Static Dictionary Features for Term Polysemy Identification

Piotr Pęzik, Antonio Jimeno, Vivan Lee, Dietrich Rebholz-Schuhmann

European Bioinformatics Institute Wellcome Trust Genome Campus Hinxton, Cambridge, CB10 1SD, UK E-mail: piotr.pezik@gmail.com

Abstract

Building a large lexical resource which pools together terms of different semantic types naturally leads to the consideration of term ambiguity issues. Both cross- and intra-domain term polysemy constitutes a formidable obstacle in tasks and applications such as Named Entity Recognition or Information Retrieval (e.g. query expansion or relevance feedback), where a single polysemous term may cause a significant concept drift. One of the biggest sources of polysemy in biomedical terminologies are protein and gene names (PGN), both those extracted from existing databases and their variants found in the literature. We provide an analysis where the effect of using static dictionary features for the detection of potential polysemy of protein and gene names (and by extension other semantic types) is clearly delineated from the contribution of applying contextual features. We argue that, although disambiguation based on static dictionary features does not outperform fully-fledged context-driven Named Entity Recognition, it does effectively filter out highly polysemous terms (increase in F-measure from 0.06 to 0.57 and from 0.21 to 0.51 as measured on two evaluation corpora). Moreover, static dictionary features are context-independent and thus more easily applicable in systems where running intense on-the-fly disambiguation for retrieved documents could be problematic.

1. Introduction

The preparation and refinement of a lexical resource constitutes an important step in many information extraction and information retrieval systems used for the biomedical domain. Unfortunately, very few of the widely available databases seem to have all the makings of a high-quality lexical resource. As an example, ontologies and taxonomies (e.g. GO, or MESH) come with little or no guarantee that the potential terms they contain are actually used in the relevant literature. Resources such as BioThesaurus 1 contain a mixture of terms and pseudo-terms extracted semi-automatically from a range of varying quality protein and gene databases. This leads to a high level of both artificial and genuine term ambiguity, which can be dealt with both at the level of the dictionary, and the context of term occurrence. The need for disambiguation techniques is recognized in Named Entity Recognition approaches, which, within the biomedical domain, have been mainly focused on the identification of protein and gene names (PGNs)². This focus is largely due to the inherent polysemy of PGNs as well as their importance as one of the fundamental entities of interest in the biomedical domain, although other semantic types such as enzymes, chemical entities and species also have to be handled in text mining applications. The recognition and normalisation of protein and gene names has received much focus in

Table 1 shows the contents of the term repository we have used to test static dictionary features.

identifiers available for each PGN.

¹ http://pir.georgetown.edu/pirwww/iprolink/biothesaurus.shtml.

We used BioThesaurus v. 2 for the experiment reported here.

BioThesaurus only contains protein and gene names mapped to Uniprot accession numbers. We have taken advantage of this

mapping and used the Uniprot annotations such as the species

ones

community-wide Information Extraction evaluation efforts such as BioCreAtivE³. The issue of term polysemy is also important in the field of Information Retrieval (i. e. query expansion or relevance feedback), for the biomedical domain, where a single polysemous term may cause a significant concept drift (e.g. Jimeno et al. 2007). In this paper we first introduce a terminological resource combining a number of semantic types imported from different biomedical databases. We show some potential advantages of having different semantic types pooled together, and how the intrinsic polysemy of such a combined resource can be dealt with. By a semantic type we mean a distinct class of biological or chemical entities, e.g. protein names as opposed to species names. Next, we focus on the task of automatic protein/gene names recognition to demonstrate how the performance of a named entity recognition system relying on static information encoded in a lexical resource can be significantly improved. Having analyzed the different usage patterns of E. coli and human PGNs, we have identified four types of polysemy affecting PGNs stored in our lexical resource and implemented a number of dictionary-filtering rules addressing each of these types as a set of decision tree features. As a result, we manage to improve the performance of a noisy dictionary-based protein name recognition system. Our experiments confirm that static dictionary features contribute significantly to the identification of highly polysemous terms. At the same time we attempt to systematize the major types of PGN polysemy and clearly delineate the influence of static dictionary features from text driven

^{2.} Term Repository

² Many NER techniques deal with resolving the ambiguity of polysemous terms and in achieving this goal they derive from the field of Word Sense Disambiguation.

