This version of pipeline uses three different Annotators to annotate gene mention in input text.

1. The first is complicated and specially sharpened for gene entities extraction tool kit is **LingPipe Core** [[1]] that uses Hidden Markov Model to extract gene chunks. A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. An HMM can be presented as the simplest dynamic Bayesian network. The mathematics behind the HMM was developed by L. E. Baum and coworkers. It is closely related to an earlier work on optimal nonlinear filtering problem by Ruslan L. Stratonovich, who was the first to describe the forward-backward procedure. [[2]]
2. Second is **ABNER** – A Biomedical Named Entity Recognizer [[3]]ABNER is a software tool for molecular biology text analysis. It began as a user-friendly interface for a system developed as part of the **NLPBA/BioNLP 2004 Shared Task** [[4]] challenge. The details of that system are described in the paper (Settles, 2004)[[5]].
   At ABNER's core is a statistical machine learning system using linear-chain conditional random fields (CRFs) with a variety of orthographic and contextual features. Version 1.5 includes two models trained on the **NLPBA** and **BioCreative** corpora, for which performance is roughly state of the art (F1 scores of **70.5** and **69.9** respectively; details at the ABNER official site [[6]]). The new version also includes a Java API allowing users to incorporate ABNER into their systems, as well as train and use models for other data.
   As ABNER has two different models there are two descriptors abnerBiocreativeAnnotatorDescriptor.xml and abnerNlpbaAnnotatorDescriptor.xml that uses the same Java class AbnerAnnotator but with different model setting.
3. Third is based on the **Entrez Gene** database [[7]]. It is a searchable database of genes, focusing on genomes that have been completely sequenced and that have an active research community to contribute gene-specific data. Information includes nomenclature, chromosomal localization, gene products and their attributes (e.g., protein interactions), associated markers, phenotypes, interactions, and links to citations, sequences, variation details, maps, expression reports, homologs, protein domain content, and external databases.
   As we access remote database every time we want to check if a gene exists in it or not it is real time consuming and significantly slows down the whole pipeline. That is why data is cached locally in RAM in HashMap and saved to hard drive periodically. Nevertheless with large amount of new uncached genes it still could be very slow. If we want that our pipeline process data more quickly we can switch off the EntrezGeneAnnotator in hw2-oiskra-aae.xml configurational file.

The main architecture of pipeline is shown on the UML Diagram 1.

[1] LingPipe Core tool kit download page at http://alias-i.com/lingpipe/web/download.html
[2] Hidden Markov Model on Wikipedia at https://en.wikipedia.org/wiki/Hidden_Markov_model
[3] ABNER - A Biomedical Named Entity Recognizer at http://pages.cs.wisc.edu/~bsettles/abner/
[4] NLPBA/BioNLP 2004 Shared Task - http://www.genisis.ch/~natlang/JNLPBA04/
[5] Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets by Burr Settles - http://burrsettles.com/pub/settles.nlpba04.pdf
[6] Details of ABNER performance - http://pages.cs.wisc.edu/~bsettles/abner/#performance
[7] Entrez Gene database - http://www.ncbi.nlm.nih.gov/gene

