

Proceedings of the Workshop

**Semantic Content Acquisition  
and Representation  
(SCAR) 2007**

Edited by:

Magnus Sahlgren    Ola Knutsson  
SICS                    KTH CSC

Workshop at NODALIDA 2007  
May 24 2007, Tartu, Estonia

SICS Technical Report T2007:06  
ISSN 1100-3154  
ISRN:SICS-T-2007/06-SE

# Workshop Programme

May 24 2007

09:45–10:00 *Introduction by the organizers*

10:00–10:30 Octavian Popescu and Bernardo Magnini: Sense Discriminative Patterns for Word Sense Disambiguation

10:30–11:00 Coffee break

11:00–11:30 Henrik Oxhammar: Evaluating Feature Selection Techniques on Semantic Likeness

11:30–12:00 Jaakko Vyrinen, Timo Honkela and Lasse Lindqvist: Towards Explicit Semantic Features using Thresholded Independent Component Analysis

12:00–12:30 *Discussion on statistical methods for semantic content acquisition (led by the organizers)*

12:30–14:00 Lunch

14:00–14:15 Demo: Ontological-Semantic Internet Search (Christian Hempelmann)

14:15–14:30 Demo: Infomat - A Vector Space Visualization Tool (Magnus Rosell)

14:30–15:00 Anne Tamm: Representing achievements from Estonian transitive sentences

15:00–15:30 *Summary of results and conclusions (led by the organizers)*

# Workshop on Semantic Content Acquisition and Representation SCAR 2007

**Magnus Sahlgren**  
SICS  
Box 1263  
SE-164 29 Kista, Sweden  
mange@sics.se

**Ola Knutsson**  
HCI-group  
KTH CSC  
100 44 Stockholm, Sweden  
knutsson@csc.kth.se

## 1 Workshop theme

Language has *aboutness*; it has meaning, or semantic content. This content exists on different levels of linguistic granularity: basically any linguistic unit from an entire text to a single morpheme can be said to have some kind of semantic content or meaning. We as human language users are incredibly adept at operating with, and on, meaning. This semantic proficiency is intuitive, immediate, and normally requires no or little processing effort. However, this ability seems to be largely unarticulated. While in normal language use questions about meaning rarely beget problems beyond definitional or referential unclearities, in linguistic studies of language the concept of meaning is one of the most problematic ones.

We as (computational) linguists are highly adept at dissecting text on a number of different levels: we can perform grammatical analysis of the words in the text, we can detect animacy and salience, we can do syntactic analysis and build parse trees of partial and whole sentences, and we can even identify and track topics throughout the text. However, we are comparatively inept when it comes to identifying and using the content or the meaning of the text and of the words. Or, to put matters in more concise terms, even though there are theories and methods that claim to accomplish this, there is a striking lack of consensus regarding both acquisition, representation, and practical utility of semantic content.

The theme of this workshop is the status of meaning in computational linguistics. In particular, we are interested in the following questions:

- Is there a place in linguistic theory for a

situation- and speaker-independent semantic model beyond syntactic models?

- What are the borders, if any, between morphosyntax, lexicon and pragmatics on the one hand and semantic models on the other?
- Are explicit semantic models necessary, useful or desirable? (Or should they be incidental to morphosyntactic and lexical analysis on the one hand and pragmatic discourse analysis on the other?)

## 2 Workshop objective

The aim of this workshop is not only to provide a forum for researchers to present and discuss theories and methods for semantic content acquisition and representation. The aim is also to discuss a common evaluation methodology whereby different approaches can be adequately compared. In comparison with the information retrieval community's successful evaluation campaigns (TREC, CLEF, and NTCIR), which have proven to be widely stimulating factors in information retrieval research, research in semantic content acquisition and representation is hampered by the lack of standardized test settings and test collections.

As a first step towards a remedy for this deficiency, we encouraged participants to apply their methods, or relate their theories, to a specific test corpus that is available in several of the Nordic languages and English. As a matter of convenience, we opted to use the Europarl corpus,<sup>1</sup> which con-

<sup>1</sup>At publication time, the Europarl corpus is freely available at: <http://people.csail.mit.edu/koehn/publications/europarl/>

sists of parallel texts from the plenary debates of the European Parliament in 11 European languages. We wanted participants to demonstrate what kind of results their methods can yield.

Our goal was that in this workshop, the relevance of an approach to meaning is judged only by what the approach can tell us about real language data. The overall purpose of this workshop is thus to put theories and models into action.

### 3 Workshop submissions

We encouraged submissions in the following areas:

- **Discussions** of foundational theoretical issues concerning meaning and representation in general.
- **Methods** for supervised, unsupervised and weakly supervised acquisition (machine learning, statistical, example- or rule-based, hybrid etc.) of semantic content.
- **Representational** schemes for semantic content (wordnets, vectorial, logic etc.).
- **Evaluation** of semantic content acquisition methods, and semantic content representations (test collections, evaluation metrics etc.).
- **Applications** of semantic content representations (information retrieval, dialogue systems, tools for language learning etc.).

We received two contributions that discuss methods for acquisition of semantic content: Jaakko Väyrynen, Timo Honkela and Lasse Lindqvist presents a method for making explicit the latent semantics of a Latent Semantic Analysis space through a statistical technique called Independent Component Analysis; Henrik Oxhammar investigates the use of feature selection techniques to extend semantic knowledge sources in the medical domain.

One contribution deals with representation of semantic content: Anne Tamm uses Lexical-Functional Grammar as a possible means to read semantics off syntax and morphology.

One contribution discusses an application of semantic content representations: Octavian Popescu

and Bernardo Magnini develops an algorithm for the automatic acquisition of sense discriminative patterns to be used on word sense disambiguation.

Finally, we received two contributions that demonstrate systems that use, or make use of, semantic content: Magnus Rosell demonstrates a visualization tool for vector space models; Christian F. Hempelmann, Victor Raskin, Riza C. Berkan and Katrina Triezenberg demonstrates a search engine that uses ontological semantic analysis.

### 4 Acknowledgement

We wish to thank our Program Committee: Peter Bruza (Queensland University of Technology, Australia), Gregory Grefenstette (CEA LIST, France), Jussi Karlgren (SICS, Sweden), Alessandro Lenci (University of Pisa, Italy), Hinrich Schütze (University of Stuttgart, Germany), Fabrizio Sebastiani (Consiglio Nazionale delle Ricerche, Italy), Dominic Widdows (MAYA Design, USA).

# Sense Discriminative Patterns for Word Sense Disambiguation

Octavian Popescu  
FBK-irst, Trento (Italy)  
popescu@itc.it

Bernardo Magnini  
FBK-irst, Trento (Italy)  
magnini@itc.it

## Abstract

Given a target word  $w_i$  to be disambiguated, we define a class of local contexts for  $w_i$  such that the sense of  $w_i$  is univocally determined. We call such local contexts *sense discriminative* and represent them with sense discriminative (SD) patterns of lexico-syntactic features. We describe an algorithm for the automatic acquisition of minimal SD patterns based on training data in SemCor.

