GeneTUKit: a software for document-level gene normalization

Minlie Huang*, Jingchen Liu and Xiaoyan Zhu
Department of Computer Science and Technology, Tsinghua University, Beijing, China

ABSTRACT
Motivation: Linking gene mentions in an article to entries of biological databases can facilitate indexing and querying biological literature greatly. Due to the high ambiguity of gene names, this task is particularly challenging. Manual annotation for this task is cost expensive, time consuming and labor intensive. Therefore, providing assistive tools to facilitate the task is of high value.

Results: We developed GeneTUKit, a document-level gene normalization software for full-text articles. This software employs both local context surrounding gene mentions and global context from the whole full-text document. It can normalize genes of different species simultaneously. When participating in BioCreAtIvE III, the system obtained good results among 37 runs: the system was ranked first, fourth and seventh in terms of TAP-20, TAP-10 and TAP-5, respectively on the 507 full-text test articles.

Availability and implementation: The software is available at http://www.qanswers.net/GeneTUKit/.

Received on October 27, 2010; revised on December 21, 2010; accepted on January 10, 2011

1 INTRODUCTION
Gene normalization is one of the most challenging tasks in bio-literature mining due to the high ambiguity of gene names as they may refer to orthologous or entirely different genes, may be named after phenotypes and other biomedical terms, or may resemble common names with non-gene entities (Hakenberg et al., 2008). It is time consuming and labor intensive to annotate full-text articles manually. Therefore, a good assistive tool for this task may facilitate the process greatly.

There has been a large body of work addressing the problem of gene mention normalization. ProMiner (Hansich et al., 2005), a strict dictionary-based approach, relies on the quality of its gene dictionaries heavily. Xu et al. (2007) proposed a method using gene profiles generated from PubMed abstracts for gene disambiguation. GNAT (Hakenberg et al., 2008) is a rule-based and machine learning (ML) based gene normalization system which used extensive background knowledge. Built from open-source libraries and publicly available resources, GENO (Wermter et al., 2008) employed a carefully crafted suite of symbolic and statistical methods. Moara (Neves et al., 2010) is a Java library for extracting and normalizing gene and protein mentions, and currently designed for four model organisms.

Our software departs from previous systems in two aspects: first, it combines local and global contexts to normalize genes at document-level. The goal of this software is not to normalize every mention correctly, but to suggest a list of normalized genes given a target document, to assist human annotators. Most previous systems are normalizing genes at mention-level and only local context surrounding a mention (e.g. the sentence where the mention was recognized) were employed. However, due to the high ambiguity of gene names, it may be insufficient to use only local context: inter-sentential or document-level context can be helpful in this task. Second, the software is designed for simultaneously normalizing genes of many different species for full-text articles. It is not limited to any specific organism, but rather deals with all species present in a gene database (Entrez Gene in this article).

2 METHODS AND SYSTEM
The workflow of our software is shown in Figure 1. The software has four main modules. The first module is for gene mention recognition, the second one for gene ID candidate generation and the third one for gene ID disambiguation. In the fourth module, the software generates confidence scores for each gene ID, where the confidence score indicates the strength of the association between a gene ID and the document.

We have used three methods for recognizing gene mentions in the first module. The first method is a conditional random field-based approach where the dictionary was compiled from Entrez Gene. The second method is ARNER (Seltjes, 2005), an open source named entity recognition system for biomedical literature. The input text is processed by these methods separately, and the resulting mentions are maintained if a mention is recognized by at least two methods. If two mentions are similar but have different boundaries, the overlapping part is taken as the final mention.

The second module generates gene ID candidates for a recognized mention. In this module, an open-source indexing package, Lucene (http://lucene.apache.org/), was used to index all the genes in Entrez Gene. Each mention was then queried and top 50 gene IDs were returned as candidates. The text of mentions and Entrez Gene entries were, respectively,

Fig. 1. The workflow of GeneTUKit. Numbers in shaded boxes are gene IDs. The real-number values in the last box are confidence scores.
We evaluated the system on the BioCreAtIvE III GN corpus (Lu and Wilbur, 2010) in terms of Threshold Average Precision (TAP-k, k = 5, 10, 20, respectively) (Carroll et al., 2010). For training, we used the 32 articles with gold-standard human annotation. For testing, the first dataset has 50 articles, each of which has gold-standard annotation, and the second one has 507 articles whose ground truth was inferred from 37 team submissions (referred as silver standard). The 507 articles also include the 50 articles from the first dataset. The results presented in Table 1 show the official evaluation results from BioCreAtIvE III. We have also tested the performance in terms of average precision. The manual error analysis has revealed that two major error types are (i) wrongly recognized gene mentions, and (ii) wrong species mapping. The Supplementary Material provide a more detailed analysis at http://www.qanswers.net/GeneTUKit/evaluation.html.

### Table 1. The evaluation results on the BioCreAtIvE III GN corpus

<table>
<thead>
<tr>
<th>Measures</th>
<th>50 articles (gold standard)</th>
<th>507 articles (silver standard)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAP-5</td>
<td>0.2973 (4/37)</td>
<td>0.4086 (7/37)</td>
</tr>
<tr>
<td>TAP-10</td>
<td>0.3125 (4/37)</td>
<td>0.4511 (4/37)</td>
</tr>
<tr>
<td>TAP-20</td>
<td>0.3248 (4/37)</td>
<td>0.4648 (1/37)</td>
</tr>
<tr>
<td>Average precision of TOP k recommendations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 5</td>
<td>0.4880 0.5764</td>
<td></td>
</tr>
<tr>
<td>k = 10</td>
<td>0.4340 0.4993</td>
<td></td>
</tr>
<tr>
<td>k = 20</td>
<td>0.3231 0.3984</td>
<td></td>
</tr>
</tbody>
</table>

The number in the bracket is the rank of our score among the 37 submissions.

### 4 CONCLUSION

GeneTUKit is a software designed for document-level gene normalization, which employs features from the local context and the global context within the whole full-text article. It can normalize genes of many different species. Given a target article, the software outputs a list of normalized genes, and each predicted gene is associated with a confidence score.

**Funding:** Natural Science Foundation of China (No. 60803075), Chinese 973 project (No. 2007CB11003).

**Conflicts of Interest:** none declared.

### REFERENCES


**CONFLICTS OF INTEREST:** none declared.

**FINANCIAL SUPPORT:** Natural Science Foundation of China (No. 60803075), Chinese 973 project (No. 2007CB11003).

**ACKNOWLEDGMENTS:** none declared.

**REFERENCES:**


