Linköping Studies in Science and Technology.

Thesis, No. 1523

Computational Terminology: Exploring Bilingual and Monolingual Term Extraction

by

Jody Foo

GRADUATE SCHOOL OF LANGUAGE TECHNOLOGY

Submitted to Linköping Institute of Technology at Linköping University in partial fulfilment of the requirements for degree of Licentiate of Philosophy

Department of Computer and Information Science
Linköping University
SE-581 83 Linköping, Sweden

Linköping 2012

ISBN 978-91-7519-944-3
ISSN 0280–7971
Printed by LiU Tryck, Linköping, Sweden. 2012

URL: http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-75243

Typeset using XeTeX and the `memoir` package
Typeset in Gentium Basic, ScalaSansOT, `Menlo`, and *Pazo Math*

Computational Terminology: Exploring Bilingual and Monolingual Term Extraction

by

Jody Foo

April 2012
ISBN 978-91-7519-944-3
Linköping Studies in Science and Technology
Thesis No. 1523
ISSN 0280–7971
LiU–Tek–Lic–2012:8

ABSTRACT

Terminologies are becoming more important to modern day society as technology and science continue to grow at an accelerating rate in a globalized environment. Agreeing upon which terms should be used to represent which concepts and how those terms should be translated into different languages is important if we wish to be able to communicate with as little confusion and misunderstandings as possible.

Since the 1990s, an increasing amount of terminology research has been devoted to facilitating and augmenting terminology-related tasks by using computers and computational methods. One focus for this research is Automatic Term Extraction (ATE).

In this compilation thesis, studies on both bilingual and monolingual ATE are presented. First, two publications reporting on how bilingual ATE using the align-extract approach can be used to extract patent terms. The result in this case was 181,000 manually validated English-Swedish patent terms which were to be used in a machine translation system for patent documents. A critical component of the method used is the Q-value metric, presented in the third paper, which can be used to rank extracted term candidates (TC) in an order that correlates with TC precision. The use of Machine Learning (ML) in monolingual ATE is the topic of the two final contributions. The first ML-related publication shows that rule induction based ML can be used to generate linguistic term selection patterns, and in the second ML-related publication, contrastive n-gram language models are used in conjunction with SVM ML to improve the precision of Term Candidate (TC)s selected using linguistic patterns.

This work has been supported by the Swedish Graduate School of Language Technology and the Swedish Research Council.

Department of Computer and Information Science
Linköping University
SE-581 83 Linköping, Sweden

Acknowledgements

Although it says on the cover of this thesis that this is my work, it would not have been possible for me to complete it without the help and support from my supervisors, colleagues, friends, family and the government.

My first contact with computational terminology related work was at Fodina Language Technology AB where I stepped in to help with the development of IView, originally developed by Michael Petterstedt. Without this beginning I do not think I would have started my PhD in this field. Fodina Language Technology and the people there have also played a key role in my research by letting me be part of the PRV[1] term extraction project.

My supervisors Magnus Merkel and Lars Ahrenberg have given me the most direct support, providing insight and helping me plan and outline my research, as well as providing editorial support in writing this thesis.

I have also been fortunate to work with great colleagues who have made my workplace a stimulating, enlightening and fun place to work. This workplace has been nurtured and curated by the prefect of the department, Mariam Kamkar, head of the division, Arne Jönsson and the head of the lab, Lars Ahrenberg. They have provided a good place to grow these last few years. For my position as PhD student at the Department of Computer and Information Science, Linköping University, I have to thank the government[2] funded National Graduate School of Language Technology (GSLT) and Swedish Research Council (Vetenskapsrådet, VR) from which I have received my funding. GSLT has also provided a great opportunity to meet other PhD students in Language Technology. I would also like to thank Jalal Makeki and Sture Hägglund for making it possible for me to work at the department while awaiting funding for my PhD student position.

Special thanks goes to my fellow PhD students, both past and present at the HCS division. In particular Maria Holmqvist, Sara Stymne, Fabian Segelström, Johan Blomkvist, Sanna Nilsson, Magnus Ingmarsson, Lisa Malmberg, Camilla Kirkegaard, Amy Rankin, Christian Smith, and Mattias Kristiansson. Thanks for great intellectual and social company, good fikas, fun PhD pubs, and for putting up with my weird discussion topics.

I have left this final paragraph to thank my family; my parents and my brother who played a great part in shaping my first chapters in this world, my beloved son Isak who brings happiness and wisdom to my life, and finally my dearest wife Catharina. I love you so much, and am so thankful for having you by my side throughout the ups and downs during these past years. Without your support I could not have written these words.

[1] Patent och Registreringsverket

[2] and by extension tax payers

Contents

1 Introduction

Terminologies are playing an increasing role in the society of today. The combination of an accelerated rate of information production and the increase in speed at which information travels has many consequences, and raises many issues. If we humans are to both produce and consume more information in less time while maintaining or even improving the content quality, we need all the help we can get.

For specialized domains, using the correct terminology plays a major part in efficient communication. *Creating* and *maintaining* a terminology however, has been, and still is, a time consuming activity. A terminology contains *definitions* of domain-specific *concepts* and the *terms* which represent these concepts. A terminology also contains information on how the different concepts are related to each other. Having a common terminology within a subject field, together with tools that integrate the terminology with e.g. document authoring activities, can, among other things, reduce the number of possible communication errors.

Terminology Work (TW) (analyzing terminology and creating a terminology), *terminography* (publishing terminology reference works), and *terminology management* are all tasks within the field of terminology which have traditionally been performed without the aid of computers. All tasks involve dealing with relatively large data sets with complex dependencies and relationships.

For example, to create and publish a domain-specific terminology, terminologists would manually extract possible terms, i.e. term candidates, either by analyzing domain-specific literature or by interviewing domain experts. The relations between term candidates would then be disseminated, and where necessary the terminologist would consult domain experts. Finally, the terms are structured into defined concepts which are then published as a work of reference.

Terminology Management (TM) is defined as “any deliberate manipulation of terminological information” (Wright & Budin, 1997, p. 1). This includes work that needs to be done when updating published terminology 1) by adding new terms and concepts, 2) by revising existing concepts e.g.

by joining or splitting them, or 3) by deprecating old terms. These tasks can be quite complex and the complexity can grow further depending on how many channels the terminology is published through, e.g. traditional book, web database, database which is integrated into authoring or translator's tools etc.

The field of *Computational Terminology (CT)* studies how computational methods can be of use when performing Terminology Work. For various reasons, the introduction and implementation of computers within terminology-related tasks have not been as fast or as aggressive as in other areas, and even today, the most common level of computerization in practice is using Microsoft Excel spread sheets, and Microsoft Word documents, and in some cases rather unsophisticated term databases. More advanced tools are becoming available, but are not widely used.

1.1 Automatic Term Extraction

Automatic Term Extraction (ATE) uses computational methods to produce term candidates for further processing when either performing terminology work or e.g. terminology maintenance tasks within TM (see 4 for more on ATE). Most ATE research is monolingual and roughly involves first extracting possible term candidates and then ranking them by degree of termness or termhood (see 4.3). There has also been work done on bilingual term extraction (see 4.5).

1.1.1 Related fields

The methods used in ATE are similar to or sometimes shared with other problem domains. *Automatic Glossary Extraction (AGE)* is very simillar to ATE, The end product, a glossary, can however be less formal than a published terminology. For example, where a terminology provides concept definitions and terms which represent them, it is fine for a glossary to provide a more less formal description to each entry. *Information Retrieval (IR)* is also a related field, and many early methods used in ATE were inspired by, or borrowed from IR research. As terms represent concepts, there is often an overlap between the terminology contained within a document, and word-units which are important to index in IR. Another related field is that of *Automatic Keyword Extraction (AKE)*, one important difference compared to ATE however, is that in AKE only the most important keywords are of interest, not all possible keywords.

1.2 Research questions

The task of *Automatic Term Extraction (ATE)* is concerned with applying computational methods to term extraction. This thesis aims to research the following topics within the area of Computational Terminology (CT):

1. Which steps are needed to integrate ATE into Terminology Work?
2. What opportunities are there for Machine Learning in ATE?

The first question is important as it puts ATE research into a context. In this context we can also try to understand what is really important for ATE to be useful. Stating that an implementation of the ATE method is not enough to perform efficient Terminology Work is relatively trivial. What is less trivial is identifying and putting the scaffolds into place that enable ATE to be useful.

In Terminology Work, terminologists work together with domain experts, examining various documents with the goal of defining concepts and their relations to each other. During this process, especially when using computational methods, much data and meta-data is produced. The second question relates to the availability of this data – Can we use this data to teach computational systems to perform valuable tasks? Being able to apply machine learning to specific tasks could reduce the development time otherwise needed to build software which solves these specific tasks.

1.3 Thesis Focus and Contributions

The work presented in this thesis focuses on the area of ATE in general, and applying machine learning and contrastive n-gram language models to ATE in particular. Besides the contributions made in the published research papers, this thesis also aims to present an overview of the different approaches to ATE that have been developed throughout the years.

The first two papers contributing to this thesis deal with bilingual automatic term extraction. More specifically, they deal with how bilingual term extraction can be used in a real life setting (Foo & Merkel, 2010), and how a metric measuring translation consistency can be used to rank term candidates (Merkel & Foo, 2007).

The bilingual extraction approach applied in Merkel and Foo (2007), Merkel et al. (2009), Foo and Merkel (2010) was the *align-extract* approach using a single selection language. After validating terms from a bilingual term extraction process we have a bilingual term list. If the extraction approach used was an align-extract approach using single sided selection, we now have terms in a second language for which we did not have a term candidate selection mechanism for. Would it not be great if we had a way of using this new data to configure a monolingual extraction environment? This is the background of the the third and fourth papers which relate to machine learning applied to monolingual term extraction.

Contributions

The research contributing to this thesis has been published as four peer-reviewed publications presented at conferences and terminology workshops, and one paper presented at a conference and published in the conference proceedings. The contributions of these and this thesis are the following;

- an overview of the existing field of Computational Terminology and an introduction to the developing field of Computational Terminology Management
- a case study from a large bilingual patent term extraction project (Foo & Merkel, 2010; Merkel et al., 2009)
- a presentation of the Q-value metric successfully used to rank bilingual term candidates, strongly indicating that terms are translated consistently (Merkel & Foo, 2007)
- a presentation of novel and successful use of a rule-induction learning system (in this case, Ripper) applied to ATE (Foo & Merkel, 2010)
- a study of how SVM machine learning can be applied to term extraction together with contrastive n-gram language models (Foo, 2011).

1.4 Thesis Outline

Before presenting the results and discussing the research performed, a background of the relevant fields is given. The main concepts in the field of terminology are presented together with how computers, software and computational methods have been applied to the field. Relevant term extraction research is also reviewed before presenting the methods used in the published work. The results from the published work is then summarized followed by a discussion. The full papers are available as appendices.

2 Terminology

The term *terminology* is ironically an ambiguous term, and can represent three separate concepts. *Terminology* can either refer to **1)** “Terminology science, [the] interdisciplinary field of knowledge dealing with concepts and their representations”, **2)** an “aggregate of terms which represent the system of concepts of an individual subject field”, or **3)** a “publication in which the system of concepts of a subject field is represented by terms” (Felber, 1984, p. 1). Analyzing, defining and naming concepts is referred to as *terminology work* and publishing the results of this work is referred to as *terminography*.

The field of terminology (Terminology Science) is a polymethodological and polytheoretical field, and methods and theories tend to differ between practitioners in different countries. Ongoing work is however being done at the International Organization for Standardization (ISO), specifically within ISO Technical Committee 37 (ISO/TC 37)[1], aimed at providing a common standard related to terminology work. The ISO history behind the creation of the ISO terminology standards originate from Eugene Wüster’s[2] work and the so called Vienna school of terminology (Felber, 1984, p. 18, 31).

2.1 Domain specificity

A terminology is always domain-specific. Some practical consequences of this is that a term may represent two different concepts in different domains. For example, the term “pipe” refers to different concepts in different domain. A terminology will only include the concept and definition relevant to one specific domain. It is also ideal for a single domain not to use a term to represent more than one concept.

[1] ISO/TC 37 is the Technical Committee within ISO that prepares standards and other documents concerning methodology and principles for terminology and language resources.

[2] Eugene Wüster, (1898-1977) was born in Wieselburg, Austria and is considered the founder of the Vienna school of terminology

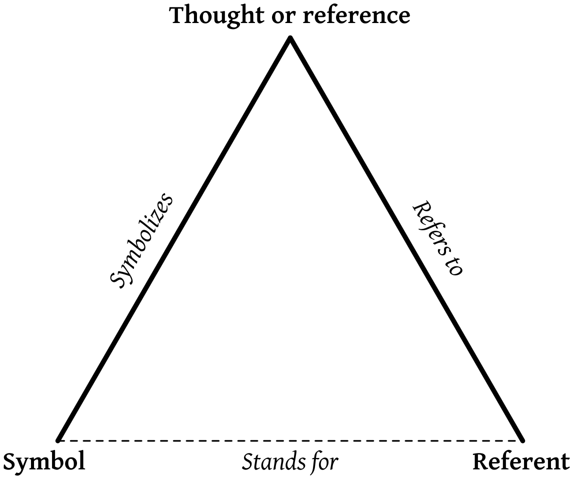

Figure 2.1: The triangle of reference. Adapted from Ogden and Richards (1972, p. 11).

2.2 Terminological structure

The semiotic triangle, or the triangle of reference (Ogden & Richards, 1972, p. 11) is a commonly used model in linguistics, semantics and semiotics which describes how linguistic symbols e.g. words, are related to actual objects in the world, referents, and thoughts. Figure 2.1 is an adaptation of the original diagram. It was used by Ogden and Richards (1972) as a tool to discuss *meaning* as conveyed using language. Various meanings of a statement can be discussed by understanding the relation between symbols, thoughts or references, and referents. Ogden also states that there is no link between *symbol* (e.g. a word) and *referent*, except for the link which connects the two through *thought*;

> Words, as everyone knows 'mean' nothing by themselves [...]. It is only when a thinker makes use of them that they stand for anything, or in one sense, have 'meaning.' (Ogden & Richards, 1972, p. 9)

The structure used in terminology is similar to that proposed by Ogden and Richards, and Sager (1990, p. 13) puts forward the following three three dimensions of terminology:

1. the cognitive dimension which relates the linguistic forms to their conceptual content, i.e. the referents in the real world;
2. the linguistic dimension which examines the existing and potential forms of the representation of terminologies;

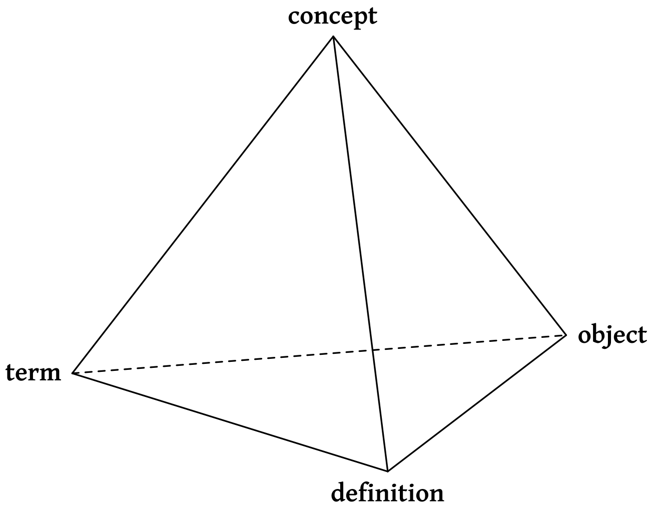

Figure 2.2: The terminology pyramid, adapted from Suonuuti (2001, p. 13)

3. the communicative dimension which looks at the use of terminologies and has to justify the human activity of terminology compilation and processing.

In essence we have a system where a sole human in the middle is the sole mediator between symbols (linguistic or other), and referents. This can be problematic when using symbols to communicate. The fact that no link exists between Ogden's symbol and referent, and the impossibility of examining other people's thoughts as argued by e.g. Nagel (1974), leads to a serious problem when we need to communicate precisely, clearly, and efficiently. This is where terminology comes in. Suonuuti (2001) presents a figure, commonly referred to as the *terminology pyramid* (fig. 2.2), of an extended version of the semiotic triangle which includes a fourth node — *definition*. The definition provides an accessible and explicit shared representation of the thought. Another way of thinking of the definition is to see it as what Clark and Brennan (1991) call *common ground*.

The three original nodes in the semiotic triangle use partially different designators that correspond to terminological structure; concept (thought), term (symbol), object. The ISO standard "ISO 704:2009" provides the following definitions:

concept 1) depicts or corresponds to objects or sets of objects; 2) is represented or expressed in language by designations or by definitions; 3) concepts are organized into concept systems;

definition defines, represents or describes the concept.

designation (terms, appellations or symbols) 1) designate or represent a concept 2) is attributed to a concept;

object an object is perceived or conceived and can be abstracted or conceptualized into concepts

2.2.1 Concepts

In terminology, the *thought* at the top of the triangle of reference is replaced with *concept*. The relation between the objects and concepts can be described as an abstraction process which takes us from the *properties* of an object to *characteristics* of a concept. The difference between properties and characteristics is similar to the difference in object oriented programming between a field in a class and its realization in an instance of that class. Concepts are classes, and objects are instances.

Delimiting characteristics

What defines a concept according to Suonuuti (2001) are its characteristicts, however listing all characteristics of a concept is not reasonable. What should be focused on instead, are the delimiting characteristics of the concept (Suonuuti, 2001).

> Delimiting characteristics are those characteristics that alone, or together with other characteristics, determine the concept and differentiate it from other concepts. (Suonuuti, 2001, p. 13).

For example, to define the concept of a bowling ball, characterizing it as spherical is not a delimiting characterization, as there are many other concepts that have that characteristic.

Concept analysis

Concepts are organized into a concept system. The system describes how concepts are related to each other. Three types of relations are possible between concepts; *generic*, *partitive*, and *associative*. Most concept systems contain concepts with all three types of relations and are called *mixed concept systems*.

The *generic relation* between concepts is a hierarchical relation where a superordinate concept is divided into one or more subordinate concepts. Each subordinate concept can in turn have their own subdivision. Subdividing a concept however, is not done arbitrarily. Each subdivision requires a specific dimension of characteristics to divide the subordinate concepts. For example, on a bicycle the superordinate concept *break* may be divided into the subconcepts *rear break* and *front break*, where the criterion for each subdivision of the concept *break* is the location of the break on the bicycle.

Partitive relations between concepts are hierarchical relations where subordinate concepts are parts of the superordinate whole. Several different forms of partitive concept systems exist (Suonuuti, 2001). For example, the *whole* in a partitive concept system may consist of similar parts (a brick wall, built from identical bricks), or the whole may constituted of different parts (a dagger which has a blade and a handle). The number of parts in a concept system may or may not be essential, a *500 piece puzzle* necessarily consists of 500 pieces, whereas a keyboard may have a varying number of keys. The system may also be *closed* or *open*, a closed system, only allowing certain parts (e.g. the year), whereas an open system (e.g. a computer) can consist of a different number of parts.

The third type of relation, the *associative relation*, is non-hierarchical and includes relation which are resemble semantic or thematic roles. Examples of such relations are cause/effect, producer/product, material/product.

2.2.2 Definitions

Traditionally, the definitions have been the core of a terminological resource and Suonuuti (2001, p. 19) says that "The overall quality of terminology work mainly depends on the quality of definitions.". The definition serves as a textual account of the concept system, can be written as an *intensional*[3] *definition* or as an *extensional*[4] *definition*. The intensional definition describes a concept's *essential* and *delimiting characteristics* . The extensional definition lists the object covered by the concept. The intensional definition can be used to directly translate general relations in a concept system (examples below taken from Suonuuti (2001))

coniferous tree tree with needle-formed leaves and exposed or naked seeds

fir coniferous tree of the genus abies

firewood wood for fuel

light-demanding tree tree preferring sunny habitat

noble gas 1) helium, neon, argon, krypton, xenon or radon 2) gas that in the natural state is chemically inactive

tolerant tree tree preferring shade

[3] intensional (n): the internal content of a concept. intentional (adj): done on purpose; deliberate (Oxford Dictionaries. April 2010)

[4] extensional (n): the range of a term or concept as measured by the objects which it denotes or contains.

tree tall plant with hard self-supporting trunk and branches that lives for many years

wood hard fibrous material obtained from the stems and branches of trees or shrubs

woodbin box for holding firewood

The above definitions are examples of intensional definitions, except for the second definition of *noble gas*. The definition of *tree* also exemplifies how partitive relations of a concept system can be incorporated into a definition. The definitions of *wood* and *woodbin*, are examples of how associative relations can be translated into definitions.

