Word Sense Disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering

Bill Andreopoulos*, Dimitra Alexopoulou and Michael Schroeder

Biotechnological Centre,
Technischen Universität Dresden, Germany
E-mail: williama@biotec.tu-dresden.de
E-mail: dimitraa@biotec.tu-dresden.de
E-mail: ms@biotec.tu-dresden.de
*Corresponding author

Abstract: With more and more genomes being sequenced, a lot of effort is devoted to their annotation with terms from controlled vocabularies such as the GeneOntology. Manual annotation based on relevant literature is tedious, but automation of this process is difficult. One particularly challenging problem is word sense disambiguation. Terms such as 'development' can refer to developmental biology or to the more general sense. Here, we present two approaches to address this problem by using term co-occurrences and document clustering. To evaluate our method we defined a corpus of 331 documents on development and developmental biology. Term co-occurrence analysis achieves an F-measure of 77%. Additionally, applying document clustering improves precision to 82%. We applied the same approach to disambiguate 'nucleus', 'transport', and 'spindle', and we achieved consistent results. Thus, our method is a viable approach towards the automation of literature-based genome annotation.

Keywords: WSD; word sense disambiguation; ontologies; GeneOntology; text mining; annotation; Bayes; clustering; GoPubMed; data mining; bioinformatics.

Reference to this paper should be made as follows: Andreopoulos, B., Alexopoulou, D. and Schroeder, M. (2008) 'Word Sense Disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering', *Int. J. Data Mining and Bioinformatics*, Vol. 2, No. 3, pp.193–215.

Biographical notes: Bill Andreopoulos received his PhD from the Department of Computer Science and Engineering of York University, Toronto, Canada (2006). He received his MSc from the Department of Computer Science of University of Toronto (2001) and his BSc from the Department of Computing and Software of McMaster University (1999). His interests include data mining, clustering, computational biology.

Dimitra Alexopoulou is currently a PhD candidate at the Technical University of Dresden, Germany. She received her MSc (2005) in Bioinformatics at the Faculty of Biology, School of Sciences, National and

Kapodistrian University of Athens, Greece. She obtained her Diploma in Biochemistry (2003) at Medical School, University of Ioannina, Greece. Her research interests include ontology engineering, data mining and structural bioinformatics.

Michael Schroeder is Professor of Bioinformatics at the Biotechnological Centre at the Technical University of Dresden, Germany. From 1998–2003 he was Lecturer and Senior Lecturer at City University, London. He received his PhD from the University of Hanover/Lisbon (1997). His interests are text mining, bio-ontologies, reasoning, protein interactions, and structural bioinformatics.

1 Introduction and motivation

Since the announcement of the Human Genome in 2000, some 200 model organisms have been sequenced and novel sequencing technologies will lead to a further increase in data generation. A prime task after sequencing a genome is the identification of genes and their annotation with relevant functions, processes, and cellular components. In order to facilitate the comparison of genomes, biologists devised a shared, species independent vocabulary, the GeneOntology (GeneOntologyConsortium, 2004), with some 20,000 terms and synonyms. Annotation of novel genomes with the GeneOntology is a manual process, in which editors read relevant literature for a gene and then decide on suitable annotation. However, manual annotation is labour-intensive: currently, there are several million genes and proteins of almost 60,000 different species represented in the public databases, but only approximately 500 of these species have had GO terms manually assigned in GOA, the GeneOntology Annotation (Camon et al., 2004, 2005). Presently, much effort is devoted to automating or aiding the annotation process (Jensen et al., 2006). In a recent text mining competition, BioCreative (Hirschman et al., 2005), one task consisted in the identification of suitable GeneOntology terms for a given gene and document. As reported by Ehrler et al. (2005) the best result in this category achieved only 20% accuracy. Identification of GeneOntology terms in literature is in general a challenging problem (Doms and Schroeder, 2005):

- *Stemming*. Often words will appear in different forms, such as ‘binding’ and ‘binds’, which can be reduced to their stem ‘bind’. However, is it valid to reduce ‘dimerization’ to ‘dimer’? The former talks about the process, the latter about the result. It is clearly not valid to reduce ‘organization’ to ‘organ’.
- *Missing words*. The text “... tyrosine phosphorylation of a recently identified STAT family member... ” should match the term ‘tyrosine phosphorylation of STAT protein’, the text “... a transcription factor that binds... ” should match the term ‘transcription factor binding’, the text “... alkalinephosphatase... ” should match the term ‘alkalinephosphatase activity’. In general, a matching can ignore words such as ‘of’, ‘a’, ‘that’, ‘activity’, etc., but obviously not ‘STAT’ or ‘alkalinephosphatase’.

- *Format of terms*. Ontology terms may contain commas, dashes, brackets, etc., which require special treatment. For 'thioredoxin-disulfide' the dash can be dropped, for "hydrolase activity, acting on ester bonds" the clause after the comma is important, but unlikely to appear as such in text. Terms containing additions such as '(sensu Insecta)' contain important contextual information, but are also unlikely to appear in text.

- *Word sense disambiguation*. Terms can have a very specific meaning in biomedical research, but mean other things in other contexts. Examples are: cell, development, envelope, spindle, death, growth, regeneration, transport, membrane, nucleus, host, reproduction, circulation, and others.

This last problem is particularly challenging (Xu et al., 2006a; Schuemie et al., 2005; Navigli and Velardi, 2005; Liu et al., 2002; Navigli et al., 2003; Schijvenaars et al., 2005; Pahikkala et al., 2005; Rebholz-Schuhmann et al., 2007; Gaudan et al., 2005b) and various approaches ranging from the use of tagged corpora, dictionaries, thesauri, supervised and unsupervised machine learning have been tried. Here, we investigate how the co-occurrence of terms and the similarity of documents can be used to infer the correct annotation.

We focus on the problem of identifying documents on biological 'development' (though we extend our performance evaluation to 'nucleus', 'transport', and 'spindle'). The GeneOntology defines 'development' as follows:

> "The biological process whose specific outcome is the progression of an organism over time from an initial condition (e.g., a zygote, a young adult or a young single celled organism) to a later condition (e.g., a multicellular animal, an aged adult or a mature single celled organism)."

On the other hand, 'development' might have other meanings in non-biological contexts. Figure 1 shows the possible co-occurrences of 'development' with other terms, illustrating its possible semantic meanings.

For training and testing of our approach we define three datasets:

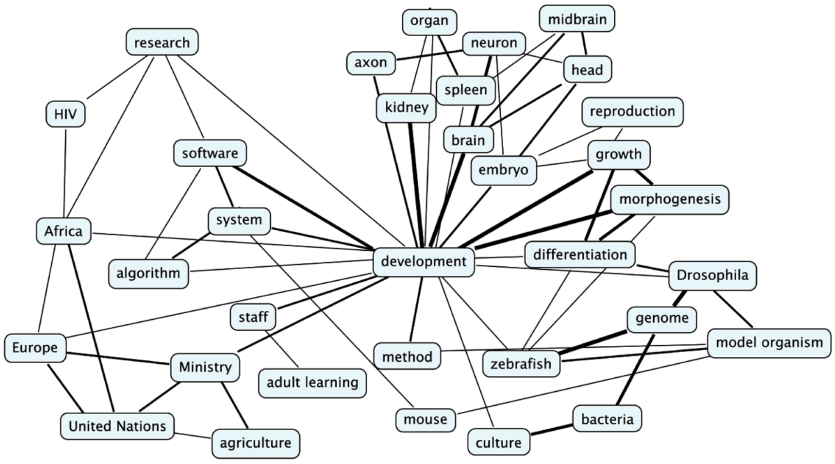
- *True Positives* (*TPs*). The documents contain the term 'development' literally and they are on developmental biology according to the curators. Examples are:
  - "Arabidopsis ribonucleotide reductases are critical for cell cycle progression, DNA damage repair, and plant *development*."
  - "Homeodomain-containing proteins are transcription factors that regulate the coordinated expression of multiple genes involved in *development*, differentiation and malignant transformation."
  - "Involvement of the TRAP220 component of the TRAP/SMCC co-activator complex in embryonic *development* thyroid hormone action."
  - "... suggesting a defect in a gene with pleiotropic effects acting during *development*."

- *False Positives* (*FPs*). These documents contain the term 'development' literally, but they are not on developmental biology according to the curators. Examples are:

- “The *development* of the Na+ gradient during illumination thus, plays an important role in energy coupling.”
- “The recent discovery of several hypothalamic factors involved in the regulation of anterior pituitary function and the *development* of sensitive immunocytochemical techniques have greatly contributed to. . . ”
- “Limited diagnostic and therapeutic interventions should be addressed as separate entities in the *development* of the patient care plan.”
- “Academic research, especially university research, tends to be substituted by *development* innovation for the production process.”

The difference of the first two False Positives from the last two is that the former will contain other GeneOntology terminology, while the latter are about general topics and thus, they do not contain any other terms.

Figure 1 The semantic meanings of ‘development’ are indicated by its co-occurrences with other terms. Besides its biological meaning, ‘development’ might also refer to software development, economic development, and others (see online version for colours)



- *False Negatives* (*FNs*). These documents do not contain the term ‘development’ literally, but they are on developmental biology according to the curators. Examples are:
  - “RYBP, a new repressor protein that interacts with components of the mammalian Polycomb complex, and with the transcription factor YY1”. The protein YY1 is annotated in the Uniprot sequence database as ‘development’.
  - “Virtual cloning and physical mapping of a human T-box gene, TBX4”. The protein TBX4 is annotated in the Uniprot sequence database as ‘development’.

- "Isolation of two novel WNT genes, WNT14 and WNT15, one of which (WNT15) is closely linked to WNT3 on human chromosome 17q21". The protein WNT14 is annotated in the Uniprot sequence database as 'development'.
- "EDF-1, a novel gene product down-regulated in human endothelial cell differentiation". The protein Endothelial Differentiation-related Factor 1 (EDF-1) is annotated in the Uniprot sequence database as 'development'.

The key problem is the identification of FPs and FNs in datasets of automatically annotated papers (e.g., GoPubMed). One key idea of this paper is that term co-occurrences in a training dataset of manually annotated papers (e.g., GOA) can help to solve the problem. With the co-occurrences, we can define term groups, which are associated with a given term. For example, does 'cell proliferation' occur frequently with 'development'? This knowledge can be put to use in two ways: If a document under examination contains 'development' but none of its co-occurring terms from the hand-curated data, then 'development' is likely to be a False Positive. If a document does not contain 'development', but some of its frequently co-occurring terms, then, it is likely to be a FN. Co-occurrence of terms can be defined in various ways. Here, we examine two approaches: First, we calculate the likelihood of co-occurrence, i.e., the number of documents in which two terms co-occur divided by the total number of documents. This likelihood does not take into account the probability of each of the terms occurring. A rarely occurring term should get a higher score than a frequently occurring term. Therefore, we define a score based on the BLOSUM (Henikoff and Henikoff, 1992) approach to substitution matrices in sequence comparison. In this context, the BLOSUM score is the logarithm of the probability of two terms co-occurring divided by the probability of the two terms occurring.

Besides the use of term co-occurrences, we also cluster documents by automatically derived annotations of the GoPubMed algorithm (Doms and Schroeder, 2005). For clustering, we use MULIC, a clustering algorithm for categorical data (Andreopoulos et al., 2006, 2007a, 2007b). The clusters are organised in layers and for each layer of documents we assign an annotation based on the likelihood method in the first step. In the same way that co-occurring terms can give a clue for the correct annotation, grouping documents with similar annotations can further improve precision and recall.

The main contributions of this paper are summarised as follows:

- We propose a methodology for finding whether an automated annotation, such as 'development', is likely to be a FP or FN. This methodology is based on co-occurrences of non-'development' annotations with 'development' in manually annotated papers of a training dataset. For this purpose we employ a co-occurrence table of all GeneOntology terms with 'development', which can be conceptualised as a co-occurrence graph, and probabilistic metrics.
- We extend our methodology with clustering of papers. Clustering deals with the issue that both the automatically and manually annotated papers are often incomplete and relevant annotations may be missing (FNs). With clustering we can aggregate annotations in a group of automatically annotated papers, rather than in a single paper. Groups of automatically annotated papers (clusters) then show which co-occurrences are more relevant to the papers in question.

The rest of this paper is organised as follows: Section 2 provides an overview of related work. Section 3 discusses the datasets and some statistics on the terms that co-occur most frequently with ‘development’. Section 4 describes the methodology we used in detail. Section 5 discusses the experimental results, precision and recall. Section 6 concludes the paper.

2 Related work

During the last years, word sense disambiguation has become a hot topic in the biomedical domain. The challenge for WSD in the biomedical domain is the rapid growth of the biomedical literature in terms of new words and their senses, with the situation getting worse with the use of abbreviations and synonyms. Quoting Ide and Véronis (1998), “WSD work has come full circle, returning most recently to empirical methods and corpus-based analyses that characterised some of the earliest attempts to solve the problem”. This illustrates the exact need in the case of the biomedical domain; the development of statistical approaches that utilise ‘established knowledge’ (like thesauri, dictionaries, ontologies and lexical knowledge bases) and require no or only some parsing of the text in order to perform the correct annotation.

Two main decision points for WSD in the biomedical domain are the selection of an appropriate corpus for training and evaluating the method and the granularity to which WSD should be performed. The first decision is still a major bottleneck due to the absence of manually annotated corpora and biomedical datasets. There exist some freely available manually annotated datasets – such as the NLM WSD test collection (http://wsd.nlm.nih.gov/), Medstract (www.medstract.org) and the BioCreAtIvE set (www.mitre.org/public/biocreative) – but they can be used only in very specific cases. In general, one needs to create own gold standard datasets depending on the task for which they will be used. The process is usually manual or semi-automatic and labour-intensive. The second decision is at what detail we need the disambiguation to happen. For example, there is a big difference between the sense of ‘bank’ as a building and the ‘BANK’ gene (B-cell scaffold protein with ankyrin repeats 1), whereas the difference between ‘BANK’ as the gene name and the gene product (BANK protein) is smaller. The discrimination between the gene name and product would be more difficult compared to the building/gene case, as the two senses would appear in the same biological context.

There are several categories of WSD algorithms that have been used so far in the biomedical domain (Ide and Véronis, 1998; Stevenson and Wilks, 2001; Schuemie et al., 2005; Edmonds and Agirre, 2006). The main distinction is between methods using established knowledge, supervised (with the use of an initial training set) and unsupervised learning methods. In the first category there have been developed some approaches, especially in the problem of gene/protein symbols’ abbreviations. Wren et al. (2005) present a collection of four databases maintaining a vast list of abbreviations together with their meaning. Schijvenaars et al. (2005) and Pahikkala et al. (2005) developed two approaches to resolve gene/protein symbols. Schijvenaars et al. (2005) achieve 92.5% accuracy on human gene symbols. The authors compare a gene’s definition compiled from a database to abstract where the gene symbol occurs. Both definition and abstract are represented as concept finger prints, i.e., vectors of biomedical terms. Both vectors are compared by a similarity measure based on cosine. Humphrey et al. (2006) used lately the Journal Descriptor Indexing (JDI) methodology to handle the ambiguity problem when trying to map free text to terms from the UMLS metathesaurus. JDI combines a statistical, corpus-based method with utilisation of pre-existing medical domain knowledge sources. For the 45 ambiguities studied, the overall average precision of the highest-scoring JDI method was 78.7% compared to 25% for their baseline method based on the frequency counts of MeSH terms in a document subset.

A lot of other approaches have used supervised Machine Learning (ML) for WSD in the biomedical domain. Hatzivassiloglou et al. (2001) developed an automated system for assigning protein, gene and mRNA labels to free text. He used three ML techniques, namely naive Bayesian learning, decision trees and inductive rule training and investigated the contribution of different features of textual information (like stopword removal, stemming, positional information of surrounding words) with final accuracy rates up to 85%. Ginter et al. (2004) worked on the disambiguation between gene and protein symbols, by introducing a new family of classifiers based on ordering and weighting of the feature vectors obtained from word counts and word co-occurrence in text. This method achieved 86.5% accuracy. Liu et al. (2002, 2004) showed that there is a need for a larger window size for disambiguation of words in the biomedical domain. Liu et al. (2002) use UMLS (Bodenreider, 2004) as ontology. Similar to our approach, they identify UMLS concepts in abstracts and analyse the co-occurrence of these terms with the term to be disambiguated. The correct sense is inferred from the majority sense associated with the co-occurring UMLS terms. Similar to our approach, co-occurrence is defined using a Bayes approach. The authors achieve a precision of 93% and a recall of 47%. Gaudan et al. (2005a) used SVMs on their algorithm to resolve abbreviations in MEDLINE and obtained a precision of 98.9% and a recall of 98.2%. Excluding rare senses (appearing in less than 40 documents) from the test set and keeping in the training set only the ambiguous short-forms that also had long-forms in the documents made the disambiguation task easier. Pahikkala et al. (2005) follow a similar approach with Schijvenaars et al. (2005). But instead of using the full abstract, they define the context of a gene symbol as a number of words before and after. The size of the context can be varied and optimised. The context is represented as a vector and a support vector machine is trained. They achieve 85% accuracy. Support vector machines are widely used in word sense disambiguation. Their performance depends on a number of parameters such as the sample size, sense distribution and degree of difficulty (Xu et al., 2006b). The authors establish that small datasets and clear or fuzzy borderline between senses impact on the classification task.

There are several cluster-based approaches using unsupervised ML. Schütze and Pedersen (1995), Schütze (1998) adapts Latent Semantic Analysis/Indexing (LSA/LSI) to represent entire contexts rather than single word types using second-order co-occurrences of lexical features. Pedersen and Bruce's (1997, 1998) work with average linkage clustering relies on a small number of first-order features to create matrices that show the pairwise similarity between contexts. These features are localised around the target word and include word co-occurrences and PoS tags. Purandare and Pedersen (2004) have tested a variety of similar algorithms obtaining an average F-measure of 44%. Yarowsky (1995) and Mihalcea (2004) have used the 'self-learning' (or 'co-learning') approach for WSD. This method is based on classifier(s) trained on a small amount of manually tagged data. The same classifiers are then used to tag new data and the most confident predictions are added to the labeled dataset. Yarowsky achieved an accuracy of 96.5% on a test set of 12 ambiguous words with an average of 4000 instances per word. Mihalcea used the same approach on the Senseval-2 generic English corpus and resulted in an improvement of 9.8% over the baseline score using a Bayesian classifier. Dorow and Widdows (2003)'s approach is based on a graph model representing words and relationships (co-occurrences) between them. Sense clusters are iteratively computed by clustering the local graph of similar words around an ambiguous word. The ambiguous words can be identified by looking at the nodes connecting otherwise unrelated clusters. These clusters represent the different senses of the word and then the labels are assigned according to WordNet (Fellbaum, 1998), a dictionary of terms and their definitions.

There are some approaches based on co-occurrences of terms and established knowledge. In the approach pursued by Navigli and Velardi (2005), and Navigli et al. (2003), the authors use WordNet (Fellbaum, 1998) to identify ambiguities in single words and in a following step of terms composed of multiple terms. Each term is represented by its interconnection to other terms. In Liu et al. (2002) a Bayes approach is pursued to weigh these co-occurrences. Then, senses are disambiguated by inferring the correct sense from unambiguous senses of co-occurring terms. Additionally to the co-occurrence, Navigli et al. (2003) uses decision tree learning to derive rules to relate concepts. An important difference between our approach and the other approaches is that we construct the co-occurrences based on GOA, which is a manually annotated dataset. Therefore, our graph contains only relations (edges) between terms (nodes) that are semantically meaningful in the context of an paper (True Positives). Dorow's graph (Dorow and Widdows, 2003) contains all the nouns that co-occur with one another, but in the case of the biological context, we are interested only in a local subgraph of Dorow's graph (i.e., 'development' only in the biomedical sense). Another discriminative difference is that we use established knowledge in the Gene Ontology to draw the nodes. Moreover, in order to account for the numerous False Negatives (FNs) in GoPubMed we perform clustering based on the available GoPubMed annotations.

3 Datasets

Our training and test datasets are GOA and GoPubMed, respectively. GoPubMed represents papers *automatically* annotated with GO terms (Doms and Schroeder, 2005), while GOA represents papers *manually* annotated with GO terms (Camon et al., 2005, 2004). GoPubMed consists of approximately 15,000,000 papers and GOA consists of approximately 34,000 papers. We map each GoPubMed paper's annotations onto the corresponding subsection of the GOA corpus.

Development. We generated three datasets containing papers from GoPubMed that are True Positives (TPs), False Positives (FPs) and False Negatives (FNs) with respect to the 'development' annotation:

- 122 FNs papers. 'Development' annotation: no in GoPubMed, yes in GOA.
- 109 TPs papers. 'Development' annotation: yes in GoPubMed, yes in GOA.
- 100 FPs papers. 'Development' annotation: yes in GoPubMed, no in GOA.

Then, we united the TPs, FPs and FNs for ‘development’ into one test dataset of 331 papers in total. Our datasets did not include True Negatives, since there are too many TNs for ‘development’ in GoPubMed and we do not want to automatically detect TNs.

Nucleus, Spindle, Transport. Similarly, we generated datasets containing papers from GoPubMed that are TPs, FPs and FNs with respect to each of the following annotations: ‘nucleus’, ‘spindle’, and ‘transport’. Then, we united the TPs, FPs and FNs for each annotation into a test dataset as Table 1 shows.

Table 1 Number of GoPubMed papers that are TPs, FPs and FNs in the test dataset for each annotation

Annotation	*TPs*	*FPs*	*FNs*	*Total*
Development	109	100	122	331
Nucleus	100	99	100	299
Spindle	48	50	7	105
Transport	91	50	156	297

Table 2 The top 10 GO annotations in GoPubMed and GOA, according to their co-occurrence with ‘development’

GoPubMed			*GOA*		
Term name	*cooc.*	*BLOSUM*	*Term name*	*cooc.*	*BLOSUM*
cell	200142	0.21	cell proliferation	25	2.40
growth	80751	0.62	transcription factor activity	23	1.29
biosynthesis	69146	0.17	regulation of transcription, DNA-dependent	22	1.95
cell development	46722	2.56	protein binding	20	−0.21
viral life cycle	45527	−0.01	nucleus	20	−0.04
antigen binding	45448	0.06	signal transduction	17	0.9
brain development	39119	0.22	integral to plasma membrane	15	0.66
cellularisation	35330	0.4	DNA binding	14	0.8
binding	35042	−0.14	cytoplasm	11	−0.21
regulation of biological process	33777	0.45	apoptosis	11	1.88
behaviour	33306	0.099	immune response	10	1.25

3.1 GoPubMed and GOA statistics

The key data for the first step of our approach are terms co-occurring with ‘development’. Tables 2 and 3 show the top ten terms associated with ‘development’ according to the number of co-occurrences and the log-odds BLOSUM score, respectively. Both tables are broken down into terms according to GoPubMed’s automated, comprehensive, but more error prone annotation and GOA’s manual, less comprehensive, but higher quality annotation. The first row in Table 2 shows that ‘cell’ is the term appearing most frequently with ‘development’ in GoPubMed.

The relatively low BLOSUM score (negative unlikely, positive likely) reflects, that the term is very general and hence not very predictive for 'development'. However, 'cell' is very effective for separating papers on cell biology from medical abstracts. The most frequently co-occurring term in GOA is 'cell proliferation'. It also has a good BLOSUM score.

Table 3 shows the top terms according to the BLOSUM score. The first line shows for GoPubMed 'petal development', which is clearly related to 'development' as it is a more specific term in the ontology, while the GOA annotation shows the extremely specific and term "3-mercaptopyruvate sulfurtransferase activity", which co-occurs only once. As GOA is limited in size, high BLOSUM scores come with low co-occurrences. The reason could be that very specific proteins like TBX4, YY1, etc. (see Section 1) are indicative of correct annotation with 'development' and these proteins are in turn correlating very well to the very detailed ontology terms listed in Table 3.

Table 3 The top 10 GO annotations in GoPubMed and GOA, according to their BLOSUM score with 'development'

GoPubMed			*GOA*		
Term name	*BLOSUM*	*cooc.*	*Term name*	*BLOSUM*	*cooc.*
petal development	2.55	78	3-mercaptopyruvate sulfurtransferase activity	4.95	1
sepal development	2.55	19	hydrolase activity, acting on acid anhydrides, catalysing transmembrane movement of substances	4.95	1
stamen development	2.55	80	intramolecular transferase activity, phosphotransferases	4.95	1
carpel morphogenesis	2.55	3	carbon-nitrogen ligase activity, with glutamine as amido-N-donor	4.95	1
sepal morphogenesis	2.55	2	lipoate-protein ligase B activity	4.95	2
stamen morphogenesis	2.55	2	transcription initiation factor activity	4.95	2
carpel structural organisation	2.55	1	sigma factor activity	4.95	1
establishment of petal orientation	2.55	2	glutamyl-tRNA(Gln) amidotransferase activity	4.95	1
meristem development	2.55	215	protein prenylation	4.95	1
gut development	2.55	578	protein amino acid prenylation	4.95	2
regulation of post-embryonic development	2.55	13	alkane 1-monooxygenase activity	4.95	1

The co-occurrence of terms can be extended to pairs frequently co-occurring with 'development'. Tables 4 and 5 summarise the joint probabilities of 'development' with the two most frequent terms. For GoPubMed (Table 4) the terms cover cell growth and differentiation in general, while for GOA (Table 5) the terms related to transcription are prominent. Both tables appear intuitively meaningful topics to indicate abstracts on cell biology.

Table 4 The top 10 pairs of Non-'development' GO annotations in GoPubMed, according to their probability of co-occurring with 'development'

GoPubMed		
Term name A	*Term name B*	*Prob.* (*A, B, development*)
cell growth	growth	0.0010374
cell	cell growth	0.001078
cell	cell surface	0.001080
cell	cell differentiation	0.001247
cell	regulation of biological process	0.001296
cell	binding	0.001296
cell	cellularisation	0.00168
cell	antigen binding	0.001722
cell	biosynthesis	0.002363
cell	growth	0.002714
cell	cell development	0.0031

Table 5 The top 10 pairs of Non-'development' GO annotations in GOA, according to their probability of co-occurring with 'development'

GOA		
Term name A	*Term name B*	*Prob.* (*A, B, development*)
signal transduction	cell proliferation	0.000145573121379
DNA binding	transcription factor activity	0.000174687745655
DNA binding	nucleus	0.000174687745655
transcription factor activity	transcription from RNA polymerase II promoter	0.000174687745655
protein binding	nucleus	0.000174687745655
protein binding	regulation of transcription, DNA-dependent	0.000174687745655
nucleus	regulation of transcription, DNA-dependent	0.000174687745655
transcription factor activity	regulation of transcription, DNA-dependent	0.00020380236993
protein binding	cytoplasm	0.00020380236993
transcription factor activity	nucleus	0.000262031618482

4 Methodology

Our objective is to find two classes of papers, those that:

- should *not* be annotated with 'Development', i.e., False Positives (FPs)
- should be annotated with 'Development', i.e., False Negatives (FNs) or True Positives (TPs).

We map the annotations in each GoPubMed paper to a graph representing co-occurrences of annotations in $\sim 34,000$ *manually* annotated papers in GOA. Based on several probabilistic metrics described below we infer the likelihood that 'development' should or should not annotate each paper.

4.1 Overview

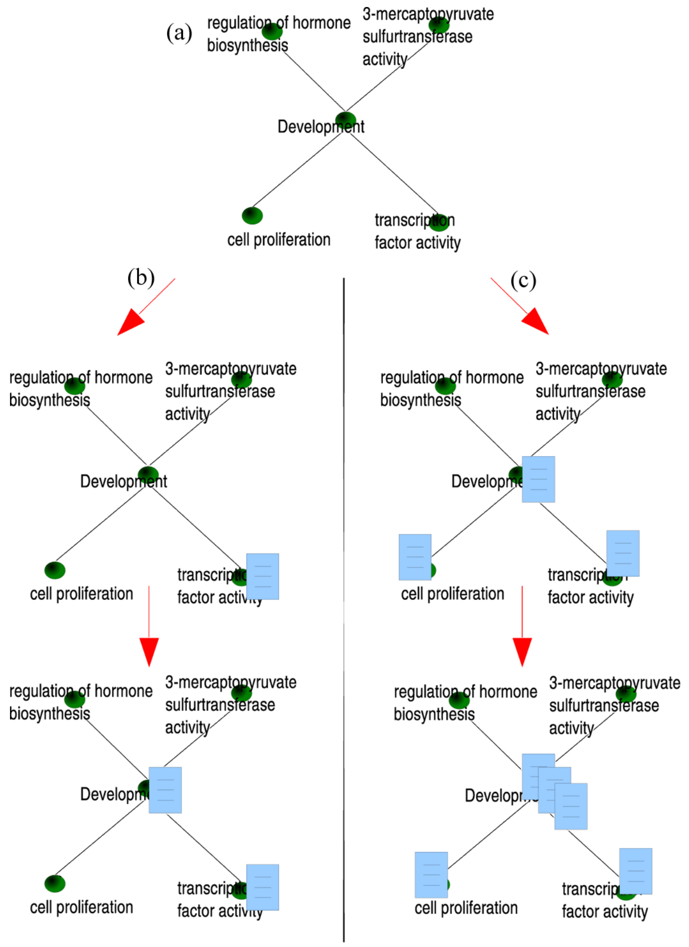
Figure 2 illustrates our methodology's main steps. Our methodology uses co-occurrences based on manually annotated GOA papers. We find co-occurring terms in all GOA papers and build a co-occurrence table, which can be conceptualised as a co-occurrence graph, representing how frequently pairs of GOA terms co-occur in all GOA papers. The nodes represent annotations and edges represent the frequency of co-occurrence of two annotations. We view each GoPubMed paper as representing co-occurring GoPubMed annotations. Our approach involves mapping each GoPubMed paper onto the co-occurrence graph of manual GOA annotations. Each GoPubMed paper is mapped to the nodes and edges of the GOA co-occurrence graph. Then, we use several metrics to estimate the likelihood of a 'development' annotation being appropriate to the GoPubMed paper, based on an n-word of n annotations that are neighbours of 'development' in the GOA co-occurrence graph.

GOA is sparsely annotated because of the effort required in assigning manual annotations. For this reason, we use the GOA co-occurrence graph such that high correlations of annotations with 'development' are considered more significant than low correlations. In the GOA co-occurrence graph an annotation a_i's low correlation with 'development' is not a very strong sign for a FP. On the other hand, an annotation a_i's high correlation with 'development' is a stronger sign for a FN or TP. With this rationale, we assign to each paper a 2-word, including 'development' and the paper's annotation most closely correlated with 'development' in the GOA co-occurrence graph. We use these 2-words with probabilistic metrics to assess which papers are most likely to be relevant to 'development' (TPs/FNs); the rest of the papers are considered more likely not to be relevant to 'development' (FPs). The use of 2-words is specific to our application, which classifies papers as TPs/FNs based on the annotation most correlated with 'development'; however, n-words for any n could potentially be used.

We also propose a clustering methodology for finding groups of GoPubMed papers (clusters) that are FPs or FNs/TPs. Our clustering methodology improves the results, since many GoPubMed papers are incomplete with missing annotations (FNs) or have wrong annotations that should be filtered out (FPs). Moreover, most annotations occur infrequently in GOA. Clustering allows to aggregate information on the occurrences of annotations over all GoPubMed papers. Clustering allows to build groups of

GoPubMed papers and to identify the union set of papers' annotations in a group. We can map the union of annotations in a group of GoPubMed papers onto the GOA co-occurrence graph, rather than each individual paper's annotations. This allows to return to the user for examination groups of papers that are likely to be FPs or FNs/TPs, rather than individual papers. For some datasets this makes the process of looking for FPs and FNs/TPs more accurate. This also makes the process faster by avoiding redundant mappings of GoPubMed papers with the same annotations to the GOA co-occurrence graph.

Figure 2 (a) An example of a GOA co-occurrence graph; (b) mapping a GoPubMed paper's annotations onto the GOA co-occurrence graph and (c) mapping a group (cluster) of GoPubMed papers annotations onto the GOA co-occurrence graph (see online version for colours)


(a)
regulation of hormone biosynthesis
3-mercaptopyruvate sulfurtransferase activity
Development
cell proliferation
transcription factor activity
(b)
(c)

4.2 *Co-occurrence graph of GOA Annotations*

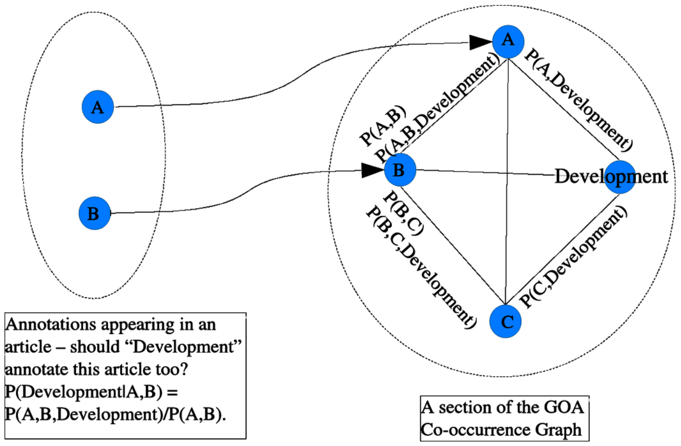
In order to formalise the notion of GO annotations' co-occurrences, we consider pairs of GO terms that appear in the same paper's abstract and we represent all such pairs of

GO terms in a GOA co-occurrence graph. We created a GOA co-occurrence graph for the set of all manually annotated papers in the GOA database. GOA papers annotated with ‘development’ are most likely to be TPs. Each node represents a GO annotation. An edge between nodes a_1 and a_2 represents:

- A probability $P(a_1, a_2, development)$ representing the likelihood of co-occurrence of the terms a_1, a_2 and ‘development’ in all GOA papers. For an edge between a_1 and the ‘development’ node this is equal to $P(a_1, development)$.
- A real number, called a BLOSUM score, representing the frequency $BLOSUM(a_1, a_2)$ of the terms’ a_1, a_2 co-occurrence over all papers. The BLOSUM score for a pair a_1 and a_2 is estimated as: $log\frac{P(a_1,a_2)}{P(a_1)P(a_2)}$.

Figure 3 shows an example of mapping a set of GoPubMed annotations for an paper onto the GOA co-occurrence graph. Using special metrics we assess the likelihood of ‘development’ being a TP/FN or FP for an paper.

Figure 3 Mapping a GoPubMed paper’s annotations onto the GOA co-occurrence graph. A GoPubMed paper point to the edges of the GOA co-occurrence graph corresponding to pairs of co-occurring GOA annotations (see online version for colours)


A
B
A
B
C
Development
P(A,B)
P(A,B,Development)
P(A,Development)
P(B,C)
P(B,C,Development)
P(C,Development)
Annotations appearing in an article – should “Development” annotate this article too? P(Development|A,B) = P(A,B,Development)/P(A,B).
A section of the GOA Co-occurrence Graph

4.3 BLOSUM metric for finding ‘development’ TPs, FPs, FNs

The first metric is the BLOSUM score, which was presented in the previous section. For an paper we find the annotation a_1 which is the most correlated to ‘development’ in the GOA co-occurrence graph. (If there is a tie, it will not affect the result.) Then, we assign to the paper the $BLOSUM(a_1, development)$ score.

The rationale for considering only one non-‘development’ annotation a_1 for each paper, as described in Subsection 4.1, is that we consider annotations most closely correlated with ‘development’ as most reliable for classifying a GoPubMed paper as FN/TP.

4.4 Naive Bayes metric for finding 'development' TPs, FPs, FNs

We are given a set of annotations $a_1, \ldots, a_n$ (for our purposes $n = 1$) that are prominent in an paper (or a group of papers derived via clustering). Suppose the GOA co-occurrence graph suggests that given these annotations $a_1, \ldots, a_n$ co-occurring in an paper, the 'development' annotation has a probability π of correctly describing this paper (TP/FN). Then, we estimate the likelihood that 'development' is a correct annotation for a GoPubMed paper, based on several manual GOA annotations $a_1, \ldots, a_n$. This pseudo-Bayesian application is a simplification of strict statistical Bayesian rules for making our method practically useful. We believe it works since:

- GOA manual annotations $a_1, \ldots, a_n$ are less likely to be FNs or FPs than automatic GoPubMed annotations
- the relatively infrequent automatic GoPubMed annotations $a_1, \ldots, a_n$, are usually less likely to be FNs or FPs than 'development'.

We consider the frequency of co-occurrence of GoPubMed papers' annotations with 'development' in the GOA co-occurrence graph. We rank as most likely FNs/TPs the GoPubMed papers with annotations that most frequently co-occur with 'development' in GOA. The GoPubMed papers ranked as most likely FPs are those with annotations that co-occur less frequently with 'development' in GOA. Our evaluation of the likelihood of co-occurrence is based on the following Naive Bayes probability:

$$P(a_1, \ldots, a_n, \text{development}) = P(a_1, \ldots, a_n \mid \text{development}) P(\text{development})$$

where $\{a_1, \ldots, a_n\}$ = set of n GoPubMed annotations that
co-occur in GOA co-occurrence graph with 'development'.

$P(development)$ remains constant and it will not affect our decision on which papers should or should not be annotated with 'development'. As we see, $P(a_1, \ldots, a_n, development)$ correlates with $P(a_1, \ldots, a_n \mid development)$; this can be interpreted as the likelihood that an paper would be annotated as $a_1, \ldots, a_n$, assuming it was annotated with 'development'.

