Overcoming Ambiguous Gene Name Synonyms in MEDLINE Searches by Context Mining

Word-vector based text classification of PubMed records

Modest von Korff, Thomas Sander
Research Information Management
Actelion Pharmaceuticals Ltd., Allschwil, Switzerland
Email: modest.korff@actelion.com, thomas.sander@actelion.com

Abstract—**Classification of ambiguous gene name synonyms is a necessity when mining PubMed Central records with gene-related queries. This work introduces the use of word-vectors for gene name disambiguation. PubMed Central was queried for gene names and their synonyms. The retrieved records were filtered and automatically separated into train- and test-data. A similarity threshold was derived from the similarity matrix of every training word-vector set. The classification performance of the word-vectors was compared to a gene name similarity classification. Both methods showed good results, but the word-vector classification was superior in terms of precision and recall.**

Keywords-Gene name disambiguation; classification; word-vectors; datamining; algorithm.

I. INTRODUCTION

Searching MEDLINE for information about genes is a common task. Retrieving results that are not related to the gene under consideration is a common experience. One reason is the existence of ambiguous gene name synonyms. Many gene name synonyms are shared by two or more genes. This means that a PubMed search for an ambiguous gene name synonym will retrieve the publications for at least two genes. If the search is performed by a scientist, he will be burdened by the additional workload to sort out the unwanted publications. Even worse, a data-mining tool, without the capability of recognizing the ambiguity, will confound the information for the gene under consideration with the information from the other gene. Problems with ambiguous gene names were already reported by Jenssen et al. [1], and were also the topic of the BioCreative 1 and 2 challenges [2] [3]. Hakenberg et al. [4] and Wermter et al. [5] undertook huge efforts to normalize gene names. More recent approaches were published by Neves et al. [6] and by Li et al. [7]. The work presented here demonstrates a solution for the gene name disambiguation problem as it was described by Li et al. and Hakenberg et al. [8]. Our method solves the issue of ambiguous gene name synonyms by context similarity classification. In Section II the applied methods and the datasets are described. Section III gives a summary of the results for the classification of ambiguous PubMed records. The conclusions for the experiments and their results are given in Section IV.

II. METHODS

A. Gene names and synonyms

Gene names and their synonyms were the starting point for our shared synonyms experiments. Two sources were used to retrieve the synonyms. A table with Human Genome Organization (HUGO) ids, gene names, approved symbols and synonyms was retrieved from HUGO Gene Nomenclature Committee (HGNC) [9]. The second source was the MEDLINE database Entrez Gene [10], which also delivered HUGO ids, gene names, and synonyms. Both these databases were used because they do not completely overlap. The combined database, **LG,** is a list of records for each gene. A single record $\mathbf{lg}_{\text{Gene}}$ from this list contains the approved symbol as the approved name and a list with all synonyms from the two data sources. A scheme for the complete algorithm is given in Fig. 1.

B. Ambiguous synonyms detection

The algorithm for detecting ambiguous synonyms consists of two parts. For the detection of ambiguous approved symbols, an approved symbol $as_{\text{Gene,query}}$ from a gene record $\mathbf{lg}_{\text{Gene,query}}$ is taken and compared to the synonyms from all other records in **LG**. This is done for every approved symbol in **LG**. If the approved symbol $as_{\text{Gene,query}}$ matches a synonym, an ambiguous approved symbol is found. Detecting ambiguous synonyms works analogously. From a record $\mathbf{lg}_{\text{Gene,query}}$, a synonym s_{query} is taken and compared to all other synonyms in **LG**. If s_{query} matches any other synonym, an ambiguous synonym has been found. This is done for every synonym in **LG**. If a record $\mathbf{lg}_{\text{Gene,query}}$ contains an ambiguous approved symbol, ambiguous approved name or an ambiguous synonym the gene record receives the label *ambiguous*.

C. Querying PubMed Central with gene name synonyms

For all ambiguous records from **LG**, queries are generated to search the PubMed Central database. One PubMed query is created for every single approved symbol, approved name or synonym. Without any further specification, all fields in the PubMed Central database are searched. Depending on the query, no records at all up to several tens of thousands are retrieved.

ISBN: 978-1-61208-409-1

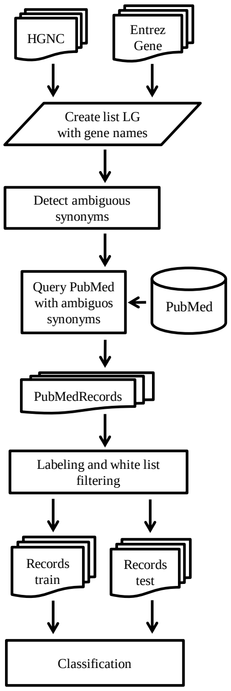
HGNC
Entrez Gene
Create list LG with gene names
Detect ambiguous synonyms
Query PubMed with ambiguos synonyms
PubMed
PubMedRecords
Labeling and white list filtering
Records train
Records test
Classification

Figure 1: Architecture of the gene name disambiguation.

