Automatic Acronym Acquisition and Term Variation Management within Domain-Specific Texts*

Goran Nenadić, Irena Spasić, Sophia Ananiadou

Computer Science, University of Salford
Newton Building, Manchester M5 4WT, UK
{G.Nenadic, I.Spasic, S.Ananiadou}@salford.ac.uk

Abstract

In this paper we present a framework for the effective management of terms and their variants that are automatically acquired from domain-specific texts. In our approach, the term variant recognition is incorporated in the automatic term retrieval process by taking into account orthographical, morphological, syntactic, lexico-semantic and pragmatic term variations. In particular, we address acronyms as a common way of introducing term variants in scientific papers. We describe a method for the automatic acquisition of newly introduced acronyms and the mapping to their 'meanings', i.e. the corresponding terms. The proposed three-step procedure is based on morpho-syntactic constraints that are commonly used in acronym definitions. First, acronym definitions containing an acronym and the corresponding term are retrieved. These two elements are matched in the second step by performing morphological analysis of words and combining forms constituting the term. The problems of acronym variation and acronym ambiguity are addressed in the third step by establishing classes of term variants that correspond to specific concepts. We present the results of the acronym acquisition in the domain of molecular biology: the precision of the method ranged from 94% to 99% depending on the size of the corpus used for evaluation, whilst the recall was 73%.

1. Introduction

Rapid changes in specialised areas, such as molecular biology (MB), computer science, telecommunications, etc. result in a vast and constantly increasing amount of knowledge represented via documents. The fast growth of electronic text collections demands innovative techniques to access, gather and systematically structure information. In order to discover knowledge, one has to identify the main concepts, which are linguistically represented by domain specific terms (Maynard et al., 2001). In the specialised fields, there is an increased amount of new terms that represent newly created concepts. Since existing term dictionaries, standardised terminologies, nomenclatures, ontologies, and other language resources offer only a partial solution to the needs of specialists, the automatic term extraction tools are indispensable for efficient knowledge discovery and dynamic update of the language resources (Ananiadou et al., 2001).

There are numerous approaches to the automatic term recognition. Some methods (Bourligault, 1992; Ananiadou, 1994) rely purely on linguistic information, namely morpho-syntactic features of term candidates. Recently, hybrid approaches combining linguistic and statistical knowledge are becoming increasingly used (Frantzi et. al, 2000; Nakagawa et al., 1998; Maynard et al., 2001). (Hatzivassiloglou et al., 2001) used a statistically based machine learning technique to acquire and disambiguate specific terms in MB (e.g. protein and gene names).

The naming conventions in many domains (especially in MB) are highly non-standardised even when it comes to the fundamental concepts. In theory, terms should be mono-referential (one-to-one correspondence between terms and concepts), but in practice we have to deal with (semantic) **ambiguities** (the same term corresponds to many concepts) and **variants** (many terms leading to the same concept). For example, the term `gene` has at least two meanings (senses) within the domain of MB (MBO, 2002):

1) *a DNA fragment transcribed and translated into a protein,* and
2) *a DNA region that carries genetic phenotype*.

On the other hand, all the following variants: `nuclear factor-kappa B`, `NF-kappaB`, `NF kappa B`, `NF(kappa)B`, `kappaB`, `NFKB factor`, `NF-KB`, `NF kB`, etc. denote the same concept described as *a transcription factor involved in immune responses, inflammation and cell proliferation* (MBO, 2002).

Dealing with term sense disambiguation is crucial for classifying terms and ontology populating (Bisson et al., 2000). The appropriate term sense is usually discovered by examining the similarity between the given term and its context (Nenadic et al., 2002). However, there is classificatory ambiguity as well, since one concept can be classified in more than one way depending on the point of view. We refer to this fact as term multi-dimensionality.

If we aim at supporting systematic acquisition and structuring of domain-specific knowledge, then handling term variation has to be treated as an essential part of terminology retrieval. Term variation ranges from simple orthographic variation to semantic variation. For example, (Jacquemin, 2001) processes morphological and syntactic variations by means of meta-rules used to describe term normalisation, while semantic variants are handled using WordNet. In this paper, we propose a framework for effective term variation management in a specific domain. In particular, we address acronyms as a common type of term variation. We present a method for the automatic acquisition of newly introduced acronyms and their mapping to the respective terms.

