Peregrine: Lightweight gene name normalization by dictionary lookup

Martijn J. Schuemie[1]
m.schuemie@erasmusmc.nl

Rob Jelier[1]
r.jelier@erasmusmc.nl

Jan A. Kors[1]
j.kors@erasmusmc.nl

[1] Biosemantics Group, Medical Informatics Department, ErasmusMC University Medical Center Rotterdam, 's-Gravendijkwal 230, 3015 CE Rotterdam, The Netherlands

Abstract

To achieve high speed with minimal effort, we created a system dubbed Peregrine that performs gene name normalization by simple dictionary lookup followed by several post-processing steps.

Keywords: gene name normalization, dictionary

1 Introduction

For molecular biologists to be able to cope with the massive amounts of information stored in scientific literature, it is not sufficient to simply have an efficient document retrieval system. For instance, to interpret a list of hundreds of up and down regulated genes in a high-throughput experiment, the required information is stored in thousands of relevant articles, too much to read. What is needed is a system that can distill the information from the literature and represent it in a compressed form.

One such system is the Anni tool, developed by the Biosemantics group (www.biosemantics.org/anni). This tool can be used for gene list annotation and knowledge discovery, and has already been applied to current biomedical problems [1].

For tools like these, it is necessary to uniquely identify gene and protein names in literature, and relate these to specific entries in molecular databases. Because the amount of literature that needs to be analyzed is large (Medline alone counts over 16 million records), the method for gene name normalization should be able to analyze large corpora in a reasonable amount of time.

We have therefore chosen to use a lightweight system we named Peregrine, which simply looks up word sequences in a dictionary that is automatically constructed from gene and protein databases. Several post processing steps are applied to reduce the number of false positives and false negatives.

The system is based on a previously published study on gene name normalization [2].

2 Methods

2.1 Tagging system

The Peregrine system translates the terms in the dictionary into sequences of tokens (i.e. sequences of words). When such a sequence of tokens is found in a document, the term, and thus the gene or protein associated with that term, is recognized in the text. Some tokens are completely ignored, since these are considered to be non-informative ("of", "the", "and", and "in"). If the term is considered a 'long form' (i.e. it contains a space and is longer than six characters), the tokens in the thesaurus and in the text are first reduced to their stem using the NLM Lexical Variant Generator program [3], to allow for small lexical variations.

2.2 Dictionary

We tested the system using two different dictionaries:

1. The dictionary provided by the BioCreAtIvE 2 organization, with 32,975 genes, and 182,989 (non-unique) gene names
2. Our own dictionary, constructed by combining five gene databases [2, 4], with 26,560 genes and 161,928 (non-unique) gene names

2.3 Manual filter

We tested the system on a random selection of 100,000 Medline records. We manually reviewed the 250 most frequently found terms, since these are most likely to be erroneous or highly ambiguous terms. We removed terms that are not really names of genes (e.g. “alternative splicing”, “open reading frame”, and “human”), or are extremely ambiguous (e.g. “CA2”, “obesity”, and “factor 1”). We removed 159 such terms from the BioCreAtIvE dictionary, 98 from our own combined dictionary.

2.4 Spelling variations

To allow for spelling variations not included in the dictionary, we applied two rules to generate new synonyms based on existing terms, as shown to be effective in a previous study [2]:

1. Arabic numbers are replaced with roman numerals and vice versa.
2. If the last part of a gene symbol consists of numbers, a word-delimiter (i.e. a hyphen or a space) is inserted. For example, “ABC1” becomes “ABC-1”. If a word delimiter is present, it is removed. (e.g. “DEF-1” becomes “DEF1”)

2.5 Automatic filter

To remove highly ambiguous terms, especially those that could have been created by the previously mentioned spelling variation generation rules, we applied an automatic filter; We removed terms that consist only of tokens that are either (a) shorter than 3 characters, (b) consist only of numbers or roman numerals, or (c) belong to a set of stopwords. Examples of terms that were removed are: “G 4”, “2.19”, and “And-1”.