2.2.3 Terms

"ISO 704:2009" defines a designation as something that represents a concept. Terms and symbols are two kinds of designations. In this thesis however, we are primarily concerned with terms — *the linguistic representations of concepts*. There is no formal restriction on the allowed length of a term. A term can consist of a single word, or several words. When publishing a term in a work of reference, they are usually written in lower case and using their un-inflected form.

Domain-specific communication is not by definition standardized or unambiguous, and misunderstandings between two parties when it comes to e.g. the specification of a product can lead to costly problems. By using a standardized terminology, the number of possible misunderstandings can be reduced, since it is clearly defined what concept is represented by which term, and what the definition of that concept is.

It is important to appreciate the distinction between concepts and terms. For example, the concept <*tree*>, is represented in the English language by the term "*tree*", and by the term "*träd*" in the Swedish language. The concept <*tree*> is language independent and could as such be referred to as concept <45649> in theory. However to facilitate communication and terminological discussion, the term is often used in real communicative situations, rather than e.g. the label <45649>. This sometimes leads to the assumption that concepts and terms are the same. Terms and concepts are equivalent, as per definition, but they are not the same. To give an example of a context where this is important, lets take the concept <C>, represented by the term "*tablet*" and the following definition: 'a computer with a built in touch sensitive screen that can be operated without using a physical keyboard'. Through various technical innovations, the kinds of devices that fit into this definition is increased and now includes e.g. smart phones, thinner devices running non-PC operating systems. The specificity of the concept has been decreased which gives rise to the need to integrate new terms into the terminology which can be used to represent new sub-concepts of <C>. The definition of the concept <C> still holds, but to reflect the technical developments, the term *tablet computer* is chosen as the *preferred* term, and the term "*tablet*" is given the *deprecated* status with regard to the concept <C>. However, the term "*tablet*" may still be chosen as the *preferred* term used to represent a subordinate concept of <C>.

There may be situations when a concept can be represented using an existing term in one language, but where a second language does not provide a term for the concept as originally defined. An example of a class of domain-specific concepts where this is relatively common, is the class of academic degrees and titles, which is a result of varying educational, and academic systems around the world. This also gives an example of a situation where it is possible for a concept to exist without there being a term which is connected to it.

Terms can either be single-word units or multi-word units. In languages such as English, where compounding is relatively rare, terms are usually multi-word units. Languages, such as Swedish, where compounds are used extensively, do not show this tendency, as most English multi-word terms are translated into a single Swedish compound word.

Apart from terms that should be used to represent a specific concept, *preferred* or *accepted* terms (use of preferred terms is preferred over use of accepted terms), a terminology can also list *forbidden* and *deprecated* terms. Forbidden terms are terms that should not be used to represent a certain concept. Deprecated terms are terms that may have been preferred or accepted in a previous version of the terminology, but are no longer recommended and should be replaced in e.g. existing documentation.

2.2.4 Objects

In common with the semiotic triangle, *objects* (terminology), or *referents* (semiotic triangle) are either concrete (e.g. apple, doll, mountain) or abstract (e.g. entertainment, service) in the real world. In terminology, objects are described as having *properties*, in contrast to concepts, which are described as having characteristics.

2.3 Terminology work

Terminology work is defined as the activities in which a terminologist compiles and creates new terms or a full terminology. There is no single defacto standard method in terminology and Wright and Budin (1997, p. 3) prefer to describe this reality as a polytheoretical and polymethodoligical reality, rather than dividing the landscape into different "schools"[5]. "ISO 704:2009" list the main activities in terminology work as follows.

- identifying concepts and concept relations;
- analyzing and modeling concept systems on the basis of identified concepts and concept relations;
- establishing representations of concept systems through concept diagrams;
- defining concepts;
- attributing designations (predominantly terms) to each concept in one or more languages;
- recording and presenting terminological data, e.g. in print and electronic media (terminography).

One distinction to take note of, is that the concept *terminology work* includes *terminography*, but there is more to terminology work than just terminography.

[5] There are traditionally three schools of terminology; 1) the Vienna School of Terminology, founded by Eugene Wüster, 2) the Prague School of Terminology developed from the Prague School of functional linguistics, and 3) the Soviet School of Terminology, headed by the terminologist Lotte. (Felber, 1984, pp. 1.091–1.093)

3 Computational Approaches

There are many approaches to augment or improve terminology-related activities using computer software and computational methods. This chapter presents some historical background of the use of computers, computer software and computational methods related to terminology.

Early visions of the use of computers in terminology focus on storage and retrieval of terminology (Felber, 1984, p. 11). In step with the general increase and presence of computers in our daily lives, the focus of computers and terminology has in later years moved past storage and retrieval to computational methods.

Previous researchers have suggested new terminology-related fields which focus on computers and terminology. Cabré (1998, p. 160) refers to Auger (1989) where *Computer-aided terminology*[1] is described as a field "located between computational linguistics (computer science applied to the processing of language data) and the industrialization of derived products (the language industries)". Cabré (1998) herself talks both about *Computerized terminology* and *Computer-aided terminology* as though they are the same concept and gives the following examples of tasks at which computers can play a highly significant role for terminologists:

- selecting documentation, prior to beginning work
- creating the corpus and isolating and extracting data
- writing the entry
- checking the information in the entry
- ordering the terminological entries.

Bourigault, Jacquemin, and Homme (2001) suggest three main problem areas for the field of *Computational Terminology*

[1] Auger's original article is in French, and he did not use the English term "Computer-aided terminology". To complicate things further, Cabre's original book is in Catalan and was translated to English by Janet DeCesaris and edited by J. C. Sager. So the question one can ask is if the term *Computer-aided terminology* actually was used by Cabré.

	Prior terminological data	**No prior terminological data**
Term discovery	Term enrichment	Term acquisition
Term recognition	Controlled indexing	Free indexing

Table 3.1: Sub-domains of Term-oriented NLP (Jacquemin & Bourigault, 2003, p. 604)

- automatic identification of terminological units and filtering of the list of term candidates
- grouping of variants and synonyms
- finding relations between terms

The lists presented by Cabré (1998) and Bourigault, Jacquemin, et al. (2001) discuss two levels of computerization. Cabré's list describes a scenario where traditional terminology work is done, but where typewriters, papers, pens and books have been partially replaced by computers. Bourigault's list describes some level of automation of complex tasks which traditionally would be performed in full by the human terminologist.

In a later publication Jacquemin and Bourigault (2003) talk about *Term-oriented NLP*. In this context, Jacquemin and Bourigault (2003) also argue that the definitions and assumptions in classical terminology are "less adapted to a computational approach to term analysis". Jacquemin and Bourigault (2003, p. 604) also argue that

> The two main activities involving terminology in NLP are *term acquisition*, the automatic discovery of new terms, and *term recognition*, the identification of known terms in text corpora.

These two activities can be performed together with *prior terminological data* or with *no prior terminological data*. The general distinction within term-oriented Natural Language Processing (NLP) as presented by Jacquemin and Bourigault is between automated term extraction in a broad sense, referred to as *Term discovery*, and automatic indexing, referred to as *Term recognition*. These two activities can be performed together with *prior terminological data* or with *no prior terminological data*. *Term discovery* includes using term extraction to create a terminology from scratch – *term acquisition*, and updating an existing terminology – *term enrichment*. The task of *term recognition* is divided into *controlled indexing* which refers to the task of finding occurrences of a given list of word-units in a text, and *free indexing*, which refers to both finding relevant word-units, and recording their location in a document. See table 3.1[2] for a tabulated view of the categories.

3.1 Computational Terminology Management

Wright and Budin (1997, p. 1) define *Terminology Management* as "any deliberate manipulation of terminological information". Computational Terminology Management (CTM) is a developing field of research that focuses on applying computational methods to facilitate *Terminology Management*. Facilitation can be provided by automatic and semi-automatic processing of data, or by providing software which allows terminologists, writers, engineers and others to manipulate terminological information in new and more powerful ways. Some examples of tasks which fall under CTM are listed below:

- terminology extraction (which includes termhood estimation)
- extracting definitions
- facilitating terminologists perfoming concept analysis
- terminology maintainence tasks (such as updating a terminology with new terms found in new documents)
- content quality assurance
- terminology support in document authoring applications

As used in this thesis, CTM differs from *Computer-aided terminology* and *Computerized terminology* by moving beyond the phase of using the computer instead of "pen and paper", to using computational methods to perform tasks which are practically impossible in a non-research context (i.e. a commercial context) for terminologists. One example of such a task is to check a large ($>$ 10,000 entries) multilingual terminology for translation inconsistencies every time a new term is added to the terminology. Such a task is too time consuming for a human to do in a real life scenario. Compared to *Computational terminology*, CTM differs by including activities beyond terminology creation, elevating e.g. *terminology use* to a key area.

[2] The terminological inconsistency between the quote and the table 3.1 is also present in Jacquemin and Bourigault (2003, p. 604).

4 Automatic Term Extraction

The previous chapter gave an overview of the possibilities in which computers and computer software can be used to facilitate and improve Terminology Management (TM). This chapter will focus on Automatic Term Extraction (ATE) methods. Here the term ATE is used to refer to term extraction specifically done using *computational methods*. The term *Manual Term Extraction (MTE)* is used when specifically discussing term extraction performed by humans (e.g. marking text in documents), and the term *Term Extraction (TE)*, is used in reference to term extraction in general, when not differentiating between who or what is performing the task. Some researchers also use the terms *Automatic Term Recognition (ATR)* (e.g. Kageura & Umino, 1996; Wong & Liu, 2008) and Automatic Term Detection (ATD) (Castellví, Bagot, & Palatresi, 2001), to represent the concept referenced here using the term ATE.

This chapter will review current research in monolingual term extraction, describe how monolingual term extraction can be extended to bilingual term extraction and finally describe how automatic term extraction performance may be evaluated. The research on term extraction is divided into three approaches: *linguistically oriented ATE*, *statistically enhanced ATE*, and *ATE methods using contrastive data*. The purpose of the survey presented here is not to cover all existing ATE research in detail, but rather to put the contributing papers in this thesis into context.

4.1 Term extraction for terminology work

The history of automatic term extraction for terminology use has its beginnings during the 1990s, beginning with research done by e.g. Damerau (1990), Ananiadou (1994), Dagan and Church (1994), Daille (1994), Justeson and Katz (1995), Kageura and Umino (1996), and Frantzi, Ananiadou, and Tsujii (1998). The development of methods and techniques in Information Retrieval (IR) and Computational Linguistics at this time seem to have reached a stage where they could be applied to the area of *Automatic Term Recognition* (which was the dominant term used to denote the area at the time).

Before we continue, it should be noted that the term "*term*" refers to different concepts depending on the field it is used in. For researchers from the field of *information retrieval*, the use of "terms" refer to *index terms*, which are the indexed units of a document. The following definitions are taken from Manning, Raghavan, and Schütze (2008, p. 22) and are valid for *the area of Information Retrieval.*

token a token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing

type a type is the class of all tokens containing the same character sequence

term a term is a (perhaps normalized) type that is included in the IR system's dictionary

Given these definitions, a *term* within IR is not required to satisfy any kind of semantic conditions, i.e. any indexed word-unit is a term. What is important when it comes to choose which word-units which should be indexed, is whether indexing a particular word-unit can increase a IR system's performance, or the precision of the system's output.

The term "*term*" in the context of terminology work, refers to a linguistic designation of *domain-specific* concept (see 2.2). When performing term extraction for use in terminology work, term extraction methods have to evaluate whether or not the units to be extracted are domain-specific or not, and if so, relevant to the domain in question.

It is therefore important that the reader takes note of the context in which the research is presented, since the goals of different fields[1] may differ on certain points. One example is the study by Zhang, Iria, Brewster, and Ciravegna (2008) which presents one of the more detailed and extensive comparative evaluations of different Automatic Term Recognition methods. Even though the evaluation includes the C-value/NC-value approach (Frantzi, Ananiadou, & Tsujii, 1998) which is presented in the context of terminology work, Zhang et al. start by saying "ATR is often a processing step preceding more complex tasks, such as semantic search (Bhagdev et al. 2007) and especially ontology engineering". For further discussion on this topic, see subsection 7.4.1.

A field related to ATE is *Automatic Keyword Extraction* or *Automatic Keyword Recognition.* In the field of Automatic Keyword Extraction, similarities with the field of ATE can be found with regard to the methods used. However, keyword extraction focuses on extracting a small set of units

[1] e.g. the task of Automatic Term Extraction in the *field of Information Retrieval* and the task of Automatic Term Extraction in the *Terminology field.*

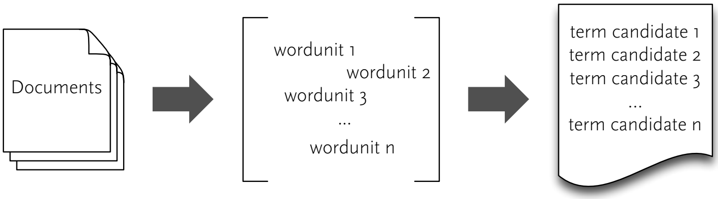

Figure 4.1: General ATE process

that can accurately describe the content of e.g. a document, whereas terminology extraction is concerned with finding all the terms in e.g. a document. Keywords are most often terms, but when performing term extraction for terminology use, we want to extract *all* terms, not only the most important or most representative ones.

ATE for use in terminology work can in many cases be divided into three sub-tasks; 1) *word-unit extraction*, 2) *ranking/categorization*, and 3) *term candidate output* (see figure 4.1).

During *word-unit extraction*, suitable units, either single words or multi-word units, are extracted from the text. The criteria used to select these units vary, but most approaches focus on noun phrases and e.g. use previously annotated part-of-speech information to find these units. The extracted phrases may also be processed to facilitate subsequent tasks. Examples of such processing are spelling correction, lemmatization and decompounding.

Termhood measurement can sometimes be integrated into word-unit extraction, but the goal of this process is either to rank or classify the extracted units according to how likely it is that they are good term candidates. Finally, the *term candidate output* phase selects which of the extracted units should be the output of the system. Depending on the situation, this may e.g. be all units classified as a certain category, all units above a certain termhood metric score, or units that match one of the previous criteria with an additional constraint on the number of occurrences of the unit in the source material. It is important to understand that the raw output from a term extraction system can *never be considered to be terms*. The output from a term extraction technique are *term candidates*. *Terms* must be validated by a human. It would be much better to talk about the task of *term candidate extraction*, but very few in the research community seem to do this.

conversion of documents to plain text
character encoding normalization
tokenization
Part-of-Speech tagging
lemmatization and stemming

Table 4.1: Common pre-processing tasks in Automatic Term Extraction

4.2 Pre-processing

Depending on the term extraction method, the input data used by the method differs. Transforming raw data to correct input data for a specific method is in the ideal case a trivial matter, but can in some cases turn out to be a rather tricky problem. Table 4.1 lists of some common pre-processing tasks, i.e. tasks that are necessary in practice, but in a sense, unrelated to the actual term extraction task. These tasks are briefly described in the following sub-sections.

4.2.1 Converting to plain text

Converting a document into plain text is the task of going from a non-plain text format to plain text. Non-plain text formats include PDF, Microsoft Word, Rich Text, and HTML documents. Problems, or non-trivial conversion decisions that can arise in during the conversion task in a broad sense, can be classified as either *layout-related* or *typography-related*.

Examples of layout-related issues

Below are some examples of layout-related issues when converting documents with layout to plain text.

multiple column layout to single column layout conversion is a problem that applies mostly to PDF documents and scanned documents where the columns must be first identified and then concatenated in the right order

figure captions can pose a problem as they can sometimes be inserted in the middle of a sentence, e.g. if a figure is placed in the middle of a paragraph.

page numbers together with associated whitespace in headers and footers should in most cases be removed as they may be wrongly inserted into sentences

table data is problematic for two reasons. First, table data can interrupt regular text. Secondly, table data seldom consists of whole sentences, which can be problematic when e.g. linguistic analysis and in many cases contain numeric data. Because of this, table data is in some cases omitted.

4.2.2 Examples of typography-related issues

Below are two examples of possible typography-related conversion issues.

subscripts and superscripts are not available in plain text and must therefore be converted in some way. One way is to omit them during the conversion. Another is to convert them to regular script. This however potentially inserts additional text in the middle of a sentence, which can disable e.g. POS pattern recognition.

use of varying font weights refers to the use of e.g. bold or italic font weights in text. This kind of variation is often semantic; e.g. in this list where the bold weighted text delimit the topic. It is also common that different font weights have different meanings in e.g. dictionaries and glossaries. It is therefore desirable that such information is retained when converting a formatted document to plain text. In most cases however, this information is discarded.

4.2.3 Character encoding normalization

Typography-related issues are in a sense related[2] to the next task preprocessing task, *character encoding normalization*. In an ideal world, there would only be one text encoding, with no platform variations. Unfortunately, we do not live in ideal world, and since components in a term extraction suite may only be compatible with a specific character encoding, this is a problem we have to deal with. When it comes to term extraction, we need to decide what to do when a glyph in character encoding used for the source text is not available in the target character encoding scheme. Typical examples include the use of trademark and copyright symbols, ellipsis, and non-latin characters.

4.2.4 Tokenization

Tokenization is the task of breaking up the continuous string of characters, which the text file consists of, into delimited tokens. The tokens are the units that will be processed during term extraction. Most tokens are

[2] Typographical information is not encoded into text, e.g. whether a string should be set as bold or regular is information above the text level. A bold letter '**b**' is the same glyph as a regular letter 'b' and a italic letter '*b*'. All are encoded in the same way in text. They are however, different weights of a font.

single words, and can be delimited by looking for 'whitespace' characters[3]. Some characters however, such as punctuation, should sometimes be treated as separate tokens, and in other cases be part of a longer token. The period character, '.', for instance is sometimes the last character of a sentence. It may also be used as a decimal denotation, as in 3.1415 and in chapter and figure enumeration. Another non-trivial example is whether or not value-unit pairs should be considered separate tokens or not. For example, should "5 km" be one token (<5 km>) or two (<5> <km>)?

4.2.5 Part-of-speech tagging

Computational part-of-speech tagging, or just Part-of-speech tagging, or Part of Speech (POS) tagging, is the task of labeling words with their correct part of speech. POS tagging is a much more limited task than parsing, as POS tagging does not involve resolving grammatical phrase structure. Research on POS tagging for English can be traced back to the work done on the Brown Corpus by Francis and Kučera during the 1960s and 1970s. The state-of-the-art POS tagging systems today e.g. TreeTagger (Schmid, 1994) (Schmid, 1995), have a precision between 96–97%.

The POS tagged data used in the studies (Foo & Merkel, 2010) (Foo, 2011) presented in thesis was tagged using Connexor Functional Dependency Grammar (Connexor FDG), which besides POS annotations, also provides the lemma form of processed words (Tapanainen & Järvinen, 1997). Connexor FDG is a commercial system developed by Connexor Oy[4].

4.2.6 Lemmatization and stemming

Morphological inflection is present in most written languages. A consequence of this is that word form variants exist in text. When performing term extraction, we want to e.g. be able to cope with the fact that *character encoding* and *character encodings* are variants of the same term, or in a language such as Swedish, which has a richer inflection scheme than English, that *teckenkodning*, *teckenkodningar*, *teckenkodningen*, *teckenkodningens*, and *teckenkodningarnas* are variants of the same term ("character encoding"). Normalization of the word form used, can be achieved through lemmatization, which gives the lemma form, or base form of a word, or through stemming, which truncates word variants into a common stem (which is not necessarily a valid word form).

4.3 Termhood

The degree of terminological relevance as measured by computational methods can be called *termhood*. The concept of termhood was defined in

[3] e.g. `<tab>` and `<space>`

[4] http://www.connexor.com/

Kageura and Umino (1996) as “The degree to which a stable lexical unit is related to some domain-specific concepts.”. However, Ananiadou (1994) does mention[5] the term, but without defining it.

The ideal goal regarding termhood is to find a metric that correlates perfectly with the concept of termhood. Such an ideal termhood metric should make it possible to achieve precision and recall scores comparable with human performance. Human performance however is not homogenous as noted by e.g. Frantzi, Ananiadou, and Tsujii (1998, p. 594):

> There exists a lack of formal or precise rules which would help us to decide between a term and a non-term. Domain experts (who are not linguists or terminologists) do not always agree on termhood.

