Available online at www.sciencedirect.com

Computational Biology and Chemistry 28 (2004) 387–391

www.elsevier.com/locate/compbiolchem

Brief communication

Retrieving definitional content for ontology development

L. Smith*, W.J. Wilbur

National Center for Biotechnology Information, NIH, NLM, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received 20 September 2004; received in revised form 22 September 2004; accepted 22 September 2004

Abstract

Ontology construction requires an understanding of the meaning and usage of its encoded concepts. While definitions found in dictionaries or glossaries may be adequate for many concepts, the actual usage in expert writing could be a better source of information for many others. The goal of this paper is to describe an automated procedure for finding definitional content in expert writing. The approach uses machine learning on phrasal features to learn when sentences in a book contain definitional content, as determined by their similarity to glossary definitions provided in the same book. The end result is not a concise definition of a given concept, but for each sentence, a predicted probability that it contains information relevant to a definition. The approach is evaluated automatically for terms with explicit definitions, and manually for terms with no available definition.

Keywords: Ontology development; Definitional content; Expert writing

1. Introduction

An ontology can be defined as "a specification of a vocabulary for a shared domain of discourse—definitions of classes, relations, functions, and other objects" (Gruber, 1993). Ontology development crosses the bridge from semantic human knowledge to some form of formal specification. An ontology developer must therefore understand the meanings and nuances of the specified terms. While many terms are "stock", and adequate definitions can be found in various dictionaries and glossaries, a developer may need to consult domain experts for actual usages that depart from dictionary definitions, or contain terms that are new, not found in dictionaries, or whose definitions are evolving. One source of domain expertise consists of electronic collections of academic and scientific writing. With the proper tools, an ontology developer could access this expertise, without being required to read and understand all of it. The goal of this paper is to demonstrate an approach to searching a database of textbooks on molecular biology and medicine which is able to find passages containing information relevant to the definition and usage of a given term.

Other approaches have been described for finding definitions that might be especially useful for ontology development. The DefScriber system (Blair-Goldensohn et al., 2003) achieves high precision in answering definitional questions "what is *X*?" by finding genus–species statements, like "*X* is a type of *Y* with properties *Z*". Their system used machine learning to first identify nonspecific definitional sentences using word frequency, punctuation, and bag-of-words features. Then, manually constructed, high precision, parse patterns were applied to match specific genus–species type sentences. In DEFINDER (Klavans and Muresan, 2000), a set of manually written rules or patterns are used to find expressions that are commonly used by authors to give definitions. Since an author's intent to provide a definition is usually not ambiguous, this approach, like DefScriber, has high precision.

For the TREC 2003 competition in question answering, Hildebrandt et al. (2004) report on a system they developed to answer definitional questions from the AQUAINT corpus. They first developed by hand 11 surface patterns likely to be found in definitional material. These were used to precompile definitional "nuggets" (fragments of text containing definition-like material) in the AQUAINT database. Given a

[*] Corresponding author. Tel.: +1 301 4358881; fax: +1 301 4802290.
E-mail addresses: lsmith@ncbi.nlm.nih.gov (L. Smith), wilbur@ncbi.nlm.nih.gov (W.J. Wilbur).

1476-9271/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2004.09.007

term for which a definition was sought they looked for precompiled nuggets as answers. They also consulted the Merriam Webster online dictionary and if the term was found they then took its definition and rated material from the AQUAINT database for similarity to the dictionary definition. These approaches gave the highest precision results when successful. If they failed, then straight information retrieval in the AQUAINT database based on the query was a last resort. Questions related to identifying famous persons seem to have played a large role in this work and some of the patterns were influenced by this emphasis.

One drawback with the approaches of DefScriber and DEFINDER is that they cannot find the definition of a term if no author has given it an explicit definition. Our approach attempts to remedy this in two ways. Instead of focusing on “definitional sentences” written intentionally to give a definition, our approach is to focus on “definitional content”, that is, sentences that contain some of the content of a definition, without necessarily being written with the intent of giving a definition. We also use patterns of language to detect definitional content, but use machine learning to discover these patterns and their relative importance. A total of 3000 phrases were explored and 847 of them were found to have a statistically significant association with definitional content ($p < 0.01$). The work of Hildebrandt et al. (2004) is related to our work in focusing on definitional content rather than formal definitions, but differs in using hand coded rules which seem to be aimed at information found in the news media. The remainder of this paper describes our approach and the results that were obtained.