³ http://biocreative.sourceforge.net/

Semantic Type	Synsets	Variants
СНЕВІ	13,473	57,581
Enzyme names	4,016	7,658
PGNs	232,258	1,931,786
Species names	367,565	441,993

Table 1 Synsets and variants in the Term Repository

Currently, the Term Repository integrates terms from different resources, including a subset of Biothesaurus (Liu et al. 2006) protein/gene names (PGNs), which as demonstrated below ensures a relatively high initial recall of PGNs. The repository also contains Chemical Terms of Biological Interest (CHEBI⁴), enzyme names ⁵, and species names imported from the NCBI species taxonomy⁶. All of these terms are organized into clusters or synsets, i.e. sets of term variants which can be mapped to the same original database sense identifier.

Although the biggest number of distinct senses is contributed by the NCBI species taxonomy, the biggest number of synonyms is introduced by PGNs. Depending on the semantic type, it is possible to annotate each cluster and each term with metadata derived from the Term Repository itself or from external resources. As an example, each orthographically distinct PGN in the repository can be annotated with relevant species identifiers, frequencies in reference corpora or the number of nodes where it appears in external references. We show how such metadata can be used to tackle the problem of term polysemy.

The PGNs imported from the Biothesaurus come from a variety of different resources. In theory, a PGN recognition tool using a comprehensive terminological resource holds a promise of a high recall value (Schuemie et al. 2007), at the expense of precision due to a higher level of polysemy introduced by orthologous gene names, as well as noisy terms coinciding with PGNs.

As a first step in the preparation of the Term Repository for a PGN recognition task, we removed the most obvious pseudo terms such as *hypothetical protein*, *putative protein* or *possible protein* and terms corrupted in the process of automatic integration. Needless to say, such a filtering method does not deal with truly polysemous terms. After this initial clean-up and some basic case normalization, we used the Term Repository as a dictionary of PGNs.

PGNs tagged as human relevant in the repository based on their Uniprot species annotation were selected from and compiled into a deterministic final state automaton-based regular expression engine. Although the recall was relatively high (0.82), the precision of 0.03 obtained for the gene normalization task on the BioCreAtivE corpus was unacceptably low. To increase precision, we first carried out a detailed analysis of the types of polysemy found in PGN name resources. For the four major types of polysemy identified, we designed six features providing indications of one or more types of PGN polysemy. While defining the features we made sure that they are either derived directly from the Term Repository itself, or assigned statically to each term, and thus easy to use for other semantic types and other terminological resources, since they do not rely on the morphology of terms. As reported in the results section of this paper, these text-independent features significantly increased the overall performance of PGN recognition.

A purely dictionary-based method cannot compete with the performance of state-of-the-art approaches which combine dictionary look-up with contextual features. Therefore, for the sake of comparison, we complement the dictionary filtering rules with a set of contextual features. It has to be stressed however, that the main focus of the experiment reported here lies in the definition of resource-independent dictionary filtering rules which address the four types of PGN polysemy identified. This is meant to address the recently recognized need for a systematic feature evaluation in PGN identification systems (Hakenberg et al. 2005). The analysis we present below depends on a clear mapping between the rules and features applied and the types of polysemy they address. Both sets of features were fed into a decision tree-based model and evaluated against the BioCreAtivE corpus. We report on the model's performance for dictionary and text features used independently of each other.

3. Major types of PGN polysemy

Having analyzed the PGNs imported from the Biothesaurus into the Term Repository, we distinguish between the following cases of genuine PGN polysemy:

- A PGN has a common English word homograph. We call this a case of domain-independent polysemy, e.g. (but, WHO). Sometimes this type of polysemy is introduced by pseudo terms by resulting from the poor quality of a lexical resource, e.g. Biothesaurus contains partial PGN terms such as human or, due to the fact that they were gathered from less trustworthy database description fields.
- 2. A PGN has a number of hyponyms and it is sometimes used synonymously with them. Examples of this type of polysemy include generic enzyme names, such as *oxidoreductase*). Sometimes a more specified case of holonymy triggers similar ambiguity, e.g. an operon name can be interpreted to denote any of the genes it contains. We call this a case of vertical polysemy (c.f. Fellbaum 1998).
- A PGN is used for a number of orthologous or otherwise homologous genes. Thus the ambiguity in the gene name results from the fact that the same

⁴ http://www.ebi.ac.uk/chebi/

⁵ http://www.chem.qmul.ac.uk/iubmb/enzyme/

⁶ http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy

- name is used for structurally identical genes found in different species.
- A PGN has a biomedical homograph, e.g. retinoblastoma. We refer to this as a case of domain-specific polysemy (Jimeno et al. 2008).

Last but not least the very use of the umbrella term PGN suggests another type of polysemy, where the same name is used to denote a gene and its product. Generally, however, gene names are not distinguished from protein names.

Type 2 and Type 3 polysemy have also been described as intra- and cross-species ambiguity (Chen et al. 2004).

4. Feature set for polysemy detection

4.1 Static dictionary features

It is possible to design a feature set for these four types of polysemy introduced above. Table 2 shows 6 features and the corresponding types of polysemy they may indicate.

#	Feature	Polysemy type	Data type
1	BNC frequency	1	Integer
2	Number of synsets	2,3	Integer
3	NCBI taxonomy ids	3	Integer
4	Generic enzyme	2	Boolean
5	Medline frequency	4,1	Integer
6	MESH nodes	4	Integer

Table 2 Static dictionary feature set addressing the four types of polysemy.

BNC frequency is the frequency of a given PGN in the British National Corpus and it is meant to give a clue about possible do-main-independent polysemy (Type 1 polysemy).

As explained above, similarly to Biothesaurus, the Term Repository organizes terms into sets of synonymous PGNs (clusters), which can be mapped to the same UniProt accession number. *Feature 2* is thus meant to provide indications of Type 2 and Type 3 polysemy at the same time; the fact that a PGN is found under different accession numbers may mean that it is used for different orthologous genes and/or that it is a generic name which can be used to refer to many different more specific PGNs.

The *number of NCBI species taxonomy identifiers* assigned to the PGN provides evidence of Type 3 polysemy.

For feature 4, we have manually tagged 37 PGNs as *highly polysemous generic enzyme names* (e.g. oxidoreductase). This Boolean flag gives indications of Type 2 polysemy.

Feature 5 is the *frequency of a PGN in Medline*. This value is meant to provide information on Type 4 polysemy,

although it does contain some information on Type 1 polysemy as well. A combination of World Street Journal and Medline frequencies has been used for PGN disambiguation by (Tanabe & Wilbur 2002).

Feature 6 is the number of times a term occurs under *non-protein MESH nodes*. This feature encodes possible indications of Type 4 polysemy.

In addition to these features, we decided to break down PGNs into their constituent word tokens. This feature does some justice to the morphological characteristics of potentially polysemous PGNs. It is possible to relate it to all of the four types of polysemy specified above, but this relation is rather latent and PGN name dependent.

We assigned each of these features to a set of terms extracted from the training data sets, labeling each as a true positive or a true negative. The annotation of PGNs in the BioCreAtivE corpus is provided at the level of a span, while in the E. coli corpus the exact strings denoting PGNs are marked-up. Any terms annotated by the dictionary look-up method as PGNs, but not in the training section of the BioCreAtivE corpus were considered as true negatives. All the other identified PGNs were considered to be true positives.

String matching techniques are often used to enable fuzzy dictionary look-up in order to take advantage of the regularities in the morphology of PGNs which may be indicative of its polysemy. (Wilbur et al. 2007). For the purposes of this experiment we do only very basic case and hyphen normalisation. Some more extensive work on fuzzy look-ups which may potentially increase the recall is in progress.

We fed the training sets into the Weka machine learning package implementation of the C4.5 decision tree algorithm (Quinlan, 1993). The different parameters for the algorithm were estimated for each dataset via 10-fold cross-validation. We show the strongest rules identified by the decision tree in the following sections.

4.2 Text-driven features

We have so far identified a number of generic rules supporting PGN recognition that can be derived from a dictionary and other static resources, and which can be relatively easily applied to other semantic types and resources. The results yielded by such dictionary filtering rules alone produce a significant improvement in the F-measure. However, they are still below the results reported for complete PGN normalization approaches evaluated on the BioCreAtivE corpus, which use both lexicon filtering and text-driven rules, with reported results in the region of 0.7-0.8 F-measure (Morgan, Hirschman 2007). To demonstrate that we have also tested the improvement gained when the dictionary filtering feature vector is complemented with a text-driven one. For that purpose we have adopted the following six corpus-specific contextual features for **PGN** identification.

- 1. The frequency of the term in the BNC Corpus.
- 2. The frequency of the term in Medline.
- Whether the term is annotated with more than one identifier.
- 4. The frequency of the term in a given passage.
- 5. The number of distinct terms linked to the same term in the passage.
- 6. Whether the term matches the boundaries of an entity identified (but not normalized) by Abner (Settles, 2005), a conditional random field-based named entity extraction tool that has shown good performance on BioCreAtivE and NLPA datasets (F-measure of 0.69 and 0.705 respectively).

A decision tree trained on this set of contextual features is applied on top of the dictionary-based approach. The improvements obtained are reported in the following sections.

5. Evaluation corpora

We have evaluated the use of static dictionary features in an NER scenario, as it is directly relevant to the problem of ambiguity resolution. The primary evaluation of the dictionary-based disambiguation features is based on the corpus used for the BioCreAtivE gene normalization task, in which abstracts have to be annotated with EntrezGene identifiers. The BioCreAtivE dataset contains 281 manually annotated abstracts together with some 5000 documents from the Gene Ontology Annotation database. The latter part of the corpus is used as noisy training data. There also are 252 abstracts provided in the test set. However, because the Term Repository contains PGNs covering other species, we thought it necessary to carry out the evaluation of the dictionary-based PGN disambiguation features on a corpus representative of non-human PGNs as well. For that purpose we have used a PGN-annotated corpus currently containing 96 abstracts related to the topic of gene regulation in E. coli. E. coli is one of the prokaryotic model organisms which is not covered by any of the BioCreAtivE data sets, so we thought it useful to perform the method described here on a novel corpus covering this species. The minimum requirement for pre-selecting an abstract was the occurrence of at least one string matching an E. coli PGN recorded in SwissProt and some keywords indicating E. coli as a species. This resulted in a set of 33,635 abstracts of which 96 abstracts relevant to the topic of gene regulation were extracted and manually verified. One important difference between this corpus and the BioCreAtivE one is that it annotates the exact occurrences of a PGN (rather than a span of text where it occurs) and assigns possibly more than one UniProt accession numbers to the PGN if it is truly ambiguous in the text (e.g. porin). A successful recognition of a protein is only counted if the exact boundaries and one of the matching Uniprot accession numbers are correctly recognized for each occurrence. The possible influence of the exact annotation schema on the performance of NER in the biomedical domain is discussed in (Alex, 2006; Shipra et

al. 2004). We have also marked up mutant genes (they might be a case of free variation in the name which is unrecorded in the existing databases⁷), and special cases of ambiguity, where multiple genes on the same operon. The complete corpus is still being revised and completed, and will soon be released to the public at http://www.ebi.ac.uk/Rebholz/software.html.

6. Results

6.1 Evaluation runs

We have carried out 6 evaluation runs for the approaches introduced above on both PGN-annotated corpora. Table 3 shows the results obtained for the three methods for the human genes on the basis of the BioCreAtivE corpus.

#	Method	Р	R	F
1	Direct look-up	0.03	0.82	0.06
2	DictFiltering	0.36	0.71	0.57
3	Text+DictFiltering	0.79	0.63	0.7

Table 3 Summary of evaluation results for human PGNs using the BioCreAtivE corpus.

Table 4 summarizes the results obtained for the three methods as they were applied to the identification of E. Coli PGNs.

#	Method	P	R	F
1	Direct look-up	0.14	0.45	0.21
2	DictFiltering	0.66	0.42	0.51
3	Text+DictFiltering	0.75	0.41	0.53

Table 4 Summary of evaluation results for E. coli PGNs based on the E. coli PGN corpus.

Method 1 (*Direct look-up*) involved applying the PGNs imported from the Biothesuarus directly on the corpora with two restrictions only: regular-expression based removal of corrupted and nonsense names, and only selecting PGNs relevant to either human or E. coli.

In Method 2 (*DictFiltering*) we used the dictionary filtering rules specified in section 2.2 above.

In Method 3 (*Text+DictFiltering*) we combined the dictionary filtering features with the text-driven ones.

6.2 Evaluation on the BioCreAtivE corpus

Figure 1 shows a significant increase of the F-measure despite the decrease of recall as direct matching of PGNs is restricted by dictionary filters and then context driven restrictions.

⁷ Incidentally, this is where NER differs from traditional WSD in that a new sense is encountered which cannot be resolved to a pre-existing sense identifier.

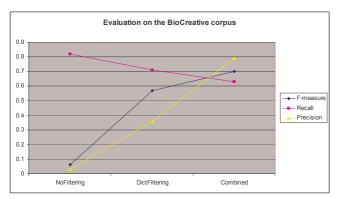


Figure 1 Precision, recall and F-measure of PGN normalization obtained based the Bio-Creative corpus.

By just applying dictionary filters based on the set of features addressing the four types of PGN polysemy introduced above, we managed to increase the precision from 0.03 to 0.36. Out of the 6 dictionary filtering features four were used by the decision tree to derive classification rules. They could be put in the following order of decreasing strength (see Table 2 for explanation of the dictionary features):

- 1. BNC frequency (Feature 1).
- 2. Medline frequency (Feature 5).
- 3. Number of clusters where a term occurs (Feature 2)
- 4. Number of distinct species taxonomy identifiers (Feature 3)

Figure 2 shows a major split in the decision space produced by the decision tree trained on static features for human PGNs.

```
bnc_freq > 15
| medlineFrequency <= 8472
| nbofSynsets <= 21
| | medlineFrequency <= 787: false (3.0)
| medlineFrequency > 787
| | bnc_freq <= 472: true (11.0/2.0)
| bnc_freq > 472: false (2.0)
| nbofSynsets > 21: false (3.0)
| medlineFrequency > 8472: false (104.0)
```

Figure 2 A fragment of the decision space split for human PGNs

Table 5 shows a number of terms which have been correctly classified as false PGNs for the BioCreAtivE corpus. *Chicken* and *alternative* are common English words as indicated by their Medline and BNC frequencies. *Tissue* and *translocation* are more of biomedical terms.

Term	Polysemy type
chicken	1
alternative	1
tissue	1,4
translocation	4
p63	3
polymerase	2

Table 5 Example true negatives filtered out with static

dictionary features in the gene name normalization task.

P63 is found in many species and therefore potentially polysemous in a PGN normalization task. *Polymerase* is an enzyme name which, in our Term Repository, has been assigned with as many as 87 UniProt accession numbers and 39 different species identifiers.

The outcome of the dictionary filtering phase was passed to the context-driven classifier, which further increased the precision to 0.79, while depressing the recall to 0.63. This is only natural, since in this experiment we treated the decisions made by the dictionary-based classifier as binding. In other words, a term marked as a negative in the dictionary-filtering stage could not be tagged by the context-driven classifier.

The most salient contextual features selected by the decision tree included intra-corpus ambiguity, Medline frequency, and the identification of a name as a PGN mention by Abner.

6.4 Evaluation on the E. coli corpus

Interestingly, the initial recall of E. coli PGNs is only 0.45, which may be partly due to the occurrences of mutant genes that have not been recorded in existing PGN resources used. Another major reason for the initially low recall is the occurrences of operon names, which we annotate with several identifiers matching all the genes on a given operon. As an example, we have assigned as many as 9 matching identifiers to the TOR (trimethylamine *N-oxide reductase*) operon. Not all of these gene names are associated with this operon in the lexical resources we have used. Yet another reason for the relatively low recall is the variability of operon names (e.g. cyoABCDE may stand for cyoA, cyoB, etc.), which occur in the corpus relatively frequently because of its gene-regulation focus. The drop in the recall as we apply the dictionary-filtering rules is rather insignificant (0.45 to 0.42) compared with the gain in the precision (from 0.15 to 0.66).

The combination of dictionary-filtering rules with the text-driven features resulted in a further increase of precision to 0.75 and a slight decrease of the recall value (0.41). The overall F-measure achieved for the E. coli PGN corpus with a combination of both methods was 0.53 (see Figure 3).

The best split of the feature space found by the decision tree for the E. coli corpus was based on a Medline frequency threshold. This criterion alone increased the precision from 0.14 to 0.66, which suggests that E. coli PGNs are much more standardized than human-relevant ones and that they do not require a complex feature set to cover specific regions of the feature space. The main split in the decision space for E. coli gene names based on Medline frequencies can be described as follows;

medlineFrequency <= 3320: true (226.0/65.0)

medlineFrequency > 3320: false (65.0/14.0)

Domain-independent polysemy as indicated by the BNC frequency feature did not play an important role in the case of E. coli PGNs, since there are very few E. coli PGNs coinciding with common English words.

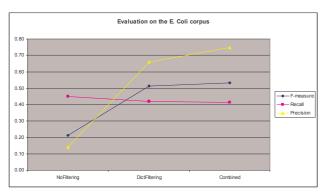


Figure 3 Precision, recall and F-measure of PGN normalization obtained based the E. coli corpus

In the decision tree generated for the context filter we find that a term's intra-corpus ambiguity constituted the strongest feature, followed by Medline frequencies and the annotations provided by the Abner tool.

Conclusions

A single lexical resource integrating terms of different semantic types contains an intrinsic representation of term polysemy. Protein and gene names are the most notoriously polysemous type of named entities found in the biomedical literature and we have identified the major types of ambiguity they can introduce. We have demonstrated how a set of features that provide indications of these polysemy types can be assigned to each PGN in our Term Repository. One advantage of the set of features we propose is that they can be assigned statically to any term in a synset-based lexicon and that they are easy to map to the types of polysemy identified. In principle, the dictionary filtering features we propose can be applied to any other semantic type in our Term Repository, although so far we only provide evaluation for PGNs. Although disambiguation based on static dictionary features does not outperform fully-fledged context-driven NER, it does effectively filter out highly polysemous terms. Additionally, static dictionary features are context-independent and thus more flexibly applicable - for instance in information retrieval systems, where running intense on-the-fly disambiguation for retrieved documents could be problematic. Once they are computed and assigned for every term in the terminological resource, static polysemy indicators can be used for ad-hoc NER, conservative query expansion or relevance feedback, independently of the context in which they are retrieved. Of course, an alternative approach to polysemy-sensitive information retrieval would involve disambiguating the collection at indexing time using fully-fledged

information extraction techniques.8

By applying dictionary filtering rules we obtained significant improvements in the F-measure of PGN identification for both corpora. As we demonstrate dictionary filtering can be further complemented with text-driven features, although the latter cannot be statically assigned to terms and need to be recomputed for the context of occurrence at hand.

