Contributed Paper

Constructing Biological Knowledge Base using Named Entities Recognition and Term Collocation

Supattanawatee Thipcharoen [a,b], Watshara Shoombuatong [c], Samerkae Somhom [b], Rattasit Sukhahuta [b], Jeerayut Chaijaruwanch* [a,b]

[a] Bioinformatics Research Laboratory, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand.
[b] Department of Computer Science, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand.
[c] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand.

*Author for correspondence; e-mail: dao546@hotmail.com

Received: 6 October 2014
Accepted: 16 March 2015

ABSTRACT

Over the last few decades, the publishing of biological literature has dramatically increased due to technological developments. Thus, a crucial process is to extract information from this large number of writings by utilizing a biological named entity (NER) approach to automatically label corresponding biological terms. It is desirable to propose an effective method to identify biological named entities and automatically establish the specific knowledge base from biological literature. Herein, we made efforts in investigating biological information extraction for establishing specific knowledge as follows: 1) proposing NER method based on the efficient conditional random fields (CRFs) model, called NER-CRF, for performing on the benchmarking data (JNLPBA2004). The proposed NER method provided a higher result with 90.42% recall, 97.74% precision, and 94.30% F-measure, compared with the existing method with 75.99% recall, 69.42% precision, and 72.55% F-measure; 2) applying the Poisson approach for constructing an interpretability biological knowledge network to give good understanding to the global properties collocation of biological terms from the literature. Our finding provided the collocations of biological terms from the literature providing some insights to the specific biological literature.

Keywords: biological information extraction, biological named entity recognition, conditional random fields, poisson collocations

1. INTRODUCTION

Over the last few decades, a large volume of biological literature was published by many resources of biological information from which automated knowledge extraction methods were developed. One of the main resources was PubMed which included many citations from MEDLINE and other scientific journals, especially in biomedical and biological articles. Biological literature had many specific domains with a wide range of technical terms including protein and gene names [1]. Previously, there were many research methods developed to handle biological extraction such as SEBIO [2], Information Hyperlinked Over Proteins (iHOP) which provides
the network of genes and proteins as a natural way of accessing the millions of abstracts [3]. Practically, an accurate identification of the word boundaries lead to accurate knowledge extraction. However, this is a complicated process [4], because one word or phrase could refer to many entities, depending on its context. Many biological named entities have various spelling forms, their compound names can be very long, and different abbreviations are frequently used in the biomedical domain, all of which make it difficult to recognize them precisely. Thus, the use of Name Entity Recognition (NER) is a fundamental task which helps to specify the best boundary of biological terms from the literature. Hence, one way that NER can be approached is by considering it as a sequence labelling task [5]. To address the issues presented above, many machine learning models have been developed that can deal with NER [6]. The benefit of these kinds of models is that they do not require a dictionary in memory and they can automatically form rule sets by learning from a large text corpus. For example, ABNER is an open source tool for automatically tagging genes, proteins and other entity names in text [7].

The Conditional Random Fields (CRFs) method is a machine learning model which has been successfully utilized to analyze biological text. Examples are forming named entity recognition and rich feature sets [8], identifying gene and protein mentions in text [9], integrating linguistic knowledge to identify biomedical named entities [10] and also identifying the result in biomedical abstracts [11]. However, the general problem of NER remains a challenging task and is still an open and active area of research. The biological terms which mutually occur in the literature can be distinguished by using the collocation approach [12]. Generally, a collocation refers to two or more words having a spatial relationship. The collocation is still a serious problem in natural language processing (NLP). An application of collocation could be considered for many research topics such as the collocations extraction for language generation and the statistical significant collocation measurements [13], automatic term recognition with a probabilistic framework, and natural language technology and query expansion [14].

The objectives for this research are to identify biological named entities with a performance labelling model and to establish the specific knowledge base from biological literature. The CRFs method, called NER-CRF, was used to recognize the biological terms from the literature. Furthermore, the Poisson approach was applied to identify the most relevant of biological terms collocating in the literature and generating an interpretability biological knowledge network to give good understanding to the collocation of biological terms from the literature. There are numerous potential applications of the Poisson approach which identify the most relevant of collocated biological terms such as miRNA-Cancer Association [15], and miRNA targets and disease genes [16].

