Terminological Acquaintance: the Importance of Contextual Information in Terminology

Diana Maynard
Dept. of Computer Science
University of Sheffield
Regent Court, 211 Portobello St
Sheffield, S1 4DP, UK
d.maynard@dcs.shef.ac.uk

Sophia Ananiadou
Computer Science, School of Sciences
University of Salford, Newton Building
Salford, M5 4WT, U.K.
s.ananiadou@salford.ac.uk

Abstract

This paper examines an idea we call **terminological acquaintance**, which considers the importance of contextual information for various applications in NLP. The importance of contextual information is not new to NLP, but it has rarely been considered in relation to terminological entities. Terminological contexts, however, are vast information sources waiting to be tapped, and are even more useful for domain-specific applications than for general language ones. We focus specifically on the relationships between these contexts and their neighbouring terms, and on how the relevant information from them can best be captured. We give examples of how such information can be clustered, and how this can be used for applications such as domain-specific lexical ontology development.

1 Introduction

Research in terminology has tended to focus largely on statistical information or, at most, very shallow sources of linguistic information. For some time, a gulf has existed between the fields of terminology and NLP, perhaps largely because work in terminology has tended to be of a very practical nature. At the other end of the scale, terminological work is rarely seen as a tool which can be used for NLP applications, but rather as a technology beneficial for fields such as Information Retrieval (Boguraev and Kennedy, 1999). Terminological research in areas such as medical informatics has been somewhat isolated from mainstream NLP. In the area of molecular biology, the importance of terminological information has been recognised only recently with the exponential growth of research literature. The integration of stronger links between all these related fields is beneficial not only for NLP applications, but for domain-specific applications requiring sublanguage textual information.

1.1 Motivation and applications

With the overwhelming increase of medical and biological textual information, there is a need to assist domain experts in gathering and classifying this information. In biology, scientific databases are valuable sources of domain information, but they suffer from terminological ambiguity. The same concept may be given different names in different databases, making it difficult for experts to access information in a consistent way (Andrade and Valencia, 1998; Schulze-Kremer, 1998; Nobata et al., 1999). Adding knowledge from texts is important for updating these databases. Nevertheless, this might only add to the existing terminological confusion. Extracting terminology automatically is only the first step towards the maintenance and updating of domain resources. To be of use, the terminology extracted must also be as unambiguous as possible.

Terminological ambiguity was not considered an issue in the past, as it was assumed that there was monoreferentiality between concept and term. Clearly, this is not the case in many multidisciplinary fields. The field of biology is especially susceptible to ambiguity because it is multi-faceted and dynamic, comprising a range of subfields varying widely in specificity. Also, since knowledge may be collated from such a wide area, problems of multi-dimensionality are inherent. Applications such as information extraction and semantic annotation of texts are some of the areas which can benefit from non-ambiguous terminology. We propose a technique (TRUCKS) which extracts terminology automatically and, by incorporating contextual information, aims at terminological disambiguation. We also propose some extensions to this work in the form of contextual clustering, which can aid in the knowledge acquisition, term disambiguation and ontology building.

2 Use of contextual information

2.1 Contextual information in general language

Contextual information has frequently been used in general language NLP applications; in particular, for word sense disambiguation, e.g. (Yarowsky, 1995; Ng and Lee, 1996). Traditional WSD methods compare the distribution of context words with typical distributions of word senses, to find the closest match (Gale et al., 1992). Contextual information is also used for purposes such as synonym identification (Grefenstette, 1994). However, automatic methods of extracting contextual information, and in particular semantic information, have not been so widespread in terminology.

2.2 Contextual information in terminology

Approaches to term extraction have largely focused on statistical methods, e.g. (Damerau, 1993; Daille et al., 1994; Cohen, 1995; Frantzi and Ananiadou, 1996), though occasionally these have some linguistic contribution (Justeson and Katz, 1995; Frantzi and Ananiadou, 1999), or are even entirely linguistic (Ananiadou, 1994). All these methods mainly make use of information about the term, except for the NC-Value (Frantzi and Ananiadou, 1999), which also considers contextual information. However, this only takes into account statistical information about frequency of occurrence of context words with candidate terms, and does not exploit any linguistic relationship between term and context. The TRUCKS system (Maynard and Ananiadou, 1999a) makes use of much deeper contextual information, considering statistical, terminological and semantic information about contexts in the field of medicine. In particular, it calculates the relative importance of different parts of the context to the term, and similarity between the term and relevant contextual information.

Extracting contextual information is a vital element of terminography, but has until recently been largely restricted to manual efforts. Some linguistic work has been carried out in this area, e.g. (Davidson et al., 1998; Meyer et al., 1999). In these cases, linguistic information about the context is used to help define the meaning of terms or to construct thesaural relations.

2.3 Frame-based acquisition of semantic information

The idea of acquiring semantic information from texts using some kind of information formatting dates back to the late 1960s with Harris' *distributional hypothesis* (Harris, 1968) and Hirschman and Sager's work in the 1970s (Sager, 1972), (Hirschman et al., 1975), which focused on determining sets of sublanguage-specific word classes using syntactic patterns from domain-specific corpora. By examining sets of lexical items found in specific syntactic environments, Hirschman and Sager established semantic word classes in the field of clinical reporting.

More recent work in the medical field has focused on using domain-specific semantic information to create frames or templates, which can then be used to structure the information found in patient records and other similar textual resources. One example of this is the *proximity processing* method (Morel-Guillllemaz et al., 1990), (Baud et al., 1992). The idea of this is to make use of pre-defined conceptual knowledge about the domain without the need for any formal parsing or grammars, although syntactic information can be used to help disambiguation if available.

Do Amaral and Satomura Do Amaral and Satomura (1995) used domain-specific semantic information from SNOMED (a medical thesaurus), in conjunction with a formal grammar, to process, structure and represent medical information in a language-independent frame. They claim that this association of medical and linguistic information is both necessary and adequate to explain the relationship between concepts and to represent the structure of the phrase. Their method involves allocating semantic classes from SNOMED to instances of terms, in conjunction with syntactic classes for verbs and various other words, to form a pattern consisting of a combination of the two. They identified the following 6 basic patterns as occurring most commonly:

```
<PROCEDURE> of the <TOPOGRAPHY> verb
  <RESULT>
<PROCEDURE> verb <RESULT>
<TOPOGRAPHY> <PROCEDURE> verb
  <RESULT>
<PROCEDURE> of the <TOPOGRAPHY> and
  <TOPOGRAPHY> verb <RESULT>
<TYPE> <PROCEDURE> verb <RESULT>
<TYPE> <PROCEDURE> of the <TOPOGRAPHY>
  verb <RESULT>
```

They then produced language-independent frames for sentences from case reports. An example frame for the sentence *An ultrasonography of the abdomen revealed hepatomegaly* would be represented (in a very simplified format) by:

```
[PROCEDURE: ultrasonography
 | of TOPOGRAPHY: abdomen
  | verb RESULT: heptomegaly]
```

Most of the work in the medical field involves creating semantic and/or syntactic frames from texts such as patient records, in an attempt to create some kind of neutral concept representation system - often for use in translation either between different languages (Do Amaral and Satomura, 1995) or between different coding systems (Rocha et al., 1994). The aim is therefore centred on identifying how the concepts in a sentence are related, and on forming a representation of the whole sentence. Any relationship between sentences tends therefore to be ignored, so that, for example, it is unimportant whether a term that occurs in two different sentences has the same meaning in each case, or whether the same concept is described in different ways in different sentences. Our approach is somewhat different in two ways. Firstly we focus on the contexts surrounding a term rather than the whole sentence. Although these contexts inevitably contain other terms, the point is that we are trying to compare contexts independently of the terms with which they appear, rather than in conjunction with them. Secondly, our aim is not to provide some kind of interlingual representation for sentences (or contexts), but rather to provide a representation which permits us to make comparisons about contexts. This means that we can make generalisations to a much greater extent. The similarity lies in the fact that in both cases, the aim is to provide a representation that will appropriately capture the meaning of the text.