This quote describes why termhood might actually be a concept which perhaps only exists in our imagination, i.e. that there is no such thing as our idealized definition of termhood. However, more importantly, the quote describes why it is not trivial to implement a metric that measures termhood. It is also important to understand that it is not always the case that a measure of termhood is needed to implement a ATE system that performs well. One example is the work described in Merkel and Foo (2007) (see section 6.3), where the translational stability of extracted bilingual word-units is measured using the Q-value. The Q-value scores are in turn used used to select term candidates. This approach is successful since terms are lexicalized to a higher degree, and have a lower lexical variability (Justeson & Katz, 1995). In other words, Q-value scores are used as indirect termhood scores for the purpose of selecting likely term candidates.

4.4 Unithood

Unithood was also introduced by Kageura and Umino (1996) and is defined as “the degree of strength or stability of syntagmatic combinations and collocations”. In the English language, many terms are multi-word units, and multi-word terms are stable collocations. Justeson and Katz (1995) performed a study of technical terminology taken from several domains (fiber optics, medicine, mathematics and physics, and psychology). From these domains, 800 terms were sampled and it was found that 70% (564 of 800) of the terms were multi-word units. Determining whether a particular word-sequence should be treated as a stable multi-word unit or not is therefore an important task.

[5] “The notion of *terminological head* of a wordform is important in this respect: this refers to the element of a complex wordform which confers term-hood on the whole wordform, which may not be the same as the morphosyntactic head (Ananiadou, 1994, p. 1037)”

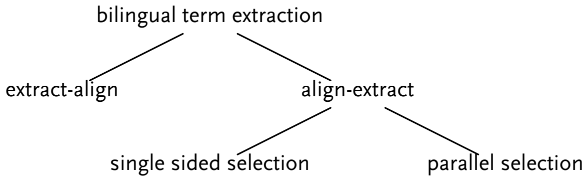

Figure 4.2: Approaches to bilingual term extraction

4.5 Bilingual Automatic Term Extraction

For this thesis we will be using the following definition of Bilingual ATE;

Bilingual Automatic Term Recognition (BATE): the task of extracting bilingual term pairs, where both source and target term candidates in a pair refer to the same concept

Currently, there are two main approaches to bilingual ATE. The most common approach is to perform monolingual ATE for each of the languages, source and target language, followed by an an alignment phase where the extracted terms in the source and target language are aligned or paired with each other e.g. Daille, Gaussier, and Langé (1994). This approach will be referred do as the *extract-align approach.*

We will refer to the second approach as the *align-extract approach.* In the align-extract approach, a pair of parallel texts are first word-aligned, followed by a pairwise extraction process. The extraction process can be performed in two ways:

1. select term candidate pairs based on one language (source or target), which we will refer to as *single sided selection*
2. select term candidate pairs based on both languages in parallel, which we will refer to as the *parallel selection*

Figure 4.2 gives an overview of the different BATE approaches and methods. The first align-extract method, single sided selection, makes it possible to extract terms in a second language as long as extraction algorithm for the first language is functional. In a sense, the single sided selection methods uses the first language as a kind of pivot language. Most align-extract approaches to bilingual term extraction use this method, e.g. Merkel and Foo (2007), Lefever, Macken, and Hoste (2009). The second variant of the align-extract approach results in dual constraints placed on the term candidate pairs. To the author's knowledge however, no research has been done using the parallel selection method.

4.6 Linguistically oriented approaches

In this section an overview of the linguistically oriented approaches is given. None of the approaches rely solely on linguistic information, as they also take into account some basic statistical data, e.g. the number of occurrences of a term candidate. However, the statistical measures used are fairly trivial compared to the measures used in section 4.7.

4.6.1 Part-of-speech-based approaches

The first term extraction approaches proposed and researched in the field of ATE were all based mostly on linguistic information. Several attempts at describing terms using a linguistic framework have been made. Justeson and Katz (1995) e.g. presents a study of a sample of English terms from four different domains (fiber optics, medicine, physics and mathematics, and psychology). The sample consisted of 200 terms from each of the four domains. The study revealed that depending on the domain, between 92.5% and 99.0% of the terms were noun phrases[6] (NPs). Of 800 terms, 35 non-NPs were found, of which 32 were adjectives and 3 were verbs. Further more, a majority of the terms, between 56.0% and 79.5%, were multi-word units.

Daille (1994), Daille et al. (1994) claim that most Multi-Word Units (MWUs) of greater length than two, are the result of a transformation of a MWU of length two. Daille (1994), Daille et al. (1994) present three different types of transformations which can produce new, and longer terms from a two word MWU. These are:

overcomposition *mechanical switch* + *keyboard* → *mechanical switch keyboard*

modification insertion of modifying adjectives or adverbs, or post-modification; e.g. *earth station* → *interfering earth station*, which in English is a insertion whereas in French, *station terienne* → *station terienne brouilleuse* is a post-modification.

coordination e.g. *package assembly* / *package disassembly* → *package assembly and disassembly*

The exact POS patterns are of course language-dependent and those which prove to be useful for English term extraction may, or may not work if used for e.g. French term extraction. Also, there may be differences depending on the domain from which the terms are to be extracted. Example of such variation can be found in the medical domain, where neoclassical term formation is common. Daille (1994) discusses common patterns which can form terms in the French language.

[6] The case may be that terms very often are NPs, but one must remember that this relation is not symmetrical, all NPs are *not* very often terms. In fact, only a small proportion of all NPs in even domain-specific texts are terms.

Linguistic approaches often use a combination of POS tagging (see subsection 4.2.5) together with POS patterns and stop-lists to extract term candidates from a corpus. POS patterns are usually described as regular expressions. Frantzi, Ananiadou, and Tsujii (1998) and others also call the use of POS patterns and stop-lists *linguistic filters* and Frantzi, Ananiadou, and Tsujii (1998) discriminate between two types of filters — *closed filters* and *open filters*. A closed filter is more "strict about which strings it permits" and have according to Frantzi, Ananiadou, and Tsujii (1998) a positive effect on precision, and a negative effect on recall. An open filter on the other hand permits more types of string and have a positive effect on recall, but a negative effect on precision. One example of a closed filter is `Noun+` which only allows noun sequences. Given the results presented in Justeson and Katz (1995), such a filter would have a higher precision than an open filter such as `((Adj|Noun)+|(Adj|Noun)*` which will include more false positives, but in turn have a higher recall.

Bourigault created the *LEXTER* system (Bourigault, 1992; Bourigault, Gonzalez-Mullier, & Gros, 1996) the purpose of which was to "give[s] out a list of likely terminological units, which are then passed on to an expert for validation". The system was built for the French language and used surface grammar analysis and a set of part-of speech patterns to pinpoint possible terminological units.

The *TERMS* system by Justeson and Katz (1995) uses a single regular expression pattern to find possible terminological units in a text. These possible candidates are then filtered based on frequency, keeping those above the frequency of two. The following regular expression was used; `((A|N)+|((A|N)(NP)?)(A|N))N`, where A is an adjective, N is a lexical noun (not a pronoun), and P is a preposition. In effect, the approach only extracts multi-word units. The frequency of a filtered multi-word unit is used as a cue to decide whether or not it should be selected as a term candidate. The frequency threshold is set manually after examining the possible term candidates.

The approach is based on term usage observations described by Justeson and Katz (1995). The main observations presented in the paper are that multi-word terms are often lexicalized noun phrases (NPs), which are not succeptible to variation in the form of omission and use of modifiers compared to non-lexicalized noun phrases. Lexicalized NPs are repeated (have a higher frequency) than non-lexicalized NPs.

The TERMS system in (Justeson & Katz, 1995) was evaluated by applying the system to three scientific papers and having the authors of the papers judge the extracted term candidates with regard to correctness. Recall was only evaluated on one of these three papers with the motivation that the task was to onerous. The type level precision in the three papers were 67%, 73% and 92% respectively. The recall of the system on the single evaluated paper was 37% on the type level.

4.6.2 A morphological approach

One motivation behind using linguistically based methods over purely statistical methods is that one may be interested not only in potential terms that have a high frequency, but also word-units which have a lower use frequency. Ananiadou (1994, p. 1034) argues that

> non-linguistic based techniques (statistical and probability based ones), while providing gross means of characterizing texts and measuring the behavior and content-bearing potential of words, are not refined enough for our purposes. In Terminology we are interested as much in word forms occurring with high frequency as in rare ones or those of the middle frequencies.

The approach proposed by Ananiadou (1994) was a morphologically oriented approach which was developed for the domain of Medicine (more specifically, Immunology). Ananiadou (1994, p. 1035) states that "Medical terminology relies heavily on Greek (mainly) and Latin neoclassical elements for the creation of terms". The proposed approach uses morphological analysis and a set of rules to determine whether a single word may or may not be a term.

Ananiadou (1994) provides a morphological description of medical terms, which centers around the use of Greek and Latin neoclassical elements for the creation of terms. Ananiadou proposes using four word structure categories (word, root, affix and comb) with a four level morphological model to classify morphological components of words. The last of the four categories, *comb* is proposed by Ananiadou to be used to hold neoclassical roots. These classifications can then be used to identify possible medical terminology that use neoclassical components. The four morphology levels are as follows:

1. Non-native compounding (neoclassical compounding)
2. Class I affixation
3. Class II affixation
4. Native compounding

Class I affixation deals with latinate morphology and Class II affixation deals with native morphology. An example analysis of the wordform 'glorious' yields `glory((cat noun))(level 1))` and `ous((cat suffix)(level 1))`. As a result of corpus analysis, suffixes were also given a annotation whether or not they were typically term forming.

4.7 Statistically enhanced approaches

Statistically enhanced approaches do not solely rely on statistical data. Rather the approaches often start with a linguistically informed selection of possible term candidates, i.e. single or multi-word units that are potential terms. These candidates have been selected using a linguistically informed approach as described in section 4.6. This set is then usually filtered by evaluating statistical metrics that are calculated for each term candidate.

4.7.1 Calculating statistics

The computational approach to term extraction can be classified as an application of corpus linguistics, and shares many of the statistical measures with that field. One way of describing the research on automatic term extraction that uses statistical measures, is to say that the goal is to find out how we can approximate our concept of what a term is, in terms of statistics.

There are different kinds of statistics which can be used to analyze potential term candidates. We have previously mentioned *unithood* (section 4.4) and *termhood* (section 4.3), and many of the statistical measures used can be directly associated with one or both of these concepts. There are however some measures, which cannot be directly associated with either termhood or unithood. This does not necessarily mean that they are of no use. One example of such a measure is the co-occurrence frequency of the components of a multi-word unit term candidate, which Wermter and Hahn (2006) observed performed similarly to statistical association measures such as the t-test and log-likelihood.

Apart from the actual measure used, different combinations of data sources can be used to obtain the statistical values for a specific potential term candidate. Table 4.2 lists different types of corpora which can be used to gather statistical data about possible terminological units in texts.

The *internal corpus* refers to the same corpus from which the word-units are extracted. *External corpus* refers to a corpus that does not contain the documents from which the phrases are extracted. External corpora are then differentiated in terms of domain. An example of an *external corpus from the same domain* is if e.g. term extraction is performed on documents 1–100 from the domain of Python programming and the documents 101–1000 are used to calculate the statistical scores. An *opposite* domain is a very dissimilar domain. In the case of documents 1–100 from the domain of *Python programming*, an *opposite* domain could for example be *book binding*. A *closely related domain* is a domain that is likely to share some concepts, in the case of Python programming, Ruby programming is a closely related domain. The notion of *other specialized domain(s)* is useful

Corpora
internal corpus
external corpus, same domain
external corpus, “opposite” domain
external corpus, closely related domain(s)
external corpus, other specialized domain(s)
external corpus, general language

Table 4.2: Different kinds of corpora used to calculate statistical scores

when e.g. doing a contrastive comparison between term distribution in the internal corpus, domain-specific texts in general, and finally, an *external, general language corpus* refers to e.g. non-technical corpora such as the British National Corpus, The Brown Corpus, etc. Such a contrastive comparison would for example find out that the term “application” is general patent document term, rather than a term belonging to a specific patent document subclass, or being a general language word.

4.7.2 Terms as lexicalized noun phrases

In Justeson and Katz (1995), terms are described as lexicalized noun phrases. Justeson and Katz (1995) write: “Terminological noun phrases (NPs) differ from other NPs because they are Lexical – they are distinctive entities requiring inclusion in the lexicon because their meanings are not unambiguosly derivable from the meanings of the words that compose them”. Justeson and Katz also continue to state that lexical NPs are “subject to a much more restricted range and extent of modifier variation on repeated references to the entities they designate, than are nonlexical NPs”. An appropriate statistical description of terms, given the findings of Justeson and Katz (1995) would be that “there is less variation among terms than non-terms”. This observation is key to many statistical measures which try to quantify termhood.

4.7.3 Mutual Information and the Loglike and Φ^2 coefficients

Daille (1994) presents a monolingual approach which uses part-of-speech patterns (POS patterns) that describe multi-word units with exactly two *main items*. Main items are defined as nouns, adjectives, adverbs, etc. but not prepositions or determiners according to Daille. Using this definition, the length of a word-unit is measured in the number of main items. Daille uses NP POS patterns to first extract a set of word-units. These word-units are all of length two (i.e. two main items) and the link between these two items is scored using the statistical measures *Mutual Information*, the *Loglike coefficient*, and the Φ^2 coefficient. The details of these scores are described in equations 4.1, 4.2, and 4.3. The units of analysis in Daille (1994) consists of exactly two main items $t = (i, j)$, the frequency f can be counted for four combinations, $a = f_{ij}$, $b = f_{ij'}$, $c = f_{i'j}$, and $d = f_{i'j'}$, where $i' \neq i$ and $j' \neq j$.

$$MI = log_2 \frac{a}{(a+b)+(a+c)} \tag{4.1}$$

$$\begin{aligned} Loglike = {} & a\, log(a) + b\, log(b) + c\, log(c) + d\, log(d) \\ & - (a+b) log(a+b) - (a+c) log(a+c) \\ & - (b+d) log(b+d) - (b+d) log(c+d) \\ & + (a+b+c+d) log(a+b+c+d) \end{aligned} \tag{4.2}$$

$$\Phi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)} \tag{4.3}$$

The method in essence uses linguistic information to extract word-units that have a part-of-speech pattern which is common among terms (noun phrases). Then, each extracted sequence, is examined with regard to stability in the corpus, i.e. unithood.

Daille (1994) briefly discusses the order of using statistics in combination with a linguistic approach — "We now face a choice: we can either isolate collocations using statistics and then apply linguistic filters, or apply linguistic filters and then statistics". The approach chosen was the latter, application of linguistic filters followed by statistics, an order which is used by all statistically enhanced approaches reviewed here. A statistics first approach would have to analyze all n-grams of size $1 <= n <= k$ where k is the maximum length of multi-word terms we want to consider. Application of linguistic filters first greatly reduces the number of items analysed, without reducing recall given the correct set of linguistic filters.

4.7.4 C-Value/NC-value

The C-Value/NC-Value (Frantzi, Ananiadou, & Tsujii, 1998; Frantzi, Ananiadou, & Miama, 2000) is a multi-word term extraction method which uses both linguistic and statistical information. First a set of POS patterns combined with a stop-list are used to select word-units from the tagged corpus. Each extracted word-unit is then scored using the C-value metric

$T_{realtime}$	$T_{floatingpoint}$
real time clock	floating point arithmetic
real time expert system	floating point constant
real time image generation	floating point operation
real time output	floating point routine
real time systems	

Table 4.3: Examples of nested terms

which is described by equation 4.4, and measures *termhood* according to Frantzi, Ananiadou, and Tsujii. Frantzi, Ananiadou, and Tsujii also state that bigrams are the smallest word unit size for which the C-value is calculated meaning that the C-value is not applicable to single word-unit term extraction.

The concept of *nested* terms is important here, and Frantzi, Ananiadou, and Tsujii define them as “We call nested terms those that appear within other longer terms, and may or may not appear by themselves in the corpus.”. Multi-word units have sub-units which may or may not be terms. Table 4.3 shows an example of nested occurrences of the terms *real time* and *floating point*, taken from Frantzi, Ananiadou, and Tsujii (1998). With the exception of *expert system*, the other nested word units in the examples are not terms, i.e. *time clock, time expert system, time image generation, image generation, time output, time systems, point arithmetic, point constant, point operation, point routine*. What the C-value equation does is that it takes into special account, candidate terms which are present in other candidate terms, i.e. candidate terms which have nested instances.

In the case of non-nested terms the C-value takes into account the length of the term candidate and the number of occurrences. In the case of nested term candidates, the C-value subtracts the average number occurrences of a as a nested term among the POS filtered possible term candidates. This means that if a occurs as a nested term candidate equally as often, or more times than it does as an independent term candidate, it will have a C-value of ≤ 0. The lower the average occurrence as a nested term candidate, the closer the C-value of nested term candidates approach that of un-nested term candidates which follows the frequency of a.

$$C - value(a) = \begin{cases} \log_2 |a| \cdot f(a) & \text{if } a \text{ is not nested} \\ \log_2 |a| \cdot \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right) & \text{otherwise} \end{cases} \tag{4.4}$$

In equation 4.4, the following notation is used

- a is the candidate string
- $\log_2 |a|$ is the number of words in a
- $f(a)$ is the frequency of a in the corpus
- T_a is the set of extracted candidate terms that contain a.
- b is a candidate term in T_a
- $P(T_a)$ is the number of candidate terms that contain a
- $f(b)$ is the frequency of the candidate term b that contains a

The C-value is used to calculate the NC-value, which also factors in "context" into the term extraction algorithm. Here context seems to mean the previous word, if the previous word was a noun, an adjective, or a verb. The C-value extracted list of words is re-ranked using information gained by looking at this "context". The motivation comes from Sager's (Sager, Dungworth, & McDonald, 1980) concept of *extended terms*. The general idea behind *extended terms* is one compatible with the discoveries of (Justeson & Katz, 1995); that terms are strict about the modifiers that they accept. Frantzi, Ananiadou, and Tsujii has interpreted and formalized this concept as follows.

$$Weight(w) = \frac{t(w)}{n} \tag{4.5}$$

The $Weight$ in equation 4.5 is the weight assigned to the context word w which is a noun, verb or an adjective. $t(w)$ is the number of terms the word w appears with, and n is the total number of terms considered. Frantzi, Ananiadou, and Tsujii argues that the weight is expressed as a probability; "the probability that the word w might be a term context word", and a way of interpreting the equation is to say that the word w can be taken as evidence that the following word is a term. The weight is combined with the C-value to construct the NC-value, described in equation 4.6.

$$NC\text{-}value(a) = 0.8 \cdot C\text{-}value(a) + 0.2 \cdot \sum_{b \in C_a} f_a(b) weight(b) \tag{4.6}$$

The second part in the equation, $\sum_{b \in C_a} f_a(b) weight(b)$, needs some explanation. C_a is the set of context words b for a, and f_a is the frequency of a with b as a context word. The sum then adds the frequency of the context words multiplied with their weight, which can be interpreted as a value which describes the variability of the context words used with the term a.

Although the C-value/NC-value has been proven to perform quite well (Zhang et al., 2008), I would like to argue that C-value as described in Frantzi, Ananiadou, and Tsujii (1998) does not measure *termhood* as defined by Kageura and Umino (1996). Rather, the C-value measures term candidate independence, or *unithood*, also defined in Kageura and Umino (1996). This is also the opinion voiced in Nakagawa and Mori (2002). The sum in the NC-value (equation 4.5), however can be argued to relate to the termhood of a term candidate as it is derived from an observed linguistic property of terms; the lack of variation with regard to their modifiers, as observed by both Sager et al. (1980) and Justeson and Katz (1995) (see also sub-section 4.7.2).