2. Instances and features of the corpus

In this section, the preparation of the text corpus used in training and evaluation is described. The parameters of machine learning are delineated, including the instances, the determination of positive instances, and the features used with naive Bayes classification (Langley, 1996).

2.1. Instances

The instances of the corpus consist of all sentences from a collection of textbooks. The NCBI provides public search access (NCBI, 2004) to several textbooks in molecular biology and medicine. Many of these textbooks have extensive glossaries, set apart in their electronic versions. Eight books were selected to form a combined corpus for this research (Brown, 2002; Cooper, 2000; Griffiths et al., 1999; Janeway et al., 2001; Lodish et al., 1999; Sachs and Brenner, 2003; Strachan and Read, 1999; Varki et al., 1999). The electronic versions contain markup that delineate the chapters, sections, and paragraphs, as well as the glossaries. Each glossary entry consists of a head term and a definition. To prepare the corpus, the text was processed by the MedPost part of speech tagger (Smith et al., 2004) which segmented paragraphs into sentences, and sentences into tokens. Each sentence was assigned an identifier that coded the book, chapter number, paragraph number within the chapter, and sentence number within the paragraph. Each sentence, as a sequence of tokens (its parts-of-speech were not used) constitutes an instance in the corpus. There were 65 664 total instances from all books.

2.2. Positive instances

A positive instance in the corpus is intended to be a sentence (instance) containing maximal definitional content relevant to a given term. There is no automated procedure that can perfectly measure the amount of definitional content in a sentence relevant to a term; the best available solution is to use human judgment. But this would require an expert to judge and rank every sentence in a book as to its definitional content for every term, and this is clearly infeasible. We therefore implemented an approximate automated procedure for selecting positive instances, described in this section, and compared it with human judgments for a sample of terms.

For each glossary head term (the phrase to be defined), at most one sentence (instance) was selected from the corresponding book and designated as a positive instance for that term in that book. To do this, the glossaries (head terms and definitions separately) were first tokenized by MedPost. A “stemmed” substring search for each glossary head term was then performed over the sentences of the corresponding book, where all words are first stemmed using the Porter stemmer (Porter, 1980). The matching sentences for a glossary term were then compared with the text of the corresponding glossary definition to obtain the similarity measure described below. The sentence with the largest similarity measure was taken as a positive instance for the corresponding term. There were 3200 glossary entries whose head terms occurred at least once in the text of the same book, which gave rise to an equal number of positive instances.

The similarity measure is intended to measure the amount of conceptual overlap between a sentence and a definition. Two similarity measures were considered, both based on the words in common between the sentence and definition (words were compared after stemming). The *word count similarity measure* (WCSM) counted the number of matching words, and the *inverse frequency similarity measure* (IFSM) summed the inverse frequency weights for each matching word. The inverse frequency for a word is defined analogous to inverse document frequency (Salton, 1998) as $\log(G/m)$, where G is the number of entries in the glossary and m is the number of times that the stemmed word occurs in the glossary. The IFSM assigns low weight to frequently occurring words and these tend to be non-content bearing words.

To evaluate whether the resulting sentences had definitional content, 10 glossary terms were randomly selected from each book for a total of 80 evaluations. The definitions were then compared with the selected sentence, and the concept overlap was graded subjectively using a scale of 1–5. In this scale, a value of 1 was used to indicate that most of

Table 1
Some glossary terms with their definitions, WCSM and IFSM sentences, and grade

Term	Grade	Definition/WCSM/IFSM
Antiport		The transport of two molecules in opposite directions across a membrane
WCSM	1	Active transport can also take place by antiport, in which two molecules are transported in opposite directions (Fig. 12.33)
IFSM	1	Same
Haploid		A nucleus that has a single copy of each chromosome
WCSM	3	Meiosis occurs only in reproductive cells, and results in a diploid cell giving rise to four haploid gametes, each of which can subsequently fuse with a gamete of the opposite sex during sexual reproduction
IFSM	3	The fact that meiosis results in four haploid cells whereas mitosis gives rise to two diploid cells is easy to explain: meiosis involves two nuclear divisions, one after the other, whereas mitosis is just a single nuclear division
Extracellular matrix		A complex array of secreted molecules including glycoproteins, proteoglycans, and/or polysaccharides and structural proteins. In plants, the extracellular matrix is also referred to as the cell wall
WCSM	5	This protein is found on Schwann cell membranes and links another membrane protein, beta-dystroglycan, to laminin in the extracellular matrix
IFSM	3	A discussion of plant glycobiology must start with a description of the structure and function of the cell wall or extracellular matrix

The grade is subjectively assigned 1 through 5 with 1 having all essential content and 5 having no content.

the definition is implied by the sentence and a 5 to indicate a complete absence of implication (additional information in the sentence was not considered negative). The two methods resulted in the same sentence selection in 60 of the 80 terms. In the 20 terms that differed, 7 were judged to have the same amount of content, in 9 of the glossary terms the IFSM sentence was judged to have more content, and in 4 glossary terms the WCSM was judged to have more content. The WCSM resulted in 5 of the sampled terms with no definitional content and the IFSM resulted in 4. Examples of the sampled terms and their evaluation are shown in Table 1. This evaluation was not exhaustive, but it does suggest that the IFSM may perform better than simple word counting. The IFSM also has the advantage of generating few ties in the selection process, whereas word counting frequently results in ties. For these reasons, the IFSM method was used to select definitional content.

2.3. Features

Each instance in the corpus was associated with a number of binary features, derived from the instance, that is, the words of the sentence. A preliminary study showed that features based on common phrases involving frequently occurring words (less frequent words replaced by underscore tokens) would perform better than single words, stemmed words, or parts of speech. The phrases were selected and applied in two ways, which we call *labeled* and *unlabeled.* To begin, the 1000 words that occurred most often in the positive sentences were retained and remaining words were replaced with an underscore (consecutive replaced words were replaced with a single underscore). In the unlabeled method, all of the subphrases from these sentences were tallied and the 3000 that occurred most often were retained as features. In the labeled method, each sentence was first searched for its corresponding glossary head term after stemming. The matching tokens were then replaced with a single token, “NPT”. Afterwards, the common words were retained and uncommon words reduced to underscore; and again, the 3000 most frequent subphrases were retained as features.

To determine the features that apply to a given sentence, it is processed the same as described above and each feature phrase that occurs in the resulting sentence (from the corresponding labeled or unlabeled set) is taken to be a feature. To determine the features that apply to a sentence using the labeled method, a term of interest must be specified in advance. If the stemmed phrase occurs in the sentence, the matching tokens are replaced with NPT before determining which feature phrases are contained in the sentence.

3. Results

The two methods of selecting features, labeled and unlabeled phrases, were compared by training on a subset of the corpus and testing on the complement and by using a form of cross-validation. For each glossary term which occurred in at least 10 sentences (there were 1233 of these), both methods were trained on all sentences that excluded the stemmed glossary term. The trained feature weights were then used to rank the held-out sentences, which contained at least one positive instance, but could contain more than one if a term appeared in the glossary of more than one book. The average relative rank (ARR) of the positive sentences was recorded for each term. If there are n sentences in the held-out set, and a positive sentence occurs with rank r, then r/n is the relative rank of that sentence. The mean of the ARR over all 1233 terms was 0.2846 for labeled features and 0.3193

Table 2
Average rank of sentences judged to have definitional content in the top 10 sentences ranked using naive Bayes classification with labeled and unlabeled phrases as features

Term	Labeled ARR	Unlabeled ARR
Gpi	0.49	0.2
Iddm	0.25	0.18
Mag	0.18	0.34
Muscle	0.26	0.27
Myelin	0.34	0.38
Parasegment	0.24	0.28
Polymerase	0.29	0.29
Ret	0.48	0.46
Splice site	0.17	0.35
Tryptophan	0.16	0.15
	0.286	0.290

for unlabeled features. The labeled features performed better 651 times, the unlabeled features 379 times and the two feature sets performed the same 203 times. This difference is highly statistically significant in favor of the labeled features, as determined using the sign test (Larson, 1982) assuming an equal probability null hypothesis.