* This research is a part of the BioPATH research project coordinated by LION BioScience (http://www.lionbioscience.com) and funded by German Ministry of Research.

The paper is organised as follows: in Section 2 we briefly discuss the term variation problems related to the automatic term recognition. In Section 3 we present our approach to the automatic acquisition of acronyms in the domain of molecular biology. Section 4 presents a framework for incorporating term variants into the term recognition process. The experiments and evaluation are given in Section 5.

2. Handling term variations

In an attempt to name the newly discovered concepts, scientists often use some predefined naming patterns, that is - term formation usually follows some planned, conscious efforts. There are even some formal naming standardisation initiatives: e.g. the Guidelines for Human Gene Nomenclature (Lander et al., 2001) establish a naming convention for new gene names, including principles such as avoiding molecular weight designations, starting a name with a lower case letter, obligatory American spelling, etc. However, this does not prevent ad hoc naming solutions (e.g. a gene known as `Bride of sevenless` or `BOSS`). Domain experts frequently introduce specific notations for terms that they use either locally (in a paper) or within a whole community.

In addition to loose naming conventions, a considerable terminological confusion arises from the following problems:

a) **Term variation**: the same concept is named differently in different sources (e.g. `TIF2`, `TIF-2`, `transcription intermediary factor-2`, and `transcriptional intermediate factor 2` all denote the same concept).

b) **Term ambiguity**: the same term may refer to different concepts, the intended sense being context dependent (e.g. the acronym `GR` is a short form for both an enzyme (`glutathione reductase`) and a nuclear receptor (`glucocorticoid receptor`)).

Term variation and ambiguity are causing problems not only for automatic knowledge acquisition but for human experts as well. Therefore, there is a genuine need for the standardisation of terminology (White et al., 1998).

In our work, we concentrate on term variation management as opposed to term ambiguity resolution. A variety of sources from which the term variation problems stem are considered. In particular, we deal with orthographical, morphological, syntactic, lexico-semantic and pragmatic phenomena. Our approach to the term variation management is based on term normalisation as an integral part of the automatic term recognition (ATR) process. Term variants are dealt with in the initial phase of ATR when term candidates are singled out. More details about the way in which the ATR process is performed and how term variations are incorporated into this process are given in Section 4.

The following subsections will briefly overview the main types of term variation. Especially, acronyms, as a special type of term variation, are discussed separately in Section 3.

2.1. Orthographical variations

This type of variation is very common in certain domains, especially in MB. The simplest orthographical variations include optional usage of hyphens, different cases (lower/upper) and different (standard) spellings (e.g. American and British English: `tumour` vs. `tumor`). In order to manage these variations all term candidates are mapped to their normalised forms, which in our approach are lower-case forms without hyphens. In addition, different Latin/Greek transcriptions and neoclassical compoundings are very frequent in the MB domain, e.g.:

`leukaemia` vs. `leukaemia`
`oestrogen` vs. `estrogen`
`amyloid beta-protein` vs. `amyloid β-protein` vs. `amyliod b-protein`

Different Latin/Greek transcriptions are converted to their normalised forms according to a manually collected mapping list (e.g. *ae* → *e, oe* → *e, β* → *B, beta* → *B,* etc.). In order to recognise specific transcriptions, we used a set of morphological constraints for the distribution of neoclassical roots and affixes as explained in the previous work (Ananiadou, 1994).

An additional orthographical phenomenon in MB is the usage of specific "mathematical" words and/or symbols (e.g. `+`, `plus` or `positive`). These symbols can be used interchangeably (e.g. `ER positive` or `ER+`). In our approach, we normalise such term candidates by replacing "mathematical" words with the corresponding symbols.

2.2. Morphological variations

This type of variation is common in many domains. The simplest morphological variations include the usage of plural, singular and possessive noun forms, e.g.:

`nuclear receptor` vs. `nuclear receptors`
`Down syndrome` vs. `Down's syndrome`

Since currently used taggers are fairly accurate in recognising such linguistic phenomena, we rely on their output and define a normalised term form as a singular form containing no possessives.

In addition to the mentioned variants, derivational variations are also present (e.g. `intermediary factor` vs. `intermediate factor`). They can be handled by stemming, but at this stage we do not normalise terms to their stemmed forms. The main reason is a difference between 'active' and 'passive' forms of modifiers (e.g. `activated factor` vs. `activating factor`), which are usually used to denote different terms.

2.3. Syntactic variations

The simplest syntactic variations include structural differences in possessive usage of nouns (with or without a preposition), e.g.:

In our approach, we handle these syntactic variants when applying linguistic filters, which are used to describe term

`clones of human` vs. `human clones`
`cancer in humans` vs. `human cancer`

candidates (see Section 4). The filters that produce normalised variants are implemented via transformations represented by unification-like LR(1) rules (Mima et al., 1995). Here is an example of a rule describing the transformation of a term candidate that contains the preposition `of`:

```
; A of B  →  B  A
Term -> (A|Noun)* Noun PREP(of) (A|Noun)* Noun
     (x0 next1) = x5
     (x0 next2) = x4
     (x0 next3) = x1
     (x0 next4) = x2
```

In our approach, the normalised forms do not contain prepositions.

The coordination of terms is also commonly used variation phenomena in MB. We differentiate between two main types of term coordination:

a) **Argument coordination**: term arguments are coordinated, thus the coordinated terms are retrieved by "multiplying" the arguments with the shared head:

`SMRT and Trip-1` **`mRNAs`**

→ { `SMRT mRNA`
  `Trip-1 mRNA`

b) **Head coordination**: term heads are coordinated, thus the coordinated terms are retrieved by "multiplying" the heads with the shared arguments:

`adrenal` `glands and gonads`

→ { `adrenal glands`
  `adrenal gonads`

We supported the two coordination types by the LR(1) rules (Mima et al., 1995), which generate candidate terms when a coordinated structure is recognised. Note that the generated term candidates are not necessarily correct terms, as the syntactic filters describing a coordinated structure are ambiguous:[1] the same filter retrieves both coordinated terms and conjunctions of terms (in which case the "multiplication" would give incorrect results) as presented in Table 1. In order to reduce the number of incorrectly recognised coordinated structures, the "multiplied" forms are not treated as term candidates unless they appear in a corpus on their own.

syntactic filter	Adj Noun and Noun
example	adrenal glands and gonads
head coordination	[adrenal [glands and gonads]]
term conjunction	[adrenal glands] and [gonads]

Table 1: Syntactic ambiguities of coordinated structures

The problem of more complex syntactic variations, involving more than one type of the described syntactic variations, remains open. For example, prepositional phrases are often combined with coordinated structures, e.g.:

```
mechanism of steroid/thyroid receptor
                  ↓
     mechanism of steroid receptor
     mechanism of thyroid receptor
                  ↓
       steroid receptor mechanism
       thyroid receptor mechanism
```

This problem demands further investigation on "precedence" between certain syntactic transformations (in the above example, coordination has a priority over preposition).

2.4. Lexico-semantic variations

Lexico-semantic variations include the usage of synonyms in the process of assigning names to concepts:

`carcinoma`	vs.	`cancer`
`eye surgery`	vs.	`ophthalmologic surgery`

In our approach, this type of variants is handled by using a dictionary of synonyms, whose entries consist of a preferred term and a list of its synonyms. The preferred forms are used to normalise terms.

We do not consider other types of semantic variants as part of the term recognition process: e.g. although `nucleic acid-binding` and `DNA-binding` are semantically related as `DNA` is a hyponym of `nucleic acids`, we do not consider the latter as a term variant of the former as we deal with synonyms only. However, this type of semantic term variation can be useful for term classification and clustering (Jacquemin, 2001).

3. Acronym acquisition

Acronyms are a very common term variation phenomenon in MB. They can be regarded both as lexico-semantic and pragmatic term variants. In the lexico-semantic sense, acronyms are used as synonyms for the corresponding terms (these terms will be referred as **expanded forms**). In the pragmatic sense, acronyms are used to facilitate the readability of scientific texts.

In the field of MB, domain experts frequently introduce specific acronyms, which are used either locally (in a paper) or within the whole community. Analogously to the problems of variants and ambiguities for terms in general, acronyms are afflicted with the similar problems arising from the following points:

a) **Acronym variation**: the same term may have several acronyms (e.g. `NF kappa B, NF kB`).
b) **Acronym ambiguity**: the same acronym may refer to different concepts (`GR` is an abbreviation for both `glucocorticoid receptor` and `glutathione reductase`).