3 Context Clustering

Clustering mechanisms are important for many NLP applications such as knowledge acquisition and machine translation, but they are also useful for terminological tasks such as term extraction and term sense disambiguation, as for example in the TRUCKS system. Clustering of **terms** is clearly useful, since it provides information about unknown terms, and can be used to help create termbanks, thesauri and other resources. The clustering of **contexts**, however, is also an important task, in that a cluster of contexts can provide more information than a single context. Context clustering can be used as a bootstrapping method for many NLP and terminology applications, such as disambiguation, knowledge acquisition, and ontology tuning.

The approach we use for context clustering is based on the creation of semantic and syntactic frames. Instead of using lexical pattern matching, which generally results in high precision but low recall, we first collect syntactic patterns from the corpus, and then, after applying a normalisation process, map these patterns to semantic frames using information from a domain-specific thesaurus, the UMLS (NLM, 1997). This method caters for the mutual interchangeability of both syntax and semantics, by allowing unknown words to be replaced both syntactically and semantically with equivalent words from the cluster.

3.1 Syntactic Information

Using a set of eye pathology records, candidate terms and their relevant contexts are extracted from the text using the SNC-Value method (Maynard and Ananiadou, 1999b). The text is then segmented and analysed using a set of analysis tools (Joshi and Srinivas, 1994). Parts of speech and more detailed syntactic information, including dependency relations, are allocated to chunks of text. Each chunk is allocated a number according to its position in the sentence, and the following three types of dependency are marked:

1. modifiers are annotated with “*”.
2. complements are annotated with “.”

Syntax	Example
NP-V(act)	fibrous_proliferation has_occurred
NP-V(pas)	fairly_extensive_haemorrhages are_seen
V(act)-NP	show follicular_arrangement
V(pas)-NP	can_be_identified peripheral_piece
NP-P-NP-V(act)	capillary_proliferation around pupil causing
NP-P-NP-V(pas)	dermis around tumour is_densely_infiltrated
V(act)-NP-P-NP	have_reached middle of sclera

Table 1: Syntactic patterns

3. uninstantiated dependencies are annotated with "?".

.

The annotations may appear a little strange linguistically, but they are reasonably consistent, and enable us to make suitable distinctions about the relations between chunks. For example, the sentence *"A substantial piece of the peripheral part of descemet's membrane and the corneo-scleral portion of the trabecular meshwork are seen in profile"* is segmented and tagged as follows:

```
0 a_substantial_piece A_NXN
1 of B_nxPnx  0*  4.
2 the_peripheral_part A_nx0N1  0.
3 of B_nxPnx  2*  4.
4 descemet's_membrane A_NXN
5 and B_nx1CONJnx2  4*  8.
6 the_corneo-scleral_portion A_NXN
7 of B_nxPnx  6*  8.
8 the_trabecular_meshwork A_NXN
9 are_seen A_nx0V  ??.
10 in B_vxPnx  9*  11.
11 profile A_NXN
12 . B_sPU  9*
...EOS...
```

where
A_NXN is a noun phrase as an argument,
A_nx0N1 is a nominal predicate,
B_nxPnx is a noun phrase modifying preposition,
B_vxPnx is a verb phrase modifying preposition,
A_nx0V is an intransitive verb[1],
B_nx1CONJnx2 is a noun phrase conjunction,
B_sPU is a sentence final punctuation mark.

[1]One quirk of the Supertagger is that it often tags passive verbs as intransitive verbs

For each context, syntactic patterns are then identified, which can be used for clustering. Seven basic patterns are identified initially, as shown in Table 1.