Given a GoPubMed paper, we consider its annotation a_1 that is most closely correlated with 'development' in the GOA co-occurrence graph. Then, we find $P(a_1, development)$; the GOA co-occurrence graph supports retrieving this value. The reason why we consider only one annotation a_1 for an paper is that GOA annotations are often incomplete and some co-occurrences with 'development' are lower than they should be. Thus, considering the GOA annotations that co-occur less frequently with 'development' might bias our evaluation of 'development' being a FN/TP or FP for an paper.

4.5 A threshold for separating likely FNs/TPs from FPs

We set a threshold for each of the two metrics described above, to separate:

- GoPubMed papers that are 'development' FNs or TPs. These papers are often manually annotated as 'development' in GOA.
- GoPubMed papers that are 'development' FPs. These papers are automatically annotated as 'development' in GoPubMed, but often do not have this manual annotation in GOA.

By comparing each GoPubMed paper's annotations to the GOA co-occurrence graph we establish a *threshold* for the values of each metric previously described. The threshold separates papers into two groups: FNs/TPs from FPs. We examine the appropriate value of *threshold* in the next section on experiments.

4.6 Clustering of GoPubMed papers

We use the MULIC clustering algorithm for partitioning the GoPubMed papers into clusters of papers with similar annotations. If a cluster of papers contains a set of GoPubMed annotations that in the GOA co-occurrence graph are correlated with 'development', then 'development' is likely a TP or FN for the papers. If the cluster's set of GoPubMed annotations are not correlated with 'development' in the GOA co-occurrence graph, then 'development' is more likely to be a FP for the papers.

The objects to be clustered are the GoPubMed papers. Clustering decisions are based on each paper's set of GoPubMed annotations. Each cluster has a *mode*, which is the union of annotations of all paper members of the cluster. The first paper is inserted into a new cluster, and the paper becomes the mode of the cluster. Then, the process continues iterating over all papers that have not been assigned to clusters yet, and each paper o is inserted into the cluster c with the most similar mode μ_c. The similarity is the intersection of o and mode μ_c.

The dissimilarity is the difference between o and the mode μ_c of the closest cluster. The dissimilarity criterion ϕ indicates how high the dissimilarity is allowed to be between an paper and the closest cluster's mode, for the paper to be inserted into the cluster. Initially ϕ equals 1, meaning that only one annotation can differ between an paper and the closest cluster's mode. If the number of different annotations between the paper and the closest cluster's mode is greater than ϕ, then, the paper is inserted into a new cluster on its own, else, the paper is inserted into the closest cluster and the mode is updated. At the end of each iteration, any cluster of size one is removed so that its paper will be re-clustered at the next iteration. The dissimilarity criterion for inserting papers in clusters is relaxed gradually; at the end of each iteration, if no paper has been assigned to a cluster of size greater than one, then the variable ϕ is incremented by 1. The iterative process stops when all papers are assigned to clusters of size greater than one.

The modes and clusters are influenced most by the annotations of the papers that are clustered first in top cluster layers. It makes more sense to cluster first the papers of low degree (with few annotations), and last the papers with the most annotations. Two papers of high degree are unlikely to have the exact same annotations, thus, it is unlikely that there will be many papers of high degree in top cluster layers. By ordering the papers and presenting them to the clustering process from low to high degree, and by gradually relaxing ϕ, the clusters get an onion-layered structure where papers in top layers have similar sets of annotations and papers in bottom layers have less similar sets of annotations.

5 Experimental evaluation

Each paper had an original classification as FP, FN or TP with respect to the 'development' annotation. Our goal was to find out whether an paper could be classified correctly as FP, FN, or TP based on its mapping to the GOA co-occurrence graph.

We are interested in papers that were erroneously automatically annotated as 'development' (FPs), or should be automatically annotated as 'development' (FNs or TPs). In order to evaluate the success of our methodology for separating likely FPs from FNs and TPs we used the precision/recall measure, as described next.

5.1 Precision (P) and Recall (R)

We think of precision and recall as indicative values (percentages) of the ability of our methodology to reconstruct the existing classes in the dataset. Without loss of generality assume that the optimal mapping assigns class c_i to the retrieved group of papers g_i.

There are two known classes in our test dataset S :

- c_1 consisting of papers known to be FNs/TPs
- c_2 consisting of papers known to be FPs.

Our result consists of two groups of papers, g_1 and g_2, the former believed to be FN/TP papers and the latter believed to be FP papers. We define precision, P_i, and recall, R_i, for a group of papers g_i, $1 \leq i \leq 2$ as follows:

$$P_i = \frac{|g_i \cap c_i|}{|g_i|} \quad \text{and} \quad R_i = \frac{|g_i \cap c_i|}{|c_i|}.$$

P_i and R_i take values between 0 and 1 and, intuitively, P_i measures the accuracy with which group g_i reproduces class c_i, while R_i measures the completeness with which group g_i reproduces class c_i. We define the precision and recall of the result as the weighted average of the precision and recall of each group of papers. More precisely:

$$P = \sum_{i=1}^{k} \frac{|c_i|}{|S|} P_i \quad \text{and} \quad R = \sum_{i=1}^{k} \frac{|c_i|}{|S|} R_i$$

$|S|$ is the size of the test dataset, i.e., the number of papers. In the case of 'development' this is 331.

5.2 Results for 'Development'

Results for classifying papers individually. In order to classify papers we define a *threshold* drawing a line that separates likely FPs from FNs and TPs. Tables 6 and 7 show the precision, recall, and F-measure ($\alpha = 1$) achieved for the two metrics and different values of threshold; these tables show how effectively each metric and threshold partition the papers into the classes of FPs and FNs/TPs.

As shown, for the BLOSUM metric the best partitioning is achieved with a threshold value of 0. This points to the significance of the results, since 0 would be the natural choice for the BLOSUM threshold value for separating FPs from FNs/TPs.

For the Naive Bayes metric (range from 0 to 1) the best partitioning is achieved with a threshold value of 0.00005. The best F-measure for the Naive Bayes metric is 0.77, slightly better than the first BLOSUM metric. The reason for the improved result may be that the BLOSUM metric is slightly biased by considering in its denominator the likelihood of individual annotations' occurrences. While the Naive Bayes metric just considers the likelihood of co-occurrences of annotations.

Results with clustering. We clustered all 331 automatically annotated GoPubMed papers in our dataset. For clustering we excluded the GoPubMed 'development' annotations, and we did not use the manually annotated GOA papers. We got 22 clusters, where each cluster had on average four layers. Most clusters had a top layer containing a set of annotations representative of corpus groups of papers.

Table 6 Results for the BLOSUM metric (without MULIC clustering of papers) on the 'development' test dataset

Threshold	*Precision*	*Recall*	*F-measure*
−1.0	0.74	0.74	0.74
−0.5	0.74	0.74	0.74
0	0.74	0.74	0.74
0.5	0.74	0.74	0.74
1.0	0.73	0.73	0.73
1.5	0.72	0.71	0.71
2.0	0.71	0.7	0.7
3.0	0.68	0.55	0.61

Table 7 Results for the Naive Bayes metric (without MULIC clustering of papers) on the 'development' test dataset

Threshold	*Precision*	*Recall*	*F-measure*
0.0000	0.74	0.74	0.74
0.00001	0.74	0.74	0.74
0.00002	0.74	0.74	0.74
0.00003	0.77	0.77	0.77
0.00004	0.77	0.77	0.77
0.00005	0.77	0.77	0.77
0.00006	0.75	0.65	0.7
0.00007	0.75	0.65	0.7
0.00008	0.75	0.65	0.7
0.00009	0.77	0.61	0.68

We consider each cluster as a distinct group of GoPubMed papers, the combined annotation set of which is mapped onto the GOA co-occurrence graph. Then, we classify each cluster as FP or FN/TP. Papers in different clusters might have dissimilar GoPubMed annotations and there is a large number of annotations in the dataset.