We have tested the effectiveness of the approach on a set of 30 highly ambiguous verbs. Results compare favourably with the ones produced by a SVM word sense disambiguation system based on bag of words.

## 1 Introduction

Leacock, Towell and Voorhes (1993) distinguish two types of contexts for a target word  $w_i$  to be disambiguated: a *local context*, which is determined by information on word order, distance and syntactic structure and is not restricted to open-class words, and a *topical context*, which is the list of those words that are likely to co-occur with a particular sense of  $w_i$ .

Several recent approaches to Word Sense Disambiguation (WSD) take advantage of the fact that the words surrounding a target word  $w_i$  provide clues for its disambiguation. A number of syntactic and semantic features in a local context  $[w_{i-n}, \dots, w_{i-1}, w_i, w_{i+1}, \dots, w_{i+n}]$  (where  $n$  is usually not higher than 3) are considered, including the token itself, the Part of Speech, the lemma, the semantic domain of the word, syntactic relations and semantic concepts. Results in supervised WSD (see, among the others, Yarowsky 1992, Pederson 1998, Ng&Lee

2002) show that a combination of such features is effective.

We think that the potential of local context information for WSD has not been fully exploited by previous approaches. In particular, this paper addresses the following issues:

1. As our main interest is WSD, we are interested in local contexts which univocally select a sense  $s_j$  of  $w_i$ . We call such contexts “sense discriminative” and we represent them as *sense discriminative* (SD) patterns of lexico-syntactic features. According to the definition, if a SD pattern matches a portion of the text, then the sense of the target word  $w_i$  is univocally determined. We propose a methodology for automatically acquiring SD patterns on a large scale.
2. Intuitively, the size of a local context should vary depending on  $w_i$ . For instance, if  $w_i$  is a verb, a preposition appearing at  $w_{i+3}$  may introduce an adjunct argument, which is relevant for selecting a particular sense of  $w_i$ . The same preposition at  $w_{i+3}$  may cause just a noise if  $w_i$  is an adjective. We propose that the size of the local context  $C$ , relevant for selecting a sense  $s_j$  of  $w_i$ , is dynamically set up, such that  $C$  is the minimal context for univocally selecting  $s_j$ .
3. An important property of some minimal SD patterns is that each element of the pattern has a specific meaning, which does not change when new words are added. As a consequence, all the words  $w_{i+n}$  are disambiguated. We call the relations that determine a single sense for each element of a minimal sense discriminative pattern *chain clarifying relationships*. The acquisition method we propose is crucially based on this property.

According to the above mentioned premises the present paper has two goals: (i) design an algorithm for the automatic acquisition of minimal sense discriminative patterns; (ii) evaluate the patterns in a WSD task.

With respect to acquisition, our method is based on the identification of the minimal set of lexico-syntactic features that allow the discrimination of a sense for  $w_i$  with respect to the other senses of the word. The algorithm is trained on a sense tagged corpus (experiments have been carried on SemCor) and starts with a dependency-based representation of the syntactic relations in the sentence containing  $w_i$ . Then, elements of the sentence that do not bring sense discriminative information are filtered out; we thus obtain a minimal SD pattern.

As for evaluation, we have tested sense discriminative patterns on a set of thirty high polysemous verbs in SemCor. The underlying hypothesis is that SD patterns are effective in particular in the case of the scarcity of the training data. We provide a comparison of the SD-based disambiguation with a simple SVM-based system, and we show that our system fares significantly higher in performance.

The paper is organized as follows. Section 2 introduces sense discriminative patterns and chain clarifying relations in a more formal way. In Section 3 we present the algorithm we have used to identify sense discriminative contexts starting from a sense annotated corpus. In Section 4 we present the results we have obtained applying SD patterns on a WSD task and we compare them against a supervised WSD system based on SVM and the bag of word approach. In section five we review related works and point out the novelty of our approach. We conclude with section six, in which we present our conclusions and directions for further research.

## 2 Chain Clarifying Relationships (CCR)

Consider the examples below:

1a) *He drove the girl to her father/to the church/ to the institute/to L.A.*

1b) *He drove the girl to ecstasy/to craziness/ to despair/ to euphoria.*

Using a sense repository, such as WordNet 1.6, we can assign a sense to any of the words in both

1a) and 1b). In 1a) the word “drive” has the sense drive#3, “cause someone or something to move by driving” and in 1b) it has the sense drive#5, “to compel or force or urge relentlessly or exert coercive pressure on”. By comparing 1a) and 1b) and by consulting an ontology, we can identify a particular feature which characterizes the prepositional complements in 1b), and which we hold responsible for the sense of “drive” in this sentence. The relationship between this feature and the sense of “drive” holds only in the common context of 1a) and 1b), namely the prepositional complement. Example 2) below shows that if this local context is not present, then the word “euphoria” does not have a disambiguating function for “drive”.

2) *He drove the girl back home in a state of euphoria.*

However, the syntactic configuration alone does not suffice, because lexical features must be taken into account, too. The particular sense combination is determined by a chain-like relationship: the sense of “girl” is determined by its function as object of the verb “drive”; the sense of “drive” is determined by the nature of the prepositional complement. We call such relationship a chain clarifying relationship (CCR). The importance of CCRs for WSD resides in the fact that by knowing the sense of one component, specific senses are forced for the others components.

In what follows we give a formal definition of the CCR, which will help us to devise an algorithm for finding CCR contexts. We start from the primitive notion of *event* (Giorgi and Pianesi, 1997). We assume that there is a set:

$$E = \{e_1, e_2, \dots, e_n\}$$

whose elements are events, and that each event can be described by a sequence of words. Let us now consider three finite sets,  $W$ ,  $S$  and  $G$ , where:

$$W = (w_1, w_2, \dots, w_w)$$

is the set of words used to describe events in  $E$ ,

$$S = (w_{11}, w_{12}, \dots, w_{1m_1}, w_{21}, w_{22}, \dots, w_{2m_2}, \dots, w_{mw_m})$$

is the set of words with senses, and

$$G = (g_1, g_2, \dots, g_{mg})$$

is the set of grammatical relations.

