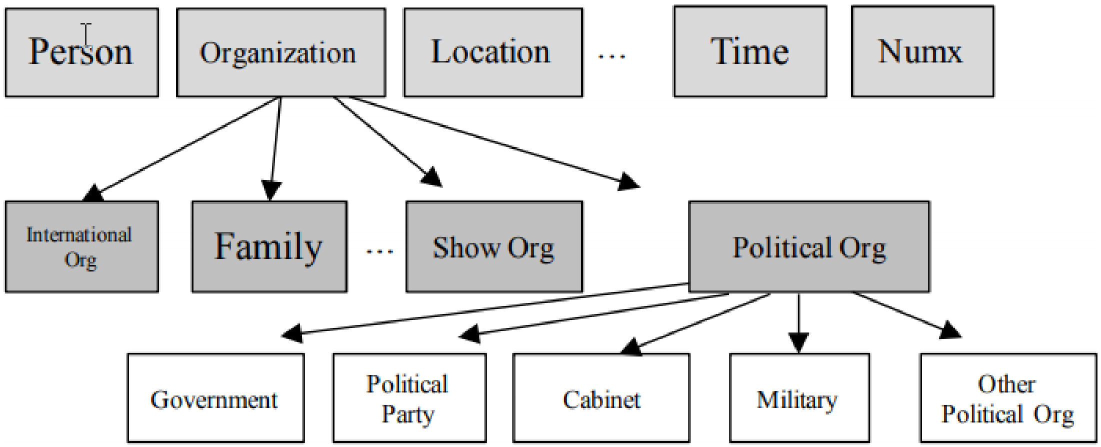
Fine-grain NER

The exercise is about Named Entity Recognition (NER). For basic NER problem definitions and evaluations, please use web resources.

We have a data set that contains 113 types of leaf-level Named Entities organized in a fine-grained structure. At the top level of the hierarchy, there are about twenty coarse-grained named entity categories, such as Person, Organization, Location, Facility, Product, Event, … Each top-level category is further divided into several second-level categories as shown in Figure 1. Each second-level category is in turn divided into several leaf-level categories.



The data (called FIGER) is split into training, develop, and test. Each split is stored as a json file. The training data is annotated by distant supervision (weak supervision), and it has about 2 million annotated tags instances, and the tags are noisy. The test set is labeled by human annotation, which contains 563 annotated tag instances.

Please download the dataset (FIGER.zip) from link.

https://storage.googleapis.com/cyberyuner/FIGER.zip

Your exercise goal is to build and train a Named Entity Recognition model, using training data and dev data of FIGER. Then, you apply the trained model on test data and report its performance.

You are free to use any type of models, such as LSTM, CNN, CRF, or even stacked transformers. We don't expect you to beat the state-of-the-art results, which can be F1 score as high as 78+. We are more interested in how you tackle the problem by using different components/architectures, and what are the reasons behind using them. A systematic ablation/progressive study of your model would be encouraged and helpful. Be careful of the time you spend on this (which can be weeks if you dig really deep), and also don't just copy existing code from github without understanding how it works.

Since the training data is huge, if you are unable to use all training data, you can downscale the model using sampled training data. However, you need to use all test data to report your performance.

Once you finish, please check-in your code in a github repository, and give necessary instructions for us to replicate your results. Please include all package requirements, or even auto-installation scripts to automate the environment setup. Please include the steps to replicate the performance on test data. Also, please attach a short report to explain how you come up with the design, and what you think important and interesting to include in the report.

There are some useful reference papers to help you coming up your model:

Xiao Ling and Daniel S. Weld. 2012. **Fine-grained entity recognition.** In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (*AAAI'12*). AAAI Press, 94–100.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. **Freebase: a collaboratively created graph database for structuring human knowledge.** In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (*SIGMOD '08*). Association for Computing Machinery, New York, NY, USA, 1247–1250. DOI:https://doi.org/10.1145/1376616.1376746

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. **Yago: a core of semantic knowledge.** In *Proceedings of the 16th international conference on World Wide Web* (*WWW '07*). Association for Computing Machinery, New York, NY, USA, 697–706. DOI:https://doi.org/10.1145/1242572.1242667

Mohamed Amir Yosef, Sandro Bauer, Johannes Hoffart, Marc Spaniol, and Gerhard Weikum. 2012. **HYENA: Hierarchical Type Classification for Entity Names.** COLING 2012, pages 1361-1370.

Dan Gillick, Nevena Lazic, Kuzman Ganchev, Jesse Kirchner, and David Huynh. 2014. **Context-Dependent Fine-Grained Entity Type Tagging.** CoRR, abs/1412.1820.

Satoshi Sekine. 2008. **Extended Named Entity Ontology with Attribute Information.** In Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2002, pages 52-57.

Khai Mai, Thai-Hoang Pham, Minh Trung Nguyen, Tuan Duc Nguyen, Danushka Bollegala, Ryohei Sasano, Satoshi Sekine. An Empirical Study on Fine-Grained Named Entity Recognition. COLING 2018, pages 711-722.