2.6 Family name filter

Some gene synonyms in the dictionary are actually family names and should therefore be removed. We used an automatic procedure to identify family names: if a term is also found in the dictionary followed by a number, roman numeral or greek letter, we considered it to be a family name. For instance, “Zinc finger protein” is also detected as a substring in “Zinc finger protein 51”, and is therefore removed as a synonym.

2.7 Simple disambiguation

Similar to Koike et al. [5], we used several simple rules to detect and possibly resolve ambiguous terms:

1. We first determined whether a term is *ambiguous*. A term is considered ambiguous if it *refers to more than one gene in the dictionary*, or when it is *shorter than six characters and does not contain a number*. A non-ambiguous term will automatically be assigned
2. An ambiguous term will only be assigned if a *synonym is found* in the same document, or the *term is the ‘preferred name’ of the gene*.

2.8 Keyword detection

Because the simple disambiguation is rather strict, we also allowed ambiguous terms to be assigned if a *keyword* was found in the same document. A keyword is a word (i.e. a token) that occurs in any of the long-form names of the gene, and appears less than n times in the dictionary as a whole. We have achieved the best results with n = 1,000. For instance, in the term “Prostate Specific Antigen” the word “Prostate” appears less than 1,000 times in the dictionary. If the ambiguous synonym “PSA” is encountered in text, and the word “Prostate” is also encountered, the gene name is recognized.

3 Results

Table 1 shows the precision and recall scores of the system on the BioCreAtIvE 2 test set, after progressive inclusion of the post-processing steps for the two different dictionaries. The highest scores for both dictionaries fall within the second quartile of scores of the BioCreAtIvE 2 competition.

We also tested the speed of the Peregrine system by analyzing a random set of 100,000 Medline records. On a Dual AMD Opteron 248 system, the tagging process and post-processing steps required 213 seconds (about 3.5 minutes).

	BioCreAtIvE dictionary		Combined dictionary	
	P	R	P	R
Tagging system	0.09	0.82	0.42	0.81
+ Manual filter	0.17	0.82	0.44	0.81
+ Spelling variations	0.18	0.84	0.43	0.83
+ Automatic filter	0.36	0.83	0.52	0.82
+ Family name filter	0.48	0.82	0.53	0.82
+ Simple disambiguation	0.77	0.65	0.79	0.68
+ Keyword detection	0.72	0.75	0.75	0.76

Table 1: Precision (P) and Recall (R) for the basic tagging system and the accumulative set of post-processing steps.

4 Discussions

The initial difference in precision between the BioCreAtIvE dictionary and our own combined dictionary appears to be primarily caused by additional highly ambiguous terms in the BioCreAtivE dictionary. Particularly, the term ‘human’ was found as a synonym of 15 genes!

Without extra steps, simple dictionary lookup of (sequences of) words in text leads to a very low precision. Several post-processing steps can be used to boost performance. Especially a set of simple disambiguation rules provide a major increase in precision, but at a loss of recall. Most of the steps described here require little or no manual effort. The resulting system is fast and robust, and can easily be applied to large corpora.

References

[1] R. Jelier, G. Jenster, L. C. Dorssers, B. J. Wouters, P. J. Hendriksen, B. Mons, R. Delwel, and J. A. Kors, "Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation," *BMC Bioinformatics*, vol. 8, pp. 14, 2007.

[2] M. J. Schuemie, B. Mons, M. Weeber, and J. A. Kors, "Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification," *Journal of Biomedical Informatics*, in press.

[3] A. McCray, S. Srinivasan, and A. Browne, "Lexical Methods for Managing Variation in Biomedical Terminologies," presented at Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 1994.

[4] J. A. Kors, M. J. Schuemie, B. J. A. Schijvenaars, M. Weeber, and B. Mons, "Combination of genetic databases for improving identification of genes and proteins in text," presented at Proceedings of BioLINK, http://www.cs.queensu.ca/biolink05/presentations/Kors.pdf, 2005.

[5] A. Koike and T. Takagi, "Gene/protein/family name recognition in biomedical literature," presented at BioLINK 2004: Linking Biological Literature, Ontologies, and Databases, 2004.