Extracting Terms from Texts with Conditional Random Fields

Li YiXuan
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
1427402002@stu.suda.edu.cn

Lu Xun*
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
luxun@suda.edu.cn

Abstract — The rapid growing of biological text has promoted the research of the text mining of various non structured documents with the emphasis on the mining of biological knowledge. At the same time, the majority of biological text mining efforts based on the identification of the name of the biological term gene and protein. Therefore, how to recognize the biological terms effectively from the text has become one of the important issues in bioinformatics. Conditional random fields, an important machine learning algorithm, is a model of the probability of a graph model to give an opinion of the label. They traditionally use a set of observations and labels to receive training. Here we use controlled release fertilizer for a class of temporal learning algorithms, in reinforcement learning. Therefore tags are operating, updated environment, the impact of the next observation. Thus, from reinforcement learning, the controlled release fertilizer provides a model of joint action in the decentralized Markov decision process, and defines how agents can communicate with each other, and choose the best way of joint action. We use the hot data corpus for training and testing. The results show that the system can effectively find out the biological terms from the text. We get along with the average accuracy of rate=90.8%, the average recall of rate=90.6%, and the average rate=90.6% F1 six category of biological terms. The results are quite good for the entity recognition system, which is named after many other biological organisms.

Keywords-machine learning; conditional random fields; national language processing; named entity recognition; bioinformatics

I. INTRODUCTION

Named entity recognition is defined as the entity with a special meaning, from the name of the person, organization and location of the text. It is the basis of other tasks, such as information extraction, analysis and machine translation. With the rapid growth of biological texts, mining related biological information, such as proteins, genes, and DNA, has become an urgent issue from a large number of direct literature.

However, it is a challenge for us to recognize the effective and efficient biological terms. With the number of named entities in the continued growth, it is difficult to build a complete list, which contains all types of biological terms. In addition, many biological terms are also multi word phrase. Therefore, it will be difficult to determine the boundaries of named entities. In addition, the same words or phrases can express different types of biological named entities in different contexts. In addition, too many different forms of writing and biological terms are abbreviated to make it worse for the project.

Term refers to the terms and conditions of a certain living creature, which has a strong connection to the organism, including protein name, gene name, cell type, and so on. Along with the progress of modern biotechnology, more and more species of genes and proteins have been proved to be a key factor in biology. It is an important task to find out the use of automation technology.

However, biometric identification is different from the general designation of entity recognition. At the time, the study of biology is still far from enough. There is no uniform nomenclature, a wide variety, and the number is increasing, so it is difficult to construct a complete glossary of terms. Most of the known biological terms are composed of long and multiple phrases that are difficult to identify.

In this work, we successfully extract relevant or certain types of information from the text. Biological named entities, such as proteins, genes, RNA, DNA, and cells, are able to find out from the text.

II. CONDITIONAL RANDOM FIELDS

A. Introduction

Conditional random field is a probabilistic labeling and structured data, such as sequence, tree, and lattice partitioning framework. The basic idea is to define conditional probability distributions in a particular observation sequence, rather than a joint distribution in a tag and an observation sequence. The main advantage of controlled release fertilizer on the hidden Markov model is the nature of its collateral conditions, which leads to the assumption that the relaxation of the hidden Markov model is necessary to ensure the independence of the thermal insulation.

In addition, the controlled release fertilizer should avoid the label bias problem, exhibiting a weakness of the Maximum Entropy Markov model (MEMMs), and other conditional Markov models, based on the directed graph model. Controlled release fertilizer is better than MEMMs and hidden Markov model in many areas of the real world, including bioinformatics, computational linguistics and speech recognition.

Conditional random field, it is a kind of differential probability model, it is a kind of random field, which is used to label or analyze sequence data.

