Enhancing automatic term recognition through recognition of variation

Goran Nenadić[*]
Department of Computation
UMIST
Manchester, UK, M60 1QD
G.Nenadic@umist.ac.uk

Sophia Ananiadou[*]
Computer Science
University of Salford
Salford, UK, M5 4WT
S.Ananiadou@salford.ac.uk

John McNaught[*]
Department of Computation
UMIST
Manchester, UK, M60 1QD
J.McNaught@umist.ac.uk

Abstract

Terminological variation is an integral part of the linguistic ability to realise a concept in many ways, but it is typically considered an obstacle to automatic term recognition (ATR) and term management. We present a method that integrates term variation in a hybrid ATR approach, in which term candidates are recognised by a set of linguistic filters and termhood assignment is based on joint frequency of occurrence of all term variants. We evaluate the effectiveness of incorporating specific types of term variation by comparing it to the performance of a baseline method that treats term variants as separate terms. We show that ATR precision is enhanced by considering joint termhoods of all term variants, while recall benefits by the introduction of new candidates through consideration of different variation types. On a biomedical test corpus we show that precision can be increased by 20–70% for the top ranked terms, while recall improves generally by 2–25%.

1 Introduction

Terminological processing has long been recognised as one of the crucial aspects of systematic knowledge acquisition and of many NLP applications (IR, IE, corpus querying, etc.). However, *term variation* has been under-discussed and is rarely accounted for in such applications.

When naming a new concept, scientists and specialists usually follow some predefined term formation patterns, a process which does not exclude the usage of term variations or alternative names for concepts. Term variations are very frequent: approximately one third of term occurrences are variants (Jacquemin, 2001). They occur not only in text, but also in controlled, manually curated terminological resources (e.g. UMLS (NLM, 2004)).

The task of an automatic term recognition (ATR) system is not only to suggest the most likely candidate terms from text, but also to correlate them with synonymous term variants. In this paper, we briefly present an analysis of term variation phenomena, whose results are subsequently incorporated into a corpus-based ATR method in order to enhance its performance.

The paper is organised as follows. In Section 2, we analyse the main types of term variation, and briefly examine how existing ATR systems treat them. Our approach to incorporating variants into ATR is presented in Section 3. In Section 4, we evaluate our approach by comparing it to a baseline method (the method without variation recognition), and we conclude the paper in Section 5.

2 Background

Terms are linguistic units that are assigned to concepts and used by domain specialists to describe and refer to specific concepts in a domain. In this sense, terms are preferred designators of concepts. In text, however, concepts are frequently denoted by different surface realisations of preferred terms, which we denote as their *term variants*. Consequently, a concept can be linguistically represented using any of the surface forms that are variants of the corresponding preferred term. We consider the following types of term variation:

(i) orthographic: e.g. usage of hyphens and slashes (*amino acid* and *amino-acid*), lower and upper cases (*NF-KB* and *NF-kb*), spelling variations (*tumour* and *tumor*), different Latin/Greek transcriptions (*oestrogen* and *estrogen*), etc.

(ii) morphological: the simplest variations are related to inflectional phenomena (e.g. singular, plural). Derivational transformations can lead to variants in some cases (*cellular gene* and *cell gene*), but not always (*activated factor* vs. *activating factor*);

(iii) lexical: genuine lexical synonyms, which may be interchangeably used (*carcinoma* and *cancer, haemorrhage* and *blood loss*);

(iv) structural: e.g. possessive usage of nouns using prepositions (*clones of human* and *human clones*), prepositional variants (*cell in blood*, *cell from blood*), term coordinations (*adrenal glands and gonads*);

---

[*] Co-affiliation: National Centre for Text Mining, Manchester, UK

(v) acronyms and abbreviations: very frequent term variation phenomena in technical sublanguages, especially in biomedicine; sometimes they may be even preferred terms (*DNA* for *deoxyribonucleic acid*).

Note that variation types *(i)* – *(iii)* affect individual constituents, while *(iv)* and *(v)* involve variation in structure of the preferred term. In any case, they do not "change" the meaning as they refer to the same concept. Daille et al. (1996) and Jacquemin (1999, 2001) further identified types of variation that modified the meaning of terms.