In our approach to acronym acquisition we deal with acronym variation problems only. The acronym acquisition is a part of ATR: each (candidate) acronym occurrence is replaced with the corresponding expanded form prior to the statistical analysis. This way, all term occurrences are considered for calculation of term-hood.

[1] We have tested coordination generation implemented in our approach, and the overall precision was 70%.

The problem of acronym ambiguity can be simply resolved by using the lastly introduced acronym definition, if there is one. If there is no definition introduced, then general methods for term disambiguation have to be used (Spasic et al., 2002).

3.1. Method

Our three-step method for acronym acquisition is based on both morphological and syntactic features of acronyms and their expanded forms. We rely on syntactic patterns that are used predominantly to introduce acronyms in scientific papers in order to locate potential acronym definitions. Once a word sequence matching such a pattern is retrieved, it is morphologically analysed with the aim of discovering the relation between the acronym and its expanded form.

3.1.1. Retrieval of acronym definitions

In the first step, we scan the text for the candidate acronym definitions. Several definition patterns have been identified manually in order to describe different contexts for introducing an acronym. We differentiate between two general types of acronym definitions:

a) **left definition**: an acronym follows its expanded form (e.g. 9-cis retinoic acid (9cRA));

b) **right definition**: an acronym is followed by its expanded form (e.g. MIBP (Myc-intron-binding peptide)).

Left definitions are far more frequent (more then 90% of all acronym definitions). In both cases, acronyms are introduced either by bracketing (e.g. tumor necrosis factor alpha (TNF-alpha) or glutathione peroxidase [GPx]) or rarely by using apposition-like format (e.g. ... enzyme-linked immunosorbent assay, ELISA,...). The definition patterns have been modelled by local grammars (Gross, 1989) and applied at the position in the text where a potential acronym is introduced in order to retrieve an acronym and its expanded form. Usually, the extracted context was wider than the actual acronym definition.

However, there are contexts seemingly introducing an acronym, although they provide no syntactic evidence (in the form of brackets or appositions) for that fact (e.g. **h**eat **s**hock **p**rotein Hsp90). Such contexts were not considered as acronym definitions in our approach.

3.1.2. Matching acronyms against expanded forms

In the second step of the acronym acquisition procedure, a set of acronym formation patterns is applied in order to match a candidate expanded form with the corresponding acronym. In general, acronyms are formed by selecting first (or first few) letters of the words from the expanded form. However, it has been noticed in the MB domain that the initial letters of **combining forms** are also used for the same purpose. Combining forms are specific affixes (mostly prefixes and infixes, e.g. acetyl, trans, di, hydro, etc.) that are regularly used in term formation patterns (e.g. **c**hloramphenicol **a**cetyl**t**ransferase (CAT)). A dictionary of domain-specific combining forms is used when matching an acronym against the expanded form.

The basic matching method between acronyms and their expanded forms is augmented by taking into account the following phenomena related to acronym definitions:

- **insertion**: a word is present in the expanded form of an acronym, but it is not used in the formation of the acronym (e.g. **t**hyroid hormone **r**eceptor (TR));
- **omission**: a word is missing from the expanded form of an acronym, although it is used when forming the acronym (e.g. [**h**uman] **e**strogen **r**eceptor (hER));
- **plural acronym**: a plural form acronym is defined (e.g. **r**etinoid **X** **r**eceptor**s** (RXRs));
- **recursive acronym**: the expanded form of an acronym already contains an abbreviation/acronym (e.g. **C**REB-**b**inding **p**rotein (CBP));
- **coordinated acronyms**: acronyms are defined within a coordinated structure (e.g. **e**strogen (ER) and **p**rogesterone (PR) **r**eceptors);
- **partial acronym**: an acronym contains a part of its expanded form, usually Greek/Latin words (e.g. **r**etinoid **X** **r**eceptor **alpha** (RXR alpha));
- **structural variation**: an acronym is defined after a morphological/structural transformation is conducted on its expanded form (e.g. **d**ay of **h**atching (HD));
- **formula-like acronym**: an acronym contains (a part of) a chemical formula (e.g. **1**alpha,**25**-**dihydroxy**vitamin **D3** [1,25(OH)2D3]).

The listed phenomena are considered when the basic matching method (matching acronym letters against constituents and combining forms) has not produced a positive result. Finally, this step results in a list of matched acronyms and their definitions.