2. MATERIALS AND METHODS

2.1 Materials

In the NER process, the NER-CRF was used to perform on the JNLPBA2004 shared task dataset, which is the updated version of the GENIA corpus [5]. The training and testing data sets contained 2,000 abstracts and 404 abstracts, respectively. The training set was obtained by selecting and manually annotating according to a small taxonomy of 48 classes based on a chemical classification. There were 36 out of 48 classes used to annotate the GENIA corpus. For testing set, the currently annotated collection of MEDLINE abstracts from the GENIA project was considered. The 404 abstracts consisted of 5,067 protein names: 1,056 DNA names,
1,921 Cell Type names, 500 Cell Line name, and 118 RNA names. The corpus was categorized into five classes: protein, DNA, cell type, cell line, and RNA. The label pattern label for each class was labeled with “B-” and gathered by its class, such as B-protein or B-DNA. Meanwhile, in the case of biological terms having two or more words, the label pattern was represented with “I-”. For example, if the class was the Human T3 protein, the class label was in the format of I-protein. In this study, a set of label patterns consisted of “B-protein”, “I-protein”, “B-cell line”, “I-cell line”, “B-cell type”, “I-cell type”, “B-DNA”, “I-DNA”, “B-RNA”, “I-RNA” or “O”.

2.2 Conditional Random Fields for Named Entity Recognition

The CRFs method was an undirected graphical model [17, 18] for utilizing in sequence analysis [19, 20] and Biomedical Named Entity Recognition (NER) [8]. The advantage of CRFs was the integration of a wide variety of arbitrary, non-independent features from an input dataset. This method could alleviate the independent assumptions of Hidden Markov Models (HMMs) [21]. Furthermore, a number of studies reported that CRFs could prevent the label bias problem and also outperformed HMMs on a number of real-world sequence labelling tasks [19, 20].

Conceptually, the conditional probability of a particular label sequence \( y = y_1, y_2, y_3, ..., y_T \) was given by observation sequence \( x = x_1, x_2, x_3, ..., x_T \) defined by a normalised product of potential functions.

\[
p(y \mid x) = \frac{1}{z(x)} \exp\left(\sum_t F(y, x, t)\right)
\]

Where \( z(x) \) is a normalisation factor, and \( F(y, x, t) \) the potential function of \( x \) and \( y \) at position \( t \). The CRFs based NER model was trained with an available known \( x \) and \( y \) dataset, from which the conditional probability distribution \( p(y \mid x) \) was inferred. Figure 1. shows the example of observation sequence \( x \) and its label \( y \). For predicting an unknown sample, the label sequence \( y \) was labelled with \( y \) having the highest conditional probability \( p(y \mid x) \). Practically, the conditional probability \( p(y \mid x) \) could be calculated by using the dynamic programming of Viterbi-like algorithm [20, 21], defined as follows:

\[
y^* = \arg \max_y \ p(y \mid x) = \exp\left(\sum_{t=1}^T F(y, x, t)\right).
\]

In this study, the conditional probability \( p(y \mid x) \) for providing the optimum label \( y^* \) was generated by applying the FlexCRFs package [22].

**Figure 1.** Example of biomedical named entity recognition.
Potential functions of CRFs for Named Entity Recognition

Potential in 1st CRFs

The potential function \( F(y, x, t) \) was the summation product of transition potential functions \( f_{i,j}^l(y_{t-1}, y_t, x) \) and state potential functions \( g_i^l(y_i, x) \) calculated as the following equation:

\[
F(y, x, t) = \sum_{l,j} \lambda_{i,j}^l f_{i,j}^l(y_{t-1}, y_t, x) + \sum_{l} \mu_i^l g_i^l(y_i, x)
\]

where \( \lambda_{i,j}^l \) and \( \mu_i^l \) were represented with \( i \) the transition function and \( i \) state function parameters, \( i \) was a index of the parameters, \( l \) were any label class. The transition potential function \( f_{i,j}^l(y_{t-1}, y_t, x) \) was the characterization of the occurrence of the observation at position \( t \) and labels \( y_{t-1} \) and \( y_t \) at position \( t \) and \( t-1 \), respectively, as defined:

\[
f_{i,j}^l(y_{t-1}, y_t, x) = \begin{cases} 
1 & \text{if } y_{t-1} = l', y_t = l \ \text{otherwise} \\
0 & \text{otherwise}
\end{cases}
\]

\[
z(x, t) = \begin{cases} 
1 & \text{if } x \text{ at position } t \pm 2 \text{ is equal to words} \\
0 & \text{otherwise}
\end{cases}
\]

where \( z(x, t) \) represented whether \( x \) (current state) at position \( t \) to \( t \) consisting of the words. For example, in Figure 1, if \( x \) was 2-M, then \( f_{i,j}^l(y_{t-1}, y_t, x) \) gave 1 when \( y_{t-1} \) and \( y_t \) were B-protein and I-protein, respectively, otherwise \( f_{i,j}^l(y_{t-1}, y_t, x) \) gave 0, as formulated with:

\[
f_{i,j}^l(y_{t-1}, y_t, x) = \begin{cases} 
1 & \text{if } y_{t-1} = B - \text{protein}, y_t = I - \text{protein} \\
0 & \text{otherwise}
\end{cases}
\]