Although many authors mention the problems related to term variation, few have dealt with linking the corresponding term variants. Also, the recognition of variants is typically performed as a separate operation, and not as part of ATR.

The simplest technique to handle some types of term variation (e.g. morphological) is based on stemming: if two term forms share a stemmed representation, they are considered as mutual variants (Jacquemin and Tzoukermann, 1999; Ananiadou et al., 2000). However, stemming may result in ambiguous denotations related to "over-stemming" (i.e. resulting in the conflation of terms which are not real variants) and "under-stemming" (i.e. resulting in the failure to link real term variants).

Other approaches to the recognition of term variants use preferred terms and known synonyms from existing term dictionaries and approximate string matching techniques to link or generate different term variants (Krauthammer et al., 2001; Tsuruoka and Tsujii, 2003).

Jacquemin (2001) presents a rule-based system, FASTR, which supports several hundred meta-rules dealing with morphological, syntactic (i.e. structural) and semantic term variation. Term variation recognition is based on the transformation of basic term structures into variant structures. However, the variants recognised by FASTR are more conceptual variants than terminological ones, as non-terminological units (such as verb phrases, extended insertions, etc.) are also linked to terms in order to improve indexing and retrieval.

3   Incorporating term variation into ATR

Our approach to ATR combines the C-value method (Frantzi et al., 2000) with the recognition of term variation, which is incorporated as an integral part of the term extraction process.

C-value is a hybrid approach combining term formation patterns with corpus-based statistical measures. Term formation patterns act as linguistic filters to a POS tagged corpus: filtered sequences are considered as potential realisations of domain concepts (*term candidates*). They are subsequently assigned *termhoods* (i.e. likelihood to represent terms) according to a statistical measure. The measure amalgamates four corpus-based characteristics of a term candidate, namely its frequency of occurrence, its frequency of occurrence as a form nested within other candidate terms, the number of candidate terms inside which it is nested, and the number of words it contains.

The original C-value method treats term variants that correspond to the same concept as separate term candidates. Consequently, by providing separate frequencies of occurrence for individual variants instead of a single frequency of occurrence calculated for a term candidate unifying all variants, the corpus-based measures and termhoods are distributed across different variants. Therefore, we aim at enhancing the statistical evaluation of termhoods through conflation of different surface representations of a given term, and through joint frequencies of occurrence of all equivalent surface forms that correspond to a single concept.

In order to conflate equivalent surface expressions, we carry out linguistic *normalisation* of individual term candidates (see examples in Table 1). Firstly, each term candidate is mapped to a specific *canonical representative* (CR) by semantically isomorphic transformations. Then, we establish an equivalence relation, where two term candidates are related *iff* they share the same CR. The partitions of this relation are denoted as *synterms*: a synterm contains surface term representations sharing the same CR.

synterm	canonical representative
human cancers *cancer in humans* *human's cancer* *human carcinoma* }	*human cancer*

Table 1: Term normalisation examples

Our aim is to form synterms prior to the syntactic estimation of termhoods for term candidates. Therefore, after the extraction of individual term candidates, we subsequently normalise them in order to generate synterms, where the normalisation is performed according to the typology of variations described in Section 2. More precisely, we consider separately the normalisation of variations that affect term candidate constituents and variations that involve structural changes. The general architecture of our ATR approach is presented in Figure 1.
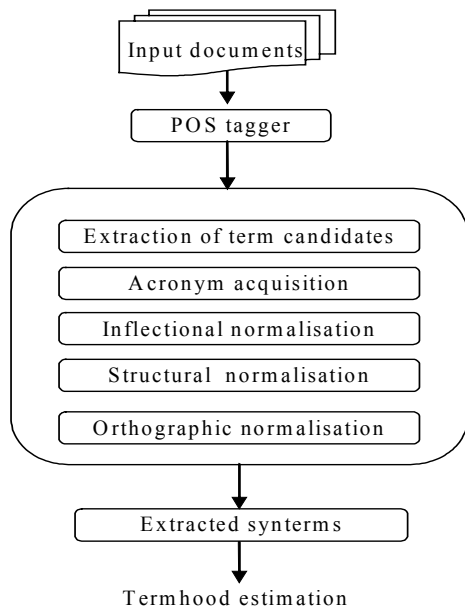
Figure 1: The architecture of the ATR process