3.1.3. Acronym clustering

In order to address the problem of acronym variation, in the third step we attempt to establish the classes of variants that correspond to the same concept. First, both acronyms and their expanded forms are normalised with respect to their orthographic, morphological, syntactic and lexico-semantic features (see subsections 2.1 — 2.4). In particular, the plural acronyms (e.g. NRs (nuclear receptors)) are matched with the corresponding singular acronym (NR (nuclear receptor)). All acronyms that share a normalised expanded form constitute an acronym cluster.

The experiments and evaluation of the acronym acquisition method are presented in Section 5.

4. Incorporating term variants into ATR

The term variation recognition and acronym acquisition have been embedded in the terminology management workbench ATRACT (Mima, 2001) as part of the term

recognition process. Our approach to the term variation management is based on term normalisation as an integral part of the ATR process. The ATR method is based on the *C/NC-value* method (Frantzi et al., 2000). It is a hybrid approach combining linguistic knowledge (term formation patterns), statistics (frequency of occurrence and frequency of occurrence within other term candidates) and contextual information. The method extracts multi-word terms and is implemented as a three-step procedure. In the first step, tem candidates are extracted by using a set of linguistic filters. The filters describe general term formation patterns and are implemented as unification-like LR(1) rules (Mima et al., 1995). In the second step, the term candidates are assigned the C-values, also referred to as term-hoods, according to formula (1). The contexts of the term candidates are examined in the third step in order to assign context factors to them and to rank the candidates accordingly.

The term variants are handled in the first step of the described procedure, i.e. during the process of the term candidate acquisition. Each term variant occurrence recognised in this step is normalised, and specifically for acronyms this means that their occurrences are mapped to their normalised expanded forms.[2] Term variants having the same normalised form are then grouped into classes in order to link each term candidate to all of its variants. This way, a list of normalised term candidate classes, rather than a list of single terms, is passed as input to the second step. The term-hood is then calculated for the whole class according to the following formula:

$$C-value(a) = \begin{cases} \log_2 |a| \cdot f(a), & a \text{ is not nested,} \\ \log_2 |a| \left(f(a) - \frac{1}{|T_a|} \sum_{b \in T_a} f(b)\right), & \text{otherwise} \end{cases} \quad (1)$$

In the previous formula, a denotes a normalised representative of a class, $f(a)$ corresponds to the cumulative frequency with which all term variants from the class occur in a given corpus, $|a|$ denotes the length of the normalised form (the number of its constituents), and T_a is a set of all classes that contain normalised form a as a nested term. After assigning context factors in the third step, term classes are ranked and the candidates having term-hoods higher than a given threshold are accepted as terms. This approach guarantees that all term variants are naturally dealt with jointly, thus supporting the fact that they denote the same concept.

Once term variants are recognised automatically, they can be managed manually within the ATRACT workbench, as users can add/remove variants or variant occurrences. For example, Figure 1 shows term variants represented in a text, while Figure 2 presents a sample resulting list of terms and their variants, which can be edited by a user.