If  $e$  is an event described with words  $w_1, w_2, \dots, w_n$  we assume that  $e$  assigns a sense  $w_{ij}$  and a grammatical relation  $g_i$  to any of these words. Therefore we consider  $e$  to be the function:

$$e: P(\{w_1, w_2, \dots, w_n\}) \rightarrow (SxG)^n$$

$$e(w_1, w_2, \dots, w_n) = (w_{1i1}xg_{i1}, w_{2i2}xg_{i2}, \dots, w_{ni n}xg_{in})$$

For a given  $k$  and  $l$ , such that  $1 \leq k \leq l \leq n$ , and  $k$  components of  $e(w_1, w_2, \dots, w_n)$  we call the *chain clarifying relation (CCR)* of  $e$  the function:

$$e_{CCR}: (SxG)^{n-k} \times (WxG)^k \rightarrow (SxG)^l$$

where  $e_{CCR}(w_{1i1}xg_{i1}, w_{2i2}xg_{i2}, \dots, w_{kik}xg_{ik}, w_{k+1i_{k+1}}xg_{i_{k+1}}, w_{k+2i_{k+2}}xg_{i_{k+2}}, \dots, w_{ni n}xg_{in}) = (w_{1i1}, w_{2i2}, \dots, w_{li l})$

The above definition captures the intuition that in certain contexts the senses of some of the words impose a restriction on the senses of other words. When  $l=n$  we have a complete sense specification, therefore the  $e_{CCR}$  function gives a sense for any of the words of  $e$ .

Let us consider two events  $e$  and  $e'$  such that they differ only with respect to two slots:

$$e(w_1, w_2, \dots, w_n) = (w_{1i1}xg_{i1}, w_{2i2}xg_{i2}, w_{kik}xg_{ik}, \dots, w_{ni n}xg_{in})$$

$$e'(w_1, w_2, \dots, w_n) = (w'_{1i1}xg_{i1}, w_{2i2}xg_{i2}, \dots, w_{kik'}xg_{ik'}, \dots, w_{ni n}xg_{in}).$$

We infer that there is a lexical difference between  $w_i$  and  $w_i'$  which is responsible for the sense difference between  $w_{kik}$  and  $w_{kik'}$ . If precisely this difference is found to be preserved for any  $e(w_1, w_2, \dots, w_n, w_{n+1}, w_{n+2}, \dots, w_m)$ , then the sequence  $(w_{1i1}xg_{i1}, w_{2i2}xg_{i2}, \dots, w_{kik-1}xg_{ik-1}, w_{k,ik+1}xg_{ik+1}, \dots, w_{ni n}xg_{in})$  is a CCR.

The examples in 1a) are local contexts having the sense constancy property in which a particular type of CCR holds. We can express a CCR under the shape of a pattern, which, by the way in which it has been determined, represents a sense discriminative (SD) pattern. A SD pattern classifies the words that fulfill its elements in classes which are valid only with respect to a particular CCR. A simple partitioning of the

nouns, for example, in semantic classes independently of a CCR may not lead to correct predictions. On the one hand, a semantic class which includes “father” and “church” may be misleading with respect to their senses in 1a), and, on the other hand, a semantic class which includes “father”, “church”, “institute”, “L.A.” is probably too vague. This suggests that rather than starting with a set of predefined features and syntactic frames, it is more useful to discover these on the basis of an investigation of sense constancy. Also, there is not a strictly one to one relationship between predicate argument structure and CCR: as our experiments showed, there are cases when only some complements or adjuncts in the sentence play an active role in disambiguation.

### 3 Acquisition of SD Patterns

The algorithm we have used for the acquisition of SD patterns consists mainly in two steps: first, for each sense of a verb, all the potential CCRs are extracted from a sense annotated corpus; second, all the patterns which are not sense discriminative are removed.

In accordance with the definition of CCRs, we have tried to find CCRs for verbs by considering only the words that have a dependency relationship with the verbs. Our working hypothesis is that we may find valid CCRs only by taking into account the external and internal arguments of the verbs. Thus we have considered the dependency chains (DC) rooted in verbs.

#### 3.1 Finding Dependency Chains

In a dependency grammar (Mel'čuk 1988) the syntactic structure of a sentence is represented in terms of dependencies between words. The dependency relationships are between a *head* and a *modifier* and are of the type one to many: a head may have many modifiers but there is only one head for each modifier. The same word may be a head or a modifier of some other words; thus the dependency relationships constitute subtrees. Here we are interested mainly in finding the subtrees rooted in predicative verbs.

After running a set of tests in order to check the accuracy of various parsers, (i.e. Lin 1998, Bikel 2004) we have decided to use the Charniak's parser which is a constituency parser. The choice was determined by the fact that the VP constituents were determined with accuracy

below 70% by the other parsers. In order to extract the dependency relationships from the Charniak's parser output we have relied on previous work on heuristics for finding the heads of the NP constituents and their types of dependency relationships (see, among others, Ratnaparkhi, 1997; Collins, 1999).

### 3.2. SD Patterns Selection

The extraction of CCRs is an iterative process that starts with the dependency trees for a particular sense of a word. The algorithm builds at each step new candidates through a process of generalization of the entities that fulfil the syntactic slots of a pattern. The candidates which are not sense discriminative are discarded and the process goes on till there are no new candidates.

We start with the dependency chains rooted in verbs extracted from a sense tagged corpus. For each verb sense, the dependency chains are clustered according to their syntactic structure. Initially, all dependency chains are considered candidates. Chains that are found in at least two clusters are removed. After this "remove" procedure, since each chain individuates a unique sense combination, in each cluster remain only the patterns which are SD patterns according to the training examples.

In order to find the minimal SD patterns we build minimal SD candidates from the existing patterns by means of a process of generalization. Inside each cluster, we search for similarities among the entities that fulfil a particular slot. For this purpose we use SUMO (Niles & all 2003), an ontology aligned to WordNet. Two or more entities are deemed to be similar if they share the same SUMO attribute. Similar entities are "generalized" by the common attribute. Then, all the patterns that have similar entities in the same slot and are identical with respect to all the other slots are collapsed into one new candidate. The algorithm repeats the remove procedure for the new candidates; the ones that pass are considered SD patterns. We stop when no new candidates are proposed.

For example the sentences in 1b) lead to to the following minimal SD pattern for the sense 3 of the verb drive:

(V=drive#3 S=[Human], O=[Human] P=to PP\_1  
=[EmotionalState])

## 4. Experiments

We have designed an experiment in order to evaluate the effectiveness of the SD patterns approach. We have chosen a set of thirty highly polysemic verbs which are listed in Table 1.

### 4.1 Training and Test Data

Since the quality of SD patterns is directly correlated with the accuracy of DCs, we have decided to extract the verb rooted DCs from a hand annotated corpus. For training, we considered the part of the Brown corpus which is also a part of the Penn Tree Bank. In this corpus verbs are annotated with the senses of WordNet and all sentences are parsed. For a part of the corpus we have annotated the senses of the nouns which are heads of the verbs' internal and external arguments and we have written a Perl script which transforms the parsed trees into dependency trees. Because in the Penn Tree bank the grammatical function is given, this transformation is accurate.