As Markov random field conditions with the airport for undirected graph model, graph whose vertices represent the random variables, the dependency relationships between the vertices of the connection on behalf of random variables, in conditions with the airports and the distribution of the random variable y for the conditional probability, given the observed value is a random variable x. In principle, the conditions with airport graph layout model is arbitrary, commonly used in the layout is chain of the resultant architecture, chain resultant architecture in both training and inference (inference), or decoding, there are efficient algorithms for calculus.

Conditional random fields with the hidden Markoff model is often mentioned, conditional random fields for the probability distribution of the input and output, as the hidden Markoff model that is a strong assumption.

Conditional random field is one of the commonly used algorithms in natural language processing field in recent years. It is commonly used in syntactic parsing, named entity recognition, and part of speech tagging. In my opinion, CRF like a reverse hidden Markov model (HMM), both of which are with the Markov chain as a hidden variable probability transition model, but HMM using hidden variables to generate observable state, the generation probability tag set statistics is a generative model; and CRF in turn through the observation state identification of latent variables, the probability by tag set statistical is a discriminative model. Due to two main models are the same, which are capable of application areas are often overlapping, but in named entities, syntactic analysis, field CRF superior.

First, the Markoff chain, here reflects the random field characteristics of the CRF (accurate to say that the Markoff random field). Here CRF and HMM assumes that the part of speech tag is satisfy the Markov property, i.e., the part of speech only and a part of speech have the transition probabilities and has nothing to do with the part of speech of other position, such as described behind the words with describe word with probability 0.5, with the "modified" probability 0.5, to word probability is 0. Therefore, we can easily obtain a probability transfer matrix, that is, the probability of B after any part of speech can be obtained by the A. This part of HMM is over, for CRF, it can be used in the two-dimensional conditional transfer matrix based on a further increase in the one-dimensional word features, such as when the AB is adjacent, A is a verb and the word length of B more than 3, B is the probability of xx". You may notice a Markov chain window is 1, it is only considered on the 1 words, this is not necessarily the most reasonable. In fact, this is a sparse eigenvalue problem of compromise, it is conceivable only to get feedback from a lot of data, the two part of speech AB statistics P B|A. And if the statistics of length 6 window, such as P (g | ABCDEF) will be encountered the problem of sparse data, because probably sequence ABCDEF is simply not in the dataset appeared. Effects of sparse data for machine learning is huge, so Markov chain actual to the global loss of information based on in exchange for the more full of data. The experimental results show that the deal in part of speech tagging is earned.

Besides the direct mapping between part of speech and word, the condition of CRF is reflected here. If the HMM is here will be conditional probability matrices for -- > direct statistical part of speech of the word, such as "generation of" verb "launch" the probability of May is 1.5%, and generate "Microsoft" the probability is 0. Then for each possible part of speech sequence binding and conditional probability multiplication can to each candidate sequence generation probability, however take the highest probability as annotation can result. CRF is reversed, the CRF through explore the word itself features (such as the length, size, matching specific vocabulary can also be including word itself), to each word into a one-dimensional feature vector (vector), then for each feature calculation features to the part of speech of the conditional probability, so that each word of the candidate POS of conditional probability that all conditional probability of feature and. For example, we assume that the feature vectors only two and P ("word length > 3" -- > noun) probability of 0.9, P ("words in at the end of the sentence" -- > noun) probability of 0.4 and a word just to meet these two characteristics, the conditional probability of term for (+ 0.9 0.4) / 2 = 0.65. Such CRF according to the transfer values combined with Markov properties of part of speech, you can use similar to HMM method to find the optimal part of speech tagging sequence.

B. Biological Term Recognition

The algorithm for chronic renal failure is considered a number of text features to make about classification of whether a word includes all or part of a malignant tumor. A number of word based features are used to determine whether a word has been labeled as a training material for a malignant tumor referred to by a manual annotation of text.

In addition to the frequency of each word, there are other features that are used to identify, as shown below.Named entity recognition is a sub task of information extraction (Extraction Information), which the Element Atomic (atomic elements) positioning and classification, and then output to a fixed format directory, for example, name, organization, location, time, number, currency, percentage, etc..