Then, we used the metrics previously described for finding whether the 'development' annotation is more likely to be a FP or FN/TP for a group of GoPubMed papers. We examined how accurately the neighbourhood of the GOA co-occurrence graph corresponding to the group's papers' annotations reflects whether 'development' is or is not appropriate for the group.

Tables 8 and 9 show that the results with clustering are improved; the best precision is 0.82 and the best F-measure is 0.78. The threshold values that give the best results are the same as without clustering; for the first BLOSUM metric the best partitioning is achieved with a threshold value of 0, while for the Naive Bayes metric with a threshold value of 0.00005.

The main reason for the improved results with clustering is that we aggregate information on annotations of clusters of related GoPubMed papers. This way, papers that are incomplete with missing annotations have a less negative effect on finding 'development' FPs, FNs and TPs.

Table 8 Results for the BLOSUM metric and with MULIC clustering of papers on the 'development' test dataset

Threshold	*Precision*	*Recall*	*F-measure*
−1.0	0.82	0.72	0.77
−0.5	0.82	0.72	0.77
0	0.82	0.72	0.77
0.5	0.82	0.72	0.77
1.0	0.79	0.73	0.76
1.5	0.78	0.74	0.76
2.0	0.79	0.75	0.77
3.0	0.71	0.7	0.7

Table 9 Results for the Naive Bayes metric and with MULIC clustering of papers on the 'development' test dataset

Threshold	*Precision*	*Recall*	*F-measure*
0.0000	0.82	0.72	0.77
0.00001	0.82	0.72	0.77
0.00002	0.82	0.72	0.77
0.00003	0.82	0.75	0.78
0.00004	0.82	0.75	0.78
0.00005	0.82	0.75	0.78
0.00006	0.77	0.75	0.76
0.00007	0.77	0.75	0.76
0.00008	0.77	0.75	0.76
0.00009	0.75	0.73	0.74
0.0001	0.75	0.73	0.74

5.3 Results for 'Transport', 'Spindle', 'Nucleus'

Table 10 shows the precision, recall and F-measure $(\alpha = 1)$ achieved for the BLOSUM and Naive Bayes metrics on the 'transport', 'spindle', and 'nucleus' datasets. These measures show how effectively each metric and threshold partition the papers in each dataset into the classes of FPs and FNs/TPs. As shown, for both the BLOSUM and Naive Bayes metrics we achieve a precision and recall of $\sim$90% in separating FPs from FNs/TPs. The partitionings of these three datasets are consistent with the 'development' dataset, pointing to the significance of the results.

In order to separate FP from FN/TP papers in the 'transport' and 'spindle' datasets, we used the same *threshold* values that gave us the best results for 'Development'; for the BLOSUM metric $threshold = 0.0$, while for the Naive Bayes metric $threshold = 0.00005$.

The BLOSUM threshold has an intuitive explanation, since greater than 0 means that an paper is more likely to be a TP/FN and less than 0 means that an paper is unlikely to be a TP/FN. The Naive Bayes threshold has to be greater than but near 0, because many annotations co-occur with 'development' and the probabilites are low. Our experiments showed that these thresholds are not sensitive to slight changes.

For the 'nucleus' dataset the co-occurrences had a different distribution, and different *threshold* values produced the best results; for the BLOSUM metric $threshold = 0.2$, while for the Naive Bayes metric $threshold = 0.0003$.

Clustering had a positive effect with disambiguating the 'development' dataset. However, clustering did not improve disambiguation of the 'transport', 'spindle', 'nucleus' datasets. The reason is that the papers with non-'development' meanings can be more clearly separated into subgroups, e.g., economic or software development. Such subgroups can be retrieved by clustering based on annotations, which helps with the disambiguation task. On the other hand, 'nucleus', 'transport', and 'spindle' have diverse meanings that could not be retrieved so easily by clustering based on annotations, thus in these cases clustering did not improve the result. Thus, clustering only helped with the disambiguation task when papers could be effectively clustered into meaningful FP subgroups on the basis of their annotations.

Table 10 Results for the BLOSUM and Naive Bayes metrics on the 'transport', 'spindle', and 'nucleus' test datasets

	BLOSUM			*Naive Bayes*		
Dataset	*Precision*	*Recall*	*F-meas.*	*Precision*	*Recall*	*F-meas.*
Transport	0.9	0.9	0.9	0.93	0.92	0.93
Spindle	0.91	0.9	0.9	0.91	0.9	0.9
Nucleus	0.81	0.8	0.81	0.89	0.89	0.89

6 Conclusion and future work

We have proposed and evaluated an approach for improving the quality of automatically annotated papers. This approach is based on co-occurrence graphs, which in our case we built on the basis of GOA. Our results are inline with other previous work, the precision/recall of which was discussed in the Introduction and Related Work sections. We initially focused on the 'development' annotation. However, our experiments showed that the method proposed in this paper is applicable to diverse annotations, such as 'nucleus' or 'transport' or 'spindle', that are often FPs or FNs.

One problem with this approach is that GOA is sparsely annotated because of the difficulty and effort required for manual annotations. For example, in GOA only 243 papers have a 'development' annotation. This may raise questions as to the statistical significance of mapping an automatically annotated GoPubMed paper onto the GOA co-occurrence graph. Even though the GOA corpus that we used for the co-occurrence graph only has 243 papers annotated with 'development', our results are still shown to be meaningful. As the GOA corpus increases, the statistical significance of the relationships in the co-occurrence graph will become stronger. Then, future experimental results will be even more meaningful for predicting FPs and FNs.

One direction worth pursuing as future work is to extend our method so that it incorporates relationships of terms in the GO hierarchy. Specifically, we would like to incorporate parent/child relationships between GO terms for predicting FP and FN annotations. For example, if an paper is annotated with a term that is a child or descendant of 'development' then the paper is likely to be a TP or FN with respect to 'development'.

Acknowledgements

The authors thank the reviewers for their valuable comments. The authors gratefully acknowledge the financial assistance of the Natural Sciences and Engineering Research Council (NSERC), the Ontario Graduate Scholarship Program (OGS), and the EU projects REWERSE and Sealife.

References

Andreopoulos, B., Aijun, A., Tzerpos, V. and Xiaogang, W. (2006) 'Clustering large software systems at multiple layers', *Information and Software Technology*, Vol. 49, No. 3, March, pp.244–254.

Andreopoulos, B., An, A., Wang, X., Faloutsos, M. and Schroeder, M. (2007a) 'Clustering by common friends finds locally significant proteins mediating modules', *Bioinformatics*, Oxford University Press, Vol. 23, No. 9, pp.1124–1131.

Andreopoulos, B., An, A. and Wang, X. (2007b) 'Hierarchical density-based clustering of categorical data and a simplification', *PAKDD*, pp.11–22.

Bodenreider, O. (2004) 'The Unified Medical Language System (UMLS): integrating biomedical terminology', *Nucleic Acids Research*, Vol. 32, pp.D267–D270.

Camon, E.B., Barrell, D.G., Dimmer, E.C., Lee, V., Magrane, M., Maslen, J., Binns, D. and Apweiler, R. (2005) 'An evaluation of GO annotation retrieval for BioCreAtIvE and GOA', *BMC Bioinformatics*, Vol. 6, Suppl. 1, p.S17.

