Homework 1

Name: Yulong Pei
Andrew id: yulongp

1. Architecture

In this homework, a named entity recognition system is implemented using UIMA SDK. A process of this system is shown in Figure 1 and the input file is processed through collection reader, NER annotator and CAS consumer, then the output file is achieved. These three components, i.e., collection reader, NER annotator and CAS consumer, form the analysis engine in UIMA framework.

In particular, (1) the collection reader reads each line from the input file and parses the line into the predefined Sentence object. (2) The NER annotator receives every Sentence object and extracts the entity mentions in this Sentence. (3) The CAS consumer writes the extracted entity mention into the output file.

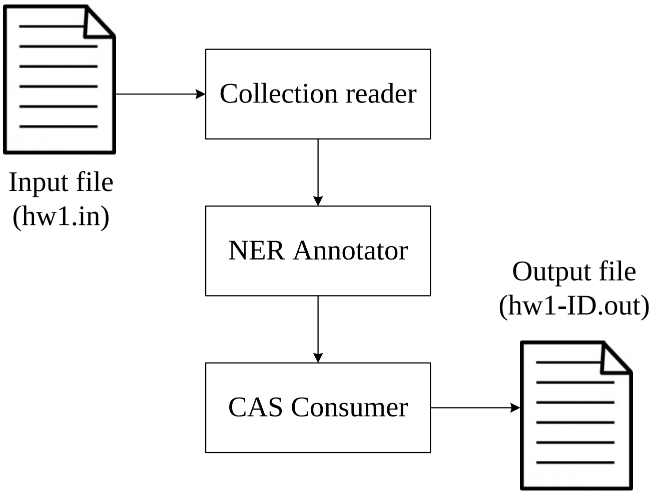

Figure 1: The process of named entity recognition system.

2. Algorithms

In this homework, the algorithmic part is based on Stanford CoreNLP toolkit[1]. Therefore in this section the related NLP and machine learning techniques are introduced and more details can be found in the corresponding document[2].

1.1 NLP techniques

Part-of-speech tagging (POS)

POS tagging aims to mark up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context. Generally, POS tagging system will assign words several different types of tags based on Penn Treebank[3], e.g., JJ (Adjective), NN (Noun, sing. or mass), PP (Personal pronoun), VB (Verb, base form), etc. Two representative models used to implement a POS tagging system are Hidden Markov

[1] http://nlp.stanford.edu/software/corenlp.shtml
[2] http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf
[3] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Model (HMM) and maximum entropy model. In the Stanford CoreNLP toolkit, the POS tagging system applies the maximum entropy model to implement the tagger.

Named entity recognition (NER)

NER is to recognize the entities in texts including names, locations, organizations, etc. NER plays an important role in information extraction, question answer, machine translation and other NLP related tasks. In general, a typical NER consists of two parts: recognize entity boundaries and determine entity categories. In this homework, the task can be viewed as a special case of NER which is to extract gene from the medical text. The NER system in the Stanford CoreNLP toolkit uses a combination of Conditional Random Field (CRF) sequence taggers and these taggers are trained on a variety of corpora.

1.2 Machine learning techniques

Maximum entropy model (MaxEnt)

Maximum entropy model is am important principle in probabilistic models. From the perspective of MaxEnt, among all the possible probabilistic models, the model with maximum entropy is the best model. Therefore, maximum entropy principle can be described as the process to select the model with maximum entropy meets the constraints. Assume the probabilistic distribution of a random variable X is P(X) and its entropy is

$$H(P) = -\sum_{x} P(x) \log P(x)$$

And the entropy meets the inequation

$$0 \leq H(P) \leq \log |X|$$

More details about maximum entropy model can be found in this report[4].

Conditional Random Field (CRF)

CRFs are a type of discriminative undirected probabilistic graphical model. It is used to encode known relationships between observations and construct consistent interpretations. It is often used for labeling or parsing of sequential data, such as natural language text or biological sequences and in computer vision. Named entity recognition belongs to sequence labeling problem and therefore linear CRF is introduced in this report. Given two random variables X and Y, P(Y|X) is the conditional probabilistic distribution of Y given X. If the random variable Y forms a Markov random field generated by a undirected graph G = (V, E), i.e.,

$$P(Y_v|X, Y_w, w \neq v) = P(Y_v|X, Y_w, w \sim v)$$

stands for any node v, we name the conditional probabilistic distribution P(Y|X) the conditional random field (CRF). More details can be found from the orginal CRF paper[5].

[4] http://repository.upenn.edu/cgi/viewcontent.cgi?article=1083&context=ircs_reports

[5] http://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers