Automatically Augmenting Terminological Lexicons from Untagged Text

George Demetriou and Robert Gaizauskas

Department of Computer Science
University of Sheffield
211 Portobello Street, Sheffield S1 4DP, United Kingdom
{G.Demetriou, R.Gaizauskas}@dcs.shef.ac.uk

Abstract

Lexical resources play a crucial role in language technology but lexical acquisition can often be a time-consuming, laborious and costly exercise. In this paper, we describe a method for the automatic acquisition of technical terminology from domain restricted texts without the need for sophisticated natural language processing tools, such as taggers or parsers, or text corpora annotated with labelled cases. The method is based on the idea of using prior or seed knowledge in order to discover co-occurrence patterns for the terms in the texts. A bootstrapping algorithm has been developed that identifies patterns and new terms in an iterative manner. Experiments with scientific journal abstracts in the biology domain indicate an accuracy rate for the extracted terms ranging from 58% to 71%. The new terms have been found useful for improving the coverage of a system used for terminology identification tasks in the biology domain.

1. Introduction

The recognition and classification of names and technical terminology in machine readable texts is important for language engineering applications such as Text Classification (TC), Information Retrieval (IR), Information Extraction (IE) and Machine Translation (MT). While approaches for the automatic extraction of names from running text can vary from rule-based (Gaizauskas et al., 1995; Justeson and Katz, 1995) to statistical ones (Bikel et al., 1997; Renals et al., 1999), in many of the IR and IE systems evaluated in the TREC (Harman, 1998), MUC (DARPA, 1998) and DARPA (DARPA, 1999) conferences, the use of specialised lexicons or gazetteers is an essential part of the name identification process. Such lexicons or lists of names are used as repositories of the entities of the domain, such as the names of persons, organisations and locations. Looking up a name in the lexicon is not computationally expensive and is often used as a first step towards the full recognition and classification of such named entities in the text.

Typically, lexicons of named entities are either handcrafted or acquired automatically from annotated corpora when sufficient text quantities have been labelled by domain experts. In either case, the acquisition of lexical knowledge represents a time-consuming and laborious process and is a hindrance to efforts of adapting existing Natural Language Processing (NLP) systems to new domains. Automating the acquisition of names from untagged texts could therefore be of great help to system developers.

There has been recently an increased interest in techniques of bootstrapping for automatic term acquisition from untagged text in relation to Named Entity (NE) recognition tasks (Collins and Singer, 1999; Cucerzan and Yarowsky, 1999; Jones et al., 1999). A bootstrapping approach to term acquisition is based on the distributional hypothesis that entities of the same semantic class usually occur in similar contextual environments. In the management succession domain (DARPA, 1995), for example, the names of persons and locations frequently occur in language patterns like `<X> succeeded <Y>` or `<X> resigned from <Z>` where `<X>` and `<Y>` represent the names of persons and `<Z>` the name of an organisation. When there is prior information about the names of interest, we can use this information as seed knowledge in order to extract co-occurrence patterns (such as `resigned from`). These patterns are then used to identify new names which are used as new seeds for extracting new patterns and so on. Using a bootstrapping method for term identification is an attractive option because there is no need to annotate the texts with labels of name classes and, usually, examples of terms that could be used as seeds can be found easily.

The approach by Jones et al. (1999) uses a bootstrapping technique for learning the names of locations in WWW pages. They first initialise a learner module with a few seed words and then run AutoSlog (an extraction system that uses heuristics in the form of domain-independent linguistic rules) to generate extraction patterns and acquire the names of locations from the unlabelled data iteratively. They report 76% accuracy for a dictionary of 250 extracted location names.

Cucerzan and Yarowsky (1999) use an Expectation-Maximisation (Dempster et al., 1977) bootstrap procedure to identify the names of places and people from untagged texts in several languages. Their algorithm learns the left and right contexts that are indicative of the semantic classes of the seed words. It then re-estimates the probabilities of contextual and morphological clues for each class and the decision to classify a name is taken by combining the different sources of evidence. The precision of the extracted names ranges from 60% to 84% depending on the language.

Collins and Singer (1999) classify the names of person, organisations and locations using an unsupervised version of the decision list method for word sense disambiguation originally proposed by (Yarowsky, 1995). A syntactic parser is used to extract the contextual patterns and the learning algorithm extracts capitalisation and contextual rules for a semantic class. These rules are used iteratively to annotate the training set with labels indicating the different semantic classes. They report results ranging from 76%

to 91% depending on the algorithm configuration and the evaluation metric used.

In this paper we describe a generic method for the automatic acquisition of scientific terminology from domain specific untagged texts. This work is related to research for the Protein Active Site Template Acquisition (PASTA) project [1] that aims at extracting protein structure information from online journal articles and abstracts. The NE identification component of the PASTA system makes extensive use of lexicons of biological information such as protein names, species names, etc. But when new protein names are introduced in the literature, the lexicons must be updated, since they can only provide information about proteins reported at the time the lexicons were compiled.

A second problem is that of spelling differences between the name of a protein in the lexicon and the name of the same protein in the texts. This may be due to variations in expression (this is especially the case of multi-word protein names), abbreviations or inconsistencies in spelling. For example, the protein entry `PI-specific phospholipase C isozyme D1` in the lexicon may be found as `phosphoinositide-specific phospholipase C-delta 1` or `phospholipase C-delta(1)`. This problem may have an impact on systems that use string pattern matching strategies for identifying the names in the texts and, although there are ways of dealing with this problem (see section 5 where the NE component of the PASTA system is discussed), it is always desirable to have entries in the lexicon that are accurate representations of the named entities that occur in the texts.

Adopting a bootstrapping approach for extracting novel terms from untagged texts is based on the idea, that when some lists of terms of the domain are available, however incomplete they may be, they can be used as prior knowledge to extract new terms.

During a first pass of the input texts, the algorithm locates the instances of the seed terms. The context around each term is explored and statistics about the co-occurrence of contextual patterns in the form of n-grams (word sequences) and the term class are computed. The patterns are scored to verify which ones satisfy certain statistical significance requirements i.e. occur significantly more in the context of a term class as compared to their frequencies of occurrence in the corpus. The patterns retained after evaluation are fed back into the system to identify new terms which are used to extract new patterns and so on. This is an iterative process which terminates when no new terms or new contextual patterns are found.

The approach is similar in principle to the bootstrapping technique by Jones et al. (1999) although the two algorithms differ in some important details such as the generation of contextual patterns, the scoring metric for evaluating the patterns, and the number of patterns applied at each iteration. This approach should also be contrasted to Vilain and Day (1996) and Bikel et al. (1997) which learn similar contextual patterns, but while they achieve high accuracy results, they rely on annotated data. In our method only untagged text, a list of seed terms and a lexicon of common English words are needed thus avoiding the cost of manually annotating a large amount of data.

We must note that the objective of acquiring terms from untagged texts is twofold:

- To investigate the feasibility of extracting terms from untagged texts in order to support the rapid adaptation of systems that use terminological lexicons to new domains.
- To evaluate the impact of the new terms on IE extraction tasks.

The next section describes the problem domain with regard to the characteristics of the biological texts and the nature of the terms to be extracted. In section 3, the technical aspects of the algorithm are described and in section 4 the experiments with the algorithm and their results are reported. In section 5, we describe the process of terminology identification within the PASTA IE system and how the augmentation of the existing lexicons with the new terms influences the results of term recognition and classification.

2. The Problem Domain

Typically, journal articles in the domain of protein structures describe details of protein composition in terms of the amino acids that take part in 3-dimensional structural arrangements. New protein structures are being reported in the literature at very high rates and the number of protein co-ordinate sets (currently about 12000) in the Protein Data Bank (PDB) (Bernstein et al., 1977) is expected to increase ten-fold in the next five years. As an example of a text from this domain, a fragment of a journal paper is shown below:

> *Results: We have determined the crystal structure of a triacylglycerol lipase from Pseudomonas cepacia (Pet) in the absence of a bound inhibitor using X-ray crystallography. The structure shows the lipase to contain an alpha/beta-hydrolase fold and a catalytic triad comprising of residues Ser87, His286 and Asp264. The enzyme shares several structural features with homologous lipases from Pseudomonas glumae (PgL) and Chromobacterium viscosum (CvL), including a calcium-binding site. The present structure of Pet reveals a highly open conformation with a solvent-accessible active site. This is in contrast to the structures of PgL and Pet in which the active site is buried under a closed or partially opened 'lid', respectively.*

Identifying protein names in biological papers is a challenging and demanding task. To our knowledge, there is no standard grammar that can fully describe the structure of protein names. Protein names can either be single words (e.g. `pectin, endonuclease`) or compounds that may consist of two or more words (e.g. `major birch pollen allergen Bet v 1`). More than 70% of the protein entries in our lexicons are multi-word names.