Camon, E.B., Magrane, M., Barrell, D.G., Lee, V., Dimmer, E.C., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. (2004) 'The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology', *Nucleic Acids Research*, Vol. 32, No. 1, pp.262–266.

Doms, A. and Schroeder, M. (2005) 'GoPubMed: exploring PubMed with the GeneOntology', *Nucleic Acid Research*, Vol. 33, Web Server Issue, pp.W783–W786.

Dorow, B. and Widdows, D. (2003) 'Discovering corpus-specific word senses', *EACL'03: Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*, pp.79–82, Morristown, NJ, USA, ISBN 1-111-56789-0.

Edmonds, P. and Agirre, E. (2006) *Word Sense Disambiguation: Algorithms and Applications*, Springer-Verlag, July, ISBN-10:1402048084.

Ehrler, F., Geissbühler, A., Jimeno, A. and Ruch, P. (2005) 'Data-poor categorization and passage retrieval for gene ontology annotation in swiss-prot', *BMC Bioinformatics*, Vol. 6, No. 1. p.S23.

Fellbaum, C. (1998) *WordNet an Electronic Lexical Database*. MIT Press, USA.

Gaudan, S., Kirsch, H. and Rebholz-Schuhmann, D. (2005a) 'Resolving abbreviations to their senses in medline', *Bioinformatics*, Vol. 21, No. 18, pp.3658–3664, URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/18/3658

Gaudan, S., Kirsch, H. and Rebholz-Schuhmann, D. (2005b) 'Resolving abbreviations to their senses in medline', *Bioinformatics*, Vol. 21, No. 18, pp.3658–3664.

GeneOntologyConsortium (2004) 'The Gene Ontology (GO) database and informatics resource', *Nucleic Acids Research*, Vol. 1, No. 32, pp.258–261.

Ginter, F., Boberg, J., J´arvinen, J. and Salakoski, T. (2004) 'New techniques for disambiguation in natural language and their application to biological text', *J. Mach. Learn. Res.*, Vol. 5, pp.605–621, ISSN 1533–7928.

Hatzivassiloglou, V., Duboue, P.A. and Rzhetsky, A. (2001) 'Disambiguating proteins, genes, and rna in text: a machine learning approach', *Bioinformatics*, Vol. 17, Suppl. 1, pp.S97–106, URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/17/suppl_1/S97

Henikoff, S. and Henikoff, J.G. (1992) 'Amino acid substitution matrices from protein blocks', *PNAS*, Vol. 89, No. 22, pp.10915–10919.

Hirschman, L., Yeh, A., Blaschke, C. and Valencia, A. (2005) 'Overview of biocreative: critical assessment *f* information extraction for biology', *BMC Bioinformatics*, Vol. 6, No. 1, p.S1.

Humphrey, S.M., Rogers, W.J. Kilicoglu, H. Demner-Fushman, D. and Rindflesch, T.C. (2006) 'Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: preliminary experiment', *Journal of the American Society for Information Science and Technology*, Vol. 57, No. 1, pp.96–113, URL http://dx.doi.org/10.1002/asi.20257

Ide, N. and Véronis, J. (1998) 'Introduction to the special issue on word sense disambiguation: the state of the art', *Comput. Linguist.*, Vol. 24, No. 1, pp.2–40, ISSN 0891-2017.

Jensen, L.J., Saric, J. and Bork, P. (2006) 'Literature mining for the biologist: from information retrieval to biological discovery', *Nature Reviews Genetics*, Vol. 7, pp.119–127.

Liu, H., Johnson, S.B. and Friedman, C. (2002) 'Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the umls', *Journal of the American Medical Informatics Association*, Vol. 9, No. 6, pp.621–636.

Liu, H., Teller, V. and Friedman, C. (2004) 'A multi-aspect comparison study of supervised word sense disambiguation', *J. Am. Med. Inform. Assoc.*, Vol. 11, No. 4, pp.320–331, URL http://www.jamia.org/cgi/content/abstract/11/4/320

Mihalcea, R. (2004) 'Co-training and self-training for word sense disambiguation', *Proceedings of CoNLL-2004*, Boston, MA, USA, pp.33–40.

Navigli, R. and Velardi, P. (2005) 'Structural semantic interconnections: a knowledge-based approach to word sense disambiguation', *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 27, No. 7, pp.1075–1086.

Navigli, R., Velardi, P. and Gangemi, A. (2003) 'Ontology learning and its application to automated terminology translation', *IEEE Intelligent Systems*, Vol. 18, No. 1, January, pp.22–31.

Pahikkala, T., Ginter, F., Boberg, J., Järvnen, J. and Salakoski, T. (2005) 'Contextual weighting for support vector machines in literature mining: an application to gene versus protein name disambiguation', *BMC Bioinformatics*, 6:157doi:10.1186/1471-2105-6-157.

Pedersen, T. and Bruce, R. (1997) 'Distinguishing word senses in untagged text', *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Providence, RI August, pp.197–207, URL citeseer.ist.psu.edu/article/pedersen97distinguishing.html

Pedersen, T. and Bruce, R. (1998) 'Knowledge lean word sense disambiguation', *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, Madison, WI, July, pp.800–805, URL citeseer.ist.psu.edu/article/pedersen98knowledge.html

Purandare, A. and Pedersen, T. (2004) 'Word sense discrimination by clustering contexts in vector and similarity spaces', *Proceedings of CoNLL-2004*, Boston, MA, USA, pp.41–48.

Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M. and Stoehr, P. (2007) 'EBIMed: Text crunching to gather facts for proteins from medline', *Bioinformatics*, Vol. 23, No. 2, pp.e237–e244.

Schijvenaars, B.J.A., Mons, B., Weeber, M., Schuemie, M.J., van Mulligen, E.M., Wain, H.M. and Kors, J.A. (2005) 'Thesaurus-based disambiguation of gene symbols', *BMC Bioinformatics*, 6:149doi:10.1186/1471-2105-6-149.

Schuemie, M.J., Kors, J.A. and Mons, B. (2005) 'Word sense disambiguation in the biomedical domain: an overview', *J. Comput. Biol.*, Vol. 12, No. 5, June, pp.554–565, doi:10.1089/cmb.2005.12.554. URL http://dx.doi.org/10.1089/cmb.2005.12.554

Schütze, H. (1998) 'Automatic word sense discrimination', *Comput. Linguist.*, Vol. 24, No. 1, pp.97–123, ISSN 0891-2017.

Schütze, H. and Pedersen, J. (1995) *Information Retrieval Based on Word Senses*, URL citeseer.ist.psu.edu/schutze95information.html

Stevenson, M. and Wilks, Y. (2001) 'The interaction of knowledge sources in word sense disambiguation', *Comput. Linguist.*, Vol. 27, No. 3, pp.321–349, ISSN 0891-2017, doi:http://dx.doi.org/10.1162/089120101317066104.

Wren, J.D., Chang, J.T., Pustejovsky, J., Adar, E., Garner, H.R. and Altman, R.B. (2005) 'Biomedical term mapping databases', *Nucleic Acids Research*, Vol. 33, pp.D289–293.

Xu, H., Markatou, M., Dimova, R., Liu, H. and Friedman, C. (2006a) 'Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues', *BMC Bioinformatics*, Vol. 7, No. 1, p.334.

Xu, H., Markatou, M., Dimova, R., Liu, H. and Friedman, C. (2006b) 'Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues', *BMC Bioinformatics*, Vol. 7, No. 1, p.334, URL http://www.biomedcentral.com/1471-2105/7/334

Yarowsky, D. (1995) 'Unsupervised word sense disambiguation rivaling supervised methods', *Meeting of the Association for Computational Linguistics*, pp.189–196, URL citeseer.ist.psu.edu/yarowsky95unsupervised.html