III. Reinforcement Learning

A. Introduction

Behavioral psychology, reinforcement learning about how agents should act, in an environment that maximizes the cumulative reward for some of the concepts and the inspired. Reinforcement learning to learn what to do and how to map the case to take action to maximize the value of the reward signal. The learners do not have to take action in most forms of machine learning, but instead have to discover which operations produce most of the rewards from try on. The most interesting and challenging cases, action may affect not only the immediate reward is also the next situation, and through, all subsequent rewards.

Reinforcement learning is defined as a learning method, which is not a representation of learning methods. Reinforcement learning is different from the standard supervised learning, and it has never been proposed to correct the correct input / output pairs, nor is it a sub optimal action to correct. Reinforcement learning is to learn how to map the scene to the action, in order to obtain the maximum value of the reward signal. Like most machine learning methods, learners are not informed of what is to be used, but by trying to find the best way to get the best reward. In the most interesting and challenging examples, the action will not only affect the direct reward, but also affect the next scene, so that all subsequent rewards. Trial and error search and delayed reward is the two most important distinguishing features in reinforcement learning.

Reinforcement learning is defined not by the description of the learning method, but by the definition of a learning problem. Any way to solve the problem of this study we all think that is a reinforcement learning method. Although we are going to be in the third chapter, we can describe a reinforcement learning problem completely according to the optimal control theory of the Markov decision process, but it is the most important basic idea to study the interaction between agent and environment. Obviously, such a agent must be able to sense the state of the environment in a certain degree, and must be able to take action that can affect the state. The agent must also have one or more targets related to the environment. So the formula for this design must include three aspects: perception ability, action and goal, which is the most simple and the possible form. Reinforcement learning is different from supervised learning (learning supervised), supervised learning is the most important research in machine learning, statistical pattern recognition and artificial neural network. Supervised learning is to learn from the knowledge provided by the external supervisor sample. This is an important learning method, but it cannot be used alone for interactive learning. The sample of the expected behavior of the interactive problem is often impractical, they are both correct and representative of the agent must take action. In some unknown areas we want to learn to be able to bring benefits, agent must be able to learn from their own experience. A challenge in reinforcement learning is to explore and utilize the balance between exploration and utilization in other types of learning. In order to get a lot of reward, reinforcement learning agent must choose the action that it has tried in the past, and it is found that the action is effective in the process of producing a reward. In order to find such an action, it must also try to try before the action has not chosen. In order to get a reward, agent must use the information it already knows, and it must be explored in order to choose the better in the future. It is difficult to make a research or only use can not guarantee that the task will not fail. Agent must try all kinds of movements, and gradually approach to the best of the performance of the action. In a random task, each action must be tried many times to obtain a reliable estimate of its expected reward. For many years, mathematicians have been exploring the dilemma of using this dilemma (see chapter second). Now, we simply believe that even in the supervision of learning the entire balance of exploration and utilization is not there as it is usually defined.

Another key feature of reinforcement learning is that it clearly puts forward the problem that the whole problem is the problem of the interaction between the agent and the uncertain environment. Compared with many other methods, the method only considers the sub task, and has not solved how to integrate these sub tasks into a larger framework. For example, many of the machine learning studies that we have previously mentioned are not explicitly stated in the machine learning related to the study of supervised learning. Other researchers have developed a general goal programming theory, but do not take into account the problem in real-time decision making in the planning task, or where the necessary predictive models will come from. Although these methods have produced many useful results, they will focus on the practice of isolating the sub problems on a serious constraint.

B. CRFs for Reinforcement Learning

Conditional random fields are the labels for the probability of a graphical model to model the probability of a given opinion. They traditionally use a set of observations and labels to receive training. All Crf is based on the assumption that the training data is independent of the same distribution.