Some of the senses of the test verbs have only a few occurrences. In order to have a better coverage of less frequent senses we added new examples, such that there are at least ten examples per each verb sense. These new examples are simplified instances of sentences from the BNC. They are made up only from the subject and the respective VP as it appears in the original sentences. The subject has been explicitly written in the cases where in the original sentence there is a trace or a relative pronoun. We parsed them with the Charniak's Parser and we extracted the dependency chains. We manually checked 140 of them and we found 98% accuracy.

The second column of Table 1 represents the number of occurrences of test verbs in the corpus common to the Brown and to the Tree Bank. The third column represents the number of examples for which we have annotated the arguments. The fourth column represents the number of the added examples. In the fifth column we list the number of patterns we found in the training corpus for each verb. In the sixth and in the seventh columns we list the minimum and the maximum number of patterns respectively. Number 0 as minimum means that there was no way to find a difference between at least two senses. The test corpus was the part of Brown corpus which is generally known as Semcor.

## 4.2 Results and Discussion

We compared the results we obtained with SD patterns against a SVM-based WSD system. For each word in a local context, features were the lemma of the word, its PoS, and its relative distance from the target word. The training corpus for the SVM was formed by all the

sentences from the common part of the Brown and the Pen Tree Bank corpora and the new added examples from the BNC. Therefore, the training corpus for the SVM includes the training corpus for SD patterns (more than 1000 examples in addition for SVM system).

verb	#occ	#tag	#add	#pat	#min	#max	verb	#occ	#tag	#add	#pat	#min	#max
begin	188	80	3	12	2	3	match	18	18	30	8	0	3
call	108	80	40	25	1	8	move	118	90	40	29	2	8
carry	68	68	40	32	1	6	play	121	80	40	29	0	5
come	317	100	30	36	1	9	pull	24	24	20	13	1	3
develop	80	60	20	17	0	3	run	97	90	50	42	0	11
draw	40	40	60	38	1	3	see	445	120	30	36	0	8
dress	10	10	30	7	1	3	serve	112	70	10	14	1	3
drive	72	40	40	14	1	5	strike	37	37	20	9	1	3
face	66	40	10	9	0	3	train	13	13	40	14	1	4
find	254	100	20	26	0	7	treat	34	34	10	11	0	4
fly	27	27	10	16	1	6	turn	85	40	40	16	1	3
go	229	100	20	35	0	12	use	291	60	40	21	2	5
keep	166	70	30	28	2	8	wander	8	8	10	4	1	3
leave	167	100	30	31	1	9	wash	1	1	30	8	0	3
live	124	70	10	11	1	3	work	120	80	30	24	1	6

Table 1: Training corpus for SD patterns.

The second column of Table 2 lists the total number of the occurrences of the test verbs in Semcor. In the third column we list the results obtained using SD patterns and in the fourth the results obtained using the SVM system. The number of senses the in corpus, which are found

by each approach, are listed in the fifth and sixth column respectively. The SD patterns approach has scored better than SVM, 49.32% vs. 42.28%.

verb	#occ	#SDP	#SVM	#senses SDPS	#senses SVM	verb	#occ	#SDP	#SVM	#senses SDPS	#senses SVM
begin	203	178	135	5	3	match	31	14	10	3	1
call	148	73	52	8	6	move	137	61	46	7	5
carry	77	41	29	10	6	play	181	87	61	11	6
come	354	184	130	9	5	pull	46	26	28	4	2
develop	114	42	28	7	4	run	131	72	30	17	5
draw	73	35	16	9	6	see	578	213	259	15	8
dress	36	18	21	3	1	serve	98	39	42	10	8
drive	68	23	21	5	3	strike	43	17	13	8	4
face	196	58	62	4	2	train	47	23	27	4	1
find	420	204	97	6	7	treat	48	13	9	3	1
fly	30	22	15	4	1	turn	130	63	74	8	3
go	256	171	125	13	4	use	439	199	356	4	1
keep	153	103	86	8	4	wander	8	3	5	2	1
leave	222	121	83	10	6	wash	39	20	21	3	2
live	120	45	57	4	3	work	344	185	79	9	5

Table 2: Comparative results for using SD patterns and SVM bag of word in WSD.

The range of the senses the SD patterns approach is able to identify is more than two times greater than the SVM system.

We also show how these two approaches perform in the cases of the less frequent senses in the corpus. Table 3, second column, reports the number of senses considered, the third, the cumulative number of occurrences in the test corpus; the fourth and the fifth columns, report the correct matching for SD patterns and for SVM. Results for SD patterns are higher than the ones obtained with SVM: 34.72% vs.13.74%.

The patterns we have obtained are generally very precise: they identify the correct sense with more than 85% accuracy. However, they are not error proof. We believe there are mainly three reasons for why the SD patterns lead to wrong predictions: (i) the approximation of CCRs with DCs, (ii) the parser accuracy, and (iii) the relative small size of the training corpus. The CCRs are determined only considering the words that have a direct dependency relationship with the target word. However, in some cases, the

information which allows word disambiguation may be beyond phrase level (Wilks&Stevenson, 1997 – 2001). The parser accuracy plays an important role in our methodology. While the method of considering only simple sentences in the training phase seems to produce good results, further improvements are required. Finally, the dimension and the diversity of sentences in the training corpus play an important role for the final result. The smaller and the more homogenous the training corpus is, the bigger the probability that a DC, which is not a SD pattern, is considered erroneously as such.

In some cases, such as semantically transparent nouns (Fillmore et al. 2002), the information which allows the correct disambiguation of the nouns that are heads of NPs, is found within the NPs. Our approach cannot handle these cases. Our estimation is that they are not very frequent, but, nevertheless, a proper treatment of such nouns contributes to an increase in accuracy.

verb	#senses	#occ	#SDP	SVM	verb	#senses	#occ	#SDP	SVM
begin	2	11	8	5	match	3	7	1	0
call	3	10	5	2	move	6	26	10	4
carry	12	30	13	4	play	13	31	16	2
come	7	20	9	2	pull	5	17	5	2
develop	10	33	13	3	run	20	46	16	6
draw	20	73	35	16	see	10	40	3	2
dress	3	13	3	2	serve	7	27	12	8
drive	5	16	4	1	strike	8	17	8	4
face	4	16	2	0	train	5	14	3	0
find	2	14	4	1	treat	1	7	2	0
fly	5	9	5	2	turn	11	31	7	4
go	14	45	14	5	use	4	19	2	2
keep	9	24	10	3	wander	2	8	4	5
leave	11	58	22	7	wash	2	9	3	3
live	3	13	2	0	work	10	34	9	3

Table 3: Results for less frequent senses.