The result is a dataset, $\mathbf{R}_{\text{Gene}}$, for each gene, containing the retrieved records. If a PubMed record contains an unambiguous approved name or an unambiguous approved symbol it receives the label *train*. If the PubMed record contains only ambiguous gene name information, approved symbol, approved name or synonym, it receives the label *ambiguous*. PubMed queries with the Entrez tool did not distinguish between lower-case and upper-case letters. Unfortunately, many letter combinations exist which differ in capitalization and are shared by different terms. Consequently, up to tens of thousands of false-positive records were retrieved for a single gene.

D. Whitelist filtering of PubMed records

A post-processing step was added to get rid of the false-positive records. If a synonym consisted of less than six characters and did not contain a space, the retrieved PubMed records were filtered for the exact upper- and lower-case pattern of the synonym. However, after this filtering process, many false-positive records still remained. These records contained terms with an identical synonym to the gene under consideration. False-positive records that contain the exact synonym can only be detected by analysing the context of the synonym. The context of the synonyms we were looking for was related to the concept 'gene'. For the record filter in G2DPubMedMiner, a gene context list of 25 terms was defined: activation, activator, allosteric, chromatin, chromosome, codon, exon, expression, gene, genome, genotype, histone, homolog, inhibitor, inhibition, intron, modulator, mutant, nucleosome, peptide, phenotype, phenotypic, polymerase, protein, target, transcript, and transposon. If a PubMed record did not contain any of these words, it was very unlikely that the record was related to a gene. Consequently, a PubMed record was only accepted if it contained at least one of the words from the context list. Furthermore all records were skipped that did not contain a disease MeSH term.

E. Test dataset with ambiguous gene records

From the list with the ambiguous genes a test dataset with eleven pairs of ambiguous gene records was selected at random (Table 1). A gene test set record $\mathbf{gtr}_{\text{Gene1,Gene2,Synonym}}$ contained two approved symbols and the ambiguous synonym they shared. For each gene test set record the corresponding train- and test-sets from PubMed Central were compiled. The training set contained all PubMed records where the approved gene name or the approved symbol was found. The test set contained the records with the ambiguous gene name synonym. All test records were manually classified and received the label *genename1*, *genename2* or *none*. None was given if the text in the PubMed record summary indicated that the gene name synonym referred to neither of the two genes.

TABLE I. TRAIN AND TEST DATA SETS.

Approved symbols		**Shared synonym**	**Number of records in data sets**		
Gene 1	***Gene 2***		***Train 1***	***Train 2***	***Test***
ANPEP	TOR1AIP 1	LAP1	24	6	72
APEX1	TEAD1	REF1	84	38	18
APEX1	TFPI2	REF1	84	59	7
CCNL2	FAM58A	cyclin M	11	4	3
CD200R1	HCRTR2	OX2R	31	21	96
CNGB1	LRRC32	GARP	29	20	101
DPYSL2	SDF2L1	dihydropyrimidinase-like 2	32	5	10
ERCC3	GTF2H1	TFIIH	90	9	51
HSD17B7	SKAP2	PRAP	22	5	22
MECOM	RUNX1	AML1-EVI-1	14	90	1
POU2F1	SLC22A1	OCT1	23	90	37
			444	347	418

F. Classification of ambiguous records

Two methods were used for the classification of the test records. A simple gene name similarity search was used as standard method. Word-vectors were used as a second classification method. A word-vector encodes a text as an integer vector. Every field in the vector corresponds to one word, and the field value is equal to the frequency count of the word in the text. Two word-vectors are compared by calculating their similarity coefficient. The method was adapted from Lewis et al. [11]. Because of their results, we decided to use the cosine similarity together with inverse-document-frequency (IDF) weighting. We changed only their formula for the similarity calculation by multiplying x^2

and y^2 with IDF_i (Eq. 1). Consequently, the similarity is scaled between zero and one:

$$\textbf{Cosine coefficient}(x) = \frac{\sum_{i=1}^{n}(x_i y_i \mathrm{IDF}_i)}{\sum_{i=1}^{n} x^2 \mathrm{IDF}_i \times \sum_{i=1}^{n} y^2 \mathrm{IDF}_i}. \qquad 1$$

For a train data set, the complete similarity matrix was calculated. This means that all pair-wise similarities were calculated between the word-vectors that were compiled from the PubMed records for one gene. The similarity values were sorted and the value at a given percentile of the sorted vector was taken as threshold value. The classification of the test data was done for different percentile values: 0.75, 9.5, 0.25, 0.05, and 0. A percentile of 0 meant that no threshold was used.