In this work, we considered a reasonable local information pattern of ±2 name entities surround the considering position \( t \). Similar to \( f_{i,j}^l(y_{t-1}, y_t, x) \), the state potential function \( g_i^l(y_i, x) \) characterizes the occurrence of the observation at position \( t \) and labels \( y_t \) at position \( t \), respectively, as defined:

\[
g_i^l(y_i, x) = \begin{cases} 
z(x, t) & \text{if } y_i = l \\
0 & \text{otherwise}
\end{cases}
\]

Potential in 2nd CRFs

In this case, \( F(y, x, t) \) was reformulated to be:

\[
F(y, x, t) = \sum_{l,j} \lambda_{i,j}^l f_{i,j}^l(y_{t-1}, y_t, x) + \sum_{l} \mu_i^l g_i^l(y_i, x)
\]

\[
+ \sum_{l,j,l} \lambda_{i,j}^l f_{i,j}^l(y_{t-2}, y_{t-1}, y_t, x) + \sum_{l,j,l} \mu_i^l g_i^l(y_{t-1}, y_t, x)
\]

where \( j \) was the index of the parameters in the 2nd CRFs:

\[
f_{i,j}^l(y_{t-2}, y_{t-1}, y_t, x) = \begin{cases} 
z(x, t) & \text{if } (y_{t-2} = l'', y_{t-1} = l', y_t = l) \\
0 & \text{otherwise}
\end{cases}
\]
A number of works reported that the 1st CRFs was suitable for labeling the short-range word pair or single words, while the 2nd CRFs could handle such problems as labeling the long-range words as biological named entity recognition with the additional summation product in Eq. 7 [19, 20]. In this study, an NER method was built with 1st and 2nd CRFs, called NER-1CRF and -2CRF, respectively, for predicting biological named entities on the testing dataset.

2.3 Using the Poisson Approach for Extracting Biological Term Collocations

Most information extracted from text corpora were shown in the form of co-occurrence counts. One of catalogues of interesting co-occurrences was collocations [12]. The collocation was the well-known challenge and problem in the field of NLP. Another most popular collocation measure in text analysis is Pointwise Mutual Information (PMI) and has been the standard association measure in collocation extraction [23]. However, the limitation of the PMI approach occurred when a dataset had low-frequency events, and the highest PMI score may not be a guarantee to indicate event importance. Thus, it is desirable to propose the measurements of collocations extraction approaches, explaining the models and providing significance value for the collocation of biological terms.

Herein, the Poisson distribution was proposed to be the measurement in collocation extraction. The Poisson distribution was a discrete probability distribution that expressed the probability of a number of events occurring in a fixed period of time. In the biological literature, it could be observed that two or more words having the relationship will be adjacent each other. Thus, the Poisson distribution [12] could be applied to a collocation of biological terms. Consider the joint occurrence of two given words $A$ and $B$ with probabilities $P_A$ and $P_B$ on an abstract. Using the Poisson distribution to measure the collocation was defined with the probability for $k$ joint occurrences in the corpus of $N$ abstracts, which could be expressed as

$$p_k = \frac{1}{k!} \lambda^k \cdot e^{-\lambda}.$$  \hspace{1cm} (11)

where

$$\lambda = nP_A \cdot P_B.$$  \hspace{1cm} (12)

In this work, the case of at least $k$ joint occurrences was calculated with

$$\sum_{l=k}^{\infty} \frac{1}{l!} \lambda^l \cdot e^{-\lambda}.$$  \hspace{1cm} (13)

The significance measure for collocations $\text{Sig}(A,B)$ can be then defined using the negative logarithm:

$$\text{Sig}(A,B) = -\log \sum_{l=k}^{\infty} \frac{1}{l!} \lambda^l \cdot e^{-\lambda} = \frac{\lambda - k \log \lambda + \log k!}{\log N}.$$  \hspace{1cm} (14)

where $\lambda$ and $k$ were a number of the probability of each word in corpus and the event of
words which collocate with each other, respectively. The values of $\lambda$ and $k$ were the degree of the relationship between the biological terms in the literature dataset. The relation of each term collocation was identified using significance value implemented by Perl programming. Finally, the collocations of biological terms are depicted in the graph network which are representing the specific topics of knowledge by the Cytoscape program [24].