4.7.5 Paradigmatic Modifiability of Terms

The *Paradigmatic Modifiability of Terms* approach by Wermter and Hahn (2005) has some similarities with C/NC-value approach by Frantzi, Ananiadou, and Tsujii. The similarity lies in leveraging the observation that terms tend not to vary much (also observed by Justeson and Katz (1995), see sub-section 4.7.2). The approach by Wermter and Hahn builds on the probability that a token in a word-unit cannot be replaced by an alternative token. Wermter and Hahn defines each position in a multi-word unit as a slot, where the token in a particular slot can be replaced by other tokens. The n-gram "*long terminal repeat*" e.g. has three slots where slot 1 is filled with "long", slot 2 by "terminal", and slot 3 by "repeat". What Wermter and Hahn want to express is a score which describes how common variations of a specific n-gram are. For a n-gram of length n, they substitute up to k tokens with a wild card, where $1 \leq k \leq n$. For each k, $n-1$ possible wild card permutations, or *selections*, sel_i are possible. The number of occurrences for each permutation is compared with the number of occurrences of the original n-gram, providing the modifiability of that particular selection, mod_{sel_i} (see equation 4.7). The modifiability for a particular value of k, the k-modifiability is calculated (defined in 4.8), which is then used to calculate the paradigmatic modifiability, $P\text{-}Mod$ (defined in 4.9).

$$mod_{sel_i} = \frac{freq(n\text{-}gram)}{freq(sel_i)} \tag{4.7}$$

$$mod_k = \prod_{i=1}^{n-1} mod_{sel_i} \tag{4.8}$$

$$P\text{-}Mod(n\text{-}gram) = \prod_{k=1}^{n} mod_k(n\text{-}gram) \tag{4.9}$$

Wermter and Hahn (2005) performed an evaluation of the *P-Mod* measure compared to the *C-value* and *t-test* score on term candidates extracted from biomedical texts and found that the *P-Mod* measure performed best.

4.8 Approaches using contrastive data

The reviewed statistically enhanced approaches in section 4.7 all used data based on an internal corpus, i.e. ranking extracted term candidates based on statistics from the corpus from which the terms were extracted. Contrastive approaches use both the internal corpus and an external corpus to feed statistics into the algorithm that ranks the term candidates. The motivation comes from the fact that terms are domain-specific, and are used as such primarily in domain-specific texts. This means that the use of these domain-specific terms should have a different usage pattern in e.g. a general language corpus compared with the internal, domain-specific corpus. This contrast in use is what contrastive approaches try to capitalize on — if a word-unit is domain-specific, there should be a difference in its use-pattern in a domain-specific corpus, in contrast to use in a balanced, general language corpus, or an out of domain corpus (e.g. a domain-specific corpus from another domain).

4.8.1 Weirdness

The *Weirdness* measure (Ahmad, Gillam, & Tostevin, 1999) was developed within the field of IR. The method was developed for use in TREC 8[7], and Ahmad et al. begin by showing that there are distributional differences between the TREC-8 corpus[8] and the British National Corpora (BNC) (BNC Consortium, 2001). The main difference pointed out is the frequency of open class words[9]. By removing the most frequently occuring words from BNC from subsets of the TREC-8 corpus and then compiling frequency ordered list from the TREC-8 subcorpora, a much more domain-specific set of words was produced. This lead to the formulation of the weirdness measure, described in equation 4.10.

$$Weirdness = \frac{w_s / t_s}{w_g / t_g} \tag{4.10}$$

[7] The Text REtrieval Conference (TREC) is an on-going series of workshops focusing on a list of different information retrieval (IR) research areas, or tracks. TREC is co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense and was started in 1992.

[8] Ahmad et al. categorize the TREC-8 corpus as a specialized text due to its composition of only financial and political news texts

[9] Word classes that accept new additions through e.g coining and derivation. Open word classes in the English language are nouns, verbs (but not auxillary verbs), adjectives, adverbs, and interjections

In equation 4.10, the following notation is used

- w_s is the word frequency in the domain-specific language corpus
- w_g is the word frequency in the general language corpus
- t_s is the total number of words in the domain-specific language corpus
- t_g is the total number of words in the general language corpus

The Weirdness measure is in other words the ratio between the normalized word frequency in a domain-specific corpus compared to the normalized word frequency in a general language corpus.

4.8.2 Contrastive weight

The contrastive weight method by Basili, Moschitti, and Pazienza (2001) builds on previous research that uses both linguistic and statistical filtering. The problem as described by Basili et al. is that when performing term extraction using previously described approaches (see sections 4.6, 4.7), the extracted term candidates include a large number of false positives. Some of these false positives fall under the category of *general language collocations*. Basili et al. continue to argue that it should be expected that these collocations are evenly distributed over all corpora, both domain-specific, and non-domain-specific. The *contrastive weight* metric is the product of a *domain-specific word-unit score*, and a cross-domain, or *general language word-unit score*. The equation supplying the general word-unit score, the *Inverse Word Frequency (IWF)* can be found in equation 4.12, where N is the "size of the corpus obtained by summing up contributions (i.e. frequencies) from all candidates in all domains". F_t (see 4.11) is the cumulative corpus frequency (in all j domains) for a term t. The IWF differs from the standard Inverse Document Frequency (IDF)[10] by counting words rather than documents.

The domain-specific component is calculated as $log(f_t^i)$ where f_t^i is the frequency of the word-unit t in the domain-specific corpus i. These components are then combined to form the *Contrastive Weight*, described in equation 4.13

$$F_t = \sum_j f_t^j \tag{4.11}$$

$$IWF(t) = log(\frac{N}{F_t}) \tag{4.12}$$

[10] $IDF = log_2 \frac{D}{df_w}$ where df_w is the number of documents where the word w occurs, and D is the total number of documents

$$w_t^i = log(f_t^i) \cdot IWF(t) \tag{4.13}$$

The IWF basically treats domains as documents if you compare it to regular IDF, i.e. a term that only occurs frequently in few domains will have a higher score than a term that occurs frequently in all domains. Finally, analogous to tf-idf weighting in IR, the term frequency for a single domain is multiplied with the IWF.

For multi-word units, Basili et al. apply the ranking only to the heads of the units. The evaluation performed by Basili et al. compares the contrastive weight metric's performance compared to domain frequency and reports this using the metric $F = \frac{1}{0{,}5/p+0{,}5/r}$ where p is the precision and r is the recall. The conclusion drawn by Basili et al. is that using contrastive weights is better for ranking term candidates than domain frequency.

4.8.3 TermExtractor

The TermExtractor system (Sclano & Velardi, 2007) is a web-based ATE system that allows multiple users to validate extracted term candidates using its web interface. TermExtractor uses POS patterns to extract a set of possible terminological units that are then filtered using multiple word lists. The remaining term candidates are then ranked using a combination of three statistical measures which are described below.

Domain Relevance describes how relevant a word unit is in the domain D_i, compared to domains D_j. The Domain Relevance for the term t in the domain D_i, $DR_{D_i}(t)$ is described in equation 4.14 where $\hat{P}$ is the Expected Value of the probability P.

$$DR_{D_i}(t) = \frac{\hat{P}(t/D_i)}{max\left(\hat{P}\left(t/D_j\right)\right)} = \frac{freq(t, D_i)}{max\left(freq\left(t, D_j\right)\right)} \tag{4.14}$$

Domain Consensus attempts measure the *consensus* a term has within a group of documents. If the term t has an even probability distribution across the documents $d_k \in D_i$, it is said to have a high consensus. The Domain Consensus for t in domain D_i is described in equation 4.15 where f_n is the normalized frequency.

$$\begin{aligned} DC_{D_i}(t) &= -\sum_{d_k \in D_i} \hat{P}\left(t/d_k\right) log\left(\hat{P}\left(t/d_k\right)\right) \\ &= -\sum_{d_k \in D_i} f_n\left(t, d_k\right) log\left(f_n\left(t, d_k\right)\right) \end{aligned} \tag{4.15}$$

Lexical Cohesion evaluates the degree of cohesion in a multi-word unit, or in practice, the ratio between using words in a collocation compared to their un-collocated use. The method is described in equation 4.16. There, w_j are the words in the multi-word unit t, and n is the length of t in number of words.

$$LC_{D_i}(t) = \frac{n \cdot freq\left(t, D_i\right) \cdot log\left(freq\left(t, D_i\right)\right)}{\sum_j freq\left(w_j, D_i\right)} \tag{4.16}$$

The three measures are weighted and then combined to produce a single score referred to by Sclano and Velardi as the *weight* of a term, $w(t, D_i)$, and is calculated according to equation 4.17.

$$w(t, D_i) = \alpha \cdot DR + \beta \cdot DC + \gamma \cdot LC \tag{4.17}$$

The coefficients α, β, γ can be set by the user, but the default value is $\alpha = \beta = \gamma = \frac{1}{3}$. As can be seen, the TermExtractor system includes statistical measures that try to capture all key components of termhood and unithood (see sections 4.3 and 4.4). The precision of the system however is around 60%, which is a bit modest. The system however brings more to the table than being a ATE algorithm run on pre-processed text files in a command line environment. The system can accommodate the whole term extraction process, going documents to validated terms, which may in many real life scenarios outweigh the advantages of using a system which performs better, but requires more configuration and manual data logistics.

4.9 Evaluation

A term extraction system can be evaluated on different levels and with different constraints imposed. Most term extraction research uses precision and recall to describe the performance of systems. However, what this precision and recall is based on varies. For example, Justeson and Katz (1995) evaluate their method by applying it to academic papers and asking domain experts to judge whether the extracted terms where domain-specific terms or not. The domain experts also performed manual term extraction on the texts to evaluate recall. On the other hand, the comparative study by Zhang et al. (2008) is performed on systems which extract terms from a large corpus. In this evaluation, recall is not evaluated and precision is measured in tiers, i.e. precision for the top 100, 1k, 5k, 10k and 20k extracted terms, also called *n-best lists*. Zhang et al. (2008) is not the first or only one to do this. Others include Frantzi and Ananiadou (1999) and Sclano and Velardi (2007).

5 Methods used

In this chapter, background to the methods used in the published word is given in a less terse manner than permitted by the format of the published papers.

5.1 Machine Learning

Some of the contributions in this thesis concern novel use of machine learning techniques in the field of term extraction. This section will give a brief overview the method involved when using Machine Learning.

Machine Learning (ML) may be defined as to include “any computer program that improves its performance at some task through experience” (Mitchell, 1997, p. 2). Machine learning can either be **1)** *supervised learning*, i.e. where software is exposed to an example, and the output from the software is evaluated and fed back to the software, or **2)** *unsupervised learning*, where it is up to the machine to fend for itself with no feedback provided by any user, or **3)** *reinforcement learning* where the machine is not presented with feedback depending on its reaction to specific examples, but rather to the actions it decides to perform. The method employed in this thesis is supervised learning.

Supervised learning uses example data to build a model of the data which can be used to predict, classify or score data. This phase is referred to as the *training phase*. The training phase is followed by a *testing phase* where the performance of the trained model is evaluated. The annotated data used when applying supervised machine learning is usually divided into at least two partitions, one used for training, and the other held back for the testing phase, so that the evaluation can use previously unseen examples.

5.1.1 Features

Features in machine learning are the components which constitute the individual examples presented to the learning system. For example, in the case of image classification, more specifically, recognition of a low resolution black and white image, the features may consist of the state of each individual pixel in the image.

In the case of ATE, features may be POS value, word-unit frequency in a corpus or the output from a termhood ranking metric to name a few. However, depending on the ML framework used, feature values may need to be encoded using e.g. only numeric values (as with a Support Vector Machine (SVM)).

Training

The phase where the system is automatically configured, i.e. where it learns to model the problem, is called the training phase. For supervised learning, this involves exposing the system to training instances where events are observed by the system, each event a collection of features and the right answer for each event given. During the training phase, the ML algorithm uses the annotated examples to build its prediction model. This model is then tested during the testing phase. An example, how a training instance used in Foo and Merkel (2010) was represented can be seen in table 5.1.

Testing

After the training phase the model is tested using the held back data. In the case of classification, the test result is usually presented as a confusion matrix which contains the number of true positives, true negatives, false positives, and false negatives. These values can then be used to calculate a precision score and a recall score, which can be combined to create an F-score.

5.1.2 Systems used in the published papers

Two different ML frameworks are used in the work presented in this thesis (Foo & Merkel, 2010; Foo, 2011). The first system, Ripper (Cohen & Singer, 1999), is a *rule induction* ML system that produces human readable rules. The second is the SVM framework. SVMs were introduced by Boser, Guyon, and Vapnik (1992) and are linear classifiers that can use kernels to also classify non-linear data. The particular implementation used in Foo (2011) was the LibSVM system (Chang & Lin, 2011).

5.2 n-gram Language Models

Another contribution of this thesis is the novel approach of using contrastive n-gram Language Models to determine termhood. This section will give an overview of the construction and use of such language models.

feature value	feature name	feature description
`låsorgan`	n-gram	the n-gram the features relate to
`n`	POS	part-of-speech tag
`pl-nom`	msd	morpho-syntactic description
`obj`	func	grammatical function
`null`	sem	semantic information
$1.39514 \cdot 10^{-5}$	nfreq	normalized n-gram frequency in text
`0.0`	zeroprobs	number of tokens with zero probability in given the language model
`6.04298`	logprob	the logistic probability value, ignoring unknown words and tokens
`1050.73`	ppl1	the geometric average of $1/probability$ of each token, i.e. perplexity
`1104040.0`	ppl2	the average perplexity per word
`yes`	key	positive or negative example

Table 5.1: Example training instance with feature names and feature descriptions. The actual training instance only consists of the actual feature values in the first column.

Language modeling is the task of predicting the next word in a sequence, given the previous words (Manning & Schütze, 1999). Language models are used to aid various language recognition tasks such as speech recognition, optical character recognition, handwriting recognition, and machine translation.

The n-gram language model approach simplifies the language prediction task from requiring to know *all* previous words to predict the next word in a sequence, to a *limited number* of previous words. This is what is referred to as the Markov Assumption.

n-gram language models are built by examining a text corpus and counting the number of occurrences of token sequences up to length n. The unit level can vary depending on the application. Applications such as handwriting recognition may for example benefit from a character level language model, rather then a word level language model. In the context of term extraction we are only concerned with word level language models. A word level unigram model consists of single word data, a bigram model consists of two word units, trigram models use three words, and four-gram models use four words.

The resulting n-gram model can be described as a conditional probabilistic model where the probability of the next word is conditioned by the previous n-1 words.

5.2.1 Smoothing

Since natural language is productive, any sample based statistical model, such an n-gram model, will suffer from data sparseness. This means that no matter how large the sample is from which we build our n-gram model, we can always encounter unknown word sequences, which either contain previously unknown words, or a previously un-encountered sequence of known words.

To be able to assign a probability to such cases, n-gram models employ smoothing, which reserves some of the total probability mass, and assigns this to events which are rare or unseen. Several smoothing techniques exist which vary in complexity and performance. The least complex smoothing method is additive smoothing, where a constant δ is added to the number of occurrences of each token, effectively raising the probability of unknown tokens by δ in proportion to the total number of occurrences (Manning & Schütze, 1999). The language models used in the published work were built using the SRILM toolkit (Stolcke, 2002). The SRILM toolkit implements several smoothing methods, among others, the modified Kneser-Ney smoothing algorithm (Chen & Goodman, 1998), which was the selected smoothing algorithm for the presented work.

6 Overview of publications and results

This chapter gives an overview of the contributions of the five publications included in this thesis.

6.1 Computer aided term bank creation and standardization

Foo, J. & Merkel, M. (2010). Computer aided term bank creation and standardization: Building standardized term banks through automated term extraction and advanced editing tools. In M. Thelen & F. Steurs (Eds.), *Terminology in Everyday Life* (13, pp. 163–180). Terminology and Lexicography Research and Practice. John Benjamins Publishing Company

The first paper in this thesis describes a semi-automatic bilingual terminology creation process applied to patent texts. The initial version of the paper was first presented at the Terminology Congress 2006 — Terminology and Society, and later revised and published as a book chapter (Foo & Merkel, 2010).

In the paper we outline a process that starts with a bilingual, parallel, patent document collection and ends with the a bilingual term list in OLIF format. The following steps are described:

1. Grammatical analysis and interactive training of word alignment
2. Full text automatic bilingual alignment
3. Type aggregation and database storage
4. Term candidate detection, filtering, categorization and ranking
5. Revision and editing
6. Multi-user support and export to various formats

After the processes leading to and performing the bilingual word-align ment are completed, aligned word-units are extracted and stored in a database, together with contextual and relevant statistics. The data is also structured so that word form variants are grouped and associated with a single lemma form. Term candidates are selected using POS patterns and then ranked using the Q-value metric (see 6.3). Using the IView software, terminologists and domain experts can then process the term candidates, before exporting them for actual use.

6.2 Automatic Extraction and Manual Validation of Hierarchical Patent Terminology

Merkel, M., Foo, J., Andersson, M., Edholm, L., Gidlund, M., & Åsberg, S. (2009). Automatic Extraction and Manual Validation of Hierarchical Patent Terminology. In B. Nistrup Madsen & H. Erdman Thomsen (Eds.), *NORDTERM 16. Ontologier og taksonomier. Rapport fra NORDTERM 2009*. København, Danmark: Copenhagen Business School

The second paper is a continuation of the work presented in Foo and Merkel (2010). This paper details the results and slightly modified processes used in a full scale bilingual patent term extraction project. The term extraction project progressed for the duration of 8 months, resulting in 750,000 term candidates which were processed manually. The final result was 181,000 unique validated term pairs which were to be used in a machine translation system for patent texts.

6.3 Terminology Extraction and Term Ranking for Standardizing Term Banks

Merkel, M. & Foo, J. (2007). Terminology extraction and term ranking for standardizing term banks. In J. Nivre, H.-J. Kaalep, K. Muischnek, & M. Koit (Eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*. (Pp. 349–354). Tartu: University of Tartu. HDL: 10062/2602

The third paper describes how the Q-value, a metric which measures *translation stability*, can be used to successfully rank bilingual word-unit pairs in a way which correlates with how likely it is that they are valid bilingual term pairs. The Q-value metric is presented in equation 6.1, where s is the source word-unit of the extracted bilingual pair, and t is the target word-unit of the extracted bilingual pair. $count(s, t)$ is the frequency of this pair in the aligned data. n_s is the number of different source word-units s_i which have the translation t, and n_t is the number of target word-units t_j which have the source word-unit s.

$$Q - value = \frac{count(s,t)}{n_s + n_t} \tag{6.1}$$

6.4 Using machine learning to perform automatic term recognition

Foo, J. & Merkel, M. (2010). Using machine learning to perform automatic term recognition. In N. Bel, B. Daille, & A. Vasiljevs (Eds.), *Proceedings of the Workshop on Methods for automatic acquisition of Language Resources and their evaluation methods held in conjunction with the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 49–54)

The fourth paper contributing to this thesis was presented at the workshop for "Methods for automatic acquisition of Language Resources and their evaluation methods" at LREC2010 (Foo & Merkel, 2010). This paper presents a novel term extraction method using Ripper (Cohen, 1995), a rule-induction learning system which uses supervised learning to produce human readable rules which can be used to classify data. Ripper was applied to Swedish patent texts using manually validated terms as to learn how to differentiate non-terms from potential terms. Skewness of training data is a problem in term extraction due to the low proportion of true terms among the n-grams of a corpus. In this paper we vary the composition of the training data to find how this effects the performance of the generated rules. The results were rules which had a maximum precision output of 78% (with a recall of 69.28%), and an optimum precision/recall of 58.86%/100.0% for single-word-units (unigrams). The results for two-word-units (bi-grams) was not usable. The reason for this was the small amount of training data available and the lack of two-word-unit terms in Swedish, as Swedish is a compounding language (1.6% of the terms were two-word-units in Swedish).

6.5 Exploring termhood using language models

Foo, J. (2011). Exploring termhood using language models. In T. Gornostay & A. Vasiljevs (Eds.), *Proceedings of the NODALIDA 2011 Workshop Creation, Harmonization and Application of Terminology Resources (CHAT2010)* (Vol. 12, pp. 32–35). NEALT Proceedings Series. Northern European Association for Language Technology (NEALT). HDL: 10062/17276

The fifth and final paper contributing to this thesis was presented at the workshop on Creation, Harmonization and Application of Terminology Resources (CHAT2011) at the 18th Nordic Conference of Computational Linguistics (NODALIDA2011) (Foo, 2011). In this paper, a novel approach using SVM machine learning and contrastive language models, was used to identify possible terminological word-units among word-sequences extracted form patent texts using linguistic patterns (POS-patterns).

The word-units were annotated using statistical data from the patent texts, as well as probability scores calculated using two n-gram language models. One model was built using the patent texts, the second built using a general language corpora (BNC). The experiment resulted in an SVM which could identify terms with a precision of 66.4% and a recall of 88.0%.