In contrast to names of persons, organisations and locations in newswire texts, capitalisation of the first letter is

[1] http://www.dcs.shef.ac.uk/research/groups/nlp/pasta/

not a standard feature of protein names. They may contain lower and upper case characters in various positions (e.g. `tRNA synthetase`), numerals and non-alphabetical characters (e.g. `1,3,8-trihydroxynaphtalene reductase`) and they can have no special prefix or suffix (e.g. `pbp-2x, c-Raf1, f1f0-atp`)[2]. In some cases, the names contain conjunctions (e.g. `Hirutonin-2 and -6`) or prepositional phrases that may indicate a subunit (e.g. `beta1-subunit of the signal-transducing G protein`) or a function (e.g. `bcl-2 inhibitors of programmed cell death`).

One way of developing an algorithm for the automatic extraction of terms would be by training statistical models (Bikel et al., 1997) or by learning phrase sequence rules (Vilain and Day, 1996) using large amounts of annotated data. However, because no corpus annotated with biological information has been available, such approaches were not feasible in our case. Furthermore, the application of syntactic taggers and parsers trained on general language texts would be questionable due to the occurrence domain specific biological terms that usually do not occur in general language. It is for these reasons that we believe a bootstrapping technique is particularly relevant for discovering new terms in this domain.

3. Description of the Bootstrapping Algorithm

A bootstrapping approach to term acquisition starts with a selection of examples of terms that may occur in the corpus. These examples are then fed into the system to identify contextual clues or patterns that frequently occur in the environments of the terms. For example, in journal articles on protein structure, authors frequently use expressions such as `the crystal structure of <P>` or `three-dimensional structure of <P> from <S>` where `<P>` is a protein name and `<S>` is the name of a species. The bootstrapping procedure would first attempt to match the seed terms in the text (i.e. `triacylglycerol lipase`) and extract its left and right contexts (i.e. `the crystal structure of a` and `from`). It will subsequently use these patterns to extract new terms which will be used as new seeds and the process will start again.

One of the considerations when applying a bootstrapping algorithm is whether the generated patterns will be reliable enough for identifying new terms. It would be reasonable to assume that from all patterns generated at each iteration of the algorithm, only a fraction would identify protein names with a high degree of reliability. In related work by Jones et al. (1999), a heuristic function scores the patterns and only the highest ranked pattern is used as seed at each iteration. In the algorithm presented in this paper, a statistical test is used to prune the list of extracted patterns and identify those which are significantly associated with protein names.

From the statistical point of view, it would be interesting to estimate the language constraint of the biology texts. Such an estimation can give an indication of the degree of the difficulty of identifying frequent language expressions in such texts. In an initial investigation, we estimated the constraint in our texts using perplexity, an information theoretic measure commonly used in language modelling (Jelinek, 1990). The perplexity for the corpus of biological texts was found to be much lower (270) than for texts of similar size drawn randomly from the British National Corpus (740). This is an indication that although the vocabulary is very rich for this domain, the language constraint is quite high and we would expect certain language patterns to occur frequently enough in the corpus so that they can be candidates for pattern generation. The outline of the bootstrapping algorithm is shown in Figure 1.

1. **Initialisation**:
   - Extract initial set of contextual patterns using the seeds
   - Identify significant patterns
   - Terminate if no significant patterns exist
2. **Name Extraction**:
   - Apply significant patterns and extract new names
   - Terminate if no new names are found
3. **Pattern Generation**:
   - Match the new names and extract new patterns
   - Identify significant patterns
   - Terminate if no significant patterns exist
4. Go to step 2

Figure 1: The term bootstrapping algorithm.

In the initialisation phase, the text is processed in a left-to-right order and a longest match procedure is used to match potential word sequences against the seed terms. For each matched term, the word sequences or n-grams of up to six words that occur immediately to the left and to the right of the term are extracted. Only the statistically significant n-grams as scored by the Pearson χ^2 test are retained as contextual patterns. In our experiments, the χ^2 significance level was set at 0.5%, and, to avoid problems with low counts, only n-grams with a frequency of 5 or more were scored.

In the name extraction phase, the patterns are applied to the text in order to identify new terms. Any pattern that indicates a left context is matched first and the algorithm will gather any word sequence to the right of the pattern as a potential new term (unless the same term has been matched before) up to the point when a right-side contextual pattern is found.

In initial versions with the algorithm, both the left-side and right-side contexts were statistically evaluated. It was found however, that the number of significant right-side contexts identified in this way was very low and in most cases there would be no correlation at all between left-side and right-side contextual patterns in our texts. For this reason, the algorithm was modified so that the right-side patterns are matched against unigrams taken from a list of common English words.

[2]For a more complete analysis of the nomenclature of protein names see Fukuda et al. (1998)

The newly identified terms are used in phase 3 to extract more contextual patterns which are in turn used to extract new terms. This iterative process terminates when no more contextual patterns or no more terms are found.

This algorithm is conceptually simple and has a number of differences from the one of Jones et al. (1999), most notable of which is the generation of contextual patterns. Jones et al. (1999) use the AutoSlog system (Riloff and Lehnert, 1993) for pattern generation which requires some sort of grammatical analysis of the sentence in order to assign noun phrases to syntactic categories such as subject, direct object or prepositional phrase. In this approach, no grammatical analysis of the text is necessary and no heuristics are used for generating the contextual n-grams. In addition, in Jones et al. (1999)'s work the patterns are scored by a heuristic scoring function that attempts to balance the frequency of a pattern with its reliability in extracting names of the same class and only the top ranked pattern is used at each iteration of the bootstrapping process. In our algorithm, the patterns are evaluated statistically and all patterns judged as significant are retained and applied at each iteration.

4. Experiments and Results

In the experiments, we used a corpus of 1500 scientific abstracts (about 350,000 words), an initial seed lexicon of 660 protein names and variable length n-grams ranging from 1 to 6 words. The bootstrapping algorithm identified 98 significant contextual patterns in total from which it extracted 984 unique new names.

For the evaluation we classified the answers into three different categories:

Correct: This category includes extractions that were true protein names in their full forms.

Partially Correct: This category includes extractions of either

- names which were not extracted in their full forms (as in the case of `lactoferricin` instead of the correct `lactoferricin b)`, or
- names that contained irrelevant items (usually common words) as part of the protein name (as in the case of `enzyme methylmalonyl-coenzyme A (CoA)` instead of the correct `methylmalonyl-coenzyme A (CoA)`).

Incorrect: This category includes incorrect answers or those that could not be put in any of the two previous categories.

The evaluation classified 58% of the answers as correct, 13% partially correct and 29% incorrect. The correct answers thus represent about 86% augmentation to the entries included in the seed lexicon. The top 20 contextual patterns identified by the algorithm are shown in table 1.

An analysis of the partially correct and incorrect answers provides some insight into the errors made by the algorithm. About half of the partially correct answers were due to the fact that the algorithm missed part of a name because it judged common English words such as `domain, subunit, type`, etc. as right patterns. For example, for a name such as `fibronectin type III`, the algorithm would extract only the constituent `fibronectin`. Such extractions were classified as partially correct only if the extracted name could be used to indicate a protein or a protein family (otherwise such names were classified as incorrect).

Significant contextual patterns
of human
structure of the
of the human
encodes a
domain of the
the cholera
of staphylococcal
solution structure of
the bacterial
first structure of a
crystal-structure of
the reaction catalyzed by
crystal structure of
members of the
structure of the human
domains of the
a member of the
3-dimensional structure of
dna-binding domain of the
rat liver
three-dimensional structure of

Table 1: Top 20 contextual patterns ranked by χ^2

A substantial proportion of the partially correct answers included common words like `the, enzyme, protein` etc. usually at the beginning of the extracted name. Such words would often follow left-side patterns such as `enzyme` in `the structure of the the enzyme 3-oxo-Delta(5)-steroid isomerase`.

It could be argued that with simple modifications to the algorithm such errors could be eliminated and the majority of the partially correct answers could be extracted as fully correct. A modification might involve, for instance, a simple checking procedure so that common English words are not included at the beginning of a protein name. Another modification would ensure that words such as `domain` or `type` (that may be part of a protein's name) are not matched as right-side contextual patterns by looking them up in a list of exceptions. However, it should be taken into account that such heuristics may not work for a different term class or a different text domain.

With regard to the incorrect extractions, a large percentage of the errors made by the algorithm were those when part of the protein name was extracted as the full name. This is is similar to the type of errors made for the partially correct answers, but in this case the extracted terms could not be protein names at all (e.g. `catalytic, copper-substituted, major beta-sheet`, etc.).

A second type of errors were due to references to proteins (or parts of proteins) that had been mentioned before in the text. For example, in the case of `the solution structure of the peptide`

`backbone was determined` the algorithm would extract `peptide backbone` as the name. There were a lot of instances in these texts where, once a protein is introduced by its name, the authors often refer back to it by using terms such as `protein, enzyme, peptide, subunit, sequence, complex,` etc. We decided to judge these referents as errors as they are general terms that cannot be used to describe a specific protein explicitly if added to the lexicon.

Finally, names referring to protein complexes were difficult to extract and were responsible for a number of errors as, for example, in the case of `the crystal structure of a stoichiometric complex between an elastose-specific inhibitor elafin and porcine pancreatic elastase (ppe)` where `stoichiometric` would be extracted as the protein name. Unfortunately, a simple statistical approach cannot capture complicated contextual relationships in such expressions, and a more sophisticated linguistic analysis would be required to identify the syntactic structures (noun phrases and prepositional phrases) that are used in many names of protein complexes.

We conducted an experiment to investigate the impact of the text size on the extraction of new terms. In this experiment, the number of the texts varied from 300 to 1500 while the seed lexicon included 660 terms.

In the bar graph of figure 2, the matched seeds represent the number of unique seed terms matched in the texts, the patterns represent the total number of significant patterns identified and the new terms indicate the total number of extracted names, either correct or incorrect.

It could be argued that with more texts there is an increased probability for finding a seed term in the texts. This hypothesis was verified practically and the probability of matching a seed term in the text was found to range from about 9% (300 texts) to 39% (1500 texts). The results suggest that with more texts, there are generally more seed terms matched and more patterns identified by the algorithm. As an effect, more new terms are extracted from the texts.

In a different experiment, we investigated the impact of the size of the seed lexicon on the discovery of patterns and the extraction of terms. In this experiment, all 1500 texts were used and the size of the seed lexicon varied from 100 to 660. On average each seed term was found to match from 0.27 times (100 seeds) to 1.9 times (660 seeds) in the texts. It can be seen from Figure 3 that with more seeds, there is an increased number of significant patterns discovered and consequently an increased number of extracted terms.

5. Evaluation within PASTA

We evaluated the contribution of the terms extracted by the bootstrapping algorithm in terminology identification tasks within the PASTA IE system using a test set of 50 scientific journal abstracts. The standard evaluation metrics of *precision* and *recall* were used for evaluation. Precision is the percentage of the system's answers that are correct. Recall is the percentage of the correct answers in the texts that the system managed to retrieve.

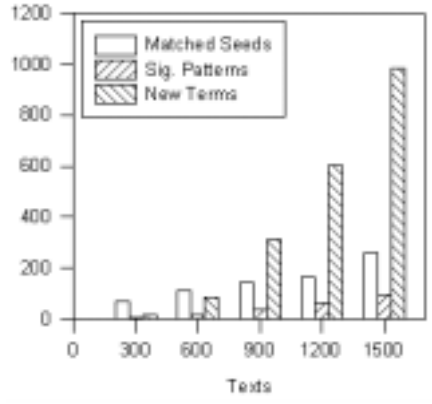

Figure 2: The impact of text size on pattern discovery and term extraction

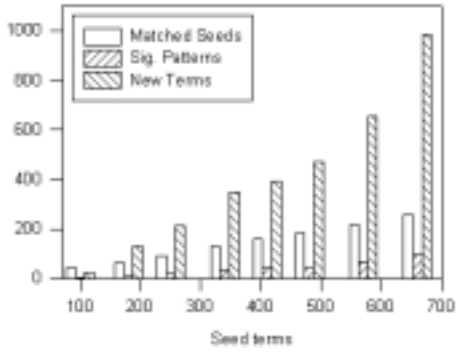

Figure 3: The impact of the seed lexicon size on pattern discovery and term extraction

It must be noted that the PASTA system has been adapted from a generic IE system LaSIE (Gaizauskas et al., 1995) that took part in the MUC competitions. PASTA has been manually tuned to the biology domain over a considerable amount of time and includes lexicons of more than 15000 biological terms split in more than 50 term classes. In previous evaluations the terminology analysis component of PASTA was found to compare favourably with NE results in the MUC conferences (88% recall and 94% precision overall for all term classes used in the system).