These are not made too specific because the aim is to capture generalities at this stage. Analysis of the text showed that these types of pattern are likely to contain similar types of context. Once the contexts have been grouped together in this way, it is easier to examine their semantic content. A distinction is made between active and passive verbs, so that these can be normalised later. For example, we would want the two contexts "*covers trabecular_meshwork*" and "*trabecular_meshwork is covered*" to be treated the same. Although their surface syntax is different, their deep structure, and hence their semantics, is the same.

3.2 Semantic Information

From these patterns, semantic frames can then be created. A set of verb classes was created manually, based on WordNet synsets (Miller et al., 1990), for the most common verbs found in the medical texts. Analysis of the text showed that a very small number of verbs occur with high frequency: the top 5 most frequently occurring verbs accounted for 66% of all verbal occurrences in our corpus, as shown in Table 2. We include, however, all the verbs occurring more than once in the corpus, since this is such a small number. The set of verb classes is shown in Table 3.

The noun phrases are then also allocated a broad semantic category, this time using UMLS, a domain-specific thesaurus in the field of medicine. Initially we have identified 10 broad noun classes, depicted with their abbreviations

Verb	Frequency	% of Total
show	223	28
contain	123	15
cut	90	11
infiltrate	70	9
bear	23	3

Table 2: Distribution of most frequent verbs in syntactic frames

Class	Abbreviation	Examples
Affect	<aff>	change, affect
Description	<desc>	occur, define
Observation	<obs>	show, feature
Position	<pos>	occupy, envelop
Procedure	<proc>	remove, sever

Table 3: Verb Classes

in Table 4.

Armed with this information, each instance of a syntactic pattern can then be automatically mapped to a semantic pattern. For each syntactic pattern, the constituents are assigned a semantic class. If the noun phrase consists of a single noun, the semantic tag allocated previously is simply transformed into its appropriate corresponding semantic class. If the noun phrase consists of a multi-word term, the same thing applies, since the term will have had a semantic tag from UMLS allocated previously. If neither of these apply, there are two remaining possibilities. Firstly, the noun phrase might contain a term but be modified by something, such that the whole does not constitute a multi-word term. For example, the noun phrase *medium_sized_artery* is not a multi-word term although the noun *artery* on its own is. In this case, *artery* will have a semantic tag (and therefore semantic class) available, and this can be extended to the whole noun phrase since we can assume that the modifiers do not change the meaning of the noun to which they are attached (i.e. that a medium-sized artery is a kind of artery).

Class	Abbrev.	Examples
Abnormality	<AB>	epithelial_oedema
Body Part	<BP>	corneal_epithelium
Body Substance	<BS>	cholesterol_crystals
Finding	<FIN>	ocular_haemorrhage
Injury	<INJ>	perforating_wound
Phenomenon	<PHEN>	spontaneous_rupture
Procedure	<PROC>	electron_microscopy
Section	<SEC>	lower_third
Size	<SI>	circumference
Virus	<VIR>	papilloma_virus

Table 4: Noun Classes

Secondly, the noun phrase might contain a group of words, none of which are terms. In this case, the noun is checked against a supplementary list to see if it may fit into one of the existing semantic classes. If so, it is assigned to this class; if not, the whole frame is rejected. This supplementary list was created manually by taking an initial set of frequently occurring words for which no class had been assigned and deciding whether they fitted into a particular category. The "Section" class is largely composed of these words, since they are not likely to be terms. If it becomes evident that other sets of words occur frequently for which there is no class established, these can be added.

