiling for belief acquisition from reported speech. Master's thesis, Concordia University, 2003.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

DG Maynard and Mark A Greenwood. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *LREC 2014 Proceedings*. ELRA, 2014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez-Hernandez. Overview of the third social media mining for health (smm4h) shared tasks at emnlp 2018. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

DS. Wishart, YD. Feunang, AC. Guo, EJ. Lo, A. Marcu, JR. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res.*, 2017.