Here we use controlled release fertilizer for a class of temporal learning algorithms, in reinforcement learning. Therefore tags are operating, updated environment, the impact of the next observation. Therefore, from reinforcement learning, controlled release fertilizer is provided to the model in a decentralized Markov decision process to take joint action. They define how the agents can communicate with each other and choose the optimal joint action.

Reinforcement learning uses the opposite strategy, which begins with a complete, interactive, target search agent. All reinforcement learning agent have a clear goal, can perceive the environment in all aspects, and can choose the action to influence the environment. In addition, it is usually assumed that the agent is a very uncertain environment, but it must be "operation" ". If reinforcement learning involves planning, it has to deal with the interaction between planning and real-time action selection, but also to solve the problem of how to obtain and improve the environment model. If reinforcement learning involves learning, it is important to determine what capacity is important for some particular reason, so it must be done. In order to improve the study, we must study the important sub problems, although we can not cover all the details of the whole agent, but it is clear that they are in complete, interactive, target search of the status of agent.

One of the major trends, including reinforcement learning, is the growing connection between artificial intelligence and engineering disciplines. Not long ago, artificial intelligence was almost entirely independent of the field of control theory and statistics. It must deal with logic and symbols rather than numbers, artificial intelligence is a large LISP program, not linear algebra, differential equations or statistics. After decades, this view has gradually weakened. Modern artificial intelligence researchers have accepted methods of statistical and control, for example, as a means of competition or simply as a means of dealing with it. The areas of artificial intelligence and conventional engineering methods that have previously been overlooked are now among the most active areas of research, including neural networks, intelligent control, and the topic we are talking about. In reinforcement learning, we extend the optimal control theory and stochastic approximation to solve more and more challenging targets in artificial intelligence.

IV. EXPERIMENT AND RESULT

A. Evaluate Criterion

We evaluate the accuracy rate, recall rate and F1 rate of the results. Accuracy is the proportion of the correct results and the overall effect of the recall is the correct results of the proportion and the number of classification should be classified. The accuracy rate and recall rate are shown in the combination evaluation of F1 speed. The following equation is exact rate, recall rate and F1 rate.

$$Precision = \frac{number\ of\ true\ positives}{number\ of\ true\ positives\ +\ number\ of\ false\ positives}$$

$$Recall = \frac{number\ of\ true\ positives}{number\ of\ true\ positives\ +\ number\ of\ false\ positives} \qquad (1)$$

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

We are a collection of data sets and test data sets from a collection of data from the hot, corpus, corpus. We divide the hot 1999 file into two parts, which, 1000 file is for training, while the rest of the test.

We get a total of 469608 after the terms of the Geniatagger tag. Our training document has three columns: the word's part of speech and the result of the standard. After the training, we get a model file containing 233300 terms. Then we can use it to predict the other 236307 terms. Test results are in Table 1. Comparison of different types of terms is shown in figure 1.

TABLE I. TESTING RESULTS.

	G#DNA	G#RNA	G#cell_line	G#cell_type	G#protein	O
P	89.4%	95.5%	91.7%	92.3%	81.1%	95.0%
R	86.4%	89.6%	89.7%	85.9%	91.7%	99.9%
F	87.9%	92.5%	90.7%	89.0%	86.1%	97.4%

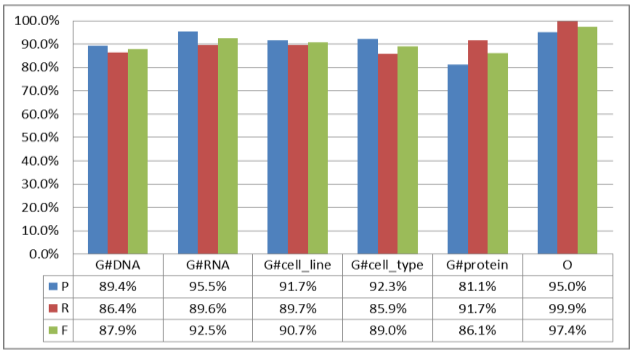

Figure 1. Comparasion of testing results.

