Sequence tagging for medical problems detection

Yacine Jernite

January 15, 2016

1 Introduction/Reminders

Our ultimate goal is to identify mentions corresponding to medical problems in text and map them to UMLS CUIs. Our initial two-steps approach was to use a simple CRF for mention detection (step 1), and focus on a more elaborate method to map mentions to CUIs, or even identify mistakes of the detection (step 2).

The reasoning behind this approach was that we felt we could make use of more interesting techniques for the second step, including leveraging MIMIC text for semi-supervised learning, and using the MRF model we developed for our ICML paper as a prior on the CUI distribution. To that end, I built a model predicting a CUI given a mention, its neighbouring words and neighbouring mentions. I first tried to do semi-supervised learning without using the MRF prior by training on MIMIC matches with different kinds of noise. I was unable to get any improvement from the MIMIC data.

While there are some things we could try to get this working, I'm very concerned that we won't get any results worth presenting even if we do. Indeed, the best F measure we are currently getting for the **detection** step is currently 0.76. To put that in perspective, the best team of the 2015 challenge got a **global** F measure of 0.757. The best **detection** accuracy of 2014 (numbers were not reported for 2015) is 0.813.

Between that and the fact that we will have to tackle step 1 more satisfactorily anyway if we want to release a tool which people can use, I grew increasingly uneasy with our approach, and finally decided to switch gears to try and improve the detection performance a few weeks ago. I'm describing the methods and current results here.

2 Related Work

3 Model description

3.1 CRF

We were previously using a CRF with the following features (and combinations thereof):

- word
- lemma
- pos (part of speech)
- normal
- word_length
- prefix
- suffix
- all_caps
- capitalized
- word_pos (word position in the sentence)
- sentence_pos (sentence position in the document)
- sentence_length
- med_prefix (when the word prefix matches a UMLS prefix)
- umls_match_tag_full
- umls_match_tag_prefix
- umls_match_tag_acro

To predict where the mentions are, I approximate the marginal log-probability of a mention by the average marginal log-probability of its component tags (and the one to the right), and threshold on this marginal log-probability. So the probability of having a contiguous mention ranging from token 2 to token 4, for example, is:

$$\log(p(mention_{(2,3,4)})) = \frac{\log(p(B_2) + \log(p(I_3) + \log(p(I_4) + \log(p(B_5 \vee O_5))}{4}$$

The model was trained with the CRF++ package, and achieved an optimal F-measure on the development set of 0.767, for a marginal probability threshold of 0.75.

3.2 New tagging models

I spent the last month or so implementing two other tagging models in TensorFlow (I wanted to learn the language, and it's nice to have something that interfaces easily with Python). The models are a neural network which predicts each tag independently(-ish) given the whole sentence (SequNN), and a hybrid neural network / CRF system (NN-CRF, the CRF potentials are the output of a neural network).

Both have a first layer which embeds each token of the sentence into a dimension d space. The embedding of a token is the sum of the embeddings of its features. I used a restricted list of features, as it did not look like it affected the performance much:

- lemma
- prefix
- suffix
- pos
- umls_match_tag_full

So the embedding of the token *complaining:* {*lemma: complain, pos: VBG, prefix:com-, suffix, -ing, umls_match_tag_full: O*} is:

$$w_{complaining} = w_{complain} + w_{VBG} + w_{com-} + w_{-ing} + w_O$$

Update: Most of those features actually do not make much of a difference, it looks like lemma + umls_match_tag_full is good enough.

3.2.1 SequNN

I then tried adding various neural network architectures on top of this first layer, and learning the model with SGD.

Convolutional layer My current network actually consists of a single convolutional layer with window size 5, followed by a ReLU. To be perfectly clear, a convolutional window takes a window $(w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2})$, concatenates them into a dimension $5d$ vector, then applies a linear transformation to dimension d'. A ReLU is then applied, followed by a linear transformation from d' to the number of tags (5 currently), followed by a Soft-Max. (A picture is on its way)

Bi-directional LSTM I also implemented a bi-directional LSTM RNN. I couldn't get it to help however (it did worse than the ConvNet on its own, and putting the ConvNet on top of the RNN layer didn't help).