In summary, the main information sources of PASTA for term recognition and classification in the biochemical domain are case-insensitive terminology lexicons, the listing of component terms of various classes, morphological cues (mainly standard biochemical suffixes) and specialised grammar rules for each terminology class.

The PASTA terminology processing component consists of a pipeline architecture of the following four processing stages:

I. Text Preprocessing

II. Morphological Analysis

III. Terminology Lookup

IV. Terminology Parsing

The text preprocessing stage provides information about the structure of the text and applies tokenisation pattern matching rules to identify the individual text units in terms of words or subword units.

The morphological analysis stage identifies the root form and suffix for each token using a general English morphological analyser that has been supplemented with 100 biomedical suffixes (such as '-ase', '-in' or 'yl') which are used as morphological cues in the terminology parsing stage.

In the terminology lookup stage, lexicons assembled from publicly available resources, are used for looking up the biochemical terms in the text. In total, the number of terms in the various lexicons exceeds 20,000 for all the terminological classes. With regard to protein names, the system currently includes more than 2000 protein names and more than 3000 component terms of protein names. Because for multi-word names there is an increased probability for different spellings or name variations, the terminology lookup component is coupled with a rule-based terminology parser.

The use of a terminology parser requires the names of proteins (or other entities in the domain) to be decomposed into constituent parts. These constituents have either been matched separately during the term lookup phase, or have specific properties that have been identified during the tokenisation or morphological analysis stages. For example, the protein name `serine carboxypeptidase II` would be recognised firstly by the classification of `serine` as a potential amino acid residue, and `II` as a protein modifier, both by being matched in the terminology lexicons. Morphological analysis would identify `carboxypeptidase` as being a potential *protein head*, due to the suffix `-ase`, and then grammar rules would apply to combine the protein head with a residue and with a known protein modifier. The set of rules is derived semi-automatically using the multi-word names in the protein lexicons and currently includes about 160 rules.

The terms discovered by the bootstrapping algorithm were added to the PASTA terminological modules both for augmenting the existing lexicons and for deriving new grammar rules. As baselines for system evaluation, we first ran the original PASTA with and without the terminology parsing module.

The results of the experiments are given in table 2. With the rule-based parser switched off, we wanted to evaluate the contribution of the new terms when only the terminology lookup component is used for term identification. The baselines for the the terminology lookup component of the PASTA system are 31% recall and 97% precision. The low recall rate highlights the insufficiency of the simple lookup procedure in matching multi-word names in the texts. When the new terms were added to the system, recall increased to 38% while precision dropped to 96%.

An investigation into the incorrect answers has revealed that nearly all of the errors can be attributed to the terms `apo` (apoprotein) and `holoenzyme` which were not tagged as proteins because they appeared as part of other protein names in the evaluation texts (although these two names were found as proteins on their own in other texts).

With the terminology parser switched on, we evaluated the overall contribution of the new terms in the PASTA terminology identification subsystem. The baseline for the original PASTA system with the grammar rules is 87% recall and 97% precision. When the new terms were used together with the old grammar rules there was an increase in terms of recall (90%) with a small decrease in precision (again 96%). In a another experiment, new rules were derived from the terms and were added to the old rules but there was no observable difference from the previous results (i.e. 90% recall, 96% precision). An explanation for this may be that the old rule set had already proved to have quite good coverage (at least for the texts used in our tests) and the new terms represent only a fraction of the total number of terms used in PASTA so that they could not make a significant contribution in deriving new rules.

The above results indicate that, overall, the new names can make a contribution to the term identification capabilities of the system but only in terms of recall. There is a small decrease in terms or precision, which is probably to be expected since with the addition of new terms may result in more erroneous entries and more ambiguities in the lexicon.

System Configuration	REC	PRE
Original lexicons	31	97
Original lexicons + new terms	38	96
Original lexicons + rules	87	97
Original lexicons + rules + new terms	90	96
Original lexicons + new terms + new rules	90	96

Table 2: Evaluation of term recognition in PASTA

6. Conclusions

In this paper, we have described an approach for the automatic acquisition of terminology from untagged texts. Based on the distributional hypothesis that terms of a semantic class occur in similar contextual patterns, a bootstrapping algorithm was developed. The algorithm uses a set of seed terms to identify contextual patterns which are in turn used to extract new terms iteratively.