V. CONCLUSION

With the rapid development of bioinformatics, more and more researchers have focused on the text of the mining, hoping to find out the biological knowledge in various unstructured documents. How to identify the biological terms and conditions effectively from the text is one of the important issues in the field of bioinformatics. The accuracy of the best biometric system has now reached more than 80%, but it is less than one of the general system. In this work we propose a novel approach to biological terms from text.

Conditional random field model has been widely used in various fields. They traditionally use a set of observations and labels to receive training. We use controlled release fertilizer for a class of temporal learning algorithm, in reinforcement learning. Therefore tags are operating, updated environment, the impact of the next observation. The test results in the state of the state show that the system can achieve the entity recognition of biological named effective and efficient.

ACKNOWLEDGMENT

Xuanyi Li(Student No. 1427402002) is a undergraduate student of School of Computer Science and Technology, Soochow University. This work is supervised by Xu Lu.Corresponding author: Xun Lu, luxun@suda.edu.cn.

REFERENCES

[1] Zarrouk E, Ayed Y B, Gargouri F. Hybrid continuous speech recognition systems by HMM, MLP and SVM: a comparative study[J]. International Journal of Speech Technology, 2014, 17:1-11.

[2] Stansfield W D. Using Crossword Puzzles to Enhance Students' Learning of Technical Biological Terms[J]. American Biology Teacher, 2014, 76(3):208-209.

[3] Rak R, Batistanavarro R T, Carter J, et al. Processing biological literature with customizable Web services supporting interoperable formats.[J]. Database the Journal of Biological Databases & Curation, 2014, 2014(26):4302-4315.

[4] Thomas P, Durek P, Solt I, et al. Computer-assisted curation of a human regulatory core network from the biological literature.[J]. Bioinformatics, 2015, 31(8):1258-1266.

[5] Žitnik S, Žitnik M, Zupan B, et al. Sieve-based relation extraction of gene regulatory networks from biological literature[J]. Bmc Bioinformatics, 2015, 16.

[6] Kuo H C, Lin K I. Extracting Protein Names from Biological Literature[J]. Advances in Computer Science An International Journal, 2014, 3(2).

[7] Sofia R, Cécile G V, Nam J L, et al. Safety of synthetic and biological DMARDs: a systematic literature review informing the 2013 update of the EULAR recommendations for management of rheumatoid arthritis.[J]. Annals of the Rheumatic Diseases, 2014, 73(3):529-535.

[8] Li C, Liakata M, Rebholzschuhmann D. Biological network extraction from scientific literature: state of the art and challenges.[J]. Briefings in Bioinformatics, 2014, 15:856-877.

[9] Paschke T, Scherer G, Heller W. Effects of Ingredients on Cigarette Smoke Composition and Biological Activity: A Literature Overview[J]. Beiträge Zur Tabakforschung, 2014, 20(3):107-247.

[10] Silveira M C. Named Entity Recognition[J]. Named Entity Recognition - ResearchGate, 2014, 50(5):807–819.

[11] Mohit B. Named Entity Recognition[J]. Theory & Applications of Natural Language Processing, 2014:221-245.

[12] Derczynski L, Maynard D, Rizzo G, et al. Analysis of Named Entity Recognition and Linking for Tweets[J]. Information Processing & Management, 2014, 51(2):32–49.

[13] Shaalan K, Mai O. A hybrid approach to Arabic named entity recognition[J]. Journal of Information Science, 2014, 40(1):67-87.

[14] Yan X, Yining W, Tianren L, et al. Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries[J]. Journal of the American Medical Informatics Association, 2014, 21(e1):e84-92.

[15] Kober J, Peters J. Reinforcement learning in robotics: A survey[J]. International Journal of Robotics Research, 2014, 32(11):1238-1274