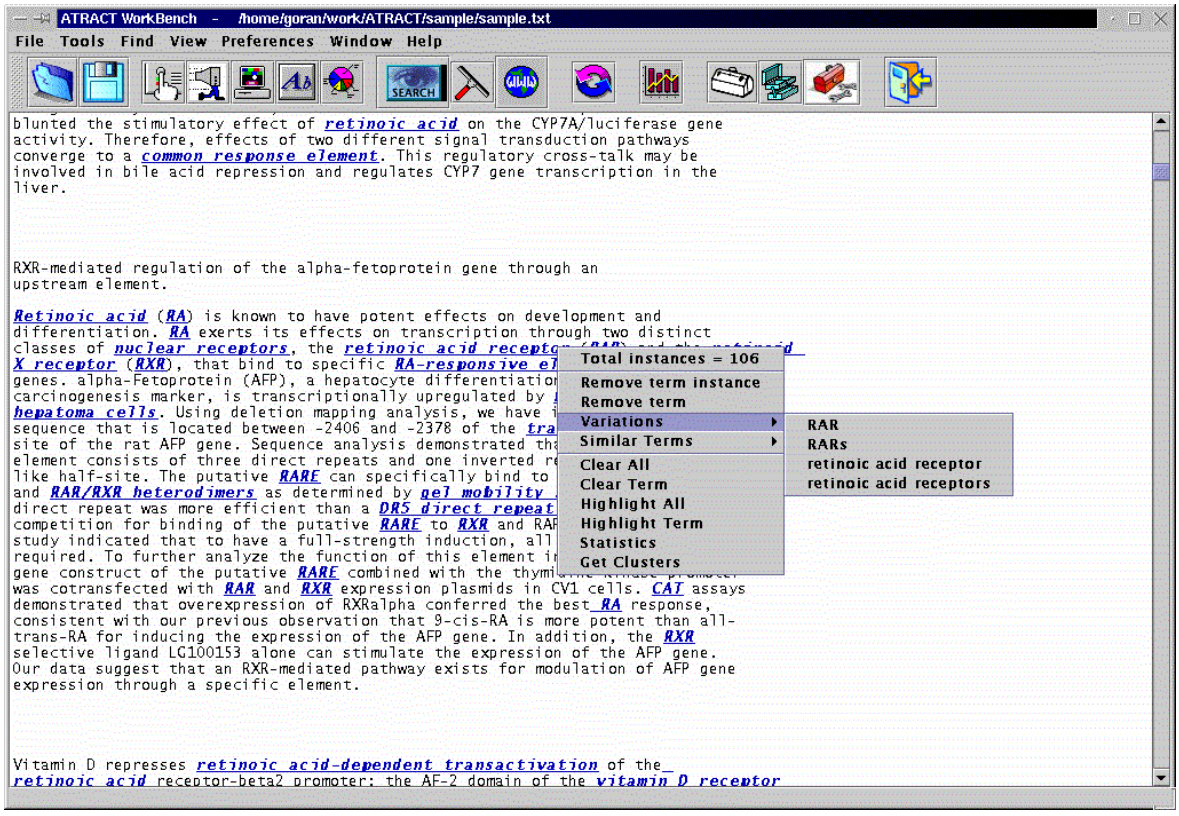

Figure 1: Sample of term variants recognised in the text

[2] Although term occurrences are normalised, all variants (i.e. their surface forms) are recorded in order to facilitate fast retrieval from a corpus.

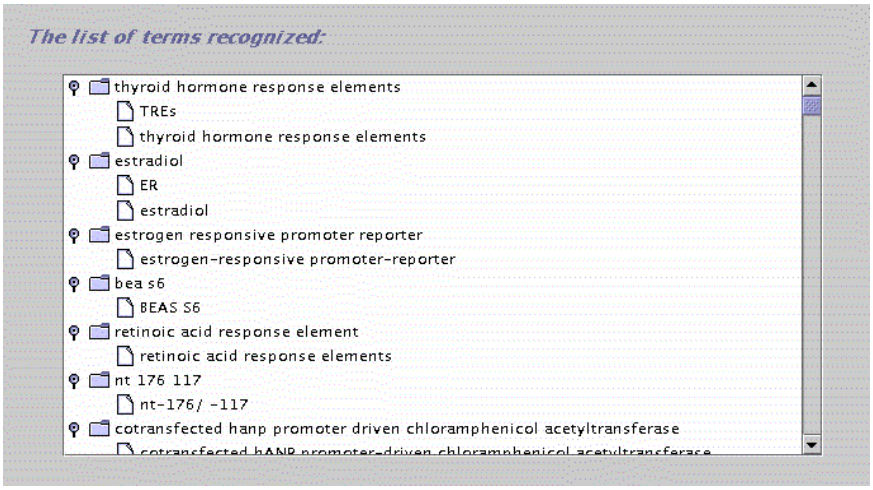

Figure 2: Sample of terms and term variants automatically recognised

5. Experiments and Evaluation

The ATRACT workbench has been used for experiments with acronym acquisition and ATR in the MB domain.