3. Results and Discussion

3.1 Performance of Named Entity Recognition Model based on CRFs

Three measurements were used to evaluate the prediction performance of the proposed NER based CRFs model (NER-CRF), namely precision, recall, and F-measure. In the evaluation of the model, we consider only exactly matched results for a chunk-based performance evaluation.

The NER model was built with the 1st and 2nd CRFs for comparing the efficient model of these two types of CRFs. The performance comparisons of NER-1CRF and -2CRF are shown in Table 1. As seen, NER-2CRF achieved the recall, precision, F-measure values higher than 90% for labelling Protein, DNA, and RNA entities. Remarkably, the NER-2CRF model afforded a precision which was greater than 95.00% and close to 99.00% for the label of Protein and RNA. Meanwhile, the prediction result of the NER-1CRF gave recall, precision, F-measure which was lower than 70% and provided the prediction result of recall as low as 46.00%. For the overall prediction result, the NER-1CRF model gave the average results with 55.92±5.95% recall, 66.89±8.80% precision, and 60.80±6.53% F-measure. Meanwhile, using the NER-2CRF model provided the average results with 90.42±5.04% recall, 97.74±0.75% precision, and

| Table 1. Performance comparison of NER based on 1st and 2nd CRFs. |
|------------------|------------------|------------------|------------------|
| Label            | Recall (%)       | Precision (%)    | F-Measure (%)    |
|                  | 1st  | 2nd  | 1st  | 2nd  | 1st  | 2nd  |
| Protein          | 56.01| 96.63| 68.16| 98.67| 61.49| 97.64|
| DNA              | 57.58| 91.10| 70.45| 97.76| 63.37| 96.36|
| Cell Type        | 58.15| 87.56| 77.95| 97.28| 66.61| 92.16|
| Cell Line        | 46.00| 83.60| 53.86| 96.76| 49.62| 89.70|
| RNA              | 61.86| 93.22| 64.04| 98.21| 62.93| 95.65|
| Mean             | 55.92±5.95 | 90.42±5.04 | 66.89±8.80 | 97.74±0.75 | 60.80±6.53 | 94.30±3.28 |

| Table 2. Performance comparison of NER-2CRF with others existing methods. |
|------------------|------------------|------------------|
| Method           | Recall (%)       | Precision (%)    | F-Measure (%)    |
| Deep knowledge resources [4] | 75.99 | 69.42 | 72.55 |
| Local and syntactic features [25] | 68.60 | 71.60 | 70.10 |
| CRF-Simple orthographic features [8] | 70.00 | 69.00 | 69.50 |
| Class-attribute stacking [6] | 75.57 | 79.68 | 77.57 |
| NER-2CRF         | 93.14 | 98.16 | 95.58 |
94.30±3.28% F-measure. It could be stated that our proposed method can accurately identify the five interested biological terms. The best prediction results of using the NER-2CRF model showed capability of 2nd CRF in biological named entity predicting. Thus, the NER-2CRF model was used as a comparison with other existing methods. The performance comparisons of our proposed NER model with four well-known NER methods are shown in Table 2. The method of the deep knowledge resources approach was established from both a closed dictionary and an open dictionary. This method provided 75.99% recall, 69.42% precision, and 72.55% F-measure. In 2005, Settles proposed the framework built from CRFs for recognizing multiple entity classes in biomedical literature by using only simple orthographic features [8]. Table 2 shows that using only simple orthographic features provide better performance values with 70.00% recall, 69.00% precision, and 69.50% F-measure, compared with using only semantic lexicons [8]. Overall, the NER based on deep knowledge resources yielded a higher performance, compared with the three related NER methods. However, our proposed NER-2CRF achieved the highest prediction result with 93.14% recall, 98.16% precision, and 95.58% F-measure. This result showed the superiority of the NER-2CRF model in biological named entity predicting.