7 Discussion

The published book chapter (Foo & Merkel, 2010) which is part of this thesis describes one of the first projects I was involved in within the field of Computational Terminology Management and was presented at the 2006 Terminology Congress held in Antwerp. Attending this conference made me conscious about various issues connected to terminology management in general, and helped me put the work we had done in perspective.

One aspect of terminology management that I was made aware of at this conference, and one that I have continued to appreciate during my research, is that the practical need for terminological resources and terminology management is very heterogeneous. The medical domain for example contains an enormous wealth of existing terminology and the terminology is laden with the history of the evolution of medical science. This historical aspect makes prescribing new terms and deprecating old ones rather complex and not without resistance. Other specialized domains, for example an emerging technical field such as tablet computing, are much smaller and texts available as source material for terminology work are both less in number and less mature/stable.

These two examples show some of the variation that can be found regarding the properties of different domains where terminology management may be needed. Terminologists and domain experts have dealt with the problems associated with these domain properties, but the field of Computational Terminology Management (CTM) has only started to map the landscape.

The research presented in this thesis is in a sense twofold. One part concerns the use of a bilingual, corpus driven, semi-automatic terminology extraction work-flow using specialized software. The second part concerns how machine learning can be used in the field of Automatic Term Extraction (ATE). In this chapter I wish to describe the link between these two parts, and why both are important for future research within CTM.

7.1 The relationship between monolingual and bilingual ATE

The first three papers deal with bilingual (English-Swedish) term extraction performed using the align-extract approach and mainly concern practical methodological issues rather than technical ATE issues. The term candidate pair selection model used in these papers is based on a single language (see 4.5. In other words, the extraction algorithm employed is technically monolingual, but applied to bilingual, word-aligned data. Using this approach, research done on monolingual term extraction can be immediately applied to bilingual term extraction. This also means that provided that parallel, bilingual data exists, it is possible to apply ATE to languages for which no monolingual ATE system is configured.

Bilingual data differs from monolingual data with regard to one central and fundamental property — it is bilingual, which also means that all annotations can also be bilingual, increasing the amount of information we have by 100%. For statistical methods, more data often means better results. However, very little research on the leverage bilingual data can provide over only having monolingual data has been done.

7.2 Multilingual ATE

Although some research claims to perform multilingual term extraction, no current research has been done on term extraction from more than two languages *in parallel*. Tufis, Barbu, and Ion (2004) perform extraction on multiple language pairs, but the extraction is always performed pairwise. What multilingual term extraction means however is a matter of interpretation. There are at least four interpretations of what multilingual can refer to:

1. the method is not language specific, i.e. applying the same extraction method to several languages, e.g. Hippisley, Cheng, and Ahmad (2005)
2. the output resource contains sections in different languages, e.g. packaging several independent resources in different languages together
3. the output units contain equivalents in more than one language, e.g. a bilingual dictionary
4. the extraction method examines more than one language at the time

In my opinion, the descriptor “multilingual” should only be applied when *three or more* languages are involved. If two languages are involved, the term *bilingual* is much more precise. Furthermore, *multilingual term extraction* should be reserved for term extraction methods that take three or more languages into account during the extraction process. If the extraction method is monolingual or bilingual and the output contains three or more languages, it is more precise to describe the process as monolingual/bilingual term extraction with multilingual output. The descriptor *multilingual* should not be attributed to resources only containing independent sections in three or more languages.

7.3 Evaluating ATE performance

A problem in the current field of term extraction is the lack of standardized test sets. The result is that different researchers use different test sets which makes comparing the performance of different systems difficult. The method used to test the performance may also differ to some degree, further complicating the matter. Comparative studies, such as Zhang et al. (2008) have provided valuable information on performance of methods in comparable testing environments. Due to the lack of available gold standard resources and standardized evaluation methods, it is difficult to evaluate the performance of a new method compared with a previous one. The lack of freely available implementations of the various methods also pose an obstacle to comparing results.

7.3.1 What should be evaluated?

Since ATE is not a single task, but rather a pipeline of tasks, there are many possibilities to what can be evaluated. Also, given that there are many uses for the output from a ATE process, there are also many ways to evaluate, i.e. different uses may consider different aspects to be more or less important.

When using ATE methods to extract the most important terms in a text, recall is not that important. However, when using ATE to create a complete terminology, recall is of the utmost importance. In the comparison performed by Zhang et al. (2008), the evaluation provided is given as precision for the n first extracted units. Such a comparison is useful when comparing a limited set of methods in a paper, but as a contribution to the research field in general, this is less useful, especially when the exact corpus and term list used to determine precision is not available to other researchers.

Also, as with all complex systems, the final result is the combined effort of all participating parts. With a system that uses a pipeline approach, the results can never be better than the weakest link (or pipe). For example, in a statistically enhanced ATE system, the recall of the final output, cannot rise above that which is possible with the results of linguistic filtering.

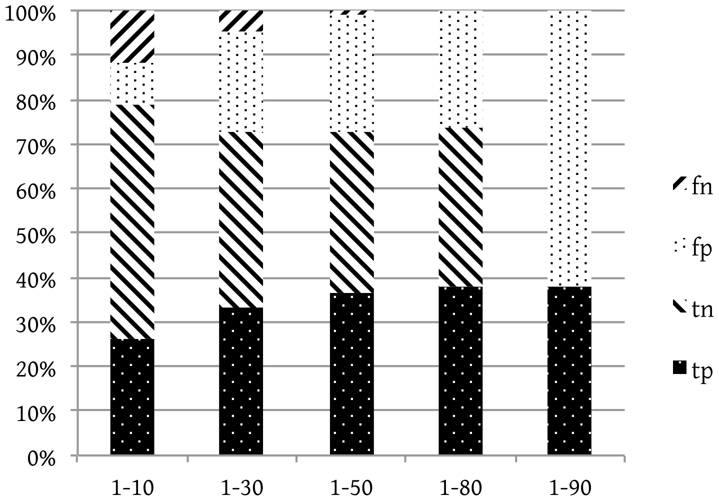

Figure 7.1: Bar chart of the series of confusion matrices for Swedish single word term extraction in sub-domain A42B (Foo & Merkel, 2010)

The precision may increase in later steps, but the recall is decided by the output from the initial linguistic filtering — which in turn depends on the tagger used to annotate the texts. An ideal evaluation method for all complex systems would be a method which illuminates the contributions of each subsystem, but as one can imagine, such an evaluation will also be very tedious.

Precision and recall

In the case of term extraction the most obvious facets of evaluation are precision and recall. However, it is necessary to also evaluate the number of false and true positives, and the number of false and true negatives. If the bulk of the precision is derived from true negatives, the method in itself is not productive enough to be of use. Also, false negatives are much less desirable then false positives, as false negatives reduce the recall of the system. In a setting of term candidate production, a higher recall with a reduced precision is more useful than a higher precision with a reduced recall.

For example in Foo and Merkel (2010), an experiment to examine the performance of Swedish term extraction rules which were generated using Ripper, a rule induction machine learning system, was conducted. The test results of the single word extraction rules can be seen in figures 7.1 and 7.2. As mentioned earlier, the "yield" of ATE is the number or percentage of true positives and the cost of precision is the amount of false

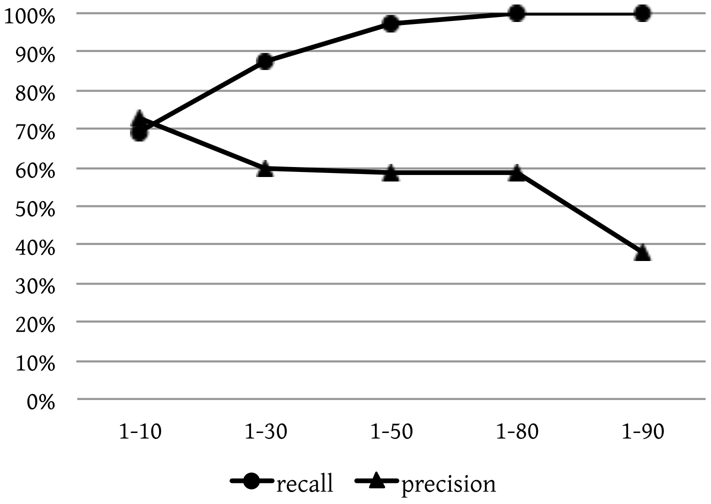

Figure 7.2: Line diagram of precision and recall for Swedish single word term extraction in sub-domain A42B (Foo & Merkel, 2010)

negatives, i.e. units which are terms but are discarded as errors.

7.4 Capturing Termhood

As described in the background (section 4.3), *termhood* is a problematic measure. What I mean by that statement is that the definition given by Kageura Kageura and Umino (1996), "The degree to which a stable lexical unit is related to some domain-specific concepts.", describes a constructed property which in the ideal case also can be measured and quantified, or at least detected. Whether or not it is actually possible to produce a 100% accurate termhood measure remains to be seen.

To this date, there has been no way to accurately define what a term is, and even domain experts can differ in opinion. No necessary and sufficient condition has been found which can be used to determine whether or not a specific single or multi-word unit is a term in a specific domain. All the properties (both quantitative and qualitative) of terms that are mentioned in conjunction with the reviewed methods, are *neither necessary, nor sufficient*.

For example, it is true that terms often are noun phrases, but not all noun phrases are terms, nor are all terms noun phrases. Likewise, it is true that high frequency noun phrases are more likely to be terms in technical texts, but all high frequency noun phrases are not necessarily terms. An even more subtle question one may ask is if e.g. noun phrases which occur significantly more often in certain domain compared to another domain by necessity are terms? A terminologist's answer would be "No.", as though a term candidate may exhibit domain-specific distribution compared with a general language corpus, the term candidate may actually belong to a parent domain or a related domain.

There is also another side to the definition of termhood which is in most cases not discussed when presenting various termhood measurements; to know whether or not a term is related to a domain-specific concept, do we not need to know which the domain-specific concepts are? That is, another way of approaching the problem would be to have existing definitions which we try to extract and map terms to.

Term extraction is an applied research field which focuses on concrete tasks such as finding candidate terms to build a terminology for a specific domain. In such a field, compared to a more theoretical field such as lexical semantics, I believe that the emergence of a more pragmatically oriented tradition is often the case. For example, none of the reviewed methods provide a combined recall and precision of 100 per cent, but they do deliver results which are usable, so even though the metric used does not model exact definition of termhood, the output of a system employing the metric corresponds well enough with the intended output to be useful.

7.4.1 Beware of the other terms

As mentioned in section 4.1, it is important to realize that all research focused on "terms" is not directly related to terminology. For example, Wong and Liu (2008) have a section where they summarize different characteristics and definitions of "terms". However, they do not distinguish between research that focuses on term extraction within *information retrieval* and research on term extraction related to *terminology work*. Among the term definitions and descriptions mentioned in their work, they include the quote "terms tend to clump together" citing Bookstein, Klein, and Raita (1998). The exact quote (Bookstein et al., 1998, p. 3) is provided below.

> The assumption underlying this paper is that occurrences of a term sensitive to content will have a greater tendency to clump, or occur in the same textual neighborhood, than those of non-content-bearing terms.

First of all, it is clear that the clumping properties of "terms" is an assumption made by Bookstein et al. Secondly, the term "clump", is a term which is also used in general language, eliciting associations of what it might mean. Bookstein et al. (1998, p. 3) describe clumping as follows;

"Traditional clustering gathers items together which have shared features. [*clumping*, or *serial clustering* can be described as] when passages containing a term appear surprisingly close together". Remember that a "term" within Information Retrieval (IR) is not necessarily connected to a domain-specific concept. However, Bookstein et al. and other researchers within IR talk about content-bearing terms which should be interpreted from an information science perspective, rather than a semantic perspective. For example, the word "computer" may be regarded as non-content-bearing within IR in a document collection from the Computer Science domain, as explained in Bookstein et al. (1998, p. 2):

> For example, the word *computer* may distinguish books on Computer Science from other books in a general collection, but may not be useful for retrieval purposes within a collection restricted to material in Computer Science.

A term extraction method for terminology creation which was constructed to exclude a word such as "computer" from the domain of Computer Science would not be a good method. However, from an IR retrieval point of view, this makes sense, since the indexing the "term" does improve the ability to distinguish between two documents in the domain. As the widely used tf-idf metric (Manning, Raghavan, et al., 2008, p. 109) from IR works in exactly this way, using the tf-idf metric to perform ATE in the context of terminology is a bad idea.

The point of this example has not been to diminish the relevance of research within IR, but rather to highlight that there is sometimes a need for modification or adaptation of methods which are to be used to facilitate e.g. terminology creation. One example of this is the *Contrastive Weight* metric as described in sub-section 4.8.2, which in fact is, an adaptation of the tf-idf metric.

7.4.2 Multi-word units in compounding languages

Results and data presented in Foo and Merkel (2010) show that terms in Swedish, which is a compounding language, does not have the same distribution of word-unit length as English. This same observation was also made in Kokkinakis and Gerdin (2010). The composition of the validated terms for the two patent sub-domains used in Foo and Merkel (2010) can be seen in table 7.1.

Single-word terms are not a problem per se, but if this feature of a compounding language results in a drop in precision compared with non-compounding languages, this is a problem.

To solve this problem, there are at least two possibilities. The first involves using bilingual term extraction. This approach, as briefly noted in

	Sub-domain A42B		**Sub-domain A61G**	
	percent	count	percent	count
1-word terms	98.45%	570	98.41%	1240
2-word terms	1.55%	9	1.59%	20
3-word terms	0.00%	0	0.00%	0

Table 7.1: Term length distribution for validated terms used in Foo and Merkel (2010)

section 4.5 uses a one-sided selection to leverage monolingual term extraction of e.g. English terms where compounds remain as separate words to find the compounded equivalents in the target language. The second possibility is to apply a de-compounding algorithm to the corpus and split the compound single-word-unit terms into multi-word-unit terms what can be found using linguistic filters. This second method however introduces the problem of combining the words that have been compounded when presenting the final product.

8 Future research

This thesis has been concerned with Computational Terminology (CT) in general, and more specifically with methods for Automatic Term Extraction (ATE) and can hopefully contribute to a better understanding of linguistic, statistic and distributional patterns of terms. The methods presented however, need to be implemented in systems which can actually facilitate e.g. terminology work and terminology management.

In this chapter, I will present the context of Computational Terminology Management where methods such as those put forward in this thesis can be used as building blocks in a larger scheme.

8.1 Computational Terminology Management

Computational Terminology Management (CTM) is a relatively new research field which can be seen as a kind of new intersection between several research areas. The roots of the field lie in *Terminology Management* which is described by Wright and Budin (1997) as “any deliberate manipulation of terminological information”.

In CTM, computational methods are applied to tasks and activities in Terminology Management in order to empower people working with terminologies to do their work in a more rewarding and efficient way.

It is my opinion that CTM-related activities can be divided further into three groups; 1) *terminology creation*, 2) *terminology maintenance* and 3) *terminology use*. CT (without the *Management*) is a very closely related field, which can be said to be a subfield of CTM. Bourigault, Jacquemin, et al. (2001) describes CT as a field mainly concerned with term extraction, and organizing and structuring the extracted term candidates. These activities are part of CTM but CTM covers a lot more.

The contribution of this thesis is mainly related to *terminology creation*, but the contribution should be put into the context of CTM.

8.1.1 Terminology Creation

The first group, *terminology creation*, concerns a) discovering concepts and their terms in a specific domain, b) defining them and c) organizing these concepts and terms into a terminological system. Wright and Budin (1997) talks about *term selection* in the context of non-computational terminology management where the humans can select terms found in a relevant, domain-specific collection of documents.

8.1.2 Terminology Maintenance

The group called *terminology maintenance* includes the tasks that have to be performed once a terminology has been created. These tasks include

- adding new *admitted* or *forbidden* terms to a concept
- checking the terminology for inconsistencies
- adding new concepts and terms to the existing terminology
- terminology harmonization

The first example, adding a new admitted or forbidden term to a concept, relates to updating an existing terminology with new data. One scenario is that during quality assurance of company documentation, a new Term Candidate (TC) is found — a possible terminological unit which does not exist in the term bank. The terminology manager or terminologist now needs to check if there are any related concepts in the term bank. If the concept which the TC refers to exists, it must now be decided what to do with the TC. If it is found to be useful to have a synonym, the TC may be given the status of *admitted*. If using a synonym is found to be confusing, the TC will be given the status *forbidden*.

The task of checking if there are related concepts can be facilitated by using computational methods such as clustering or latent semantic analysis together with software applications which display the output of these methods in an accessible way.

8.1.3 Terminology Use

Once there exists a terminology, it needs to be made available in a adequate way with regard to the context in which it is needed. Examples of such contexts are

- writing technical documentation
- learning a new specialized domain
- terminology lookup for a non-domain-expert
- quality assurance of terminology use in texts

- updating existing documents to reflect changes in the terminology

The specific requirements for these contexts are very different, as are the tools used in the specific situations. To be able to provide the best user experience with regard to the value of having a terminology, it is essential to consider the conditions under which the terminology is to be used.

Terminology use when producing technical documentation

In technical documentation, it is important to use a non-ambiguous language, i.e. strive to use one term for each concept. This task is greatly facilitated by having integrated terminology support in the software the writer is using. Only having the terminology available for lookup in e.g. printed form, is a great hindrance to actually using it when writing.

Another use case for technical documentation is quality assurance of existing documentation. This is in principle the same undertaking as when providing support while writing, but applied to a set of existing documents. What sets the two tasks apart is the interface and level of interactivity. In the case of author assistance, we should provide an interactive experience. In the case of quality assurance, we should provide a report and a way of acting on the issues reported.

Updating documentation with terminology revisions

Another context where computational methods can facilitate terminology use is when a terminology has been revised and certain terms have been deprecated. When terms have been deprecated, it is necessary to update any documentation which uses these terms to use admitted terms instead. This can in the worse case be a very tedious task, requiring all changes to be made manually, but with the proper integration between the term bank and the Content Management System (CMS), this can become an automated process.

8.2 Possible directions

This section outlines possible directions for future research with regard to the results presented in this thesis and the broader context of CTM.

Refinement of machine learning and language model use

The experiments using Machine Learning (ML) to automatically configure ATE systems has shown promise given the early stages of the research. The obvious next step is to enhance the features available to the ML framework to include scores calculated using existing termhood, unithood, and contrastive measures such as those reviewed in chapter 4. Another possible expansion of the feature space is to include bilingual data to perform monolingual term extraction.

Regarding the novel techniques using language models, one direction for future research is to interpret existing metrics such as the Weirdness metric (sub-section 4.8.1) or the C-value/NC-value (sub-section 4.7.4) to use values derived from a language model rather than raw frequency counts. Smoothed language models provide an additional level of flexibility as they can also accommodate unseen word-units to a certain extent.

Also, as ATE methods start to use an increasing amount of contrastive data, n-gram language models could serve as a way share corpus data without having to provide access to the full text of the corpus in the case that this is problematic. The language models themselves also are smaller in size, which is in it self not a critical issue in this day and age, but could be beneficial in certain cases.

Term extraction for compounding languages using using decompounding strategies

One issue which is highlighted by the study presented in Foo and Merkel (2010) was that in a compounding language, such as Swedish, there are more single-word terms than multi-word terms. A brief analysis of the terms used in Foo and Merkel (2010) revealed that 27.95% of the single-word terms were compounds. What would be interesting to examine is whether compounded words are more probable to be terms or not. Lefever et al. (2009) tried to increase recall by decompounding Dutch words when performing French-Dutch[1] bilingual term extraction. The results indicated that the approach was marginally successful. Decompounding compound words to be able to apply unithood measures is not really necessary since the compound itself is evidence of unithood.

Tools for terminology creation

Besides providing ATE algorithms and tools, use of computational methods and computers can be used to aggregate and present contextual and relational information together with e.g. term candidates. For example the tool developed and used in Foo and Merkel (2010), Merkel et al. (2009) also displayed the textual context in which the potential terms were extracted from. Using such information, the domain experts and terminologists working with the data could make better informed decisions regarding the term status of the word unit compared with only viewing the word unit out of context.

[1] Compound words are relatively rare in French, but common in Dutch.

Another issue in terminology creation is finding relationships between term candidates. The most common output from ATE algorithms is a plain list of ordered term candidates. In such a list, it is difficult for the terminologist to examine related terms which must be considered when performing concept analysis. A list is also an ill-suited representation when it is necessary to move between different levels of information granularity for large data sets.

