Name Normalization

Anoop Sarkar – anoop@cs.sfu.ca

The task of name normalization is to take an input document which contains many named entities, and associate that document with a list of unique name identifiers (from some pre-existing database of unique names). Assume we have a list of unique list of companies: each company has a unique identifier. The list also contains a list of possible variants (including acronyms) for the company name. Typically the document will contain versions of the name that do not match any of the previously listed variants in the database. In addition, many name variants across different identifiers are likely to be identical.
Note that if the original document is no longer available, and we are merging names in a database, then the problem reduces to string matching and is often referred to as *record linkage* (see *http://secondstring.sf.net/*). This task has attracted special attention in the BioNLP area where it is referred to as *gene (name) normalization*.

Gene Normalization

The input is a document (a PubMed/Medline abstract) and the output is a sequence of Entrez Gene identifiers. The training data includes the mentions in the text that is the evidence for each identifier.
From *bc2GNtestdocs/10491763.txt*:

> The transcription factor hepatocyte nuclear factor (HNF)-6 is an upstream regulator of several genes involved in the pathogenesis of maturity-onset diabetes of the young. We therefore tested the hypothesis that variability in the HNF-6 gene is associated with subsets of Type II (non-insulin-dependent) diabetes mellitus and estimates of insulin secretion in glucose tolerant subjects. METHODS: We cloned the coding region as well as the intron-exon boundaries of the HNF-6 gene. We then examined them on genomic DNA in six MODY probands without mutations in the MODY1, MODY3 and MODY4 genes and in 54 patients with late-onset Type II diabetes by combined single strand conformational polymorphism-heteroduplex analysis followed by direct sequencing of identified variants. An identified missense variant was examined in association studies and genotype-phenotype studies. . . . *(rest of document deleted)*

From *grep 10491763 bc2GNtest.genelist*:

document identifier	Entrez Gene identifier	mention in the text
10491763	3172	'MODY1'
10491763	3175	'Hepatocyte nuclear factor-6'
		'hepatocyte nuclear factor (HNF)-6'
		'HNF-6'
10491763	3630	'insulin C-peptide'
10491763	3651	'MODY4'
10491763	6927	'MODY3'

Entrez Gene

From (Maglott et. al. 2005):

> Entrez Gene is the gene-specific database at the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH) in Bethesda, MD, USA. Entrez Gene provides unique integer identifiers for genes and other loci for a subset of model organisms.

The task the biologists would like to solve is to automatically associate a list of Entrez Gene identifiers with each article in PubMed. In general, this may involve gene names from different species such as Mouse (52,494 genes),

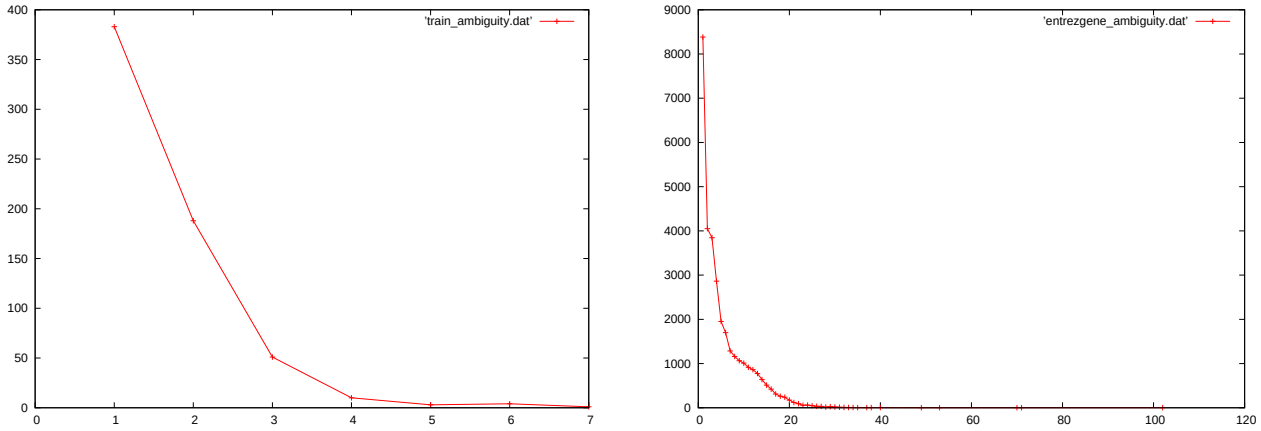
Fly (27,749 genes), Yeast (7,928 genes), or Human (32,975 genes). Due to the evolutionary processes that have moulded the DNA of these species, many gene names are very similar across species since they perform very similar functions in the organism.