III. Results

A total of 35,631 gene names were extracted from HUGO. The ambiguous synonyms detector found 7166 pairs of genes that shared at least one synonym. From this set of gene pairs, eleven were selected for the test dataset. The processing time for a dataset, including querying PubMed and the consequent processing of the results, strongly depended on the number of retrieved PubMed records and took up to 30 minutes.

The results for the classification experiments are given in Table 2 with precision and recall as figures of merit.

TABLE II. Results for the classification experiments.

Method	Result		
	Precision	Recall	Harmonic mean
GenenameSim	0.83	0.28	0.42
WVSim 0	0.52	1	0.68
WVSim 0.05	0.63	0.87	0.73
WVSim 0.25	0.68	0.57	0.62
WVSim 0.5	0.88	0.29	0.44
WVSim 0.75	0.91	0.19	0.31

In the last column of the table the harmonic mean combines precision and recall. For the simple approach with the gene name similarity classification a precision of 0.83 and a recall of 0.28 was reached. The next five rows show the results for the classification using word-vectors and the five different threshold percentiles. The maximum harmonic mean was reached for a threshold of 0.05 (WVSim 0.05). To compare our results with other approaches like those of Li at al. [7], or Xu at al. [12] et al. is difficult, because gene name normalization and disambiguation are often done together. Or, supervised methods are used, with the disadvantage of being successful only in the training domain.

IV. Conclusions

Identification of more than 7,000 gene pairs sharing at least one synonym demonstrated that the classification of ambiguous gene names is a worthwhile undertaking. With a test data set, compiled from eleven pairs of ambiguous gene names, it was shown that word-vector classification reduced the ambiguity significantly. A similarity threshold value, which was automatically derived from the similarity matrix of the training data, increased the precision of the classification results. The entire process, starting with querying PubMed Central, followed by filtering and train- and test-set generation, and the classification is unsupervised and can be fully automated. Word-vector classification for gene name disambiguation is a valuable addition to every data-mining tool working on PubMed records with gene-related queries.

References

[1] T. K. Jenssen, A. Laegreid, J. Komorowski and E. Hovig, "A literature network of human genes for high-throughput analysis of gene expression," Nature Genetics, vol. 28, 2001, pp. 21-28, doi: 10.1038/ng0501-21.

[2] L. Hirschman, M. Colosimo, A. Morgan and A. Yeh, "Overview of BioCreAtIvE task 1B: normalized gene lists," BMC Bioinformatics, vol. 6 Suppl 1, 2005, pp. S11.11-S11.10, doi: 10.1186/1471-2105-6-S1-S11.

[3] A. A. Morgan et al., "Overview of BioCreative II gene normalization," Genome Biology, vol. 9 Suppl 2, 2008, pp. S3.1-S3.19, doi: 10.1186/gb-2008-9-s2-s3.

[4] J. Hakenberg et al., "Gene mention normalization and interaction extraction with context models and sentence motifs," Genome Biology, vol. 9 (Suppl 2) , S14, 2008, doi: 10.1186/gb-2008-9-S2-S1.

[5] J. Wermter, K. Tomanek and U. Hahn, "High-performance gene name normalization with GeNo," Bioinformatics, vol. 25, 2009, pp. 815-821, doi: 10.1093/bioinformatics/btp071.

[6] M. L. Neves, J. M. Carazo and A. Pascual-Montano, "Moara: a Java library for extracting and normalizing gene and protein mentions," BMC Bioinformatics, vol. 11, 2010, pp. 157, doi: 10.1186/1471-2105-11-157.

[7] L. Li, S. Liu, W. Fan, D. Huang and H. Zhou, "A multistage gene normalization system integrating multiple effective methods," PLoS One, vol. 8, 2013, pp. e81956, doi: 10.1371/journal.pone.0081956.

[8] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder and G. Gonzalez, "Inter-species normalization of gene mentions with GNAT," Bioinformatics, vol. 24, 2008, pp. i126-132, doi: 10.1093/bioinformatics/btn299.

[9] HUGO Gene Nomenclature Committee at the European Bioinformatics Institute. [retrieved: Mar. 2015]. Available from: http://www.genenames.org

[10] D. Maglott, J. Ostell, K. D. Pruitt and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," Nucleic Acids Res, vol. 33, 2005, pp. D54-58, doi: 10.1093/nar/gki031.

[11] J. Lewis, S. Ossowski, J. Hicks, M. Errami and H. R. Garner, "Text similarity: an alternative way to search MEDLINE," Bioinformatics, vol. 22, 2006, pp. 2298-2304, doi: 10.1093/bioinformatics/btl388.

[12] H. Xu et al., "Gene symbol disambiguation using knowledge-based profiles," Bioinformatics, vol. 23, 2007, pp. 1015-1022, doi: 10.1093/bioinformatics/btm056.