The two methods were also compared manually. For each of 10 selected terms, a selection of sentences were presented to a knowledgeable reader who identified those sentences containing relevant definitional content for the corresponding term. The terms were selected using the results of the textbook project (NCBI, 2004) based on a statistical test for a term to be associated with a concept discussed specifically in sections in the book (as opposed to occurring randomly throughout the book). The highest scoring terms that did not appear in any of the glossaries were selected. The sentences that were selected for each term consisted in the top 10 sentences ranked by the labeled method and the top 10 ranked by the unlabeled method, and then combined and placed in a random order. The average rank of the marked sentences for the two methods were compared, with results summarized in Table 2. This comparison did not show a statistically significant advantage for the labeled method, but the tendency was in that direction.

The effectiveness of the ranking algorithm for the labeled method was evaluated manually. The goal of this evaluation is to determine if sentences with high rank were more likely to contain definitional sentences than those with low rank. An additional 15 terms were chosen as in the previous paragraph, and the top and bottom 10 sentences (as ranked by the labeled method) were placed in random order and presented to a knowledgeable reader who again identified those sentences containing relevant definitional content for the corresponding term. The number of identified sentences in each group is shown in Table 3. The difference between the two groups (top and bottom) were statistically significant compared to random sentence ranking. The top 10 ranked sentences contained more definitional content than the bottom 10 in 14 terms out of 15, and the opposite in one case. Furthermore, the top 10 sentences from the 15 terms contained twice the number with definitional content than the bottom 10 (82 versus 41).

Table 3
Fifteen random terms not found in any glossary, and the number of sentences judged to have definitional content in the top 10 and bottom 10 as ranked by naive Bayes classification with labeled phrases as features

Term	Top 10	Bottom 10	No. of sentences
Bicoid	7	6	38
Cancer cells	3	4	135
Cyclin b	5	2	73
Deltag	4	3	49
Diabetes	3	1	57
Glycogenin	6	1	17
Mendel	6	1	135
Mitochondrial DNA	3	2	47
Mother cell	5	0	21
Profilin	8	3	27
Psii	6	3	40
Srp	8	3	28
Sxl	8	6	43
Twins	3	0	72
Vkappa	7	6	20
	82	41	

4. Discussion

The use of machine learning to compute weights for phrases generalizes the approach of manually enumerating linguistic patterns of expression associated with making definitions. For the labeled method, we investigated 3000 phrases which were the most frequent phrases in the positive sentences (sentences in the text most similar to the glossary definition). For the unlabeled method, we also investigated 3000 phrases which were the most frequent phrases in the positive sentences. The two methods yielded somewhat different sets of phrases because 184 of the phrases in the labeled method included the term label (NPT). In order to examine the significance of the 3000 phrases in each case, we formed a 2×2 contingency table based on positive or negative sentences versus presence or absence of the phrase and a χ^2-test was applied. There were 847 labeled phrases and 1147 unlabeled phrases found to be statistically significant ($p < 0.01$) in correlating with definitional content.

Of the 3000 phrases used as features in the labeled case, only 539 had weakly negative Bayesian weights. The most negative weight phrase was the single word *studies* and the most positive weight phrase was *is the NPT _ which*. The phrases with largest positive weights are shown in Table 4, together with some representative sentences containing them, and the corresponding term. Note how most of these involve the word “called” which is a preferred word for indicating definition. The other phrases are less predictable, but still understandable, illustrating the advantage of machine learning for this task. Also note how all of these phrases explicitly refer to NPT, and are therefore related to the immediate context of the term.

Table 4
Some positive weight phrases from labeled features, with examples sentences quoted from the textbooks used in this study (see Section 2.1) – all sentences had an overall positive Bayes score for the corresponding term, which appears in bold

Phrase	Sentence
is the NPT _ which	The most common measure of variation around the center is the **variance**, which is defined as the average squared deviation of the observations from the mean, or . . . *formula*. . .
called an NPT _	In general, external and internal surfaces of tissues and organs are covered by a layer of epithelial cells called an **epithelium** (Fig. 6-4)
_ called the NPT _	This protein, called the **mannan-bindinglectin** (MBL), is a collectin, like C1q
is the NPT _	A typical *E. coli* example is the **lactoseoperon**, the first operon to be discovered (Jacob and Monod, 1961), which contains three genes involved in conversion of the disaccharide sugar lactose into its monosaccharide units—glucose and galactose (Fig. 2.20A)
are called NPT	Most protein kinases phosphorylate either serine and threonine or tyrosine residues: these enzymes are called **protein-serine**/**threoninekinases** or protein-tyrosine kinases, respectively

These sentences all had positive Bayes score for the corresponding term.