3.1 Normalising term constituent variation

In the case of variations that do not affect the structure of terms, the formation of CRs is based on a POS tagger (for inflectional variation) and simple heuristics (for orthographic normalisation). For example, different transcriptions of neoclassical combining forms are treated by replacements of specific character combinations ($ae \rightarrow e$, $ph \rightarrow f$) in such forms (and only in such forms). Inflectional normalisation is based on POS tagging: a canonical term candidate form is a singular form containing no possessives (*Down's syndrome* $\rightarrow$ *down syndrome*).

In order to address lexical variants, one can use dictionaries of synonyms where the preferred terms are used for normalisation purposes ({*hepatic microsomes, liver microsomes*} $\rightarrow$ *liver microsomes*). In experiments reported here, we did not attempt to normalise lexical variation.

3.2 Normalising term structure variation

Variations affecting term structure are less frequent but more complex. Here we consider two types of term variation: prepositional term candidates and coordinated term candidates (for a detailed analysis of these variations see (Nenadic et al., 2004)).

Prepositional term candidates are normalised by transformation into corresponding expressions without prepositions. Using prepositions *of, in, for* and *by* as anchors, we generate semantically isomorphic CRs by inversion. For example, the candidate *nuclear factor of activated T cell* is transformed into *activated T cell nuclear factor*. Here is a simplified example of a rule describing the transformation of a term candidate that contains the preposition *of*:

> **if** structure of *term candidate* **is**
> $(A|N)_1$* N_1 Prep(**of**) $(A|N)_2$* N_2
> **then** CR = $(A|N)_2$* N_2 $(A|N)_1$* N_1

In order to address the problems of determining the boundaries of term constituents in text (to the right and left of prepositions), for each prepositional term candidate we generate all possible nested candidates[†] and their corresponding CRs. For example, for the candidate *regulation of gene expression*, we generate both *gene regulation* and *gene expression regulation*. Since this approach also generates a number of false candidates, additional heuristics are used to enhance precision, such as removing adverbials and determiners, using a stop list of terminologically irrelevant prepositional expressions (e.g. *number of …, list of …, case of …, in presence of …*), etc.

A similar approach is used for the recognition of coordinated term candidates: coordinating conjunctions (*and, or, but not, as well as*, etc.) are used as anchors, and when a coordinating structure is recognised in text, the corresponding CRs of the candidate terms involved are generated.

We differentiate between head coordination (where term heads are coordinated, e.g. *adrenal glands and gonads*) and argument coordination (where term arguments/modifiers are coordinated, e.g. *SMRT and Trip-1 mRNAs*).

The recognition and extraction of coordinated terms is highly ambiguous even for human specialists, since coordinated terms and term conjunctions share the same structures (see Table 2). Also, similar patterns cover both argument and head coordinations, which makes it difficult to extract coordinated constituents (i.e. terms). Not only is the recognition of term coordinations and their subtypes ambiguous, but also internal boundaries of coordinated terms are blurred. In a separate study, we have shown that morphosyntactic features are insufficient both for the successful recognition of coordinations and for the extraction of coordinated terms: in many cases, the correct interpretation and decoding of term coordinations is only possible with sufficient background knowledge (Nenadic et al., 2004).