A reasonable solution to these problems related to navigating and analyzing ATE output data is do research how to design software which can help terminologists in these tasks. For example, when dealing with bilingually extracted data, a list representation of extracted terms such as the one used in IView (Foo & Merkel, 2010; Merkel et al., 2009), cannot visualize the translation graph which a certain term is part of. These kinds of relationships are important to consider when deciding e.g. on which candidates should be preferred and which candidates should be forbidden.

Tools for terminology maintenance

Finally, better tools and methods for performing terminology maintenance need to be researched and developed. The ATE methods presented here are tailored towards large scale term extraction processes which are typically performed to gather data for the terminology creation phase. However, a terminology must be maintained, and eventually, more time will be spent on maintaining a terminology compared to creating it.

The typical maintenance scenario is to update an existing terminology with terms from documents not present in the original term extraction corpus. One problem which we are faced with here is that there is less data to calculate statistics on. Another problem is that we need to find how the newly discovered terms are related to the existing concepts in the terminology. What computational methods can we use to perform such a search?

8.3 Summary

This thesis has presented the area of both monolingual and bilingual ATE. The contributions have shown that for ATE to be useful, it has to be part of a workflow, i.e. it is necessary to have access to the right tools to process the output of an ATE algorithm. The development of such tools, which use computational methods to e.g. analyze relationships between output data is therefore one possible direction for future research. Furthermore, ML techniques have been successfully applied to two subtasks within ATE prompting further studies to investigate how the approach can be refined. Finally, the use of smoothed language models as a source for statistical measures can also be investigated further.

Bibliography

Ahmad, K., Gillam, L., & Tostevin, L. (1999). Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In *Proceedings of the Eighth Text REtrieval Conference (TREC 8)*.

Ahrenberg, L. (2007). LinES 1.0 Annotation: Format, contents and guidelines. **retrieved from** `http://www.ida.liu.se/~lah/transmap/Corpus/guidelines.pdf`

Ahrenberg, L., Merkel, M., & Petterstedt, M. (2003). Interactive word alignment for language engineering. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics (EACL '03)* (Vol. 2, pp. 49–52). Budapest, Hungary: Association for Computational Linguistics. doi:`10.3115/1067737.1067746`

Ananiadou, S. (1994). A methodology for automatic term recognition. In *Proceedings of the 15th conference on Computational linguistics (COLING '94)* (pp. 1034–1038). Kyoto, Japan. doi:`10.3115/991250.991317`

Auger, P. (1989, Sept.). La terminotique et les industries de la langue. *Meta: journal des traducteurs / Meta: Translators' Journal*, *34*(3), 450–456.

Basili, R., Moschitti, A., & Pazienza, M. (2001). A contrastive approach to term extraction. In *Proceedings of Terminology and Knowledge Acquisition from Texts (TIA 2001)*. Nancy, France.

BNC Consortium. (2001). The British National Corpus, version 2 (BNC World). Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Bookstein, A., Klein, S. T., & Raita, T. (1998). Clumping properties of content-bearing words. *Journal of the American Society for Information Science*, *49*(2), 102–114. doi:`10.1002/(SICI)1097-4571(199802)49:2<102::AID-ASI2>3.0.CO;2-5`

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for optimal margin classifiers. In D. Haussler (Ed.), *Proceedings of the fifth annual workshop on Computational learning theory (COLT '92)* (pp. 144–152). New York, New York, USA: Association for Computing Machinery. doi:`10.1145/130385.130401`

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th conference on Computational linguistics (COLING '92)* (pp. 977–981).

Bourigault, D., Gonzalez-Mullier, I., & Gros, C. (1996). LEXTER, a Natural Processing Tool for Terminology Extraction. In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström & C. R. Papmehl (Eds.), *Proceedings of Euralex '96: Papers submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg, Sweden* [Part II](pp. 771–779). Göteborg, Sweden: Göteborg University, Department of Swedish.

Bourigault, D., Jacquemin, C., & Homme, M.-C. L. (2001). Introduction. In D. Bourigault, C. Jacquemin & M.-C. L. Homme (Eds.), *Recent Advances in Computational Terminology* (Vol. 2, pp. iix–xviii). Natural Language Processing. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Cabré, M. T. (1998). *Terminology* (J. C. Sager, Ed.). Terminology and Lexicography Research and Practice. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Castellví, M. T. C., Bagot, R. E., & Palatresi, J. V. (2001). Automatic term detection - A review of current systems. In D. Bourigault, C. Jacquemin & M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology* (pp. 53–88). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*, 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chen, S. F., & Goodman, J. (1998). *An empirical study of smoothing techniques for language modeling* (Technical Report No. TR-10-98). Center for Research in Computing Technology, Harvard University, Cambridge, Massachusetts.

Clark, H. H., & Brennan, S. E. (1991). Grounding in Communication. In L. B. Resnik, J. M. Levine & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Amarican Psychological Association.

Cohen, J. D. (1995). Highlights: language- and domain-independent automatic indexing terms for abstracting. *Journal of the American Society for Information Science*, *46*, 162–174. doi:`10.1002/(SICI)1097-4571(199504)46:3<162::AID-ASI2>3.0.CO;2-6`

Cohen, W. W. (1995). Fast Effective Rule Induction. In A. Prieditis & S. J. Russell (Eds.), *Proceedings of the Twelfth International Conference on Machine Learning* (pp. 115–123). Morgan Kaufmann Publishers.

Cohen, W. W., & Singer, Y. (1999). A simple, fast, and effective rule learner. In *Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence (AAAI '99/IAAI '99)* (pp. 335–342). M: American Association for Artificial Intelligence.

Dagan, I., & Church, K. (1994). Termight: identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied natural language processing (ANLC '94)* (pp. 34–40). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/974358.974367

Daille, B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *Proceedings of The Balancing Act: Combining Symbolic and Statistical Approaches to Language. Workshop at the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 29–36).

Daille, B. (2000). Morphological Rule Induction for Terminology Acquistion. In *Proceedings of the 18th conference on Computational Linguistics (COLING 2000)* (Vol. 1, Vols. 2, pp. 215–221). Morgan Kaufmann Publishers.

Daille, B., Gaussier, É., & Langé, J.-M. (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proceedings of the 15th conference on Computational Linguistics (COLING 94)* (pp. 515–521). doi:10.3115/991886.991975

Damerau, F. J. (1990). Evaluating computer-generated domain-oriented vocabularies. *Information Processing & Management*, *26*(6), 791–801. doi:10.1016/0306-4573(90)90052-4

Deléger, L., Merkel, M., & Zweigenbaum, P. (2006). Enriching Medical Terminologies: an Approach Based on Aligned Corpora. In A. Hasman, R. Haux, J. van der Lei, E. D. Clercq & F. H. R. France (Eds.), *Ubiquity: Technologies for Better Health in Aging Societies - Proceedings of MIE2006, The XXst International Congress of the European Federation for Medical Informatics* (Vol. 124, pp. 724–752). Studies in Health Technology and Informatics. IOS Press.

Dice, L. R. (1945, July). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*(3), 297–302. doi:10.2307/1932409

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2009). Package e1071. R Software package, avaliable at http://cran.r-project.org/web/packages/e1071/index.html.

Dyvik, H. (2004). Translations as semantic mirrors: from parallel corpus to wordnet. In K. Aijmer & B. Altenberg (Eds.), *Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)* [Advances in Corpus Linguistics](Vol. 49, *1*, pp. 311–326). Language and Computers - Studies in Practical Linguistics. Amsterdam/New York, NY: Rodopi.

EAGLES. (2002). XCES Corpus Encoding Standard for XML. **retrieved from** Expert Advisory Group on Language Engineering (EAGLES): http://www.xces.org/

Felber, H. (1984). *Terminology manual.* Unesco : Infoterm. Available at http://unesdoc.unesco.org/Ulis/cgi-bin/ulis.pl?catno=62033. Paris: General Information Programme and UNISIST / International Information Centre for Terminology.

Foo, J. (2011). Exploring termhood using language models. In T. Gornostay & A. Vasiljevs (Eds.), *Proceedings of the NODALIDA 2011 Workshop Creation, Harmonization and Application of Terminology Resources (CHAT2010)*

(Vol. 12, pp. 32–35). NEALT Proceedings Series. Northern European Association for Language Technology (NEALT). HDL: 10062/17276

Foo, J., & Merkel, M. (2010). Computer aided term bank creation and standardization: Building standardized term banks through automated term extraction and advanced editing tools. In M. Thelen & F. Steurs (Eds.), *Terminology in Everyday Life* (13, pp. 163–180). Terminology and Lexicography Research and Practice. John Benjamins Publishing Company.

Foo, J., & Merkel, M. (2010). Using machine learning to perform automatic term recognition. In N. Bel, B. Daille & A. Vasiljevs (Eds.), *Proceedings of the Workshop on Methods for automatic acquisition of Language Resources and their evaluation methods held in conjunction with the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 49–54).

Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999). Domain-Specific Keyphrase Extraction. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI 99)* (Vol. 2, pp. 668–671).

Frantzi, K. T., & Ananiadou, S. (1999). The C-value/NC-value domain-independent method for multi-word term extraction. *Journal of Natural Language Processing*, *6*(3), 145–179. doi:10.5715/jnlp.6.3_145

Frantzi, K. T., Ananiadou, S., & Miama, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, *3*(2), 115–130. doi:10.1007/s007999900023

Frantzi, K. T., Ananiadou, S., & Tsujii, J. (1998). The C-value/NC-value Method of Automatic Recognition for Multi-word Terms. In G. Goos, J. Hartmanis & J. van Leeuwen (Eds.), *Research and Advanced Technology for Digital Libraries: Proceedings of the Second European Conference, ECDL'98* (Vol. 1513, pp. 585–604). Lecture Notes in Computer Science. Berlin / Heidelberg: Springer. doi:10.1007/3-540-49653-X_35

Frantzi, K. T., Ananiadou, S., & Tsujii, J. (1999). Classifying Technical Terms: Proceedings of an iccc/ifip conference. In J. W. T. Smith, A. Ardö & P. Linde (Eds.), *Electronic Publishing ´99. Redefining the Information Chain - New Ways and Voices* (p. 326). ELPUB:1999. ICCC/IFIP third Conference on Electronic Publishing'99. University of Karlskrona/Ronneby, Sweden. Washington D.C.: ICCC Press. urn:nbn:se: elpub-9915

Hippisley, A., Cheng, D., & Ahmad, K. (2005). The head-modifier principle and multilingual term extraction. *Natural Language Engineering*, *11*(02), 129–157. doi:10.1017/S1351324904003535

Hulth, A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In M. Collins & M. Steedman (Eds.), *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)* (pp. 216–223). Association for Computational Linguistics.

Hulth, A. (2004). *Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction.* (PhD Thesis, Department of Computer Systems Sciences, Stockholm University, Stockholm, Sweden).

Isabelle, P. (1993). Bi-textual Aids for Translators. In *Proceedings of the Eight Annual Conference of the UW Centre for the New OED and Text Research.*

ISO 704:2009. (2009). Terminology Work – Principles and methods. (2009). ISO Technical Committee 37 (TC 37).

Itagaki, M., Aikawa, T., & He, X. (2007). Automatic Validation of Terminology Translation Consistency with Statistical Method. In *Proceedings of Machine Translation Summit XI* (pp. 269–274).

Jacquemin, C., & Bourigault, D. (2003). Term Extraction and Automatic Indexing. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (Chap. 33, pp. 599–615). Oxford University Press.

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*(1), 9–27. doi:`10.1017/S1351324900000048`

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition. *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, *3*, 259–289. doi:`10.1075/term.3.2.03kag`

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL '03)* (Vol. 1, pp. 48–54). S: Association for Computational Linguistics. doi:`10.3115/1073445.1073462`

Kokkinakis, D., & Gerdin, U. (2010). A Swedish Scientific Medical Corpus for Terminology Management and Linguistic Exploration. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).*

Lefever, E., Macken, L., & Hoste, V. (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09).* Association for Computational Linguistics.

LISA. (2005). LISA Terminology Survey for the Localization Industry. **retrieved from** Localization Industry Standards Association (LISA): `http://www.lisa.org/sigs/terminology/termsurvey2001preresults.html`

Lombard, R. (2006). A practical case for managing source-language. In K. J. Dunne (Ed.), *Perspectives on Localization* (13, p. 356). American Translators Association Scholarly Monograph Series. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval* (First edition). Cambridge University Press.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts / London, England: The MIT Press.

Merkel, M., & Foo, J. (2007). Terminology extraction and term ranking for standardizing term banks. In J. Nivre, H.-J. Kaalep, K. Muischnek & M. Koit (Eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007.* (Pp. 349–354). Tartu: University of Tartu. HDL: 10062/2602

Merkel, M., Petterstedt, M., & Ahrenberg, L. (2003). Interactive Word Alignment for Corpus Linguistics. In D. Archer, P. Rayson, A. Wilson & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 conference (CL2003)* (16, pp. 533–542). UCREL technical paper. Lancaster, UK: UCREL, Lancaster University.

Merkel, M., Foo, J., Andersson, M., Edholm, L., Gidlund, M., & Åsberg, S. (2009). Automatic Extraction and Manual Validation of Hierarchical Patent Terminology. In B. Nistrup Madsen & H. Erdman Thomsen (Eds.), *NORDTERM 16. Ontologier og taksonomier. Rapport fra NORDTERM 2009.* København, Danmark: Copenhagen Business School.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Series in Computer Science. McGraw-Hill.

Morin, E., Daille, B., Takeuchi, K., & Kageura, K. (2007). Bilingual Terminology Mining - Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)* (pp. 664–671).

Nagel, T. (1974, Oct.). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435–450.

Nakagawa, H., & Mori, T. (2002). A Simple but Powerful Automatic Term Extraction Method. In *Proceedings of the Second International Workshop on Computational Terminology (COMPUTERM 2002)*.

Nyström, M., Merkel, M., Petersson, H., & Åhlfeldt, H. (2007). Creating a medical dictionary using word alignment: The influence of sources and resources. *BMC Medical Informatics and Decision Making*, *7*(1). doi:10.1186/1472-6947-7-37

Och, F. J., & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, *29*, 19–51. doi:10.1162/089120103321337421

Ogden, C. K., & Richards, I. A. (1972). *The Meaning of Meaning* (Tenth edition). L: Routledge & Kegan Paul Ltd.

PAROLE Corpus at the Swedish Language Bank, Språkbanken, University of Gothenburg. (2010).

Patry, A., & Langlais, P. (2005). Corpus-Based Terminology Extraction. In *Proceedings of the 7th International Conference on Terminology & Knowledge Engineering* (pp. 313–321). Copenhagen, Denmark.

Pazienza, M. T., Pennacchiotti, M., & Zanzotto, F. M. (2005). Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. In S. Sirmakessis (Ed.), *Knowledge Mining* (Vol. 185, pp. 255–279). Studies in Fuzziness and Soft Computing. Berlin / Heidelberg, Germany: Springer. doi:10.1007/3-540-32394-5_20

Priss, U., & Old, L. J. (2005). Conceptual Exploration of Semantic Mirrors. In B. Ganter & R. Godin (Eds.), *Formal Concept Analysis* (Vol. 3403, pp. 21–32). Berlin / Heidelberg: Springer. doi:10.1007/978-3-540-32262-7_2

Quinlan, J. R. (1993). *C4.5: programs for machine learning* (J. Hammet, Ed.). Morgan Kaufmann series in machine learning. S: Morgan Kaufmann Publishers.

Sager, J. C. (1990). *A Practical Course in Terminology Processing*. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Sager, J. C., Dungworth, D., & McDonald, P. F. (1980). *English special languages: Principles and practice in science and technology*. Wiesbaden, Germany: Oscar Brandstetter Verlag.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Schmid, H. (1995). Improvements In Part-of-Speech Tagging With an Application To German. In *Proceedings of the Workshop "From Texts to Tags: Issues in Multilingual Language Analysis" organized by the ACL special interest group for linguistic data and corpus-based approaches to NLP (sigdat) held in conjunction with the meeting of the European Chapter of the Association of Computational Linguistics (EACL)*. Dublin, Ireland.

Sclano, F., & Velardi, P. (2007). TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. In *Proceedings of the 9th Conference on Terminology and Artificial Intelligence (TIA 2007)*.

Stolcke, A. (2002). SRILM - an Extensible Language Modeling Toolkit. In J. H. L. Hansen & B. Pellom (Eds.), *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP2002 - INTERSPEECH 2002)* (pp. 901–904). ISCA Archive.

Suonuuti, H. (2001). *Guide to Terminology* (2nd edition). NORDTERM. The Finnish Centre for Technical Terminology.

Täger, W. (2007, July 11). *The European Machine Translation Programme*. Presentation at MT Summit XI Workshop on Patent Translation. Copenhagen, Denmark.

Tapanainen, P., & Järvinen, T. (1997). A non-projective dependency parser. In *Proceedings of the fifth conference on Applied Natural Language Processing* (pp. 64–71). Washinton, DC, USA. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/974557.974568

Tiedemann, J. (2003). Combining Clues for Word Alignment. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 339–346).

Tufis, D., Barbu, A., & Ion, R. (2004). Extracting multilingual lexicons from parallel corpora. *Computers and the Humanities*, *38*, 163–189. doi:10.1023/B:CHUM.0000031172.03949.48

Turney, P. D. (2000). Learning Algorithms for Keyphrase Extraction. *Information Retrieval*, *2*, 303–336. doi:10.1023/A:1009976227802

Wermter, J., & Hahn, U. (2005). Paradimatic Modifiability Statistics for the Extraction of Complex Multi-Word Terms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)* (pp. 843–850). Association for Computational Linguistics (ACL).

Wermter, J., & Hahn, U. (2006). You Can't Beat Frequency (Unless You Use Linguistic Knowledge): A Qualitative Evaluation of Association Measures for Collocation and Term Extraction. In *Proceedings of 21st International Conference on Computational Linguistics (COLING 2006) and 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006)* (Vol. 1, Vols. 1, pp. 785–792). Association for Computational Linguistics (ACL). doi:10.3115/1220175.1220274

WIPO. (2005). International Patent Classification (IPC): Eight Edition (2006). World Intellectual Property Organization (WIPO).

Wong, W., & Liu, W. (2008). Determination of unithood and termhood for term recognition. In M. Song & F. B. Wu (Eds.), *Handbook of Research on Text and Web Mining Technologies* (pp. 500–529). IGI Global. doi:10.4018/978-1-59904-990-8.ch030

Wright, S. E. (2006, May 20). The role of terminology management in localization. Terminology seminar given on Internet (webinar). SDL.

Wright, S. E., & Budin, G. (Eds.). (1997). *Handbook of Terminology Management* (Vols. 2). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Wu, F. B., Li, Q., Bot, R. S., & Chen, X. (2005). Domain-specific keyphrase extraction. In O. Herzog, H.-J. Schek, N. Fuhr, A. Chowdhury & W. Teiken (Eds.), *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (pp. 283–284). N: Association for Computing Machinery (ACM). doi:10.1145/1099554.1099628

Zhang, Z., Iria, J., Brewster, C., & Ciravegna, F. (2008). A Comparative Evaluation of Term Recognition Algorithms. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis & D. Tapias (Eds.), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA).

Linköpings universitet

Avdelning, Institution
Division, Department

HCS,
Dept. of Computer and Information Science
581 83 Linköping

Datum
Date

April 2012

Språk
Language

☐ Svenska/Swedish
☒ Engelska/English
☐ ________________

Rapporttyp
Report category

☒ Licentiatavhandling
☐ Examensarbete
☐ C-uppsats
☐ D-uppsats
☐ Övrig rapport
☐ ________________

ISBN
978-91-7519-944-3

ISRN
LiU-Tek-Lic–2012:8

Serietitel och serienummer
Title of series, numbering

ISSN
0280–7971

Linköping Studies in Science and Technology

Thesis No. 1523

URL för elektronisk version

http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-75243

Titel
Title

Computational Terminology: Exploring Bilingual and Monolingual Term Extraction

Författare
Author

Jody Foo

Sammanfattning
Abstract

Terminologies are becoming more important to modern day society as technology and science continue to grow at an accelerating rate in a globalized environment. Agreeing upon which terms should be used to represent which concepts and how those terms should be translated into different languages is important if we wish to be able to communicate with as little confusion and misunderstandings as possible.

Since the 1990s, an increasing amount of terminology research has been devoted to facilitating and augmenting terminology-related tasks by using computers and computational methods. One focus for this research is Automatic Term Extraction (ATE).