Biocreative 2: Gene normalization task

In the Biocreative 2 task, we will focus on a single organism: Human. Thus, we ignore Entrez Gene identifiers for other species even if their named entities occur in the document.
A list of synonyms was provided for each Entrez Gene identifier.
A set of 640 training examples is provided. Each training example includes the PubMed/Medline article abstract, the named entities in the document and their associated Entrez Gene identifiers.
Here is the most ambiguous entry with respect to the number of named entity variants in the same document in the (clean) training data; document number 2365818, Entrez Gene identifier 594:

E1 beta subunit of human branched chain alpha-ketoacid dehydrogenase complex
E1 beta subunit of the branched chain alpha-ketoacid dehydrogenase (BCKDH) complex
BCKDH-E1 beta
E1 beta subunit of BCKDH complex
lambda hBE1 beta-1
E1 beta subunit of human pyruvate dehydrogenase complex
E1 beta

Here is a comparison of the named entity variants in training vs. the lexicon:



How difficult is this problem?

Avg synonym length in words	2.17
Avg synonyms per identifier	5.55
Avg identifiers per synonym (ambiguity)	1.12

Also provided is a set of 9322 noisy training documents, where each document is associated with a noisy list of Entrez Gene identifiers. However, no named entities are provided.
The test set has 785 documents. For each document the system provides a list of Entrez Gene identifiers which is then compared to the gold-standard in the following way:
Each document has multiple identifiers that must be detected by the system. For evaluation, the system output of the list of identifiers per document is matched with curated data. Identifiers that match are true positives (TP), identifiers found but not matched are false positives (FP). Gold standard identifiers that were not matched are false negatives (FN).

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$f\text{-}measure = \frac{2 \cdot P \cdot R}{P + R}$$

This is computed per document, and then combined in two ways for all documents:

- Micro-average: combine the results for P and R across identifiers in all documents.
- Macro-average: compute P and R per document and then average across documents.

The macro-average is used to compute statistical significance using a one-sided *t*-test. The Biocreative 2 organizers also ran a *bootstrap resampling* significance test where each system was tested on 10,000 random sets of 250 documents (with replacement) from a smaller test set of 262 documents.
Best reported result on this task has P = 78.9%; R = 83.3%; F-measure micro-average 81.0%; F-measure macro-average 81.1%.

Steps in this task

The following steps try to flesh out the problems:

1. *Creating a synonym list for each gene identifier*. A synonym list is provided by Entrez Gene. But this is expanded further by using different techniques and other data sources. This is typically done via a variety of pattern matching into the databases and creating a standardized way of storing each variant in the synonym list.

2. *Identifying gene mentions in the text*. This is done using standard means of finding names in text, using standard software that exists for this purpose. This can be achieved by using the synonym list which is then used to match as many mentions as possible in the documents sometimes using approximate matches or acronym handling. This can also be achieved using gene named entity taggers. Also, a combination of the two techniques has been tried.

   Multiple matches creates a massive ambiguity problem where multiple identifiers are selected for each document (which improves recall, but trashes precision).

3. *Predicting the unique gene identifier for each gene mention*. Even a single mention is useful to provide the gene identifier. Commonly confused gene identifiers should not be predicted simultaneously in this stage. This happens often due to the overlap in their respective synonym lists or matching method. Each gene mention, and the entire document is used to create a sequence of positive and negative examples for each Entrez gene identifier (compatible with the document) along with a Y/N label (training data). This training data is then used to build a classifier in order to find only the valid identifiers for new documents.