3.2 Construction of Networks of Biological Terms Collocations

The 3,939 biological terms were used to construct networks of biological term collocations derived from the testing data set. Herein, The Poisson approach was utilized to estimate the significance value for each of the biological terms collocations. The significance value was calculated by using Eq 14. Table 3 shows the 10 top-ranked collocation of biological terms having the highest significance value. The largest value of significance value was the strongest collocation. As seen, 3 out of 10 top-ranked collocations consisted of cytokines and its collocated biological terms, i.e. cytokine - transcription factors (16.15), cytokines – mRNA (13.58), and cytokines – NfkappaB (13.10). Meanwhile, the lymphoblasts and its collocated biological terms are found three times in Table 3, consisting of lymphoblasts – cytokines (17.83), lymphoblasts – tumor (14.31), and lymphoblasts - glucocorticoid receptors (15.20). The collocation of lymphoblasts – cytokines gave the highest significance value with 17.83. lymphoblasts [Cell

Table 3. The 10 top-ranked biological terms collocations.

<table>
<thead>
<tr>
<th>Collocation of biological terms</th>
<th>Sig(A,B)</th>
<th>k</th>
<th>(\lambda)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>lymphoblasts[Cell type] -- cytokines[protein]</td>
<td>17.83</td>
<td>2.00</td>
<td>49.55</td>
<td>[26, 27, 31]</td>
</tr>
<tr>
<td>cytokines[protein] -- transcription factors[protein]</td>
<td>16.15</td>
<td>26.00</td>
<td>0.26</td>
<td>[32, 33, 34]</td>
</tr>
<tr>
<td>lymphoblasts[Cell type] -- glucocorticoid receptors</td>
<td>15.20</td>
<td>1.00</td>
<td>44.98</td>
<td>[35, 36, 37]</td>
</tr>
<tr>
<td>lymphoblasts[Cell type] -- tumor[Cell type]</td>
<td>14.31</td>
<td>1.00</td>
<td>38.88</td>
<td>[38, 39, 40]</td>
</tr>
<tr>
<td>NfkappaB[protein] -- transcription factors[protein]</td>
<td>13.86</td>
<td>23.00</td>
<td>0.26</td>
<td>[41, 42, 43]</td>
</tr>
<tr>
<td>monococytes[Cell type] -- transcription factors[protein]</td>
<td>13.72</td>
<td>26.00</td>
<td>0.46</td>
<td>[44, 45, 46]</td>
</tr>
<tr>
<td>cytokines[protein] -- mRNA[RNA]</td>
<td>13.58</td>
<td>20.00</td>
<td>0.14</td>
<td>[47, 48, 49]</td>
</tr>
<tr>
<td>GC-rich region[DNA] -- promoters[DNA]</td>
<td>13.47</td>
<td>1.00</td>
<td>36.66</td>
<td>[50, 51, 52]</td>
</tr>
<tr>
<td>promoters[DNA] -- kinase[protein]</td>
<td>13.16</td>
<td>20.00</td>
<td>0.16</td>
<td>[53, 54, 55]</td>
</tr>
<tr>
<td>cytokines[protein] -- NfkappaB[protein]</td>
<td>13.10</td>
<td>19.00</td>
<td>0.13</td>
<td>[56, 57, 58]</td>
</tr>
</tbody>
</table>
type] – cytokines [protein] relate with “on|in” words. lymphoblasts [Cell type] – cytokines [protein] occur together in the corpus such as Effect of testicular cytokines on proliferation of rat T-leukaemic lymphoblasts in vitro [26]; expression of inflammatory cytokines altered in lymphoblasts of autistic subjects[27]. Semantic–Relation Extraction helps to extract more about the relation. Part of speech: Parse Tree is used for Semantic-Relation Extraction such as Effect on NN of IN testicular NN cytokines NNS on IN proliferation NN of IN rat NN T-leukaemic NN lymphoblasts NNS in IN vitro NN [26]; expression NN of IN inflammatory JJ cytokines NNS altered VBD in IN lymphoblasts NNS of IN autistic JJ subjects NNS[27]. Part of Speech: Parse tree can extract relation between biological terms in the semantic way and some relations correspond with biological semantic-relation extraction existed such as Gene Ontology. However, the relation extraction should consider in state or event. Bio-molecular event extraction will be considered for future work in semantic-relation extraction to generate the completely graph network.