5.1. Experiments with acronym acquisition

The experiments with the acronym acquisition have been conducted on two separate corpora containing 2008 and 6323 MEDLINE abstracts (MEDLINE, 2002), and a random sample of 50 abstracts taken from the first corpus has been used for the evaluation. Table 2 shows some examples of automatically recognised acronyms, and the evaluation is presented in Table 3.

acronym(s)	expanded form(s)
RAR alpha RAR-alpha RARA RARa	retinoic acid receptor alpha
RARs RAR	retinoic acid receptor retinoic acid receptors
RT-PCR	reverse transcription PCR
TR TRs	thyroid hormone receptor thyroid hormone receptors thyroid receptor
9-c-RA 9cRA	9-cis-retinoic acid 9-cis retinoic acid
ES	Ewing sarcoma Ewing's sarcoma Ewings sarcoma

Table 2: Sample of acronyms acquired

The precision of the acronym acquisition method was very high: it ranges from 94% to almost 99% depending on the size of a corpus used. Among the incorrect acronyms, the majority were acronyms acquired from the coordinated acronym patterns. The main source for this was the fact that some of the coordinated structures were not acquired correctly, which means that corresponding patterns have to be more specific.

corpus / acronyms	2008 abstracts	6323 abstracts	50 abstracts
number of (distinct) acronyms recognised	1015	2343	66
number of correct acronyms recognised	992	2314	62
number of acronyms introduced	-	-	85
precision	**97.73%**	**98.76%**	**93.94%**
recall	-	-	**72.94%**

Table 3: Acronym acquisition results

Although the recall level of 73% is respectable, we believe that the recall can be further improved, since additional patterns were identified during the manual evaluation phase.

term (and term variants)	term-hood
COUP-TF II	8.00
retinoic acid receptor retinoic acid receptor retinoic acid receptors RAR, RARs	6.33
nuclear receptor nuclear receptor nuclear receptors NR, NRs	6.00
all-trans retionic acid all trans retionic acid all-trans-retinoic acids ATRA, at-RA, atRA	4.75
nuclear receptor co-repressor nuclear receptor co-repressor NCoR	4.25

Table 4: Sample of recognised terms and their term-hoods

5.2. Experiments with ATR

The ATR experiments with the term variation management incorporated were conducted on the corpus containing 2008 abstracts from Medline (MEDLINE 2002). Table 4 presents a sample of automatically recognised terms and their variants. Figure 3 shows the distribution of the precision for three sets of terms grouped by their term-hoods: precision for the top ranked terms (with a term-hood above 6.00) was 98%. The recall and precision of the method are presented in Figure 4.

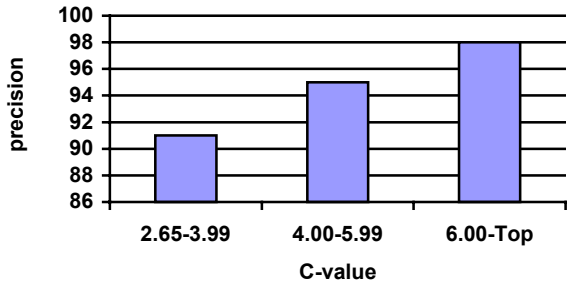

Figure 3: The distribution of precision of the C/NC-value method over three groups of terms

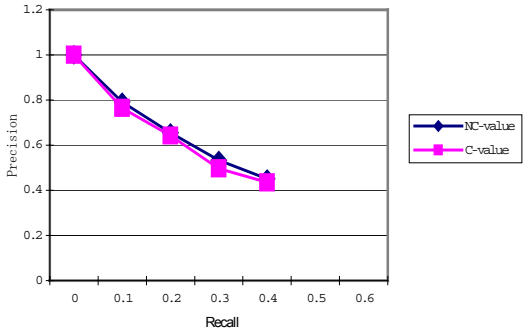

Figure 4: Precision and recall of the C/NC-value method

6. Conclusion

In this paper we have presented a framework for the effective management of terms and their variants automatically acquired from domain-specific texts. In our approach, the term variant recognition has been incorporated into the ATR process by taking into account orthographical, morphological, syntactic, lexico-semantic and pragmatic term variations. All term variants are considered jointly for the calculation of term-hood, that way improving the precision of the ATR method and providing more systematic knowledge acquisition.

We have paid special attention to acronyms as a common way of introducing term variants in scientific papers. A method for the automatic acquisition of newly introduced acronyms and the mapping to their expanded forms has been developed. The problems of acronym variation and acronym ambiguity have been addressed by establishing classes of term variants that correspond to specific concepts. The approach has been tested in the MB domain: for a small corpus, the system achieved 94% precision and 73% recall. For a larger corpus, precision rose to 99%.