Another conclusion emerges from the different precision and recall values obtained for the two corpora used for evaluation. One explanation of these differences could be that E. coli PGNs are more normalized than human PGNs (hence higher precision), but they are not as well represented in existing PGN databases (hence lower recall). Also, it seems that fewer rules are needed for the identification of truly polysemous E. coli PGNs than in the case of the more ambiguous human relevant PGNs.

Acknowledgements and funding

This research was sponsored by the EC STREP project "BOOTStrep" (FP6-028099, www.bootstrep.org).

Many thanks to Jung-jae Kim for his contribution to the preparation of the corpus.

References

Beatrice Alex, Malvina Nissim and Claire Grover, The Impact of Annotation on the Performance of Protein Tagging in Biomedical Text, In: Proceedings of LREC 2006, Genoa, Italy.

Chen L, Liu H, Friedman C., (2004), Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics. 2005 Jan 15;21(2):248-56. Epub 2004 Aug 27.

Fellbaum Ch., (1998). Wordnet - An Electronic Lexical Database, MIT Press, Cambridge, Massachusetts, London England.

Hakenberg J, Bickel S, Plake C, Brefeld U, Zahn H, Faulstich L, Leser U, Scheffer T. (2005). Systematic feature evaluation for gene name recognition. BMC Bioinformatics. 2005;6 Suppl 1:S9.

Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga-Llavori, R., Rebholz-Schuhmann, R. (2008) Assessment of disease named entity recognition on a corpus of annotated sentences, BMC Bioinformatics (to appear)

Jimeno A, Pezik P., Rebholz-Schuhmann D., (2007) Information Retrieval and Information Extraction in TREC Genomics 2007. The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings. NIST Special Publication: SP 500-274. Washington, USA.

Kim JJ., Pezik P., Rebholz-Schuhmann D. (2008) MedEvi: Retrieving textual evidence of relations between biomedical concepts from Medline. Bioinformatics Advance Access published online on April 9, 2008, Oxford, England.

Liu H., Hu Z., Zhang J. and Wu C., (2006). BioThesaurus:

⁸ One example of a retrieval engine which combines both techniques is MedEvi (Kim et al. 2008).

- a web-based thesaurus of protein and gene names. *Bioinformatics* 2006 22(1):103-105.
- Morgan A., Hirschman L. (2007) Overview of BioCreAtivE II Gene Normalization. *Proceedings of the Second BioCreAtivE Challenge Evaluation Workshop*.
- Quinlan, J. R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers
- Schuemie MJ, Mons B, Weeber M, Kors JA. (2007) Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. *Journal of Biomedical Informatics*. 2007 Jun;40(3):316-24.
- Settles B. (2005) ABNER: an open source tool for automatically tagging genes, pro-teins and other entity names in text. *Bioinformatics*. 2005 Jul 15;21(14):3191-2.
- Shipra Dingare, Jenny Finkel, Christopher Manning, Malvina Nissim, and Beatrice Alex. 2004. Exploring the Boundaries: Gene and Protein Identification in Biomedical Text. Proceedings of the BioCreative Workshop, Granada.
- Shipra Dingare, Jenny Finkel, Malvina Nissim, Christopher Manning and Claire Grover. 2004. A System For Identifying Named Entities in Biomedical Text: How Results From Two Evaluations Reflect on Both the System and the Evaluations. In The 2004 BioLink meeting: Linking Literature, Information and Knowledge for Biology at ISMB 2004. Republished as Shipra Dingare, Malvina Nissim, Jenny Finkel, Christopher Manning and Claire Grover. 2005. Comparative and Functional Genomics 6: 77-85.
- Tanabe L, Wilbur WJ. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics*. 2002 Aug;18(8):1124-32.
- Wilbur J., Smith L., Tanabe L., (2007) BioCreAtivE 2. Gene Mention Task. Proceedings of the Second BioCreAtivE Challenge Evaluation Workshop