The relations are divided into two classes consisting of Lexical-Relation Extraction and Semantic-Relation Extraction. Herein, Table 3 demonstrates the Lexical-Relation Extraction based on the relationship among biological terms that could be used to calculate significance values. As seen in Table 3, the relationship between lymphoblasts [Cell type] – cytokines [protein] gave the highest value. It could be well reflected in many experimental reports (column 3). More details of other collocation of biological terms are shown in Figure 2. The Poisson approach was further applied to construct the networks on the NCBI corpus (275 abstracts). Figure 3 shows the network of collocations of Cyanobacteria that give some information between Photosystem II (PS II) and other associated proteins. The top 3 collocations of PS II consisted of Oxygen Evolving Complex (OEC), alpha [Protein], and CP47 [Protein]. The strongest collocation was PS II – OEC with a 25.63 significance value. Many recent publications in various fields have reported the collocation of PS II – OEC in the works of artificial photosynthesis [28], improving catalytic reaction [29, 30].

In this study, the well-known statistical parameter ($t$-test) was also used to compare with our proposal. The in-depth detail of this parameter was derived from the work [59]. Table 4 shows

Table 4. The 10 top-ranked biological terms collocations ranked by Poisson approach and $t$-test.

<table>
<thead>
<tr>
<th>Collocation of biological terms</th>
<th>Poisson (Rank)</th>
<th>$t$-test (Rank)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lymphoblasts [Cell type] -- cytokines [protein]</td>
<td>17.83 (1)</td>
<td>0.07 (12)</td>
</tr>
<tr>
<td>cytokines [protein] -- transcription factors [protein]</td>
<td>16.15 (2)</td>
<td>0.25 (1)</td>
</tr>
<tr>
<td>lymphoblasts [Cell type] -- glucocorticoid receptors [protein]</td>
<td>15.20 (3)</td>
<td>0.03 (15)</td>
</tr>
<tr>
<td>lymphoblasts [Cell type] -- tumor [Cell type]</td>
<td>14.31 (4)</td>
<td>0.04 (14)</td>
</tr>
<tr>
<td>NFkappaB [protein] -- transcription factors [protein]</td>
<td>13.86 (5)</td>
<td>0.24 (3)</td>
</tr>
<tr>
<td>monocytes [Cell type] -- transcription factors [protein]</td>
<td>13.72 (6)</td>
<td>0.25 (1)</td>
</tr>
<tr>
<td>cytokines [protein] -- mRNA [RNA]</td>
<td>13.58 (7)</td>
<td>0.22 (4)</td>
</tr>
<tr>
<td>GC-rich region [DNA] -- promoters [DNA]</td>
<td>13.47 (8)</td>
<td>0.05 (13)</td>
</tr>
<tr>
<td>promoters [DNA] -- kinase [protein]</td>
<td>13.16 (9)</td>
<td>0.22 (4)</td>
</tr>
<tr>
<td>cytokines [protein] -- NFkappaB [protein]</td>
<td>13.10 (10)</td>
<td>0.21 (6)</td>
</tr>
</tbody>
</table>
Figure 2. Networks of biological terms collocations.

Figure 3. Proteins co-occurrences in cyanobacteria derived from Poisson approach.
the ranks of the 10 top-ranked collocations obtained from the Poisson approach and were also ordered using t-test. As seen, 6 out of 10 top-ranked collocations consists of cytokines[protein] -- transcription factors[protein] (2,1), NFkappaB[protein] -- transcription factors[protein] (5,3), monocytes[Cell type] -- transcription factors[protein] (6,1), monocytes[Cell type] -- transcription factors[protein] (7,4), promoters[DNA] -- kinase[protein] (9,4) and cytokines[protein] -- NFkappaB[protein] (10,6). These were obtained from both Poisson approach and t-test, where (n, m) is the n and m order obtained from Poisson approach and t-test, respectively. The results from Poisson approach and t-test were slightly different. Since, Lexical-Relation Extraction was designed using computational approach; it may incur the incorrect associations. However, we can mitigate such a problem by confirming with the previous reports.

4. CONCLUSION

We have proposed a biological named entity recognition (NER) approach automatically establishing an interpretability biological knowledge network. In this work, we aim to propose a computational model NER-CRF for automatically extracting and establishing functional genomics knowledge. The NER approach was established with the efficient graphical model CRF, called NER-CRF. The NER-CRF performed well in predicting biological named entity compared with existing NER methods. Furthermore, the Poisson approach was utilized for constructing an interpretability biological knowledge network. The biological knowledge network gave good understanding to the global properties collocation of biological terms from the literature. Our finding provided the collocations of biological terms from the literature, affording some insights to the interested biological literature.

ACKNOWLEDGEMENTS

We thank R. Cutler for helpful comments and suggestions on this manuscript. This project was supported by The Graduate School of Chiang Mai University, National Center for Genetic Engineering and Biotechnology, and The Far Eastern University, Thailand.

REFERENCES