Let us take as an example the set of contexts which fit the syntactic pattern Verb + NP + Prep + NP, and where the verb is of the *observation* category. The results of the initial mapping procedure are illustrated in Table 5. Each noun phrase is allocated a noun class derived from its semantic category in UMLS. Some of these, e.g. *fibrous_tissue* and *muscle*, can be immediately allocated their class, since they appear in the text exactly as found in UMLS, i.e. they are classified by UMLS as a term. Others are not classed as terms in UMLS, but contain terms within the noun phrase which are. For example *superficial_lamellae* contains the term *lamellae*, and *visual_cell_layer* contains the term *cell_layer*. Others do not contain UMLS terms at all, but are nevertheless possible to classify, such as *mid-zone*. Finally, there may remain noun phrases which are impossible to classify and which are therefore likely to be either words or terms unknown by UMLS and by the user, or which are simply errors, for example *pilary_complex*. These would be rejected if no category can be allocated, although they may be stored for later manual intervention.

The final stage is the creation of semantic frames which can be filled by normalised lexi-

Verb	NP	Prep	NP
shows<obs>	disruption <FIN>	of	lens <BP>
shows<obs>	atrophy <AB>	of	mid_zone <SEC>
shows<obs>	atrophy <AB>	of	muscle <BP>
containing<obs>	elements <BS>	of	pilary_complex
containing<obs>	crystals <BS>	in	subretinal_space <SEC>
shows<obs>	erosion <AB>	of	superficial_lamellae <BP>
shows<obs>	vascularization <AB>	of	superficial_stroma<BP>
showing<obs>	acanthosis <AB>	with	typical_molluscum_bodies<BP>

Table 5: Strings matching the pattern V(act)<obs> N Prep N

cal forms. The semantic frame incorporates the syntactic pattern from which it was derived, and is produced using a set of rules, such as the following:

```
1. NP + V(Act) --> sem(NP) + sem(V)

2. V(Act) + NP --> sem(V) + sem(NP)

3. NP + V(Pas) --> sem(V) + sem(NP)
```

This can be interpreted as meaning that where we have a syntactic pattern consisting of an NP followed by an active verb, as in (1), a frame is created consisting of the semantic category for the NP followed by the semantic category for the verb (with no special feature to indicate whether it is active or passive). Where we have a syntactic pattern consisting of an active verb followed by an NP, as in (2), a frame is created consisting of the semantic category for the verb followed by the semantic category for the NP. Where we have a syntactic pattern consisting of an NP and a passive verb, as in (3), a semantic frame is created consisting of the semantic category for the verb followed by the semantic category for the NP. This has the effect of collapsing the syntactic patterns (2) and (3) into one semantic frame.

Once the frames have been created, the surface lexical forms can be transformed into slot fillers. First the verbform is normalised to its root form, and then any passive forms are reordered with the same set of rules as above. Table 6 depicts examples of the original syntactic pattern, the semantic frames created and sample fillers for the frames (i.e. the normalised contexts).

3.3 An example

Let us illustrate the whole procedure with an example. Consider the following contexts:

```
crosses cornea
vitreous_chamber traversed
traverses_tumour
optic_nerve surrounded
envelops iris
```

After allocating syntactic and semantic categories we get the following:

```
crosses_cornea               V(Act) + NP <pos><BP>
vitreous_chamber traversed   NP + V(Pas) <BP><pos>
traverses_tumour             V(Act) + NP <pos><BP>
optic_nerve surrounded       NP + V(Pas) <BP><pos>
envelops iris                V(Act) + NP <pos><BP>
```

Using Rules 2 and 3, we produce the following semantic frames:

```
crosses cornea              <pos><BP>
vitreous_chamber traversed  <pos><BP>
traverses tumour            <pos><BP>
optic_nerve surrounded      <pos><BP>
envelops iris               <pos><BP>
```

3.4 Results and Discussion

From an initial unordered set of contexts, we can therefore produce clusters of contexts which have similar semantic and syntactic properties.

The method we have described for context clustering is only a prototype system, which is currently semi-automatic. Its aim is to demonstrate the idea of context clustering, and to show how the contexts identified in a term extraction process can be put to further use, given the semantic information already acquired, and some deeper syntactic information.