The unlabeled features also consist of mostly positive weights, 447 of 3000 had negative weight with the most negatively weighted phrase being the single word *studies* as in the labeled case. The unlabeled phrases with largest positive weight were *process called _*, *a process called*, *process called*, *often called*, *called an*, *called _ or _*, *called _ or*, *_ also called*, *is defined*, and *a form of*. These also are clearly understandable as being correlated to definitions, and also show the tendency to use the word *called* when making a definition. The main difference between the labeled and unlabeled phrases is that all of the highest weight phrases in the labeled case involve the label NPT itself, whereas the unlabeled phrases are nonspecific. In other words, the unlabeled features are able to find and rank definitional content, but without term specificity.

The lack of specificity of the unlabeled method will lead it to rank sentences with definitional content, regardless of the term to which the content relates. Despite the nonspecific nature of the unlabeled phrases, there are still circumstances where the unlabeled method could be better able to rank definitional content for a term than the labeled method. For example, where a term of interest is referenced anaphorically or elliptically, definitional content may be found at a distance from the explicit reference.

5. Conclusions

We have given a method which employs textbooks with glossaries as a resource from which information in the form of indicator phrases may be extracted and used to identify sentences likely to have definitional content. We believe such methods provide a useful approach to investigate the meaning of terms in biology. At its current level, the results are useful to a human operator seeking to construct definitions. This is an important step in ontology construction and may be useful in a less formal setting where a user simply wants to gain a better understanding of terminology.

In future work, we hope to apply the same method to identify passages longer than sentences as useful reference points. We also believe that anaphor resolution methods can be usefully combined with this approach to improve results.

References

Blair-Goldensohn, S., McKeown, K.R., Schlaikjer, A.H., 2003. A hybrid approach for QA track definitional questions. In: The 12th Text Retrieval Conference, TREC 2003, pp. 185–192.

Brown, T.A., 2002. Genomes, second ed. BIOS Scientific Publishers Ltd., Oxford, UK.

Cooper, G.M., 2000. The Cell—A Molecular Approach, second ed. Sinauer Associates Inc., Sunderland, MA.

Griffiths, A.J.F., et al., 1999. Introduction to Genetic Analysis. Freeman, New York.

Gruber, T., 1993. A translation approach to portable ontology specifications. Knowledge Acquisition 5, 199–220.

Hildebrandt, W., Katz, B., Lin, J., 2004. Answering definition questions with multiple knowledge sources. In: Proceedings of HLT-NAACL 2004, pp. 49–56.

Janeway, C.A., et al., 2001. Immunobiology, fifth ed. Garland Publishing, New York and London.

Klavans, J.L., Muresan, S., 2000. DEFINDER: rule-based methods for the extraction of medical terminology and their associated definitions from on-line text. In: Proceedings of AMIA Symposium, p. 1096.

Langley, P., 1996. Elements of Machine Learning. Morgan Kaufmann, San Francisco.

Larson, H.J., 1982. Introduction to Probability Theory and Statistical Inference. Wiley, New York.

Lodish, H., et al., 1999. Molecular Cell Biology, fourth ed. Freeman, New York.

National Center for Biotechnology Information, 2004. http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=books. Also, for a discussion of the statistical weighting of terms, see a discussion at http://ncbi.nlm.nih.gov/entrez/query/Books.live/Help/back.html.

Porter, M.F., 1980. An algorithm for suffix stripping. Program 14 (3), 130–137.

Sachs, R., Brenner, D., 2003. Chromosome Aberrations Produced by Ionizing Radiation: Quantitative Studies. NCBI, National Library of Medicine, Bethesda, MD.

Salton, G., 1998. Automatic Text Processing. Addison-Wesley, Reading, MA.

Smith, L., Rindflesch, T., Wilbur, W.J., 2004. MedPost: a part of speech tagger for biomedical text. Bioinformatics 20 (14), 2320–2321.

Strachan, T., Read, A.P., 1999. Human Molecular Genetics, second ed. BIOS Scientific Publishers Ltd., Oxford, UK.

Varki, A., et al., 1999. Essentials of Glycobiology, first ed. Cold Spring Harbor Laboratory Press, Plainview, NY.