## 5. Related Works

Based on the Harris' Distributional Hypothesis (HDH), many approaches to WSD have focused on the contexts formed by the words surrounding the target word. With respect with verb behaviour, selectional restrictions have been used in WSD ( see among others Resnik 1997,

McCarthy, Carroll, Preis 2001, Briscoe 2001). Also, Hindle (Hindle 1990) has tried to classify the English nouns in similarity classes by using a mutual information measure with respect to the subject and object roles. Such information is very useful only in certain cases and, as such, it might not be used directly for doing WSD.

Lin and Pantel (Lin, Pantel 2001) transpose the HDH from words to dependency trees. However, their measure of similarity is based on a frequency measure. They maintain that a (slotX, he) is less indicative than a (slotX, sheriff). While this might be true in some cases, the measure of similarity is given by the behaviour of the other components of the contexts: both “he” and “sheriff” act either exactly the same with respect to certain verb meanings, or totally different with respect to some others. A classification of these cases is obviously of great importance for WSD. However, this classification problem cannot be addressed by employing the method the authors present. The same arguments are also valid in connection with the method proposed by Li&Abe, based on MDL (Li&Abe 1998). Another limitation of these methods, which our proposal overcomes, is that they only consider subject and object positions. However, in many cases the relevant entities are complements, and/or prepositions and particles. It has been shown that closed class categories, especially preposition and particles, play an important role in disambiguation and wrong prediction are made if they are not taken into account. (see, among others, Collins and Brooks 1995, Stetina&Nagao 1997). Our results have shown that only a small fraction (27%) of SD patterns include just the subject and/or the object.

Zhao, Meyers and Grishman (Zhao, Meyers and Grishman 2004, Zhao) proposed a SVM application to slot detection, which combines two different kernels, one of them being defined on dependency trees. Their method tries to identify the possible fillers for an event, but it does not attempt to treat ambiguous cases; also, the matching score algorithm makes no distinction between the importance of the words, considering equal matching score for any word within two levels.

Pederson and al. (1997-2005) have clustered together the examples that represent similar contexts for WSD. However, given that they adopt mainly the methodology of ordered pairs of bigrams of substantive words, their technique works only at the word level, which may lead to a data sparseness problem. Ignoring syntactic clues may increase the level of noise, as there is no control over the relevance of a bigram.

Many of the purely syntactic methods have considered the properties of the subcategorization frame of verbs. Verbs have

been partitioned in semantic classes based mainly on Levin’s classes of alternation. (Dorr&Jones 1996, Palmer&all 1998-2005, Collins, McCarthy, Korhonen 2002, Lapata&Brew 2004). These semantic classes might be used in WSD via a process of alignment with hierarchies of concepts as defined in sense repository resources (Shin&Mihalcea 2005). However the problem of the consistency of alignment is still an open issue and further research must be pursued before applying these methods to WSD.

## 6. Conclusion and Further Research

We have presented a method for determining a particular type of local context, within which the relevant entities for WSD can be discovered. Our experiment has shown that it is possible to represent such contexts as Sense Discriminative patterns. The results we obtained applying this method to WSD compare favourably with other results.

One of the major limitations in achieving higher results is the small size of the training corpus. The quality of SD patterns depends to a great extent on the variety of examples in the training corpora.

The CCR property of some local context allows a bootstrapping procedure in the acquisition of SD patterns. This remains an issue for further research.

The SD patterns for verbs, characterize the behaviour of words which constitute a VP phrase with respect to the word senses. In fact, to each pattern corresponds a regular expression. Thus a decision list algorithm could be implemented in order to optimize the matching procedure.

## References

- Brew, L., 2004, “Verb class disambiguation using informative priors Computational Linguistics”, Volume 30, pages: 45 – 73.
- Briscoe, T., 2001, “From Dictionary to Corpus to Self-Organizing Dictionary: Learning, Valency Associations in the Face of Variation and Change”, In Proceedings of Corpus, Linguistics. Lancaster University, UK.
- Carroll J., Briscoe T., 2001 “High precision extraction of grammatical relations”, Workshop on Parsing Technologies, Beijing.

- Collins M., Brooks J., 1995. "Prepositional phrase attachment through a backed-off model". In Proceedings of the Third Workshop on Very Large Corpora, pages 27--38, Cambridge.
- Collins, M. 1999, "Head-Driven Statistical Models for Natural Language Parsing" Ph.D. thesis, University of Pennsylvania.
- Dorr, B., Jones, D., 1999 "Acquisition of Semantic Lexicons in Breadth and Depth of Semantic Lexicons, edited by Evelyne Viegas. Kluwer Press. .
- Hindle, D., 1990, "Noun classification from predicate argument structures", In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 268--275.
- Fillmore, C., Baker, C. and Sato, Hiroaki, 2002: "Seeing Arguments through Transparent Structures". In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC). Las Palmas. 787-91
- Korhonen, A., 2002. "Subcategorization Acquisition", PhD thesis published as Technical Report UCAM-CL-TR-530. Computer Laboratory
- Leacock, C., Towell, G., & Voorhes, E., "Towards Building Contextual Representations of Word Senses Using Statistical Models", In Proceedings, SIGLEX workshop: Acquisition of Lexical Knowledge from Text, ACL., 1993.
- Lee, Y., Ng, H., 2002, "An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation", In Proceedings of EMNLP'02, pages 41--48, Philadelphia, PA, USA.
- Li, D., Abe, N. 1998, "Word Clustering and Disambiguation Based on Co-occurrence Data". COLING-ACL : 749-755.
- Lin, D., Pantel, P., 2001, "Discovery of Inference Rules for Question Answering", Natural Language Engineering 7(4):343-360.
- McCarthy, D., Carroll, J. and Preiss, J. (2001) "Disambiguating noun and verb senses using automatically acquired selectional preferences", In Proceedings of the SENSEVAL-2 Workshop at ACL/EACL'01 , Toulouse, France.
- Ratnaparkhi, A., 1997. A Linear Observed Time Statistical Parser Based on Maximum Entropy Models. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing.
- Dang, T., Kipper, K., Palmer, K., Rosenzweig, J., "Investigating regular sense extensions based on intersective Levin classes",. Coling-ACL98 , Montreal CA, August 11-17, 1998.
- Pederson, T., 1998, "Learning Probabilistic Models of Word Sense Disambiguation ", Southern Methodist University, 197 pages (PhD Dissertation)
- Pederson T., 2005, "SenseClusters: Unsupervised Clustering and Labeling of Similar Contexts" , Proceedings of the Demonstration and Interactive Poster Session of the 43rd Annual Meeting of the Association for Computational Linguistics.
- Resnik, P. 1997, "Selectional Preference y Sense Disambiguation, in Proceedings of the SIGLEX WorkShop Tagging Text with Lexical Semantics: Why, What y How?." Washington.
- Shi, L., Mihalcea, R., 2005, "Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing", in Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico.
- Stevenson K., Wilks., Y., 2001 "The interaction of knowledge sources in word sense disambiguation", Computational Linguistics, 27(3):321--349.
- Zhao, S., Meyers A., and Grishman, R., 2004 , Proceedings of the 20th International Conference on Computational Linguistics Geneva, Switzerland.
- Stetina J, Nagao M 1997 "Corpus based PP attachment ambiguity resolution with a semantic dictionary." In Zhou J, Church K (eds), Proc. of the 5th Workshop on very large corpora, Beijing and Hongkong, pp 66-80.
- Yarowsky, D. 1992. "Word-sense disambiguation using statistical models of Roget's categories trained on large corpora". In COLING-92.