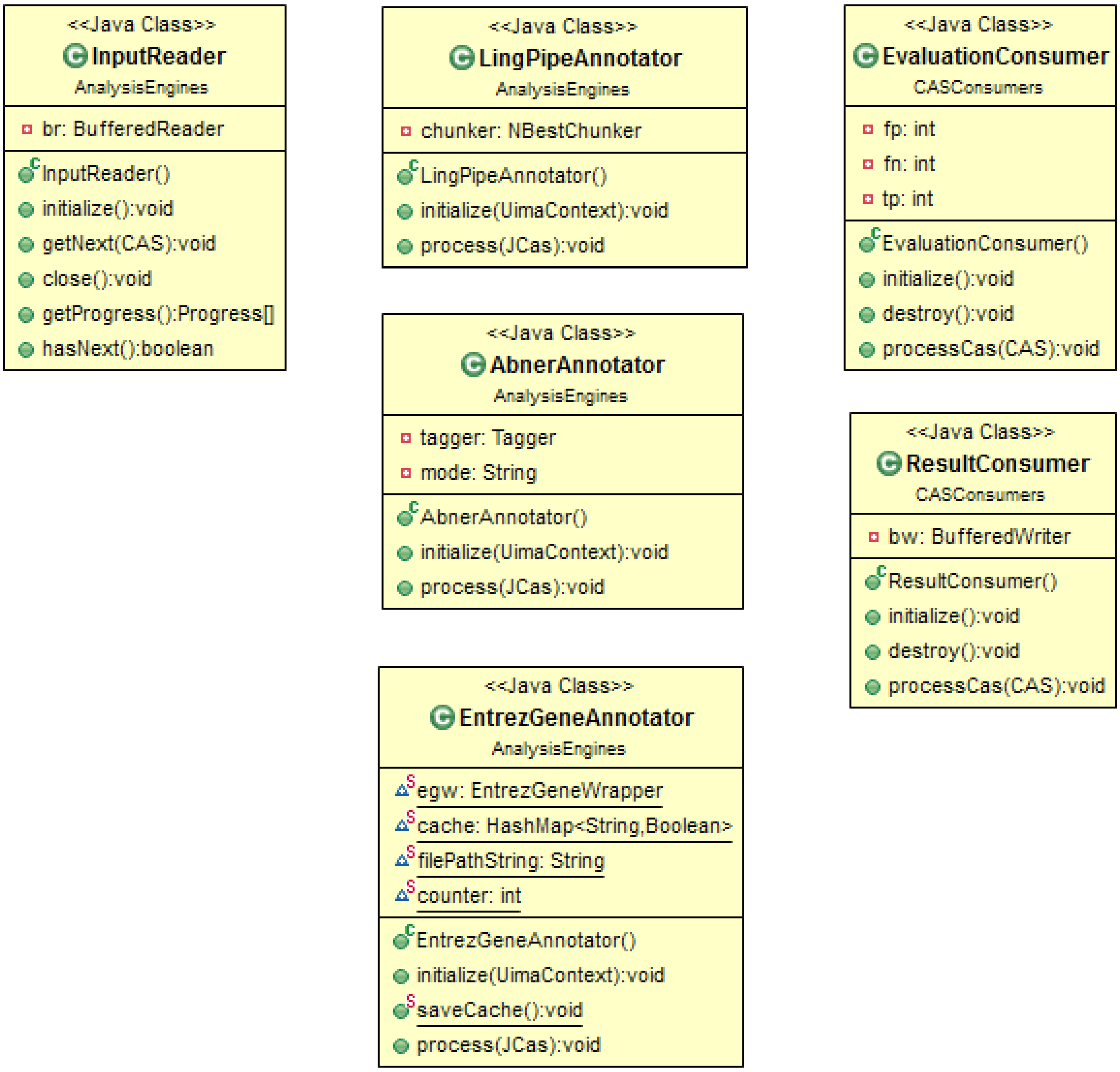

Diagram 1. Main parts of pipeline. Data flows from left to right.

Data is loaded from text file with Input Reader and then sequentially processed by LingPipeAnnotator. It extracts gene mentions and gives confidence rate to each annotation in range from -1 till -10000. Farther data sequentially goes through AbnerNlpbaAnnotator and AbnerBiocreativeAnnotator (AbnerAnnotator). If those two annotators mark the same piece of sentence as the gene mention we increase confidence of annotation. If it is first mention of this annotation we give it confidence level -10000. After all we select gene mentions that were found by only one annotator and look for them in Entrez Gene Database. If finally there is no such term in database we delete this annotation from the UIMA index. This strategy in theory has to increase the precision level of pipeline but probably will decrease recall. Resulting F1-measure will increase or decrease depending on the input data. As we have only one source of data (GeneTag) given that LingPipeAnnotator based on the tool that was trained using the same corpora (GeneTag) it is hard to measure real performance of the pipeline.

In case of decreasing of performance using above mentioned technique we can use another algorithm and save annotation from all 4 annotators. It probably will decrease precision but have to increase recall. Given the flexible nature of UIMA and pipeline developed it is very easy to reconfigure it to use one or another technique.