The output of the system could provide input to a semi-automatic process for extending the terminological lexicons for the domain and therefore assist in the portability or adaptability of natural language processing systems to new domains. The advantage of this method is that it does not depend for training on tagged corpora or linguistic tools such as syntactic and semantic taggers which usually are costly and time-consuming to produce.

The results of the experiments with biological text resources demonstrate the viability of the proposed method for acquiring new terms from domain restricted texts. For our texts, the accuracy of the method ranged from 58% to 71%, depending on the interpretation of the results.

This approach may not be limited to augmenting existing lexicons of terms but could also be profitable in term recognition and classification tasks. On the other hand, we must be aware of the limitations of such a simple method which is based just on statistical information. It was found

that the algorithm will often pick common words or expressions as domain specific terms (mainly those used for coreference in the texts) and it is not clear what the best strategy can be for detecting and filtering out such terms during bootstrapping.

The new terms were found to make a positive contribution to the recall rate of our IE system albeit with a minimal negative effect on its precision. It is unlikely though, even with a large augmentation of the existing lexicons, that the precision and recall rates would approach 100% since new terms will continue to be added to the vocabulary and any lexicon will almost always include occasional errors.

There are opportunities for further development and testing of the technique. We plan to test its applicability with more term classes and new domains. It would be interesting to investigate whether such an approach can help in the discovery of patterns that describe domain specific relations. One possibility is to employ a bootstrapping algorithm for identifying representative patterns that may describe, for instance, that a protein comes from a species or that a residue is found in a protein. The generation of such patterns automatically can be useful for discourse modelling purposes and the rapid adaptation of NLP systems to new applications.

7. Acknowledgments

The PASTA project is funded under the UK BBSRC/EPRSC BioInformatics Programme (50/BIF08754) and is a collaboration between the Departments of Computer Science, Information Studies and Molecular Biology and Biotechnology at the University of Sheffield. The authors would like to thank Dr. Peter Artymiuk and Prof. Peter Willett of the University of Sheffield for supplying their expertise in the biology domain.

8. References

F. Bernstein, T. Koetzle, G. Williams, E. J. Meyer, M. Brice, J. Rodgers, O. Kennard, M. Shimanouchi, and M. Tasumi. 1977. The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, (112):535–542. Available at `http://www.rcsb.org.pdb`.

D. Bikel, S. Miller, R. Schwartz, and R. Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 194–201.

M. Collins and Y. Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Coneference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–110.

S. Cucerzan and D. Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence. In *Proceedings of the Joint SIGDAT Coneference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 90–99.

DARPA, editor. 1995. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Defense Advanced Research Projects Agency, Morgan Kaufmann.

DARPA, editor. 1998. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Defense Advanced Research Projects Agency, Morgan Kaufmann. Available at `http://www.saic.com`.

DARPA, editor. 1999. *Proceedings of the DARPA Broadcast News Workshop*. Defense Advanced Research Projects Agency, Morgan Kaufmann.

A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society*, B(39):1–38.

K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. 1998. Information extraction: Identifying protein names from biologi cal papers. In *Proceedings of the Pacific Symposium on Biocomputing '98 (PSB' 98)*, pages 707–718.

R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham, and Y. Wilks. 1995. Description of the LaSIE system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 207–220. Morgan Kaufmann.

D. Harman. 1998. The text retrieval conferences (TRECs) and the cross-language track. In *Proceedings of the First International Conference on Language Resources & Evaluation*.

F. Jelinek. 1990. Self-organised language modeling for speech recognition. In A. Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*, pages 450–503. Morgan Kaufmann.

R. Jones, A. McCallum, K. Nigam, and E. Riloff. 1999. Bootstrapping for text learning tasks. In *IJCAI'99 Workshop on Text Mining: Foundations, Techniques and Applications*, pages 52–63, Stockholm, Sweden.

J. Justeson and S. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Journal of Natural Language Engineering*, 1(1):9–27.

S. Renals, Y. Gotoh, R. Gaizauskas, and M. Stevenson. 1999. Baseline IE-NE experiments using the SPRACH/LaSIE system. In *Proceedings of the DARPA Broadcast News Workshop*. Morgan Kaufmann.

E. Riloff and W. Lehnert. 1993. Automated dictionary construction for information extraction from text. In *Proceedings of Ninth IEEE Conference on Artificial Intelligence for Applications*, pages 93–99.

M. Vilain and D. Day. 1996. Finite-state parsing by rule sequences. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*.

D. Yarowsky. 1995. Unsupervised word-sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL95)*, pages 189–196.