In this compilation thesis, studies on both bilingual and monolingual ATE are presented. First, two publications reporting on how bilingual ATE using the align-extract approach can be used to extract patent terms. The result in this case was 181,000 manually validated English-Swedish patent terms which were to be used in a machine translation system for patent documents. A critical component of the method used is the Q-value metric, presented in the third paper, which can be used to rank extracted term candidates (TC) in an order that correlates with TC precision. The use of Machine Learning (ML) in monolingual ATE is the topic of the two final contributions. The first ML-related publication shows that rule induction based ML can be used to generate linguistic term selection patterns, and in the second ML-related publication, contrastive n-gram language models are used in conjunction with SVM ML to improve the precision of Term Candidate (TC)s selected using linguistic patterns.

Nyckelord
Keywords

terminology, automatic term extraction, automatic term recognition, computational terminology, terminology management

Department of Computer and Information Science
Linköpings universitet

Licentiate Theses

Linköpings Studies in Science and Technology
Faculty of Arts and Sciences

No 17 **Vojin Plavsic:** Interleaved Processing of Non-Numerical Data Stored on a Cyclic Memory. (Available at: FOA, Box 1165, S-581 11 Linköping, Sweden. FOA Report B30062E)

No 28 **Arne Jönsson, Mikael Patel:** An Interactive Flowcharting Technique for Communicating and Realizing Algorithms, 1984.

No 29 **Johnny Eckerland:** Retargeting of an Incremental Code Generator, 1984.

No 48 **Henrik Nordin:** On the Use of Typical Cases for Knowledge-Based Consultation and Teaching, 1985.

No 52 **Zebo Peng:** Steps Towards the Formalization of Designing VLSI Systems, 1985.

No 60 **Johan Fagerström:** Simulation and Evaluation of Architecture based on Asynchronous Processes, 1985.

No 71 **Jalal Maleki:** ICONStraint, A Dependency Directed Constraint Maintenance System, 1987.

No 72 **Tony Larsson:** On the Specification and Verification of VLSI Systems, 1986.

No 73 **Ola Strömfors:** A Structure Editor for Documents and Programs, 1986.

No 74 **Christos Levcopoulos:** New Results about the Approximation Behavior of the Greedy Triangulation, 1986.

No 104 **Shamsul I. Chowdhury: S**tatistical Expert Systems - a Special Application Area for Knowledge-Based Computer Methodology, 1987.

No 108 **Rober Bilos:** Incremental Scanning and Token-Based Editing, 1987.

No 111 **Hans Block:** SPORT-SORT Sorting Algorithms and Sport Tournaments, 1987.

No 113 **Ralph Rönnquist:** Network and Lattice Based Approaches to the Representation of Knowledge, 1987.

No 118 **Mariam Kamkar, Nahid Shahmehri:** Affect-Chaining in Program Flow Analysis Applied to Queries of Programs, 1987.

No 126 **Dan Strömberg:** Transfer and Distribution of Application Programs, 1987.

No 127 **Kristian Sandahl:** Case Studies in Knowledge Acquisition, Migration and User Acceptance of Expert Systems, 1987.

No 139 **Christer Bäckström:** Reasoning about Interdependent Actions, 1988.

No 140 **Mats Wirén:** On Control Strategies and Incrementality in Unification-Based Chart Parsing, 1988.

No 146 **Johan Hultman:** A Software System for Defining and Controlling Actions in a Mechanical System, 1988.

No 150 **Tim Hansen:** Diagnosing Faults using Knowledge about Malfunctioning Behavior, 1988.

No 165 **Jonas Löwgren:** Supporting Design and Management of Expert System User Interfaces, 1989.

No 166 **Ola Petersson:** On Adaptive Sorting in Sequential and Parallel Models, 1989.

No 174 **Yngve Larsson:** Dynamic Configuration in a Distributed Environment, 1989.

No 177 **Peter Åberg:** Design of a Multiple View Presentation and Interaction Manager, 1989.

No 181 **Henrik Eriksson:** A Study in Domain-Oriented Tool Support for Knowledge Acquisition, 1989.

No 184 **Ivan Rankin:** The Deep Generation of Text in Expert Critiquing Systems, 1989.

No 187 **Simin Nadjm-Tehrani:** Contributions to the Declarative Approach to Debugging Prolog Programs, 1989.

No 189 **Magnus Merkel:** Temporal Information in Natural Language, 1989.

No 196 **Ulf Nilsson:** A Systematic Approach to Abstract Interpretation of Logic Programs, 1989.

No 197 **Staffan Bonnier:** Horn Clause Logic with External Procedures: Towards a Theoretical Framework, 1989.

No 203 **Christer Hansson:** A Prototype System for Logical Reasoning about Time and Action, 1990.

No 212 **Björn Fjellborg:** An Approach to Extraction of Pipeline Structures for VLSI High-Level Synthesis, 1990.

No 230 **Patrick Doherty:** A Three-Valued Approach to Non-Monotonic Reasoning, 1990.

No 237 **Tomas Sokolnicki:** Coaching Partial Plans: An Approach to Knowledge-Based Tutoring, 1990.

No 250 **Lars Strömberg:** Postmortem Debugging of Distributed Systems, 1990.

No 253 **Torbjörn Näslund:** SLDFA-Resolution - Computing Answers for Negative Queries, 1990.

No 260 **Peter D. Holmes:** Using Connectivity Graphs to Support Map-Related Reasoning, 1991.

No 283 **Olof Johansson:** Improving Implementation of Graphical User Interfaces for Object-Oriented Knowledge- Bases, 1991.

No 298 **Rolf G Larsson:** Aktivitetsbaserad kalkylering i ett nytt ekonomisystem, 1991.

No 318 **Lena Srömbäck:** Studies in Extended Unification-Based Formalism for Linguistic Description: An Algorithm for Feature Structures with Disjunction and a Proposal for Flexible Systems, 1992.

No 319 **Mikael Pettersson:** DML-A Language and System for the Generation of Efficient Compilers from Denotational Specification, 1992.

No 326 **Andreas Kågedal:** Logic Programming with External Procedures: an Implementation, 1992.

No 328 **Patrick Lambrix:** Aspects of Version Management of Composite Objects, 1992.

No 333 **Xinli Gu:** Testability Analysis and Improvement in High-Level Synthesis Systems, 1992.

No 335 **Torbjörn Näslund:** On the Role of Evaluations in Iterative Development of Managerial Support Systems, 1992.

No 348 **Ulf Cederling:** Industrial Software Development - a Case Study, 1992.

No 352 **Magnus Morin:** Predictable Cyclic Computations in Autonomous Systems: A Computational Model and Implementation, 1992.

No 371 **Mehran Noghabai:** Evaluation of Strategic Investments in Information Technology, 1993.

No 378 **Mats Larsson:** A Transformational Approach to Formal Digital System Design, 1993.

No 380 **Johan Ringström:** Compiler Generation for Parallel Languages from Denotational Specifications, 1993.

No 381 **Michael Jansson:** Propagation of Change in an Intelligent Information System, 1993.

No 383 **Jonni Harrius:** An Architecture and a Knowledge Representation Model for Expert Critiquing Systems, 1993.

No 386 **Per Österling:** Symbolic Modelling of the Dynamic Environments of Autonomous Agents, 1993.

No 398 **Johan Boye:** Dependency-based Groudness Analysis of Functional Logic Programs, 1993.

No 402 **Lars Degerstedt:** Tabulated Resolution for Well Founded Semantics, 1993.

No 406 **Anna Moberg:** Satellitkontor - en studie av kommunikationsmönster vid arbete på distans, 1993.

No 414 **Peter Carlsson:** Separation av företagsledning och finansiering - fallstudier av företagsledarutköp ur ett agent-teoretiskt perspektiv, 1994.

No 417 **Camilla Sjöström:** Revision och lagreglering - ett historiskt perspektiv, 1994.

No 436 **Cecilia Sjöberg:** Voices in Design: Argumentation in Participatory Development, 1994.

No 437 **Lars Viklund:** Contributions to a High-level Programming Environment for a Scientific Computing, 1994.

No 440 **Peter Loborg:** Error Recovery Support in Manufacturing Control Systems, 1994.

FHS 3/94 **Owen Eriksson:** Informationssystem med verksamhetskvalitet - utvärdering baserat på ett verksamhetsinriktat och samskapande perspektiv, 1994.

FHS 4/94 **Karin Pettersson:** Informationssystemstrukturering, ansvarsfördelning och användarinflytande - En komparativ studie med utgångspunkt i två informationssystemstrategier, 1994.

No 441 **Lars Poignant:** Informationsteknologi och företagsetablering - Effekter på produktivitet och region, 1994.

No 446 **Gustav Fahl:** Object Views of Relational Data in Multidatabase Systems, 1994.

No 450 **Henrik Nilsson:** A Declarative Approach to Debugging for Lazy Functional Languages, 1994.

No 451 **Jonas Lind:** Creditor - Firm Relations: an Interdisciplinary Analysis, 1994.

No 452 **Martin Sköld:** Active Rules based on Object Relational Queries - Efficient Change Monitoring Techniques, 1994.

No 455 **Pär Carlshamre:** A Collaborative Approach to Usability Engineering: Technical Communicators and System Developers in Usability-Oriented Systems Development, 1994.

FHS 5/94 **Stefan Cronholm:** Varför CASE-verktyg i systemutveckling? - En motiv- och konsekvensstudie avseende arbetssätt och arbetsformer, 1994.

No 462 **Mikael Lindvall:** A Study of Traceability in Object-Oriented Systems Development, 1994.

No 463 **Fredrik Nilsson:** Strategi och ekonomisk styrning - En studie av Sandviks förvärv av Bahco Verktyg, 1994.

No 464 **Hans Olsén:** Collage Induction: Proving Properties of Logic Programs by Program Synthesis, 1994.

No 469 **Lars Karlsson:** Specification and Synthesis of Plans Using the Features and Fluents Framework, 1995.

No 473 **Ulf Söderman:** On Conceptual Modelling of Mode Switching Systems, 1995.

No 475 **Choong-ho Yi:** Reasoning about Concurrent Actions in the Trajectory Semantics, 1995.

No 476 **Bo Lagerström:** Successiv resultatavräkning av pågående arbeten. - Fallstudier i tre byggföretag, 1995.

No 478 **Peter Jonsson:** Complexity of State-Variable Planning under Structural Restrictions, 1995.

FHS 7/95 **Anders Avdic:** Arbetsintegrerad systemutveckling med kalkylprogram, 1995.

No 482 **Eva L Ragnemalm:** Towards Student Modelling through Collaborative Dialogue with a Learning Companion, 1995.

No 488 **Eva Toller:** Contributions to Parallel Multiparadigm Languages: Combining Object-Oriented and Rule-Based Programming, 1995.

No 489 **Erik Stoy:** A Petri Net Based Unified Representation for Hardware/Software Co-Design, 1995.

No 497 **Johan Herber:** Environment Support for Building Structured Mathematical Models, 1995.

No 498 **Stefan Svenberg:** Structure-Driven Derivation of Inter-Lingual Functor-Argument Trees for Multi-Lingual Generation, 1995.

No 503 **Hee-Cheol Kim:** Prediction and Postdiction under Uncertainty, 1995.

FHS 8/95 **Dan Fristedt:** Metoder i användning - mot förbättring av systemutveckling genom situationell metodkunskap och metodanalys, 1995.

FHS 9/95 **Malin Bergvall:** Systemförvaltning i praktiken - en kvalitativ studie avseende centrala begrepp, aktiviteter och ansvarsroller, 1995.

No 513 **Joachim Karlsson:** Towards a Strategy for Software Requirements Selection, 1995.

No 517 **Jakob Axelsson:** Schedulability-Driven Partitioning of Heterogeneous Real-Time Systems, 1995.

No 518 **Göran Forslund:** Toward Cooperative Advice-Giving Systems: The Expert Systems Experience, 1995.

No 522 **Jörgen Andersson:** Bilder av småföretagares ekonomistyrning, 1995.

No 538 **Staffan Flodin:** Efficient Management of Object-Oriented Queries with Late Binding, 1996.

No 545 **Vadim Engelson:** An Approach to Automatic Construction of Graphical User Interfaces for Applications in Scientific Computing, 1996.

No 546 **Magnus Werner :** Multidatabase Integration using Polymorphic Queries and Views, 1996.

FiF-a 1/96 **Mikael Lind:** Affärsprocessinriktad förändringsanalys - utveckling och tillämpning av synsätt och metod, 1996.

No 549 **Jonas Hallberg:** High-Level Synthesis under Local Timing Constraints, 1996.

No 550 **Kristina Larsen:** Förutsättningar och begränsningar för arbete på distans - erfarenheter från fyra svenska företag. 1996.

No 557 **Mikael Johansson:** Quality Functions for Requirements Engineering Methods, 1996.

No 558 **Patrik Nordling:** The Simulation of Rolling Bearing Dynamics on Parallel Computers, 1996.

No 561 **Anders Ekman:** Exploration of Polygonal Environments, 1996.

No 563 **Niclas Andersson:** Compilation of Mathematical Models to Parallel Code, 1996.

No 567 **Johan Jenvald:** Simulation and Data Collection in Battle Training, 1996.

No 575 **Niclas Ohlsson:** Software Quality Engineering by Early Identification of Fault-Prone Modules, 1996.

No 576 **Mikael Ericsson:** Commenting Systems as Design Support—A Wizard-of-Oz Study, 1996.

No 587 **Jörgen Lindström:** Chefers användning av kommunikationsteknik, 1996.

No 589 **Esa Falkenroth:** Data Management in Control Applications - A Proposal Based on Active Database Systems, 1996.

No 591 **Niclas Wahllöf:** A Default Extension to Description Logics and its Applications, 1996.

No 595 **Annika Larsson:** Ekonomisk Styrning och Organisatorisk Passion - ett interaktivt perspektiv, 1997.

No 597 **Ling Lin:** A Value-based Indexing Technique for Time Sequences, 1997.

No 598 **Rego Granlund:** C^3Fire - A Microworld Supporting Emergency Management Training, 1997.

No 599 **Peter Ingels:** A Robust Text Processing Technique Applied to Lexical Error Recovery, 1997.

No 607 **Per-Arne Persson:** Toward a Grounded Theory for Support of Command and Control in Military Coalitions, 1997.

No 609 **Jonas S Karlsson:** A Scalable Data Structure for a Parallel Data Server, 1997.

FiF-a 4 **Carita Åbom:** Videomötesteknik i olika affärssituationer - möjligheter och hinder, 1997.

FiF-a 6 **Tommy Wedlund**: Att skapa en företagsanpassad systemutvecklingsmodell - genom rekonstruktion, värdering och vidareutveckling i T50-bolag inom ABB, 1997.

No 615 **Silvia Coradeschi**: A Decision-Mechanism for Reactive and Coordinated Agents, 1997.

No 623 **Jan Ollinen:** Det flexibla kontorets utveckling på Digital - Ett stöd för multiflex? 1997.

No 626 **David Byers:** Towards Estimating Software Testability Using Static Analysis, 1997.

No 627 **Fredrik Eklund:** Declarative Error Diagnosis of GAPLog Programs, 1997.

No 629 **Gunilla Ivefors:** Krigsspel och Informationsteknik inför en oförutsägbar framtid, 1997**.**

No 631 **Jens-Olof Lindh:** Analysing Traffic Safety from a Case-Based Reasoning Perspective, 1997

No 639 **Jukka Mäki-Turja:**. Smalltalk - a suitable Real-Time Language, 1997.

No 640 **Juha Takkinen**: CAFE: Towards a Conceptual Model for Information Management in Electronic Mail, 1997.

No 643 **Man Lin**: Formal Analysis of Reactive Rule-based Programs, 1997.

No 653 **Mats Gustafsson**: Bringing Role-Based Access Control to Distributed Systems, 1997.

FiF-a 13 **Boris Karlsson:** Metodanalys för förståelse och utveckling av systemutvecklingsverksamhet. Analys och värdering av systemutvecklingsmodeller och dess användning, 1997.

No 674 **Marcus Bjäreland:** Two Aspects of Automating Logics of Action and Change - Regression and Tractability, 1998.

No 676 **Jan Håkegård**: Hierarchical Test Architecture and Board-Level Test Controller Synthesis, 1998.

No 668 **Per-Ove Zetterlund**: Normering av svensk redovisning - En studie av tillkomsten av Redovisningsrådets rekommendation om koncernredovisning (RR01:91), 1998.

No 675 **Jimmy Tjäder**: Projektledaren & planen - en studie av projektledning i tre installations- och systemutvecklingsprojekt, 1998.

FiF-a 14 **Ulf Melin**: Informationssystem vid ökad affärs- och processorientering - egenskaper, strategier och utveckling, 1998.

No 695 **Tim Heyer**: COMPASS: Introduction of Formal Methods in Code Development and Inspection, 1998.

No 700 **Patrik Hägglund:** Programming Languages for Computer Algebra, 1998.

FiF-a 16 **Marie-Therese Christiansson:** Inter-organisatorisk verksamhetsutveckling - metoder som stöd vid utveckling av partnerskap och informationssystem, 1998.

No 712 **Christina Wennestam:** Information om immateriella resurser. Investeringar i forskning och utveckling samt i personal inom skogsindustrin, 1998.

No 719 **Joakim Gustafsson:** Extending Temporal Action Logic for Ramification and Concurrency, 1998.

No 723 **Henrik André-Jönsson:** Indexing time-series data using text indexing methods, 1999.

No 725 **Erik Larsson:** High-Level Testability Analysis and Enhancement Techniques, 1998.

No 730 **Carl-Johan Westin:** Informationsförsörjning: en fråga om ansvar - aktiviteter och uppdrag i fem stora svenska organisationers operativa informationsförsörjning, 1998.

No 731 **Åse Jansson:** Miljöhänsyn - en del i företags styrning, 1998.

No 733 **Thomas Padron-McCarthy:** Performance-Polymorphic Declarative Queries, 1998.

No 734 **Anders Bäckström:** Värdeskapande kreditgivning - Kreditriskhantering ur ett agentteoretiskt perspektiv, 1998.

FiF-a 21 **Ulf Seigerroth:** Integration av förändringsmetoder - en modell för välgrundad metodintegration, 1999.

FiF-a 22 **Fredrik Öberg:** Object-Oriented Frameworks - A New Strategy for Case Tool Development, 1998.

No 737 **Jonas Mellin:** Predictable Event Monitoring, 1998.

No 738 **Joakim Eriksson:** Specifying and Managing Rules in an Active Real-Time Database System, 1998.

FiF-a 25 **Bengt E W Andersson:** Samverkande informationssystem mellan aktörer i offentliga åtaganden - En teori om aktörsarenor i samverkan om utbyte av information, 1998.

No 742 **Pawel Pietrzak:** Static Incorrectness Diagnosis of CLP (FD), 1999.

No 748 **Tobias Ritzau:** Real-Time Reference Counting in RT-Java, 1999.

No 751 **Anders Ferntoft:** Elektronisk affärskommunikation - kontaktkostnader och kontaktprocesser mellan kunder och leverantörer på producentmarknader, 1999.

No 752 **Jo Skåmedal:** Arbete på distans och arbetsformens påverkan på resor och resmönster, 1999.

No 753 **Johan Alvehus:** Mötets metaforer. En studie av berättelser om möten, 1999.

No 754 **Magnus Lindahl:** Bankens villkor i låneavtal vid kreditgivning till högt belånade företagsförvärv: En studie ur ett agentteoretiskt perspektiv, 2000.

No 766 **Martin V. Howard:** Designing dynamic visualizations of temporal data, 1999.

No 769 **Jesper Andersson:** Towards Reactive Software Architectures, 1999.

No 775 **Anders Henriksson:** Unique kernel diagnosis, 1999.

FiF-a 30 **Pär J. Ågerfalk:** Pragmatization of Information Systems - A Theoretical and Methodological Outline, 1999.

No 787 **Charlotte Björkegren:** Learning for the next project - Bearers and barriers in knowledge transfer within an organisation, 1999.

No 788 **Håkan Nilsson:** Informationsteknik som drivkraft i granskningsprocessen - En studie av fyra revisionsbyråer, 2000.

No 790 **Erik Berglund:** Use-Oriented Documentation in Software Development, 1999.

No 791 **Klas Gäre:** Verksamhetsförändringar i samband med IS-införande, 1999.

No 800 **Anders Subotic:** Software Quality Inspection, 1999.

No 807 **Svein Bergum**: Managerial communication in telework, 2000.

No 809 **Flavius Gruian:** Energy-Aware Design of Digital Systems, 2000.