# Evaluating Feature Selection Techniques on Semantic Likeness

**Henrik Oxhammar**

Stockholm University

henrik.oxhammar@ling.su.se

## Abstract

In this paper, we describe a first in a series of experiments for determining the usefulness of standard feature selection techniques on the task of enlarging large semantic knowledge sources. This study measures and compares the performance of four techniques, including odds ratio, chi-square, and correlation coefficient. We also include our own procedure for detecting significant terms that we consider as a baseline technique. We compare lists of ranked terms extracted from a medical corpus (OHSUMED) to terms in a medical vocabulary (MeSH).

Results show that all four techniques tend to rank significant terms higher than less significant terms, although chi-square and correlation coefficient clearly outdo the other techniques on this test. When comparing the order of terms with their semantic relatedness to particular concepts in our gold standard, we notice that our baseline technique suggests orderings of terms that conform more closely to the conceptual relations in the vocabulary.

## 1 Introduction

Controlled vocabularies<sup>1</sup> are records of cautiously elected terms (single words or phrases) symbolizing concepts (objects) in a particular domain. Con-

trolled vocabularies are typically structured hierarchically, and explicitly represent various conceptual relations, such as the broader- (generic), narrower- (specific) and synonymy (similar) relations. Furthermore, each concept is typically associated with a distinctive code that bestows each concept with a unique sense. The unique sense of a concept, in combination with the concept's relationship to others, makes available a clearer and more harmonized understanding about its meaning. Controlled vocabularies exist for many domains including, the procurement- (e.g., UNSPSC, ecl@ss, CPV), patent- (e.g., IPC) and medical domain (e.g., UMLS, MeSH).

As these vocabularies are available in machine-readable format, we can use them as resources in computer applications to reduce some of the ambiguity of natural language by associating pieces of information (e.g., documents) to concepts in these vocabularies. This can allow heterogeneous information to become homogenous information and can ultimately lead to intelligent organization, standardization (interoperability), and visualization of unstructured textual information. However, these resources have a clear weakness. As trained professionals typically construct and maintain these resources by hand, their content (terms denoting concepts), and representation (relations among concepts) can quickly be out-dated. Recognizing that large quantities of electronic text are available these days, it is advantageous to acquire significant terms from these collections (semi-) automatically, and to update the concepts in controlled vocabularies with this additional information. It is essential that such a technique discrimi-

---

<sup>1</sup> Also referred to as taxonomies, nomenclatures, thesauri or (light-weight) ontologies

nate well between concept-related and concept-neutral terms.

Statistical- and information-theoretic *feature selection techniques* have proved useful in the areas of information retrieval and text categorization. In information retrieval, feature selection techniques such as *document frequency* and *term frequency/inverse document frequency* (tfidf) are often adopted for sorting out relevant documents from irrelevant ones given a certain query. In text categorization, techniques like *chi-square*, *information gain*, and *odds ratio* are applied to reduce the feature set, as to allow the classifier to learn from smaller sets of relevant terms. Interestingly, despite their known ability to identify significant and discriminative terms for categories, it seems that no extensive study has been made that empirically establishes the suitability of the same techniques for the task of enhancing the content of large (semantic) knowledge sources such as controlled vocabularies.

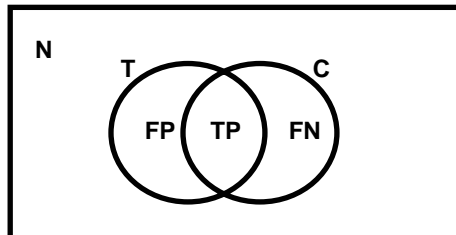
This study evaluates and compares the performance of four well-known feature selection techniques when applied to the task of detecting concept-significant terms in texts. We describe a preliminary experiment where we let each of these techniques weight and rank terms in a collection of manually labeled medical literature (OHSUMED), and we evaluate these lists of terms by comparing them against terms symbolizing 2317 concepts in the Medical Subject Headings (MeSH) vocabulary.

## 2 Feature Selection Techniques

In text categorization, feature selection is the task of selecting a small number terms from a set of documents that best represents the meaning of these documents. (Galavotti et al., 2000) Many techniques have been developed for this task (see Sebastiani, 2002 for an overview), and we report on four such techniques in this study. The techniques we evaluated were *chi-square*, *odds ratio* and *correlation coefficient*. We also included a metric we proposed ourselves, which we name *category frequency*.

In the following formulas,  $t_j$  denotes a term and  $c_i$  a concept, where each function assigns a score to

that term, indicating how significant that term is for that particular concept. Below,  $T$  represents documents containing  $t_j$ , and  $C$  corresponds to all documents that a professional indexer has assigned to concept  $c_i$ .  $TP$  stands for documents shared by both  $c_i$  and  $t_j$ , and  $FN$  for the set of documents belonging to  $c_i$  but not including  $t_j$ .  $FP$  represents documents that do not belong to  $c_i$  but contain  $t_j$ .  $N$  represents all documents in the text collection.



### 2.1 Category Frequency ( $cf$ )

We computed the category frequency as:

$$cf(t_j, c_i) = \frac{|TP|}{|C|}$$

That is, we compute the fraction of documents shared by  $t_j$  and  $c_i$ , and the total number of concept-relevant documents. We base category frequency on the notion that the significance of a term can be determined simply by establishing its distribution among the relevant documents of a concept. With this technique, we do not take the additional distributional behavior of the term into consideration. That is, this technique will not penalize terms that have a wide distribution in a text collection and it will rank terms occurring frequently among concept-relevant documents higher than terms that occur rarely in this set. We regard category frequency as baseline technique.