[†] Each constituent extracted from a nested prepositional term candidate has to follow a pattern used for the extraction of individual candidate terms.

example	*adrenal glands <u>and</u> gonads*
head coordination	[*adrenal* [*glands <u>and</u> gonads*]]
term conjunction	[*adrenal glands*] <u>*and*</u> [*gonads*]

Table 2: Ambiguities within coordinated structures

In order to address the problems of structural ambiguities and boundaries of coordinated terms, we also generate all possible nested coordination expressions and corresponding term candidates. For example, from a candidate coordination *viral gene expression and replication* we generate two pairs of coordinated term candidates:

viral gene expression and *viral gene replication*
viral gene expression and *viral replication*

Patterns for the extraction of term candidates from coordinations have been acquired semi-manually for a subset of term coordinations. For each pattern, we define a procedure for the extraction of coordinated term candidates and generation of the corresponding CRs (see Table 3 for examples). The generated candidates from coordinated structures are subsequently treated as individual term candidates.

3.3 Normalising acronym variation

We treat acronym extraction as part of the ATR process (see Figure 1). In (Nenadic et al., 2002) we suggested a simple procedure for acquiring acronyms and their expanded forms (EFs), which was mainly based on using orthographic and syntactic features of contexts where acronyms were introduced. The model is based on three types of patterns: *acronym* patterns (defining common internal acronym structures and forms), *definition* patterns (based on syntactic patterns which describe typical contexts where acronyms are introduced in text) and *matching* patterns (the set of matching rules between acronyms and their corresponding EFs).

Acronyms also exhibit variation (e.g. *RAR alpha, RAR-alpha, RARA, RARa, RA receptor alpha* etc. are all acronyms for *retinoic acid receptor alpha*). Therefore, in addition to extracting acronyms, we further gather all acronym variants and their EFs, and we map them into a single CR. Since in this paper acronyms are taken as term variants, we "replace" acronym occurrences by the CR of their EFs. In order to bypass the problem of acronym ambiguity, we replace/normalise only acronyms that are introduced in a given document.

$(N|A)_1$ **&** $(N|A)_2$ $(N^+)_3$
candidate$_1$ = $(N|A)_2$ $(N^+)_3$
candidate$_2$ = $(N|A)_1$ *nested*(N^+_3)

e.g. *B and T cell antigen*
     T cell antigen
     B cell antigen, B antigen

N_1 **&** N_2 A_3 N^+_4
candidate$_1$ = N_2 A_3 N^+_4
candidate$_2$ = N_1 A_3 N^+_4

e.g. *function or surface antigenic profile*
     surface antigenic profile
     function antigenic profile

N^+_1 N_2 **&** $(N|A)_3$
candidate$_1$ = N^+_1 N_2
candidate$_2$ = *nested*(N^+_1) $(N|A)_3$

e.g. *breast cancer therapy and prevention*
     breast cancer therapy
     breast caner prevention, breast prevention

N^+_1 $(A^+)_2$ A_3 **&** A_4
candidate$_1$ = N^+_1 $(A^+)_2$ A_3
candidate$_2$ = N^+_1 $(A^+)_2$ A_4

e.g. *RNA polymerases II and III*
     RNA polymerasis II
     RNA polymerasis III

Table 3: Examples of patterns used for the extraction of term candidates from coordinations (*nested* denotes the generation of all possible linearly nested substrings)

3.4 Calculating termhoods with variants

Term variants sharing the same CR are grouped together into synterms, and the calculation of C-values (i.e. termhoods) is performed for the whole synterm rather than for individual term candidates. The main reason for doing this is to avoid the distribution of frequencies of occurrence of term candidates across different variants, as these frequencies have a significant impact on estimating termhoods. Instead of providing separate frequencies of occurrence and obtaining termhoods for individual term candidates, we provide a single frequency of occurrence and joint termhood calculated for a synterm, which unifies all variants.

Similarly to the estimation of C-values for individual term candidates (Frantzi et al., 2000), the formula for calculating the termhoods for synterms is as follows:

$$C\text{-}value(c) = \begin{cases} \log_2 |CR| \cdot (f(CR) - \dfrac{1}{|T_{CR}|} \sum_{b \in T_{CR}} f(b)), & CR \text{ is nested} \\ \log_2 |CR| \cdot f(CR), & CR \text{ is not nested} \end{cases}$$

where *c* denotes a synterm whose elements share a canonical representative (denoted as *CR* in the formula), *f(CR)* corresponds to the cumulative frequency with which all term candidates from the synterm *c* occur in a given corpus, |*CR*| denotes the average length of the term candidates (the number of constituents), and T_{CR} is a set of all synterms whose CRs contain the given *CR* as a nested substring.