The automatic support for handling term variations can be used for semi-automatic update of the existing language resources. For example, the recognised term variants can be used to populate term dictionaries. Term variants unification and normalisation also provides a broader basis for IR and IE tasks, as queries can be expanded by referring to a class of synonymous terms as opposed to a single term. Any other term-centred tasks, such as classification and clustering of terms, can rely on the unified term variants in order to enhance statistically based procedures or to provide the wider context for specific term analysis.

Although our work is placed within the MB domain, the approach can be easily adapted and applied to other domains. The future work on this subject will include a study on derivational term variants, the improvement of recall by covering additional types of acronym definitions, and possibly the application of the method to other domains. In addition to improving the automatic handling of term variations, we also plan to investigate the problems of term sense disambiguation.

Acknowledgment

We would like to thank Dr Hideki Mima (University of Tokyo) and Kostas Manios (University of Salford) for software support, and Dr. Sylvie Albert (LION BioScience) for the evaluation of results.

7. References

Ananiadou, S., 1994. A Methodology for Automatic Term Recognition. *Proceedings of COLING-94*. Kyoto, Japan.

Ananiadou, S., Nenadic, G., Mima, H., 2001. A Terminology Management Workbench for Molecular Biology. In *Proceedings of Workshop on Information Extraction in Molecular Biology*, Twente, The Netherlands.

Bisson, G., Nedellec, C., Canamero, D., 2000. Designing Clustering Methods for Ontology Building - The Mo'K Workbench. In S. Staab, A. Maedche, C. Nedellec, and P. Wiemer Hastings (Eds.): *Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00*, Berlin, Germany.

Bourligault, D., 1992. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In *Proceedings of 14th International Conference on Computational Linguistics*. Nantes, France, pp. 977-981.

Gross M., Perrin D. (Eds.), 1989. Electronic Dictionaries and Automata in Computational Linguistics. *Lecture Notes in Computer Science*. Berlin, Springer Verlag.

Hatzivassiloglou, V., Duboue, P., Rzhetsky, A., 2001. Disambiguating Proteins, Genes, and RNA in Text: a Machine Learning Approach. In *BIOINFORMATICS*, Vol. 17:1, pp. S97-S106.

Frantzi, K. T., Ananiadou, S., Mima, H., 2000. Automatic Recognition of Multi-Word Terms: the C-value/NC-value method. *International Journal on Digital Libraries,* Vol. 3:2, pp.115-130.

Jacquemin, C., 2001. *Spotting and Discovering Terms through NLP*. MIT Press, Cambridge MA.

Lander ES, et al. (International Human Genome Sequencing Consortium), 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822), pp. 860-921.

Maynard, D., Ananiadou, S., 2001. Terms Extraction Using a Similarity-based Approach. In D. Bourigault, C. Jacquemin, and M.C. L'Homme (Eds): *Recent Advances in Computational Terminology*. John Benjamin Publishing Company, Amsterdam, Philadelphia, pp. 262-278.

MBO, 20002. Shulze-Kremer's Ontology for Molecular Biology. Available at: http://igd.rz-berlin.mog.de/~www/oe/mbo.html

MEDLINE, 2002. *National Library of Medicine*. Available at: http://www.ncbi.nlm.nih.gov/PubMed/

Mima, H., Ando, K., Aoe, J., 1995. Incremental Generation of LR(1) Parse Tables. *Proceedings of NLPRS'95*, Pacific-Rim Symposium, Seoul, Korea.

Mima, H., Ananiadou, S., Nenadic, G., 2001. The ATRACT Workbench: Automatic Term Recognition and Clustering for Terms. In *Lecture Notes in Artificial Intelligence*, 2166, Springer-Verlag.

Nakagawa, H., Mori, T., 2000. Nested Collocation and Compound Noun for Term Recognition. In *Proceedings of the First Workshop on Computational Terminology COMPUTERM 98*, pp. 64—70.

Nenadic, G., Spasic, I., Ananiadou, S., 2002. Automatic Discovery of Term Similarities Using Pattern Mining. Submitted.

Spasic, I., Nenadic, G., Ananiadou, S., 2002. Tuning Context Features with Genetic Algorithms. In *Proceedings of LREC-3*, Las Palmas, Spain.

White, J., Maltais, L., Nebert, D., 1998. An Increasingly Urgent Need for Standardized Gene Nomenclature. *Natural Genetics*.