### 2.2 Odds ratio ( $odds$ )

The odds of some event taking place, is the probability of that event occurring divided by the probability of that event not taking place. (Freedman et al. 1991)

The rationale behind Odds ratio is that a term is distributed differently among relevant and non-relevant documents to a concept, and Odds ratio

determines whether it is equally probable that we find that term in both these sets of documents. We computed the Odds ratio according to the formula given by Mladenic (1998):

$$odds(t_j, c_i) = \frac{\frac{|TP|}{|C|} \bullet \frac{1 - \frac{|FP|}{|N-C|}}{1 - \frac{|TP|}{|C|} \bullet \frac{|FP|}{|N-C|}}}{\frac{|TP|}{|C|} \bullet \frac{1 - \frac{|FP|}{|N-C|}}{1 - \frac{|TP|}{|C|} \bullet \frac{|FP|}{|N-C|}}}$$

To be more precise, Odds ratio computes the ratio between the probability of term  $t_j$  occurring in the relevant document set of concept  $c_i$ , and the probability of  $t_j$  occurring in documents that are not relevant to  $c_i$ . Therefore, in contrast to category frequency, Odds ratio additionally considers the distribution of  $t_j$  in those documents that are not relevant to  $c_i$ , and will thereby decrease the significance of those terms that occur frequently in that set.

### 2.3 Chi-square (*chi*)

Chi-square measures the difference between *observed* values in some sample and values we can *expect* to observe in this sample. (Freedman et al. 1991). When we apply chi-square to perform feature selection, we assume that a term  $t_j$  and a concept  $c_i$  are independent of each other. Next, we test this hypothesis by measuring the difference between those co-occurrence relations between  $t_j$  and  $c_i$  we have observed in our text collection, and those co-occurrence relations we can expect to happen by chance. If chi-square determines that those values we have observed are significantly different from those expected values, we reject initial hypothesis and conclude that some significant relationship exists between term  $t_j$  and concept  $c_i$ .

We computed the chi-square according to the definition by given by Yang and Pedersen (1997):

$$chi(t_j, c_i) = \frac{|N| \left[ \frac{|TP| \bullet |N - (T - C + TP)| - |TP| \bullet |FN|}{|T| \bullet |N - A| \bullet |C| \bullet |N - C|} \right]^2}{|N| \left[ \frac{|TP| \bullet |N - (T - C + TP)| - |TP| \bullet |FN|}{|T| \bullet |N - A| \bullet |C| \bullet |N - C|} \right]^2}$$

If we detect no difference between observed and expected values, then  $t_j$  and  $c_i$  are truly independent and we obtain a value of zero for  $t_j$ . Moreover, chi-square regards terms as less significant when smaller differences are obtained, while considering

terms as more significant when bigger differences are observed.

### 2.4 Correlation Coefficient (*cc*)

Ng et al. (1997) offer a variant to the chi-square metric. In contrast to chi-square, correlation coefficient assigns a negative value to a term  $t_j$  when a weaker correspondence between  $t_j$  and concept  $c_i$  has been observed. Ng et al. motivate their proposed technique by saying that, if there is some suggestion that a term is significant in the relevant document set then that term is preferred over terms that are significant in both relevant and non-relevant documents. This technique diminishes the significance of terms occurring in non-relevant documents considerably, while more drastically promoting terms that frequently occur in relevant documents to a concept  $c_i$ .

$$cc(t_j, c_i) = \frac{\sqrt{|N|} \left[ \frac{|TP| \bullet |N - (T - C + TP)| - |TP| \bullet |FN|}{|T| \bullet |N - A| \bullet |C| \bullet |N - C|} \right]^2}{\sqrt{|T| \bullet |N - A| \bullet |C| \bullet |N - C|}}$$

## 3 Experimental Setup

In this section, we explain our data and experimental methodology.

### 3.1 Controlled Vocabulary

Medical Subject Headings (MeSH)<sup>2</sup> is one of the more famous controlled vocabularies to date. MeSH's primary purpose is as a tool for indexing medical-related texts and it is an essential aid when searching for biomedical and other health-related literature in the Medline Database<sup>3</sup>.

MeSH is designed and updated (once-a-year) by trained professionals and it represents a large assortment of concepts from the medical domain. The latest version (2007) contains a total of 22,997 so-called *descriptors* which are terms that symbolize these concepts. Accompanying each concept is a unique identification code (so called *tree number*). This code determines the precise location of each concept in the hierarchy, and from it, we can resolve which terms give a more general definition of a particular concept (i.e., the descriptors of its ancestral concepts), which terms describe similar

<sup>2</sup> <http://www.nlm.nih.gov/mesh/>

<sup>3</sup> <http://medline.cos.com/>

concepts (i.e., siblings concepts) and which terms denote more specific cases of a particular concept (i.e., descendant descriptors). MeSH arranges concepts in an eleven level deep hierarchical structure, defining highly generic to very specific concepts. For instance, at the second level<sup>4</sup>, we find 16 broad concepts, including “*Diseases*”, “*Health Care*” and “*Organisms*”. As we navigate further down the tree structure, we find increasingly more specific concepts, such as “*Respiratory Tract Diseases*” >> “*Lung Diseases*” >> “*Atelectasis*” and >> “*Middle Lobe Syndrome*”. Additionally, many of the concepts in MeSH have *entry terms* associated with them. These are additional terms being synonyms (or quasi-synonyms, such as different spellings and plural forms) to the descriptor. E.g., we find that *cancer*, *tumor*, *neoplasms* and *benign neoplasm* are all entry terms for the concept “*Neoplasm*”.

The OHSUMED collection, that we describe in the next section, included relevance judgments for 4904 MeSH concepts. We included 2317 of these concepts in our experiment, each with a unique location in MeSH, and their descriptors became our gold standard. We considered each descriptor (e.g., “*Lung Diseases*”) of a concept as a significant term for that concept, composed with the descriptors of its descendants (e.g., “*Atelectasis*” and “*Middle Lobe Syndrome*”). If a concept was a leaf (e.g., “*Middle Lobe Syndrome*”), instead we additionally regarded each (possible) entry term (e.g., *brock syndrome*, *brocks syndrome*, *brock's syndrome*) as significant for that particular concept.

### 3.2 Text Collection

The textual resource used in these experiments was the OHSUMED collection (Hersh, 1994). OHSUMED is a subset of the Medline Database and includes 348.566 references to 270 medical journals collected between 1987 and 1991. Most of these texts are references to journal articles, but some are references to conference proceedings, letters to editors and other medical reports. While many references include only a title, the majority also include an abstract, truncated at 250 words. We set the content of a document to include the title and (possibly) the abstract of a reference. In

---

<sup>4</sup> We added a root node in these experiments to connect all branches.

view of the fact that OHSUMED includes references from Medline, each reference consequently came with a number of manually assigned MeSH concepts. That is, for each of the 2317 concepts previously selected, we knew their relevant and non-relevant document sets.