This approach ensures that all term variants are naturally dealt with jointly, thus supporting the fact that they denote the same concept. As a consequence, we expect that precision would be enhanced by considering joint frequencies of occurrence and termhoods for all variants of candidate terms, while recall would benefit by the introduction of new candidates through consideration of different variation types.

4   Evaluation and discussion

In order to assess the effectiveness of incorporating specific types of term variation into ATR, we compared the performance of the baseline C-value method (without considering variations) with the approach including recognition and conflation of term variants. Here we are not interested in an absolute measure of the ATR performance, but rather in the comparison of results obtained through handling different variation types.

We conducted two sets of experiments: in the first experiment, we analysed the incorporation of term candidates resulting from considering term variations individually, while, in the second, we experimented with the integration of combined variations in the ATR process.

The evaluation was carried out using the GENIA corpus (GENIA, 2004), which contains 2,000 abstracts in the biomedical domain with 76,592 manually marked occurrences of terms. These occurrences (which include different term variants) correspond to 29,781 different, unique terms. Each occurrence of a term in the corpus (except occurrences of acronyms) is linked to the corresponding "normalised" term (typically a singular form), while coordinated terms are identified, marked and normalised within term coordinations. A third of occurrences of GENIA terms are affected by inflectional variations, and almost half of GENIA terms have inflectional variants appearing in the corpus. On the other hand, only 0.5% of terms contain a preposition, while 2% of all term occurrences are coordinated, involving 9% of distinct GENIA terms (for a detailed analysis of GENIA terms see (Nenadic et al., 2004)).

We used the list of GENIA terms as a gold standard for the evaluation. Since our ATR method produces a ranked list of suggested synterms, we considered precision at fixed rank cut-offs (intervals): precision was calculated as the ratio between the number of correctly recognised terms and the total number of entities recognised in a given interval (where an interval included all terms from the top ranked synterms).‡ The baseline method (original C-value) was treated in the same way, as term candidates suggested by the original C-value could be seen as singleton synterms. In order to estimate the influence on recall, we also used all variants from suggested synterms.

The incorporation of individual variations affecting term constituents into ATR had considerable positive effects, especially on the most frequently occurring terms (see Figures 2a and 2b): for some intervals, inflectional variants, for example, improved precision by almost 50%. Similarly, the integration of acronyms improved precision, in particular for frequent terms (up to 70%), as acronyms are typically introduced for such terms. As one would expect, the combined constituent-level variations further improved interval precisions compared both to the baseline method and individual variations (see Figure 2c).

However, the incorporation of structural variants (in particular for prepositional terms) negatively influenced precision compared to the baseline method, as many false candidates were introduced.

In order to assess the quality of extracted prepositional term candidates, we evaluated a set of the 117 most frequently occurring candidates with prepositions: 80% of suggested expressions were deemed relevant by domain experts, although they were not included in the gold GENIA standard (such as *expression of genes* or *binding of NF kappa B*). Still, the recognition of prepositional term candidates is difficult as they are infrequent and there are no clear morphosyntactic cues that can differentiate between terminologically relevant and irrelevant prepositional phrases.

The incorporation of coordinated term candidates had only marginal influence on precision, mainly because they were not frequent in the GENIA corpus. Furthermore, simple term conjunctions

---

‡ It was an open question whether to count the recognition of each term form (e.g. singular and plural forms, an acronym and its EF, prepositional and non-prepositional forms) separately (i.e. as two positive "hits") or as one positive "hit" (see also (Church, 1995)). Since the evaluation of the baseline method (original C-value) typically counts such hits separately, we decided to follow this approach, and consequently count all positive hits from synterms.

were far more frequent than term coordinations, which made their extraction highly ambiguous. Still, using only the patterns from Table 3, we have correctly extracted 35.76% of all GENIA coordinated terms, with more than a half of all suggested candidates being found among those that appeared exclusively in coordinations. However, these patterns also generated a number of false coordination expressions, and consequently a number of false term candidates.