   Here is one method to create the training data (Crim et. al. 2005): Every synonym in the lexicon was matched to each training document. For each match, extract the matching text plus the immediate context (e.g. 3 words around the match). In the (clean) training set, if the mention was provided with an identifier (see example above) then the match was labeled positive; otherwise, it was labeled negative.

Difficult Cases

In the (Maglott et. al. 2005) article there is a discussion of difficult cases:

- Is there a recall problem? Are there mentions that are hard to map to their full forms (to the synonym list) or their identifiers? The answer seems to be 'no'. The 'pooled maximum recall' by combining all systems in Biocreative 2 was a surprising 97.2% (at 23.2% precision). Highest single system recall was 87.5%.
- Missing synonyms: The gene name 'Ly9' does not occur in the Entrez Gene synonym list, but 'Humly9' does. Including the suffix 'ly9' makes things worse because it then also matches the mouse gene called 'Ly9' which sometimes occurs in documents where 'Humly9' is not mentioned. Here is an example where they co-occur:

  > PubMed ID 8537117
  > Ly9 is a mouse cell membrane antigen found on all lymphocytes and coded for by a gene that maps to chromosome 1. We previously described the isolation and characterization of a full-length cDNA clone for mouse Ly9. Using cross-species hybridization we isolated cDNA clones encoding the human homologue Humly9.

- Short, highly ambiguous symbols: 'AMP'
- Conjoined expressions and range expressions: 'freac-1 to freac-7' implies mention of 'freac-2' ; 'protein kinase C isoforms alpha, epsilon, and zeta'
- Descriptions of the gene: document has 'receptor for the ciliary neurotrophic factor' while lexicon has 'ciliary neurotrophic factor receptor'

Opportunities

- Flexible loss functions (trading recall for precision for matching; while the reverse for the identifier classifier).
- Noisy data.
- Using all of Medline as unlabeled data.
- Name normalization in a different domain – company names.

References

- Data sets downloaded to: /cs/natlang-e/data/biocreative
- D. Maglott et. al. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. January 1; 2005.
- T. Sandler, A. I. Schein, and L. H. Ungar. Automatic Term List Generation for Entity Tagging. Bioinformatics 2006, 22(6): 651. http://bioinformatics.oxfordjournals.org/cgi/content/abstract/22/6/651
- J. Crim, R. McDonald, and F. Pereira. Automatically Annotating Documents with Normalized Gene Lists. BMC Bioinformatics 2005, 6(Suppl 1):S13. http://www.biomedcentral.com/content/pdf/1471-2105-6-S1-S13.pdf
- K. Ganchev et. al. Penn/UMass/CHOP Biocreative II systems. manuscript. http://papers.ldc.upenn.edu/BioCreative2007/Ganchev+2007-Penn-UMass-CHOP-Biocreative-II-Systems.pdf
- A. Morgan et. al. Overview of BioCreative II gene normalization. Genome Biology 2008, 9(Suppl 2):S3. doi:10.1186/gb-2008-9-s2-s3. http://genomebiology.com/content/pdf/gb-2008-9-s2-s3.pdf
- L. Hirschman et. al. Overview of BioCreAtIvE task 1B: normalized gene lists. BMC Bioinformatics 2005, 6(Suppl 1):S11. doi:10.1186/1471-2105-6-S1-S11. http://www.biomedcentral.com/1471-2105/6/S1/S11
- Overview of BioCreative II gene normalization. http://genomebiology.com/2008/9/S2/S3
- Overview of BioCreative task 1B: normalized gene lists. http://www.biomedcentral.com/1471-2105/6/S1/S11
- Biocreative 2 gene normalization task. http://biocreative.sourceforge.net/biocreative_2_gn.html
- BioNLP corpora. http://compbio.uchsc.edu/ccp/corpora/obtaining.shtml