Naturally, we cannot assume that all the clusters will be perfect. A certain amount of noise

Syntax	Semantic Frame	Slot Filler Example
NP-V(Act)	<AB><aff>	basaloid_proliferation affect
	<BP><obs>	corneal_epithelium show
NP-V(Pas)	<desc><BP>	attenuate corneal epithelium
	<obs><BP>	represent retina
	<proc><SEC>	take sections
	<proc><BP>	cut optic_nerve
	<proc><AB>	remove basal_cell_papilloma
	<pos><SEC>	separate small_sections
	<pos><BP>	traverse vitreous_chamber
	<pos><AB>	surround foamy_histiocytes
V(Act)-NP	<aff><BP>	affect epidermis
	<obs><AB>	show laminar_splitting
	<proc><SEC>	cut edges
	<pos><AB>	traverse tumour
	<pos><BP>	cross cornea

Table 6: Examples of semantic frames with fillers

and omissions is only to be expected. We currently aim for accuracy over recall in this respect, since erroneous classification tends to be more harmful than incomplete classification, and it is easier for us to refine our system if it is overgenerative rather than undergenerative at this point.

Table 7 shows some examples of different context clusters created. The asterisked lexical strings denote errors. Three main reasons can be identified for these: syntactic tagging errors, chunking errors and semantic classification errors. Syntactic tagging errors result in a word or chunk being allocated the wrong syntactic tag: for example, *somewhat* is identified as an NP. Chunking errors result in a word or group of words being incorrectly identified as a chunk: for example, the phrase *both globes cut* is incorrectly divided into the two chunks *both* and *globes_cut*, instead of *both_globes* and *cut*. Similarly, *nasal_side_the_retina* is wrongly identified as a chunk. Semantic classification errors result in a chunk being allocated an incorrect noun or verb class. For example, *cholesterol_crystals* is tagged as *Body Part*, but it should really be a *Phenomenon* (in line with terms such as *calcium_deposit*, which are tagged as *Phenomenon*). In this case, the error lies with the UMLS classification. Errors are also possible when a multiword term is not found in UMLS. In this case, if the head noun is found to be present, the semantic tag for this is used for the whole term. This assumes that a term will have the same sematnic category as its head noun, but this is not always the case. We did not find any examples where this type of error actually occurred, but with a more comprehensive system or a larger corpus this is certainly a possibility.

One advantage of our method, however, is that because we do not rely on lexical pattern matching, some types of errors such as typographical errors can be ignored without being rejected. For term recognition, this kind of error is problematic because misspelled terms will still be recognised but will be treated as separate terms from their correctly spelled versions, thereby leading to statistical differences. Although this is not such a problem where only one occurrence of a frequently occurring term is misspelled, some terms suffer from frequent misspellings or simply have variant spellings, and this can be more problematic. This is often the case with hyphenated terms and terms containing apostrophes, which are not always used consistently. The clustering method has the advantage of ignoring these differences. Sometimes it may ignore a misspelled word because it cannot allocate a semantic class to it, but as we have seen, there are various safety mechanisms in place to deal with precisely such cases. For example, the term *bullous_separation* occurs in the text misspelled as *nullous_separation*. Al-

Pattern	Lexical String
$<obs><AB>$	show chronic_inflammatory_changes
	see atrophic_changes
	show vacuolar_changes
	show gross_liquefactive_changes
	see hyaline_degenerative_changes
	show hyaline_degenerative_changes
$<BP><pos>$	track cross
	large_vessel enter
	track cross
	adventitia overlie
	vessel enter
$<proc><BP>$	cut optic_nerve
	shed cells
	fold retina
	take membrane
	open globe
	displace lens
	remove cornea
$<proc><AB>$	remove basal_cell_papilloma
$<BP>\ Prep<BP><aff>$	nerve behind cribriform_plate affect
	**considerable_traction on ciliary_processes affect
$<aff><AB>\ Prep<BP>$	show degenerative_changes in collagen
	show bullous_separation in axial_region
	**show nullous_separation of epithelium
	show bullous_separation in central_region
$<aff><SEC>\ of<BP>$	affect lower_half of retina
	affect all_layers of cornea
	affect pigmented_epithelium of peripheral_retina
	affect adjacent_skin to small_extent
$<pos><SEC>\ of<BP>$	occupy posterior_nasal_portion of globe
	traverse horizontal_axis of cornea
	overlie external_surface of tumour
	occupy temporal_third of anterior_chamber
	envelop lower_half of iris
	separate bundles of cells
$<pos><BP>\ Prep<LOC>$	enter eye at peripheral_border
	invade angle in planes
	**traverse cornea in somewhat