The integration of term variants was also useful for re-ranking of true positive term candidates: the combined rank was typically higher than the separate ranks of term variants. Furthermore, some terms, not suggested by the baseline method at all, were ranked highly when variants were conflated (for example, the term *T-lymphocyte* was recognised only as a coordinated term candidate, while *replication of HIV-1* was extracted only by considering prepositional term candidates). In order to estimate the overall influence on recall of ATR, we used all terms from the respective synterms (see Table 4 for the detailed results). In general, the incorporation of inflectional variants increased recall by ¼, while acronyms improved recall by almost ⅔ when only the most frequent terms were considered. It is interesting that acronym acquisition can further improve recall by extracting variants that have more complex internal structures (such as EFs containing prepositions (*REA = repressor of estrogen activity*) and/or coordinations (*SMRT = silencing mediator of retinoic and thyroid receptor*)). Prepositional and coordination candidate terms had some influence on recall, in particular as they increased the likelihood of some candidates to be suggested as terms. Low recall of term coordinations may be increased by adding more patterns (which would probably negatively affect precision).

Summarising, experiments performed on the GENIA corpus have shown that the incorporation of term variations into the ATR process resulted in significantly better precision and recall. In general, acronyms and inflectional unification are the most important variation types (at least in the domain of biomedicine). Individually, they increased precision by 20–70% for the top ranked synterm intervals, while recall is generally improved, in some cases up to 25%. Other term variations had only marginal influence on the performance, mainly because they were infrequent in the test corpus (compared to the total number of term occurrences, and not only with regard to specific individual candidates, but also in general). For these variations, larger-scale corpora may show their stronger influence.
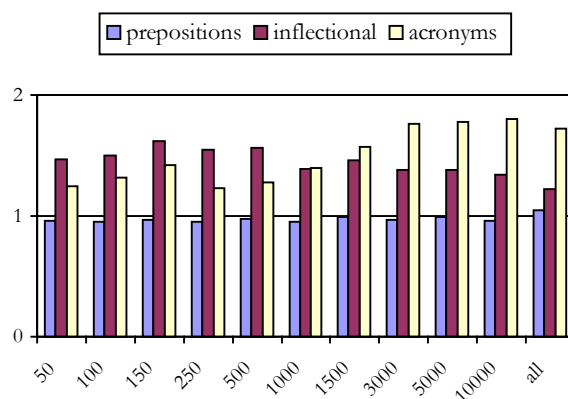

Figure 2a: Comparison of interval ATR precision of the baseline method (=1) to ATR precisions with integrated recognition of individual term variants (terms with frequency > 5)
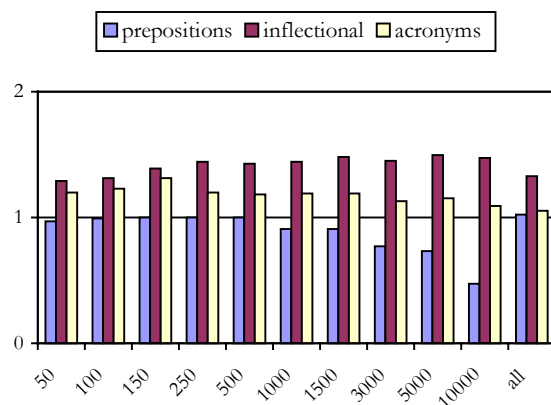

Figure 2b: Comparison of interval ATR precision of the baseline method (=1) to ATR precisions with integrated recognition of individual term variants (terms with frequency > 0)
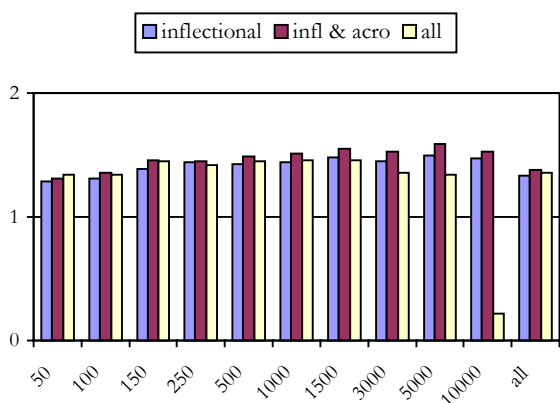

