**A machine learning system to predict species labels of gene mentions in medical abstracts**

Under the guidance of Dr. Lyle Ungar,
Rushin Shah
University of Pennsylvania
July 11th, 2006

*Problem*

Our task was to design a system that takes as input a gene mention in a medical abstract, and returns the species that the gene mention in question belongs to. Input to this system is supplied in the form of Medline abstracts. All the gene mentions in a particular abstract should be recognized, and tagged for the correct species.

*Larger Goal*

Our task can be thought of as an instance of Entity Reference Resolution i.e. for a token, recognizing its semantic category, in a sense, answering the question, what is the token? Our current task can be thought of as a precursor to creating a wiki system for genes, which automatically creates a web page for every gene, and automatically marks up the Medline abstracts with pointers to these genes.

*Approach*

We used a machine learning approach to implement this system. Our problem can be thought of as choosing the correct labels for a given set of unlabelled data, i.e. obtaining the most probable sequence of labels for a given sequence of tokens. Here, labels correspond to species names, and tokens correspond to gene mentions. The central task is, therefore, to find for each gene mention, a set of features which are helpful in identifying the correct tag. Such features can be based on:

- Content of the gene mention

e.g.: Is the gene name in a lexicon of human genes?

- Context of the gene mention

e.g.: Is there a species name mentioned close by to the gene name? If so, which species and how close?

Discrete as well as continuous features can be used in solving this problem.

A sufficiently descriptive set of features is essential to be able to perform our tagging with a high degree of accuracy. Once we have obtained such a feature set, a system that utilizes this set and learns the correct weights for these features, can be easily implemented using mathematical models such as Conditional Random Fields (CRFs) or Maximum Entropy Markov Models.

CRFs are probabilistic tagging models that give the conditional probability of a possible tag sequence, given an input token sequence. If we use a conditional random field (CRF), the task of training the CRF can be thought of as obtaining a set of weights for the features that we include in our CRF, which maximizes the log likelihood of our training set. The primary advantages of CRFs over other models are:

- Features that rely on the tag values of other nearby gene instances can be used, i.e., in a sense, simultaneous tagging can be performed.

- A large set of atomic features can be compiled, and feature induction can be applied to obtain the relevant compound features, i.e. features that rely on more than 1 attribute of the gene instance.
- The label bias problem that occurs in Maximum Entropy models can be avoided.

Maximum Entropy Markov Models (MeMMs) are variations on Hidden Markov Models (HMMs). These models attempts to characterize a string of tokens (such as words in a sentence, or sound fragments in a speech signal) as a most likely set of transitions through a Markov model. Each of the states corresponds to a conceptual stage in the sentence or speech signal, and each state can emit certain tokens (i.e. words or speech fragments). The most likely path through the HMM or MEMM would be defined as the one that is most likely to generate the observed sequence of tokens. The idea behind maximum entropy models is that instead of trying to train a model to simply emit the tokens from the training data, one can instead create a set of Boolean features, and then train a model to exhibit these features in the same proportions that they are found in the training data.

MeMMs are independently normalized at each position in a tag sequence, while CRFs have a single combined normalizing denominator for the entire tag sequence. This independent normalization in MeMMs prevents decisions at different positions from being weighed against each other. This is known as the label bias problem, and it adversely affects the accuracy of MeMMs. As a result, they give less accuracy than CRF's, which were specifically designed to alleviate the label bias problem. However, MeMMs are faster to test and train than CRFs, which allows us to train them on a larger set, than could be possible using CRFs.

In order to train and test our CRF/Maximum Entropy model, we firstly require a set of pre-annotated Medline abstracts, i.e. abstracts for which we know the correct species tag for each gene mention in that abstract. Such abstracts can be obtained, either by taking all Medline abstracts that have only 1 species name in their MeSH headings, since this implies that only one species is mentioned in that abstract, and hence all gene mentions in it must necessarily belong to that species, or by manually tagging abstracts, with the help of domain experts. This is necessary to obtain abstracts which have genes of more than 1 species mentioned. We then extract the values of all features for all gene mentions in these abstracts, and use the MATLAB implementation of CRFs and Maximum Entropy classifiers, along with the values of all gene mentions and their features, to create our model.

*Implementation*

A unique aspect of our implementation is that all data, i.e. all Medline abstracts, gene mentions, gene lexicons, species lists, are stored in a MySQL database instead of a flat file hierarchy.

The process of implementing this system can be thought of as having the following phases:

- Tagging the entire Medline for gene mentions, using the BioTagger developed at the University of Pennsylvania.
- Creating our MySQL database. This involves compiling Medline abstracts, gene lexicons, species lists, gene mentions, and transferring them into a MySQL database. The two main issues here are having an efficient and comprehensive database schema, and extracting data from various disparate formats, such as XML, flat file, etc. and putting it into the database. The advantage of using a MySQL database is the availability of options such as clustering, indices, to improve the accuracy and speed of our system.
- Obtaining a sufficiently large random set of tagged abstracts from our database for training and testing our model, and obtaining all gene mentions from this set. Tagged abstracts can be

obtained as described earlier, by:
    - Extracting all abstracts that have exactly 1 species mentioned in their MeSH headings
    - Tagging abstracts which many contain gene mentions from multiple species, with the help of domain experts
- Deciding an informative, useful and sufficient feature set, consisting of both content- and context-based features, for our problem, i.e. a set of useful features that describes each gene mention, and gives useful clues as to which species that gene mention might belong to.
- Extracting values of all the above-mentioned features for each gene mention.
- Using the machine learning package MATLAB, to implement a CRF/Maximum Entropy model, with the training portion of the set of gene mentions and their features.
- Evaluating our model on the testing portion of the set of gene mentions and their features, and checking the results for accuracy, speed, etc.

*Features*

The current version of the feature set that we have used for this problem consists of the following features:

- Content based:
    - Is the gene name in any of the species-specific gene lexicons?
    - Is any species name/code a substring of the gene name?

- Context based:
    - Is any species name mentioned in the article title?
    - Is any species name mentioned in the journal title?
    - Is any species name mentioned in the MeSH headings of the abstract?
    - Is any species name mentioned in the abstract itself?
    - Which species is mentioned closest to the gene mention?
    - How many times is each species mentioned in the abstract?

Some more capabilities that we want to add include:

- We want to use Feature Induction, so that the system can create and use the most useful feature conjunctions, i.e. features that depend on multiple predicates, and discard those features that prove to be least relevant.
- We want to consider the correct species tags for adjacent species mentions simultaneously, thus taking into account the value of the species tag for the previous gene mention, while calculating that for the current gene mention.

*Results*

We ran a CRF model, and a Maximum Entropy model, on randomly collected tagged abstracts, which contain 1 species each, only. Since we haven't been able to obtain multi-species abstracts yet, we haven't been able to train our models on those. The proportion of training data out of the total data was kept at 0.6 The Maximum Entropy model gave a mean accuracy of 89%, whereas the CRF model, after incorporating continuous features into it yet, gave a mean accuracy of 91%.