Table 7: Examples of final clustered contexts

though this cannot be allocated a semantic tag initially (since it does not occur in the UMLS Metathesaurus), the term *separation* does occur and can thus be allocated a tag. As an untagged NP chunk, the whole term *nullous_separation* takes on the class allocated to the noun *separation*, and the whole context in which it occurs will be clustered into the same set as any occurrences of the correct term *bullous_separation*.

3.5 Granularity of clustering

The level of granularity of the clustering mechanism is also an issue which needs to be considered. This is mostly dependent on what the clusters are required for. Some applications may require a very broad grouping, while others may require more fine-grained distinctions. It may not even be dependent on the type of application but on the particular instance of application. For example, a clustering technique which is used for ontology building might require a different level of granularity according to the domain or text type, or even for the ultimate use of the ontology. So a general mechanism for context clustering ideally needs to be flexible in its level of granularity, so that it can be used for different purposes with the minimum of tuning. The method we have described is initially set at a low level, such that clustering is in broad groups. If necessary, it can be made more fine-grained by the addition of finer semantic distinctions between classes. This would be achieved by tightening restrictions on which semantic categories constitute a semantic class, for the noun phrases, and for the verbs by subdividing the semantic classes into finer distinctions. For example, the verb class *Procedure* could be subdivided into *Cutting*, *Removing*, *Inserting* and *Altering* Procedures.

4 Conclusions

In this paper, we have examined the idea of terminological acquaintance, and discussed the importance of contextual information in terminology. We focused particularly on the idea of forming clusters of contextual information for use in terminological applications in the field of medicine. In contrast to the more usual statistical techniques, our approach is primarily linguistic, which may be more suitable for small corpora. We have given some examples of how the clustered contexts can be used. In particular, the building and tuning of domain-specific ontologies could benefit enormously from this type of work. One possibility is to combine this kind of method with statistical algorithms such as those discussed in the last section. Other possibilities include combining a similarity measure, such as that used in the TRUCKS system, with this method, in order to tune existing ontologies.

5 Acknowledgements

We would like to thank A. Joshi and B. Srinivas at the University of Pennsylvania for allowing us to make use of their Supertagger and other analysers.

References

S. Ananiadou. 1994. A methodology for automatic term recognition. In *Proc. of 15th International Conference on Computational Linguistics (COLING)*, Kyoto, Japan.

M. Andrade and A. Valencia. 1998. Automatic extraction of keywords from scientifc text: application to the knowledge domain of protein families. *Bioinformatics*, 14(7):600–607.

R.H. Baud, A-M. Rassinoux, and J-R. Scherrer. 1992. Natural language processing and medical records. In *Proc. of Medinfo '92*, pages 1362–1367.

B. Boguraev and C. Kennedy. 1999. Applications of term identification technology: domain description and content characterisation. *Natural Language Engineering*, 5(1):17–44.

J.D. Cohen. 1995. Highlights: Language- and domain-independent automatic indexing terms for abstracting. *Journal of the American Society for Information Science*, 46(3):162–174.

B. Daille, E. Gaussier, and J.M. Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proc. of 15th International Conference on Computational Linguistics (COLING)*, pages 515–521.

Fred J. Damerau. 1993. Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management*, 29(4):433–447.

