Comparison of different strategies for utilizing two CHEMDNER corporation

METGen-04-

I.R. Rocha[I,II,III,IV], J.T. Toschack[V,VI,VII,VIII], **H. Miyazaki**[VIII,IX]

[I]Center Seville, Sevilla, Spain , [II]Centro, Sevilla, Spain , [III]Senter, Barcelon, Spain , [IV]Alicant Center, Barcelon, Spain , [V]BioAnd, Málaga, Spain , [VI]ColCiencias, Bogota, [VII]Neuroscience Solutions to Cancer Research Group, Imperial College, London, United Kingdom, [VIII]Centro Velázquez, Sevilla, Spain , [IX]CENTRO DE INVESTIGACION Y DESARROLLO PASCUAL VILA, Barcelona, Spain

Machine *learning based chemical* named entity recognition (CNER) systems use annotated chemical information corpora as training data to generate rules to identify chemical named entities. For making effective rules for a particular chemical named entity recognition task, it is desirable to have a large training data that covers wide varieties of chemical named entity examples for the task. For the CHEMDNER patent task, there are two corpora available for training. One is the corpus for the patent task and the other is the CHEMDNER corpus for PubMed abstract constructed for CHEMDNER task in BioCreative IV. Even though these corpora were constructed based on the same annotation guideline, the style of writing for patent is different from the one for abstract of the research paper. In this research, we implement CNER tools based on different strategies for utilizing these two

CHEMDNER corpora and compare results for clarifying the issues related to these strategies. Our basic system uses conditional random field (CRF) as a machine learning technique. For the CRF features, we use linguistic features in addition to domain knowledge feature produced by ChemSpot, a common chemical named entity recognition tool. We implemented the system with the following three different strategies. 1) Use CHEMDNER patent corpus only for training, 2) Merge the two CHEMDNER corpora for training to enlarge training examples, 3) Use output of basic system trained on CHEMDNER PubMed corpus as an additional feature of the CRF that uses CHEMDNER patent corpus for training. This can help learning any consistent differences between annotation schemas of both corpora. We compare the results of each system by using simple system performance measure (e.g., recall, precision, and F-score) and analysis on the unique findings of each system.