Type system: **NamedEntityAnnotation** contains named entity positions as well as named entity itself

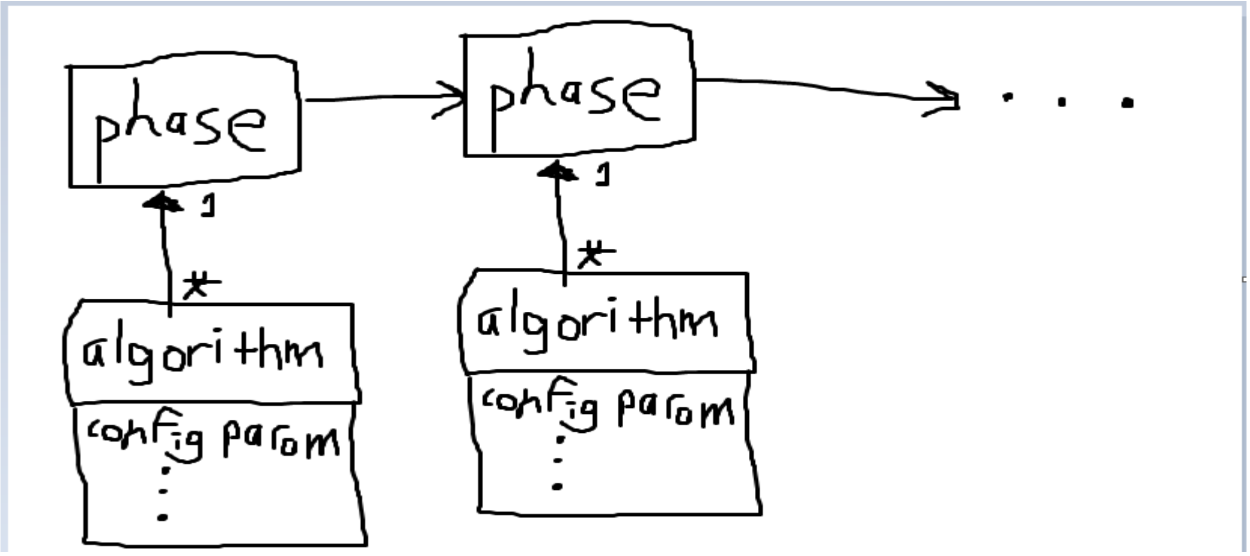
ShittySentenceID is a shitty container for the sentence ID for this CAS

Basically I'm just using lingpipe to do the chunking. The collection reader reads the input, the analysis engine uses lingpipe to chunk the sentences, and the CAS consumer writes out the genes to output files.

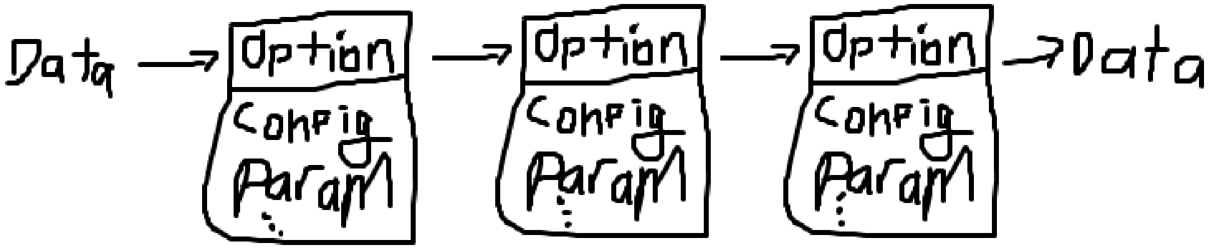
Performance was compared to sample output file. It looked good enough.

1. We used an HMM chunker. HMMs are ML techniques.
2. We used an HMM chunker. HMMs are also NLP techniques, for doing NER
3. None
4. The GeneTag model file that contains names of genes
5. None
6. Uses GeneTag model file trained from GENETAG corpus
7. Lingpipe
8. Input file read line-by-line into CAS's, where NER is then done to get chunks of gene names, and then written out gene-by-gene into the output file
9. These tools all suck. So many errors happen that get in the way of doing actual, useful programming.

UML Diagram 1



UML Diagram 2


Data
Option
Config
Param
Option
Config
Param
Option
Config
Param
Data