Prediction targets I then first replaced the marginals of output by the CRF in Section 3.1 by the probabilities output by the NN. However, I could not get past an F-measure of 0.7 with this system.

My next step was then to replace the tag prediction layer by a window prediction layer: instead of giving the probability of a tag, the last layer gives the probability of a tag and its left and right neighbour (125 possibilities). This got me back close to the same performance as the CRF (0.763).

3.2.2 NN-CRF

Model I also trained a neural network to output the potentials of a CRF on the tags. The current architecture is basically the same as for the per-tag (per-window) prediction: embedding layer to window 5 convolution to ReLU to pair-wise potentials (5 x 5 matrices).

Learning The model is trained by SGD on the joint likelihood of the sequence of tags (inference is pretty fast), and back-propagating from the potentials to the NN parameters. The best F-measure obtained is about the same as the vanilla CRF. I'm hoping to play a bit with the structure and learning parameters to improve that.

I also tried pseudo-likelihood learning of the model, but this performed terribly: the prediction of a tag given its neighbours is near perfect (0.997 on training, 0.993 on validation), but the best F-measure is around 0.4.

3.2.3 Other implementation details

I use the Adam optimizer with a learning rate of 0.02 for NN-CRF and 0.001 for SequNN (higher leads to gradient explosion), and gradient clipping (I'm not sure it's woking). I have L1 regularization on the lemma embeddings, and L2 regularization on all the embeddings. I need to add L2 regularization to the other embedding. No dropout for now.

4 Next steps

4.1 New tagging scheme

The standard mention identification methodology consists in tagging tokens with 'B' (first token of the mention), 'I' (other tokens of the mention) or 'O' labels (everything else). This reduces the problem to sequence tagging, which can be solved linearly in the length of the sentence. For example, "the patient suffers from a broken jaw ." will have tags O-O-O-O-O-B-I-O. About 10% of the mentions we are dealing with are dis-contiguous. While those can technically be represented with a B-I-O scheme, there is a concern that a model such as a CRF will have difficulties propagating information from one part of a mention to another. We remedy this by adding two new labels: 'OD' for tokens which are within the scope of a mention but not part of the mention itself, and 'ID' for tokens which are part of a dis-contiguous mention. For example, "the pain is stongest in the arm ." will have tags O-B-OD-ODOD-OD-ID-O. About 95% of the sentences in our data can be represented with this tagging scheme. It fails however any time two or more mentions overlap. It is worth noting that the vast majority of cases of overlapping mentions in our data are of the form "A and B are C" or "A is B and C". While any per-token tagging scheme which covers all possible configurations would require a prohibitively large tag set, it is easy enough to come up with one which covers all of the aforementioned cases without making our scheme much more complex. The idea is the following: when mentions overlap, we identify each by its "first proper tag". In the example "left arm and shoulder are swollen", the mention "left arm swollen" will be identified by "arm", and the mention "left shoulder swollen" by "shoulder". We introduce a new tag 'In' to denote such an identifying token (the sequence will then be labeled B-In-OD-In-OD-ID). To make the model fully coherent however, we actually need two more tags: 'Bn', which acts as 'In' but for the first word of a mention ("elbow and wrist broken" is labeled Bn-OD-Bn-ID), and 'Ip' for a token which is part of only one of two overlapping mention, but is not the identifying token ("inflamation of left kidney and spleen" will be labeled B-OD-In-Ip-OD-In). This scheme uniquely defines all but 0.5% of our sentences (it still fails in the case of nested mentions, which exist but are quite rare).

4.2 Combining methods

Taking the average of the marginals output by the two models consistently gets a better max F measure than either of the two, by 1-3 points ()

4.3 Better representation

We could add a language modelling objective to get better embeddings.

We could train the first layer(s) (representation and maybe convolution) jointly.

Adding pre-trained word embeddings does not help. I have tried both using them as is, and adding a trainable projection.

References

[1] Matthieu Labeau, Kevin Löser, Alexandre Allauzen, and Rue John von Neumann. Non-lexical neural architecture for fine-grained pos tagging. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 232–237, 2015.

[2] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip Torr. Conditional random fields as recurrent neural networks. *arXiv preprint arXiv:1502.03240*, 2015.