Figure 2c: Comparison of interval ATR precision of the baseline method (=1) to ATR precisions with integrated recognition of combined term variants (terms with frequency > 0)

term sets	prep.	coord.	infl.	acro.
freq. ≥ 5	+5.30%	+12.42%	+17.52%	+60.49%
freq. > 0	+2.36%	+2.53%	+25.25%	+8.52%

Table 4: Improvement in recall when variations are considered as an integral part of ATR

5   Conclusion

In this paper we discussed possibilities for the extraction and conflation of different types of variation of term candidates. We demonstrated that the incorporation of treatment of term variation enhanced the performance of an ATR system, and that tackling term variation phenomena was an essential step for ATR. In our case, precision was boosted by considering joint frequencies of occurrence and termhoods for all candidate terms from candidate synterms, while recall benefited from the introduction of new candidates through consideration of different variation types. Although we experimented with a biomedical corpus, our techniques are general and can be applied to other domains.

Variations affecting single term candidate constituents are the most frequent phenomena, and also straightforward for implementation as part of an ATR process. The conflation of such term candidate variants can be further tuned for a specific domain by using lists of combining forms and affixes. The incorporation of acronyms had a significant high positive effect, in particular on more frequent terms (since acronyms are introduced for terms that are used more frequently).

However, more complex structural phenomena had a moderate positive influence on recall, but, in general, the negative effect on precision. The main reason for such performances is structural and terminological ambiguity of these expressions, in addition to their low frequency of occurrence (compared to the total number of term occurrences). For handling such complex variants, a knowledge-intensive and domain-specific approach is needed, as coordinated term candidates or candidates with prepositions need to be additionally semantically analysed in order to suggest more reliable term candidates, and to introduce fewer false candidates.

Apart from being useful for boosting precision and recall, the integration of term variation into ATR is particularly important for smaller corpora (where linking related occurrences is vital for successful terminology management) as well as for many text-mining tasks (such as IR, IE, term or document clustering and classification, etc.).

Finally, as future work, we plan to investigate more knowledge intensive, domain-specific treatment of prepositional and coordinated terms, as well as pronominal term references.

6   Acknowledgements

References

S. Ananiadou, S. Albert and D. Schuhmann. 2000. Evaluation of Automatic Term Recognition of Nuclear Receptors from Medline. *Genome Informatics Series*, vol. 11.

K.W. Church. 1995. One Term or Two? Proc. of SIGIR-95, pp. 310-318.

B. Daille, B. Habert and C. Jacquemin. 1996. Empirical Observation of Term Variation and Principles for Their Description. *Terminology* 3(2), pp. 197–258.

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

C. Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. Proc. of 37th Annual Meeting of ACL, pp. 341–348.

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

C. Jacquemin and E. Tzoukermann. 1999. NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon and Syntax, in T. Strzalkowski (ed.), Natural Language Information Retrieval, Kluwer, pp. 25-74

M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman. 2001. Using BLAST for identifying gene and protein names in journal articles. *Gene*, 259(1–2): pp. 245–52.

GENIA. 2004. GENIA resources. Available at http://www.tsujii.is.u-tokyo.ac.jp/~Genia/

G. Nenadic, I. Spasic and S. Ananiadou. 2002. Automatic Acronym Acquisition and Term Variation Management within Domain-Specific Texts. Proc. of LREC 2002, pp. 2155–2162.

G. Nenadic, I. Spasic and S. Ananiadou. 2004. Mining Biomedical Abstracts: What's in a Term? Proc. of IJC-NLP, pp. 247-254.

NLM. 2004. National Library of Medicine, Unified Medical Language System.

Y. Tsuruoka and J. Tsujii. 2003. Probabilistic Term Variant Generator for Biomedical Terms. Proc. of 26th Annual ACM SIGIR Conference.