Proceedings of the Australasian Language Technology Workshop 2007, pages 143-145

Exploring Extensions to Machine Learning-based Gene Normalisation

Benjamin Goudey,† **Nicola Stokes**†‡ and **David Martinez**†
† Computer Science and Software Engineering
and
‡ NICTA VRL
University of Melbourne
VIC 3010, Australia
{bgoudey,nstokes,davidm}@csse.unimelb.edu.au

1 Introduction

One of the foundational text-mining tasks in the biomedical domain is the identification of genes and protein names in journal papers. However, the ambiguous nature of gene names means that the performance of information management tasks such as query-based retrieval will suffer if gene name mentions are not explicitly mapped back to a unique identifier in order to resolve issues relating to synonymy (i.e. many different lexical forms representing the same gene) and ambiguity (i.e. many distinct genes sharing the same lexical form). This task is called gene name normalisation, and was recently investigated at the BioCreative Challenge (Hirschman et al., 2004b), a text-mining evaluation forum focusing on core biomedical text processing tasks. In this work, we present a machine learning approach to gene normalisation based on work by Crim et al. (2005). We compare this system with a number of simple dictionary lookup-based methods. We also investigate a number of novel features not used by Crim et al. (2005). Our results show that it is difficult to improve upon the original set of features used by Crim et al. We also show that for some organisims gene name normalisation can be successfully performed using simple dictionary lookup techniques.

2 Data

The experiments described in this paper were performed on the data provided by the first BioCreative workshop for gene normalisation. For each abstract in the test collection the system must create a list of normalised gene names mentioned in the text. Three distinct organism datasets were investigated: *yeast*, *mouse*, *fly*. Systems are provided with a gene synonym list for each organism containing a comprehensive list of gene identifiers and many of their related gene mentions, together with a set of training instances (abstracts with corresponding gene lists) for each organism. The gene lists used as training data were created by filtering down pre-existing manually compiled lists that applied to whole documents. This automatic filtering process added noise to the training data by lowering the recall of gene lists to 86%, 80% and 55% for the yeast, fly and mouse data respectively. More information on the data for this task can be found in (Hirschman et al., 2004a).

3 System Description

As already mentioned, we use a machine learning approach similar to that used by Crim et al. (2005), one of the top performing systems at the BioCreative I Challenge. There are 3 main stages in our system: first, the document is run through a high recall gene identification system; each candidate mention is then used to create a series of instances for each possible gene identifier related to the mention, extracting a variety of contextual features based on the surrounding text; and finally, instances are passed to a maximum entropy classifier. We use the training data to build our model, where each instance in the testing part is classified and a confidence value is returned. Finally, the gene identifiers with the highest confidence value for a particular gene mention are added to the gene list for that abstract.

4 Results and Conclusions

One of the limitations of the system by Crim et al. (2005) is that the use of exact matching and the in-

	Yeast			Mouse			Fly		
	Prec.	Rec.	F-sc.	Prec.	Rec.	F-sc.	Prec.	Rec.	F-sc.
BioTagger - Basic	94.0	59.2	72.6	73.8	70.4	71.8	49.5	39.3	40.5
LU - Basic	89.0	90.9	89.9	1.9	89.7	3.7	2.0	95.3	3.8
LU - Entrez/Filtered	94.0	88.7	91.3	73.7	74.8	**74.3**	45.8	91.6	61.0
LU - Variations and Entrez/Filtered	93.5	89.6	**91.5**	60.4	77.9	68.1	41.8	92.1	57.5
ML - Basic	95.4	77.3	85.4	84.4	56.8	67.9	75.1	72.5	73.8
ML - Entrez/Filtered	95.2	87.4	91.2	82.2	66.2	73.3	74.0	81.6	**77.6**
ML - Variations and Entrez/Filtered	94.7	88.3	91.4	78.7	68.6	73.3	71.8	82.5	76.8

Table 1: BioTagger results and synonym list expansion over our machine learning (ML) and lookup-based (LU) systems (best f-scores shown in bold)

	Yeast			Mouse			Fly		
	Prec.	Rec.	F-sc.	Prec.	Rec.	F-sc.	Prec.	Rec.	F-sc
Crim et al. (2005)	95.6	88.1	91.7	78.7	73.2	75.8	70.4	78.3	74.2
ML - Crim Features	94.9	88.9	91.8	82.2	66.2	73.3	74.0	81.6	77.6
ML - All Features	95.1	88.1	91.4	79.4	71.0	75.0	69.5	82.8	75.5
ML - Optimal Features	94.4	90.0	**92.2**	78.8	73.7	**76.2**	75.6	81.5	**78.5**

Table 2: Comparison of (Crim et al., 2005) and our ML system with different feature sets (best f-scores shown in bold)

completeness of the synonym list limits the ability of system to achieve high recall. To address this issue, we experiment with a variety of synonym list expansion and filtering methods including:

- **Lexical Variations** - the creation of gene name variations with different hyphenation and spacing patterns.
- **Entrez Gene** - the expansion of the original synonym list with information from Entrez Gene (Maglott et al., 2005).
- **Conditional Probability** - a conditional probability filter which was used by (Crim et al., 2005) in their pattern matching system.

We tested two different approaches to this task: the first performs no explicit disambiguation, but adds all possible gene identifiers to the gene list for each gene mention; the second by using the maximum entropy classifier as outlined in Section 3. We also compared our results to those of the BioTagger (McDonald and Pereira, 2005), a well-known "out of the box" gene identification system.

For our two main systems (lookup and machine learning), we ran different combinations of the synonym list expansion, with the top two performing results and the baseline shown in Table 1. We can see that the recall of the BioTagger is very low, which suggests that we would need some tuning to apply it to this specific dataset. These results illustrate that the yeast data needs almost no explicit disambiguation, i.e. the the lookup-based system performs best. While the fly data, which contains very ambiguous gene mentions, needs a machine-learning approach to identify the correct identifier. Surprisingly, the mouse, which also has a reasonable degree of ambiguity, performs at its best with a lookup based system. It must be noted, that the noise of the training data may have contributed to the poor performance of the classifier, yet this is still an interesting result.

In the next experiment, our aim was to improve classifier performance by increasing the original feature set (from 5 to 21) with different features derived from linguistic information (POS tags) and information from external resources (e.g. whether the target word is defined in WordNet). Using these features we compare our gene normalisation system to that of (Crim et al., 2005) using all new features as well as an optimal subset of these. The results are shown in Table 2. While the results of using the entire extended feature set tend to degrade performance compared to the basic set (except for mouse), using a subset of features unique to each organism does lead to some performance improvements. This implies that a one-classifier fits all approach is not suitable for gene normalisation, and that individual classifiers must be created for each organism.

References

J. Crim, R. McDonald, and F. Pereira. 2005. Automatically annotating documents with normalised gene lists. *BMC Bioinformatics*, 6 (Supplement I)(13).

L. Hirschman, M. Colosimo, A. Morgan, and A. Yeh. 2004a. Overview of biocreative task 1b: Normalized gene lists. *Journal of Bioimedical Informatics*, 37.

L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. 2004b. Overview of biocreative: Critical assessment of information extraction for biology. *Journal of Bioimedical Informatics*, 37.

Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. 2005. Entrez gene: Gene-centered information at ncbi. *Nucleic Acids Resarch*, 33 (Database Issue).

R. McDonald and F. Pereira. 2005. Identifying gene and protein mentions in text using conditional random fields. *Journal of Bioimedical Informatics*, 37.