L. Davidson, J. Kavanagh, K. Mackintosh, I. Meyer, and D. Skuce. 1998. Semi-automatic extraction of knowledge-rich contexts from corpora. In *Proc. of 1st Workshop on Computational Terminology (Computerm '98)*, pages 50–56, Montreal, Canada. COLING-ACL.

M.B. Do Amaral and Y. Satomura. 1995. Associating semantic grammars with the SNOMED: Processing medical language and representing clinical facts into a language-independent frame. In *Proc. of 8th World Congress on Medical Informatics (Medinfo)*, pages 18–22, Edmonton, Alberta.

K.T. Frantzi and S. Ananiadou. 1996. Extracting nested collocations. In *Proc. of the 16th International Conference on Computational Linguistics (COLING)*, pages 41–46, Copenhagen, Denmark.

K.T. Frantzi and S. Ananiadou. 1999. The C-Value/NC-Value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3):145–179.

W.A. Gale, K.W. Church, and D. Yarowsky. 1992. Work on statistical methods for word sense disambiguation. In *Probabilistic Approaches to Natural Language: Papers from the 1992 Fall Symposium*, pages 54–60, Cambridge, Massachussetts. AAAI Press.

G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers.

Z.S. Harris. 1968. *Mathematical Structures of Language.* Wiley (Interscience), New York.

L. Hirschman, R. Grishman, and N. Sager. 1975. Grammatically based automatic word class formation. *Information Processing and Retrieval*, 11:39–57.

A. Joshi and B. Srinivas. 1994. Disambiguation of super parts of speech (or supertags): Almost parsing. In *Proc. of 15th International Conference on Computational Linguistics (COLING)*, Kyoto, Japan.

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

D.G. Maynard and S. Ananiadou. 1999a. Identifying contextual information for term extraction. In *Proc. of 5th International Congress on Terminology and Knowledge Engineering (TKE '99)*, pages 212–221, Innsbruck, Austria.

D.G. Maynard and S. Ananiadou. 1999b. Term extraction using a similarity-based approach. In *Recent Advances in Computational Terminology.* John Benjamins. to appear.

I. Meyer, K. Mackintosh, C. Barrière, and T. Morgan. 1999. Conceptual sampling for terminographical corpus analysis. In *Proc. of 5th International Congress on Terminology and Knowledge Engineering (TKE '99)*, pages 256–267, Innsbruck, Austria.

G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1990. Five papers on WordNet. Technical Report CSL Report 43, Cognitive Science Laboratory. Princeton University.

A.M. Morel-Guilllemaz, R.H. Baud, and J.R. Scherrer. 1990. Proximity processing of medical texts. In *Proceedings of MIE 1990*, pages 625–630. Springer-Verlag.

H.T. Ng and Hian-Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense. In *Proc. of ACL '96.*

NLM, 1997. *UMLS Knowledge Sources.* National Library of Medicine, U.S. Dept. of Health and Human Services, 8th edition, January.

C. Nobata, N. Collier, and J.-I. Tsujii. 1999. Automatic term identification and classification in biology texts. In *Proc. of NLPRS-99*, pages 369–375, Beijing, China.

R.A. Rocha, B. Rocha, and S.M. Huff. 1994. Automated translation between medical vocabularies using a frame-based interlingua. In *Proc. of SCAMC '94*, pages 690–694.

Naomi Sager. 1972. Syntactic formatting of scientific information. In *Proc. of 1972 Fall Joint Computer Conference*, volume 41 of *AFIPS Conf. Proc.*, pages 791–800, Montvale, NJ.

S. Schulze-Kremer. 1998. Ontologies for molecular biology. In *Proc. of 3rd Pacific Symposium on Biocomputing (PSB)*, pages 695–706.

J.-I. Tsujii. 2000. The GENIA Project. http://www-tsujii.is.s.u-tokyo.ac.jp/ nigel/GENIA/index.html. Dept. of Information Science, University of Tokyo.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of 33rd Annual Meeting of the Association for Computational Linguistics, ACL'95*, pages 189–196.