Gene normalization as a problem of information retrieval

Cheng-Ju Kuo[1], Maurice HT Ling[2,3] and Chun-Nan Hsu*[1,4]

[1]Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
[2]School of Chemical and Life Sciences, Singapore Polytechnic, Republic of Singapore
[3]Department of Zoology, The University of Melbourne, Parkville, Victoria, Australia
[4]Information Sciences Institute, University of Southern California, Marina del Rey, California, USA

Email: Cheng-Ju Kuo - clarkkuo@iis.sinica.edu.tw; Maurice HT Ling - mauriceling@acm.org; Chun-Nan Hsu*- chunnan@isi.edu;

*Corresponding author

The goal of gene normalization (GN) task is to link the names of gene or gene products mentioned in the literature to standard database identifiers [1]. In the BioCreative II GN task, we realized that finding out gene mention candidates as many as possible may be a key to implement a GN system with high recall. We propose to use bi-directional parsing models of Conditional Random Fields (CRFs) to tag gene mentions (any possible composition of tokens likely to be a gene mention) in a region of text. It is useful to alleviate the requirement to manually create rule sets which may change with time. The GM module [2], which was ranked second among twenty-one participants in BioCreative II challenge, was built on BioCreative II gene mention tagging (GM) training corpus. In order to finding all candidates of gene mentions, we collected twenty prediction sequences from two parsing models and transform each sequence to one set of gene mentions from a sentence. That is, there are at most fourty sets of gene mentions that might be produced from a single input sentence. We merged them into one set, tested on BioCreative II GM test corpus and achieved a recall of 0.9419 in the internal test. The resulting fourty sets of gene mentions were merged as input to the GN system.

We used BIOADI [3] to identify all pairs of abbreviation and its long form in the article. If the long form of an abbreviation is not tagged as a gene mention in the sentence where the abbreviation pair is extracted, the abbreviation is marked as an invalid gene mention, then has to be removed from gene mentions extracted in GM step. We took advantage of contextual information among sentences to remove mentions which are not referred to gene or gene products. For example, "NAA" is a candidate gene mention tagged from a text of "...metabolites in 5 patients: N-acetyl-aspartate (NAA), creatine...". We can link the definition of "NAA" to "N-acetyl-asparate" and know that "N-acetyl-asparate" is not been tagged as a gene mention, we can ignore all "NAA"s tagged in the sentences of the same article. It is useful to improve

GM performance by reducing false-positives.

The rest of gene mentions will be resolved into its species and assigned a taxonomy identifier (taxid) of NCBI Taxonomy database. We combined the Taxonomy database and LINNAEUS species dictionary [4] to a taxid-name dictionary. We used the dictionary to resolve species word to unique taxonomy identifier of Taxonomy database by applying a heuristic approach described in [5]. There are three rules for species assignment. Firstly, assign the nearset taxid which is precending the gene mention. Secondly, assign the taxid which locates in the same sentence. Lastly, assign the taxid which has the highest occurrence of the literature.

After taxid assignment, we obtain a list of gene mentions with its taxid of each input sentence which will be used to query Entrez Gene database to know whether there is a similar text in Entrez Gene. If no record is found, the gene mention will be removed from the list. This step can remove most of non-gene mention records. As a rule, we choose the longest gene mention of any region where gene mentions were found by the GM system. This can avoid mention overlapping which might reduce the precision of GN module. After the process, we took an expensive matching process to resolve each gene mention for its Entrez Gene identifier candidates. We applied Lucene, with a customized tokenizer and analyzer, to speed up the matching process. Each name of retrieved Entrez Gene record will be compiled to a regular expression pattern for fuzzy match to the query text of gene mention. If the gene mention matches the pattern, the gene mention is assigned to the identifier. If more than one identifier are assigned to a gene mention, we set a heuristic rule to select one as the output.

We propose a system that can evaluate the quality of gene mention. Each GN result is assigned a confidence score. The system is built based on logistic regression and trained on BioCreative II GM training and test data. Each GN result will be pass to the model to evaluate the quality of its gene mention. The logisitic regression model will return a decision value between 0 and 1. We directly used the output value as a significance for the GN result. In the training stage, we achieved a precision of 0.643 with 0.588 recall (F-score = 0.614) on BioCreative III GN 32 full annotated articles. TAP-5, TAP-10 and TAP-15 scores on the training articles are shown in Table 1. Table 2 is the results for three submitted runs on the most difficult 50 articles selected by the task organizers.

References

1. Morgan A, Lu Z, Wang X, Cohen A, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J, Sun C, Liu Hh, Torres R, Krauthammer M, Lau W, Liu H, Hsu CN, Schuemie M, Cohen KB, Hirschman L: **Overview of**

Table 1: TAP-5, -10 and -20 on 32 training articles in inside test

Run	Unweighted Average TAP		
	TAP-5	TAP-10	TAP-20
1	0.3123	0.4151	0.4151

Table 2: TAP-5, -10 and -20 of three submitted runs on 50 test articles

Run	Unweighted Average TAP		
	TAP-5	TAP-10	TAP-20
1	0.2099	0.2447	0.2447
2	0.2048	0.2420	0.2420
3	0.2061	0.2432	0.2432

BioCreative II gene normalization. *Genome Biology* 2008, **9**(Suppl 2):S3, [http://genomebiology.com/2008/9/S2/S3].

2. Hsu CN, Chang YM, Kuo CJ, Lin YS, Huang HS, Chung IF: **Integrating high dimensional bi-directional parsing models for gene mention tagging**. *Bioinformatics* 2008, **24**(13):i286–i294.
3. Kuo CJ, Ling M, Lin KT, Hsu CN: **BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature**. *BMC Bioinformatics* 2009, **10**(Suppl 15):S7, [http://www.biomedcentral.com/1471-2105/10/S15/S7].
4. Gerner M, Nenadic G, Bergman C: **LINNAEUS: A species name identification system for biomedical literature**. *BMC Bioinformatics* 2010, **11**:85, [http://www.biomedcentral.com/1471-2105/11/85].
5. Wang X, Tsujii J, Ananiadou S: **Disambiguating the species of biomedical named entities using natural language parsers**. *Bioinformatics* 2010, **26**(5):661–667, [http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/5/661].

Figures

Figure 1 - System flowchart