FiF-a 32 **Karin Hedström:** Kunskapsanvändning och kunskapsutveckling hos verksamhetskonsulter - Erfarenheter från ett FOU-samarbete, 2000.

No 808 **Linda Askenäs:** Affärssystemet - En studie om teknikens aktiva och passiva roll i en organisation, 2000.

No 820 **Jean Paul Meynard:** Control of industrial robots through high-level task programming, 2000.

No 823 **Lars Hult:** Publika Gränsytor - ett designexempel, 2000.

No 832 **Paul Pop:** Scheduling and Communication Synthesis for Distributed Real-Time Systems, 2000.

FiF-a 34 **Göran Hultgren:** Nätverksinriktad Förändringsanalys - perspektiv och metoder som stöd för förståelse och utveckling av affärsrelationer och informationssystem, 2000.

No 842 **Magnus Kald:** The role of management control systems in strategic business units, 2000.

No 844 **Mikael Cäker:** Vad kostar kunden? Modeller för intern redovisning, 2000.

FiF-a 37 **Ewa Braf**: Organisationers kunskapsverksamheter - en kritisk studie av ”knowledge management”, 2000.

FiF-a 40 **Henrik Lindberg:** Webbaserade affärsprocesser - Möjligheter och begränsningar, 2000.

FiF-a 41 **Benneth Christiansson:** Att komponentbasera informationssystem - Vad säger teori och praktik?, 2000.

No. 854 **Ola Pettersson:** Deliberation in a Mobile Robot, 2000.

No 863 **Dan Lawesson**: Towards Behavioral Model Fault Isolation for Object Oriented Control Systems, 2000.

No 881 **Johan Moe:** Execution Tracing of Large Distributed Systems, 2001.

No 882 **Yuxiao Zhao:** XML-based Frameworks for Internet Commerce and an Implementation of B2B e-procurement, 2001.

No 890 **Annika Flycht-Eriksson:** Domain Knowledge Management in Information-providing Dialogue systems, 2001.

FiF-a 47 **Per-Arne Segerkvis**t: Webbaserade imaginära organisationers samverkansformer: Informationssystemarkitektur och aktörssamverkan som förutsättningar för affärsprocesser, 2001.

No 894 **Stefan Svarén:** Styrning av investeringar i divisionaliserade företag - Ett koncernperspektiv, 2001.

No 906 **Lin Han**: Secure and Scalable E-Service Software Delivery, 2001.

No 917 **Emma Hansson:** Optionsprogram för anställda - en studie av svenska börsföretag, 2001.

No 916 **Susanne Odar:** IT som stöd för strategiska beslut, en studie av datorimplementerade modeller av verksamhet som stöd för beslut om anskaffning av JAS 1982, 2002.

FiF-a-49 **Stefan Holgersson:** IT-system och filtrering av verksamhetskunskap - kvalitetsproblem vid analyser och beslutsfattande som bygger på uppgifter hämtade från polisens IT-system, 2001.

FiF-a-51 **Per Oscarsson:** Informationssäkerhet i verksamheter - begrepp och modeller som stöd för förståelse av informationssäkerhet och dess hantering, 2001.

No 919 **Luis Alejandro Cortes:** A Petri Net Based Modeling and Verification Technique for Real-Time Embedded Systems, 2001.

No 915 **Niklas Sandell:** Redovisning i skuggan av en bankkris - Värdering av fastigheter. 2001.

No 931 **Fredrik Elg:** Ett dynamiskt perspektiv på individuella skillnader av heuristisk kompetens, intelligens, mentala modeller, mål och konfidens i kontroll av mikrovärlden Moro, 2002.

No 933 **Peter Aronsson:** Automatic Parallelization of Simulation Code from Equation Based Simulation Languages, 2002.

No 938 **Bourhane Kadmiry**: Fuzzy Control of Unmanned Helicopter, 2002.

No 942 **Patrik Haslum**: Prediction as a Knowledge Representation Problem: A Case Study in Model Design, 2002.

No 956 **Robert Sevenius:** On the instruments of governance - A law & economics study of capital instruments in limited liability companies, 2002.

FiF-a 58 **Johan Petersson:** Lokala elektroniska marknadsplatser - informationssystem för platsbundna affärer, 2002.

No 964 **Peter Bunus:** Debugging and Structural Analysis of Declarative Equation-Based Languages, 2002.

No 973 **Gert Jervan:** High-Level Test Generation and Built-In Self-Test Techniques for Digital Systems, 2002.

No 958 **Fredrika Berglund:** Management Control and Strategy - a Case Study of Pharmaceutical Drug Development, 2002.

FiF-a 61 **Fredrik Karlsson:** Meta-Method for Method Configuration - A Rational Unified Process Case, 2002.

No 985 **Sorin Manolache:** Schedulability Analysis of Real-Time Systems with Stochastic Task Execution Times, 2002.

No 982 **Diana Szentiványi:** Performance and Availability Trade-offs in Fault-Tolerant Middleware, 2002.

No 989 **Iakov Nakhimovski:** Modeling and Simulation of Contacting Flexible Bodies in Multibody Systems, 2002.

No 990 **Levon Saldamli:** PDEModelica - Towards a High-Level Language for Modeling with Partial Differential Equations, 2002.

No 991 **Almut Herzog:** Secure Execution Environment for Java Electronic Services, 2002.

No 999 **Jon Edvardsson:** Contributions to Program- and Specification-based Test Data Generation, 2002.

No 1000 **Anders Arpteg:** Adaptive Semi-structured Information Extraction, 2002.

No 1001 **Andrzej Bednarski:** A Dynamic Programming Approach to Optimal Retargetable Code Generation for Irregular Architectures, 2002.

No 988 **Mattias Arvola:** Good to use! : Use quality of multi-user applications in the home, 2003.

FiF-a 62 **Lennart Ljung:** Utveckling av en projektivitetsmodell - om organisationers förmåga att tillämpa projektarbetsformen, 2003.

No 1003 **Pernilla Qvarfordt:** User experience of spoken feedback in multimodal interaction, 2003.

No 1005 **Alexander Siemers:** Visualization of Dynamic Multibody Simulation With Special Reference to Contacts, 2003.

No 1008 **Jens Gustavsson:** Towards Unanticipated Runtime Software Evolution, 2003.

No 1010 **Calin Curescu:** Adaptive QoS-aware Resource Allocation for Wireless Networks, 2003.

No 1015 **Anna Andersson:** Management Information Systems in Process-oriented Healthcare Organisations, 2003.

No 1018 **Björn Johansson:** Feedforward Control in Dynamic Situations, 2003.

No 1022 **Traian Pop:** Scheduling and Optimisation of Heterogeneous Time/Event-Triggered Distributed Embedded Systems, 2003.

FiF-a 65 **Britt-Marie Johansson:** Kundkommunikation på distans - en studie om kommunikationsmediets betydelse i affärstransaktioner, 2003.

No 1024 **Aleksandra Tešanovic:** Towards Aspectual Component-Based Real-Time System Development, 2003.

No 1034 **Arja Vainio-Larsson**: Designing for Use in a Future Context - Five Case Studies in Retrospect, 2003.

No 1033 **Peter Nilsson:** Svenska bankers redovisningsval vid reservering för befarade kreditförluster - En studie vid införandet av nya redovisningsregler, 2003.

FiF-a 69 **Fredrik Ericsson:** Information Technology for Learning and Acquiring of Work Knowledge, 2003.

No 1049 **Marcus Comstedt:** Towards Fine-Grained Binary Composition through Link Time Weaving, 2003.

No 1052 **Åsa Hedenskog:** Increasing the Automation of Radio Network Control, 2003.

No 1054 **Claudiu Duma:** Security and Efficiency Tradeoffs in Multicast Group Key Management, 2003.

FiF-a 71 **Emma Eliason:** Effektanalys av IT-systems handlingsutrymme, 2003.

No 1055 **Carl Cederberg:** Experiments in Indirect Fault Injection with Open Source and Industrial Software, 2003.

No 1058 **Daniel Karlsson:** Towards Formal Verification in a Component-based Reuse Methodology, 2003.

FiF-a 73 **Anders Hjalmarsson:** Att etablera och vidmakthålla förbättringsverksamhet - behovet av koordination och interaktion vid förändring av systemutvecklingsverksamheter, 2004.

No 1079 **Pontus Johansson:** Design and Development of Recommender Dialogue Systems, 2004.

No 1084 **Charlotte Stoltz:** Calling for Call Centres - A Study of Call Centre Locations in a Swedish Rural Region, 2004.

FiF-a 74 **Björn Johansson:** Deciding on Using Application Service Provision in SMEs, 2004.

No 1094 **Genevieve Gorrell:** Language Modelling and Error Handling in Spoken Dialogue Systems, 2004.

No 1095 **Ulf Johansson:** Rule Extraction - the Key to Accurate and Comprehensible Data Mining Models, 2004.

No 1099 **Sonia Sangari:** Computational Models of Some Communicative Head Movements, 2004.

No 1110 **Hans Nässla:** Intra-Family Information Flow and Prospects for Communication Systems, 2004.

No 1116 **Henrik Sällberg:** On the value of customer loyalty programs - A study of point programs and switching costs, 2004.

FiF-a 77 **Ulf Larsson:** Designarbete i dialog - karaktärisering av interaktionen mellan användare och utvecklare i en systemutvecklingsprocess, 2004.

No 1126 **Andreas Borg:** Contribution to Management and Validation of Non-Functional Requirements, 2004.

No 1127 **Per-Ola Kristensson:** Large Vocabulary Shorthand Writing on Stylus Keyboard, 2004.

No 1132 **Pär-Anders Albinsson:** Interacting with Command and Control Systems: Tools for Operators and Designers, 2004.

No 1130 **Ioan Chisalita:** Safety-Oriented Communication in Mobile Networks for Vehicles, 2004.

No 1138 **Thomas Gustafsson:** Maintaining Data Consistency in Embedded Databases for Vehicular Systems, 2004.

No 1149 **Vaida Jakoniené:** A Study in Integrating Multiple Biological Data Sources, 2005.

No 1156 **Abdil Rashid Mohamed**: High-Level Techniques for Built-In Self-Test Resources Optimization, 2005.

No 1162 **Adrian Pop:** Contributions to Meta-Modeling Tools and Methods, 2005.

No 1165 **Fidel Vascós Palacios:** On the information exchange between physicians and social insurance officers in the sick leave process: an Activity Theoretical perspective, 2005.

FiF-a 84 **Jenny Lagsten:** Verksamhetsutvecklande utvärdering i informationssystemprojekt, 2005.

No 1166 **Emma Larsdotter Nilsson:** Modeling, Simulation, and Visualization of Metabolic Pathways Using Modelica, 2005.

No 1167 **Christina Keller:** Virtual Learning Environments in higher education. A study of students' acceptance of educational technology, 2005.

No 1168 **Cécile Åberg:** Integration of organizational workflows and the Semantic Web, 2005.

FiF-a 85 **Anders Forsman:** Standardisering som grund för informationssamverkan och IT-tjänster - En fallstudie baserad på trafikinformationstjänsten RDS-TMC, 2005.

No 1171 **Yu-Hsing Huang:** A systemic traffic accident model, 2005.

FiF-a 86 **Jan Olausson:** Att modellera uppdrag - grunder för förståelse av processinriktade informationssystem i transaktionsintensiva verksamheter, 2005.

No 1172 **Petter Ahlström**: Affärsstrategier för seniorbostadsmarknaden, 2005.

No 1183 **Mathias Cöster**: Beyond IT and Productivity - How Digitization Transformed the Graphic Industry, 2005.

No 1184 **Åsa Horzella**: Beyond IT and Productivity - Effects of Digitized Information Flows in Grocery Distribution, 2005.

No 1185 **Maria Kollberg**: Beyond IT and Productivity - Effects of Digitized Information Flows in the Logging Industry, 2005.

No 1190 **David Dinka**: Role and Identity - Experience of technology in professional settings, 2005.

No 1191 **Andreas Hansson**: Increasing the Storage Capacity of Recursive Auto-associative Memory by Segmenting Data, 2005.

No 1192 **Nicklas Bergfeldt:** Towards Detached Communication for Robot Cooperation, 2005.

No 1194 **Dennis Maciuszek:** Towards Dependable Virtual Companions for Later Life, 2005.

No 1204 **Beatrice Alenljung:** Decision-making in the Requirements Engineering Process: A Human-centered Approach, 2005.

No 1206 **Anders Larsson:** System-on-Chip Test Scheduling and Test Infrastructure Design, 2005.

No 1207 **John Wilander:** Policy and Implementation Assurance for Software Security, 2005.

No 1209 **Andreas Käll:** Översättningar av en managementmodell - En studie av införandet av Balanced Scorecard i ett landsting, 2005.

No 1225 **He Tan:** Aligning and Merging Biomedical Ontologies, 2006.

No 1228 **Artur Wilk:** Descriptive Types for XML Query Language Xcerpt, 2006.

No 1229 **Per Olof Pettersson:** Sampling-based Path Planning for an Autonomous Helicopter, 2006.

No 1231 **Kalle Burbeck:** Adaptive Real-time Anomaly Detection for Safeguarding Critical Networks, 2006.

No 1233 **Daniela Mihailescu:** Implementation Methodology in Action: A Study of an Enterprise Systems Implementation Methodology, 2006.

No 1244 **Jörgen Skågeby:** Public and Non-public gifting on the Internet, 2006.

No 1248 **Karolina Eliasson:** The Use of Case-Based Reasoning in a Human-Robot Dialog System, 2006.

No 1263 **Misook Park-Westman:** Managing Competence Development Programs in a Cross-Cultural Organisation - What are the Barriers and Enablers, 2006.

FiF-a 90 **Amra Halilovic:** Ett praktikperspektiv på hantering av mjukvarukomponenter, 2006.

No 1272 **Raquel Flodström**: A Framework for the Strategic Management of Information Technology, 2006.

No 1277 **Viacheslav Izosimov:** Scheduling and Optimization of Fault-Tolerant Embedded Systems, 2006.

No 1283 **Håkan Hasewinkel:** A Blueprint for Using Commercial Games off the Shelf in Defence Training, Education and Research Simulations, 2006.

FiF-a 91 **Hanna Broberg:** Verksamhetsanpassade IT-stöd - Designteori och metod, 2006.

No 1286 **Robert Kaminski:** Towards an XML Document Restructuring Framework, 2006.

No 1293 **Jiri Trnka:** Prerequisites for data sharing in emergency management, 2007.

No 1302 **Björn Hägglund:** A Framework for Designing Constraint Stores, 2007.

No 1303 **Daniel Andreasson**: Slack-Time Aware Dynamic Routing Schemes for On-Chip Networks, 2007.

No 1305 **Magnus Ingmarsson:** Modelling User Tasks and Intentions for Service Discovery in Ubiquitous Computing, 2007.

No 1306 **Gustaf Svedjemo**: Ontology as Conceptual Schema when Modelling Historical Maps for Database Storage, 2007.

No 1307 **Gianpaolo Conte**: Navigation Functionalities for an Autonomous UAV Helicopter, 2007.

No 1309 **Ola Leifler**: User-Centric Critiquing in Command and Control: The DKExpert and ComPlan Approaches, 2007.

No 1312 **Henrik Svensson**: Embodied simulation as off-line representation, 2007.

No 1313 **Zhiyuan He:** System-on-Chip Test Scheduling with Defect-Probability and Temperature Considerations, 2007.

No 1317 **Jonas Elmqvist:** Components, Safety Interfaces and Compositional Analysis, 2007.

No 1320 **Håkan Sundblad:** Question Classification in Question Answering Systems, 2007.

No 1323 **Magnus Lundqvist**: Information Demand and Use: Improving Information Flow within Small-scale Business Contexts, 2007.

No 1329 **Martin Magnusson:** Deductive Planning and Composite Actions in Temporal Action Logic, 2007.

No 1331 **Mikael Asplund:** Restoring Consistency after Network Partitions, 2007.

No 1332 **Martin Fransson:** Towards Individualized Drug Dosage - General Methods and Case Studies, 2007.

No 1333 **Karin Camara:** A Visual Query Language Served by a Multi-sensor Environment, 2007.

No 1337 **David Broman:** Safety, Security, and Semantic Aspects of Equation-Based Object-Oriented Languages and Environments, 2007.

No 1339 **Mikhail Chalabine:** Invasive Interactive Parallelization, 2007.

No 1351 **Susanna Nilsson:** A Holistic Approach to Usability Evaluations of Mixed Reality Systems, 2008.

No 1353 **Shanai Ardi:** A Model and Implementation of a Security Plug-in for the Software Life Cycle, 2008.

No 1356 **Erik Kuiper:** Mobility and Routing in a Delay-tolerant Network of Unmanned Aerial Vehicles, 2008.

No 1359 **Jana Rambusch**: Situated Play, 2008.

No 1361 **Martin Karresand:** Completing the Picture - Fragments and Back Again, 2008.

No 1363 **Per Nyblom:** Dynamic Abstraction for Interleaved Task Planning and Execution, 2008.

No 1371 **Fredrik Lantz:** Terrain Object Recognition and Context Fusion for Decision Support, 2008.

No 1373 **Martin Östlund:** Assistance Plus: 3D-mediated Advice-giving on Pharmaceutical Products, 2008.

No 1381 **Håkan Lundvall:** Automatic Parallelization using Pipelining for Equation-Based Simulation Languages, 2008.

No 1386 **Mirko Thorstensson:** Using Observers for Model Based Data Collection in Distributed Tactical Operations, 2008.

No 1387 **Bahlol Rahimi:** Implementation of Health Information Systems, 2008.

No 1392 **Maria Holmqvist:** Word Alignment by Re-using Parallel Phrases, 2008.

No 1393 **Mattias Eriksson**: Integrated Software Pipelining, 2009.

No 1401 **Annika Öhgren:** Towards an Ontology Development Methodology for Small and Medium-sized Enterprises, 2009.

No 1410 **Rickard Holsmark:** Deadlock Free Routing in Mesh Networks on Chip with Regions, 2009.

No 1421 **Sara Stymne:** Compound Processing for Phrase-Based Statistical Machine Translation, 2009.

No 1427 **Tommy Ellqvist**: Supporting Scientific Collaboration through Workflows and Provenance, 2009.

No 1450 **Fabian Segelström:** Visualisations in Service Design, 2010.

No 1459 **Min Bao**: System Level Techniques for Temperature-Aware Energy Optimization, 2010.

No 1466 **Mohammad Saifullah**: Exploring Biologically Inspired Interactive Networks for Object Recognition, 2011

No 1468 **Qiang Liu**: Dealing with Missing Mappings and Structure in a Network of Ontologies, 2011.

No 1469 **Ruxandra Pop**: Mapping Concurrent Applications to Multiprocessor Systems with Multithreaded Processors and Network on Chip-Based Interconnections, 2011.

No 1476 **Per-Magnus Olsson**: Positioning Algorithms for Surveillance Using Unmanned Aerial Vehicles, 2011.

No 1481 **Anna Vapen**: Contributions to Web Authentication for Untrusted Computers, 2011.

No 1485 **Loove Broms:** Sustainable Interactions: Studies in the Design of Energy Awareness Artefacts, 2011.

FiF-a 101 **Johan Blomkvist:** Conceptualising Prototypes in Service Design, 2011.

No 1490 **Håkan Warnquist:** Computer-Assisted Troubleshooting for Efficient Off-board Diagnosis, 2011.

No 1503 **Jakob Rosén:** Predictable Real-Time Applications on Multiprocessor Systems-on-Chip, 2011.

No 1504 **Usman Dastgeer:** Skeleton Programming for Heterogeneous GPU-based Systems, 2011.

No 1506 **David Landén:** Complex Task Allocation for Delegation: From Theory to Practice, 2011.

No 1507 **Kristian Stavåker**: Contributions to Parallel Simulation of Equation-Based Models on Graphics Processing Units, 2011.

No 1509 **Mariusz Wzorek:** Selected Aspects of Navigation and Path Planning in Unmanned Aircraft Systems, 2011.

No 1510 **Piotr Rudol:** Increasing Autonomy of Unmanned Aircraft Systems Through the Use of Imaging Sensors, 2011.

No 1513 **Anders Carstensen:** The Evolution of the Connector View Concept: Enterprise Models for Interoperability Solutions in the Extended Enterprise, 2011.

No 1523 **Jody Foo:** Computational Terminology: Exploring Bilingual and Monolingual Term Extraction, 2012.