

An Automata Based Approach to Biomedical Named Entity Identification

James Dowdall¹, Bill Keller¹, Lluís Padró², and Muntsa Padró²

¹ School of Cognitive and Computing Sciences
The University of Sussex

{j.m.dowdall,bill}@sussex.ac.uk

² TALP Research Center

Universitat Politècnica de Catalunya
{padro,mpadro}@lsi.upc.edu

In this work, we present a new approach to Biomedical NE identification using an automata learning algorithm: Causal-State Splitting Reconstruction [1]. This algorithm has previously been applied to Named Entity Recognition [2] obtaining good results given the simplicity of the approach.

The same approach has been applied to Biomedical NE identification, using GENIA corpus 3.0, with 10-fold cross-validation. Our system attained $F1 = 73.14\%$.

These results can be compared directly to [3] and [4], which used the same data. First system obtains $F1 = 57.4\%$ using ME Models, and the second one reports $F1 = 79.2\%$ using SVMs. Both improve their results using post-processing techniques, reaching $F1 = 76.9\%$ and $F1 = 79.9\%$ respectively.

Our system does not use any post-processing techniques, and takes into account few features, so the results are considered very promising. In future work some post-processing will be developed to improve the results.

References

1. Shalizi, C.R., Shalizi, K.L.: Blind construction of optimal nonlinear recursive predictors for discrete sequences. In: Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference. (2004)
2. Padró, M., Padró, L.: Applying a finite automata acquisition algorithm to named entity recognition. In: Proceedings of FSMNLP'05, Helsinki, Finland (2005)
3. Lin, Y.F., Tsai, T.H., Chou, W.C., Wu, K.P., Sung, T.Y., Hsu, W.L.: A maximum entropy approach to biomedical named entity recognition. In: BIODDD. (2004)
4. Lee, K.J., Hwang, Y.S., Kim, S., Rim, H.C.: Biomedical named entity recognition using two-phase model based on svms. J. of Biomedical Informatics 37 (2004)