Using Information Extractors with the Neural ElectroMagnetic Ontologies

Fernando Gutierrez, Daya C. Wimalasuriya, and Dejing Dou

Computer and Information Science Department
University of Oregon
Eugene, Oregon 97403, USA
{fernando,dayacw,dou}@cs.uoregon.edu

Abstract. In this paper, we present a study to identify instances in research papers from PubMed for the ontology developed by NEMO through the use of a component-based approach for information extraction.

1 Introduction

In this paper, we present how a novel component-based approach for information extraction named OBCIE (Ontology-Based Components for Information Extraction) [2] has been used to extract instances from PubMed[1] papers for the ontology developed in the Neural ElectroMagnetic Ontologies (NEMO) [1] project.

For the study, the OBCIE platform of *two-phase classification* [2] has been selected to develop an information extractor. The two-phase classification consists of a first phase to identify sentences containing the required information, and of a second phase to identify words within sentences that carry the actual information. The version 1.4 of the ontology developed by the NEMO project was used in the study. From the ontology, 10 classes were randomly selected for information extraction. The corpus consists in 500 abstracts that are divided into 10 sets. The abstracts come from research papers related to one of the ten concepts analyzed in this study. The research papers are obtained from the web-based archive of biomedical literature, PubMed.

2 Experiments and Results

For each class a set of words is picked as features of the class. The features are selected by contrasting their presence in sentences with and without the class (lift). If the word is relevant enough, then it is selected as feature. These features are used in the classification to help identify instances of the class.

Table 1 shows the results for *internal* and *external* performance measures. The internal performance measure compares parts of the extracted string with

[1] http://www.ncbi.nlm.nih.gov/pubmed

R. Meersman, T. Dillon, and P. Herrero (Eds.): OTM 2011 Workshops, LNCS 7046, pp. 31–32, 2011.

parts of the key string, whereas the external performance measure compares the whole extracted string with the key string. This makes the external measure more accurate since it compares the actual extraction with the key.

Table 1. Results for the ten classes (%)

Concept	Internal			External		
	Precision	Recall	F1	Precision	Recall	F1
EEG data set	69.40	65.95	67.63	45.52	41.79	43.57
MEG data set	80.00	62.33	70.07	66.10	57.35	61.41
Anatomical region	80.39	47.67	59.85	78.57	40.24	53.22
Anatomical surface	0.0	0.0	0.0	0.0	0.0	0.0
Brodmann area	80.24	58.03	67.35	71.79	50.00	58.94
Electrical conductivity	65.62	42.00	51.21	50.00	27.08	35.13
Electromagnetic field	74.00	58.73	65.48	68.88	60.78	64.58
ERP topography	27.54	79.31	40.88	32.51	81.53	46.49
Multimodal part of cortical surface region	75.80	62.66	68.61	79.62	58.90	67.71
Unimodal part of cortical surface region	64.00	62.99	63.49	45.45	42.01	43.66
All (except Anatomical surface)	68.56	59.97	61.62	59.83	51.08	52.75

3 Discussion and Conclusion

With an average F1 measure of 52.75%, the results of the present work can be considered acceptable when compared with OBCIE's previous study [2] that recorded F1 measures slightly below 40%, while information extraction systems such as Kylin [3] recorded F1 measures in the range of 40% - 50%. The results also show that depending on how the selection of features is done the performance of the information extraction may vary.

The application of the information extractors of OBCIE on the NEMO ontology has been successfully completed as evidenced by the acceptable performance measures obtained. Making use of such techniques in the development and enrichment of biomedical ontologies becomes increasingly important as the size and scope of such ontologies increase rapidly.

References

1. Dou, D., Frishkoff, G., Rong, J., Frank, R., Malony, A., Tucker, D.: Development of NeuroElectroMagnetic Ontologies (NEMO): A Framework for Mining Brainwave Ontologies. In: Proceedings of the 13th ACM International Conference on Knowledge Discovery and Data Mining (KDD), pp. 270–279 (2007)
2. Wimalasuriya, D.C., Dou, D.: Components for information extraction: ontology-based information extractors and generic platforms. In: CIKM, pp. 9–18 (2010)
3. Wu, F., Weld, D.S.: Autonomously semantifying wikipedia. In: CIKM, pp. 41–50. ACM, NY (2007)