Before indexing this collection, we performed inflectional stemming and NP chunking, and we omitted all terms not identified as single nouns or noun phrases. Once the indexing was complete, we applied each feature selection technique to the 2317 features sets. We setup this process as follows: Given a MeSH concept, we retrieved all of its associated documents from the document collection, and collected the complete feature set of (unique) terms. In order to contrast these terms with those terms we had in our gold standard, we kept only the ones that were already present (or parts of descriptors) in MeSH. While these lists typically included 1400 terms, for some concepts we obtained over 5000 terms, while for others we obtained less than 100. Next, we applied each feature selection technique to weight and rank each of these terms. Once this process was complete, we obtained four lists for each of the 2317 concepts, where each list included the same set of terms, while varying only in respect to the ordering of those terms. Next, we evaluated each feature selection technique by comparing the lists they had produced with terms in our gold standard we knew where significant.

## 4 Evaluation Metrics

We evaluated the performance of each feature selection technique based on the ordered feature lists previously obtained. Essentially, a technique was performing well if it ranked significant terms higher than less significant terms. We employed three evaluation metrics: the *Wilcoxon rank-sum test*, *precision at n*, and the *Spearman rank correlation*.

### 4.1 Wilcoxon Rank-Sum Test

Using the Wilcoxon Rank-Sum test<sup>5</sup> (Mann and Whitney, 1947), we measured the overall tendency of each technique ranking significant terms either

---

<sup>5</sup> Alternatively, Mann-Whitney U test.

high or low. This metric took an ordered list of terms for a given concept, and verified whether significant terms normally appeared at the beginning or at the end of this list. The rank sum becomes low when significant terms exist near the beginning of the list and high when insignificant terms precede relevant terms in the list. We considered the ordering of terms as non-random when the sum of the ranks varied more than we could expect by chance.

## 4.2 Precision at $n$

Precision at  $n$  also provides a mean for measuring the quality of rankings. In contrast to the previous metric, we can inspect the precision at certain positions in this ranking. Precision at  $n$  gives the accuracy obtained for the first  $n$  terms that we know from our gold standard to be significant. A perfect technique therefore places all significant terms at the beginning of the list, while positioning less significant terms at the lower end of the list. We computed precision at  $n$  ( $p(n)$ ) according to:

$$p(n) = \frac{rel_n}{n}$$

where  $n$  is some ranking position and  $rel_n$  the number of relevant terms found among the first  $n$  terms suggested. We computed the precision at rank positions 5, 10, 15, 20, 30, 100, 200, 500 and 1000, and by averaging the precision values for each technique over all 2317 concepts.

## 4.3 Spearman's Rank Correlation

Semantic similarity<sup>6</sup> measures are metrics for computing the relatedness in meaning between concepts (or terms denoting them) based on their distance to each other in a hierarchy. (Budanitsky and Hirst, 2004). They all build upon the assumption that concepts (or terms denoting them) situated closely in the hierarchical space are more similar in meaning than concepts (or terms denoting them) that are separated farther away. E.g., in WordNet (Fellbaum, 1998), we find that *wolf* and *dog* are more related than *dog* and *hat*, since, in WordNet, *wolf* and *dog* share the same parent (i.e., *Canine*).

<sup>6</sup> Also known as semantic distance or relatedness.

The idea was to compare the ordering of terms decided by each feature selection technique, with the order these terms obtained based on their *semantic distance* to respective concepts in our experiment. That is, let's suppose that some technique determined '*hypoglycemia*' to be insignificant for the concept "*Diabetes Mellitus*", and thereby giving it a low rank. However, if we compute the distance between '*hypoglycemia*' and "*Diabetes Mellitus*", in MeSH, we find that '*hypoglycemia*' gets a high relatedness value, as this term symbolizes one of two siblings of "*Diabetes Mellitus*" and thereby receives a high rank. If cases like this were frequent, it would indicate that this particular technique was unable to detect significant terms.

Spearman's Rank Correlation (*rho*) is a metric for comparing ordering of items. When two lists come in the same order, they are identical, and the rank correlation becomes one (1). Conversely, if one is the inverse of the other, then the correlation becomes -1. We obtain a correlation value of zero when there is no relation between the two. The rank correlation is computed using:

$$rho = \frac{6 \sum d_i^2}{n \bullet (n^2 - 1)}$$

where  $d_i$  is the difference between each entry pair, and where  $n$  equals the number of entry pairs.

Using Leacock-Chodorow's measure of path length (Leacock and Chodorow, 1994), we computed the distance between each term in our feature lists and a concept in question. We now had two orderings with the identical set of terms, which we could compare. Specifically, one list including the ordering of terms decided by some feature selection technique, and the other being a list based on the semantic distance between each term and a certain concept.

In hierarchies such as MeSH, relatedness rapidly decreases as distance increases. This is especially true when a path between a term and a concept leads through the root of the hierarchy. These are cases when a term and a concept are positioned in separate branches of the 16 main concepts at the second level. Recognizing this fact, we (additionally) normalized the path length metric by setting a

threshold, such that the relatedness value of became zero if the path from a term to a concept included the root concept.

## 5 Results

The Wilcoxon Rank Sum test gave us a clear indication that, for a large majority of concepts, each of the four feature selection techniques ranked significant terms before less significant terms. Further, Table 1 illustrates the precision that each feature selection technique obtained at each of the nine ranking positions, where these values are averaged over all 2317 concepts. We observe that Odds ratio (*odds*) scores the lowest precision values at all cut-off points on this test. Both the Chi-square (*chi*) and Correlation Coefficient (*cc*) metrics perform better than the rivaling techniques. In fact, their performances are identical. Our baseline technique (*cf*) performs slightly lower than *chi* and *cc*.

Rank position	Feature Selection Technique			
	<i>cf</i>	<i>odds</i>	<i>chi</i>	<i>cc</i>
5	0,32	0,23	0,39	0,39
10	0,19	0,17	0,25	0,25
15	0,14	0,13	0,19	0,19
20	0,12	0,11	0,16	0,16
30	0,09	0,08	0,12	0,12
100	0,04	0,04	0,05	0,05
200	0,02	0,02	0,03	0,03
500	0,01	0,01	0,01	0,01
1000	0,009	0,009	0,009	0,009

**Table 1:** Precision at rank position 5 --1000. Values are averaged over 2317 experiments.

In Table 2, we see the average correlation in rankings between the lists of terms ordered by each technique, and the ordering of the same set of terms based on their semantic distance to respective concepts included in these experiments. Here, we assigned the real distance value of a term even if its path to a concept included the root concept. Again, values are averaged over all 2317 concepts.

Feature Selection Technique	Rank Correlation
<i>cf</i>	0,30
<i>odds</i>	-0,19
<i>chi</i>	-0,06
<i>cc</i>	-0,13

**Table 2:** Rank Correlations averaged over 2317 concepts. Path via root node allowed.

This tells us that, *chi*, *cc* and *odds* all have a tendency toward ranking terms in *contradictory* order to the Leacock-Chodorow's measure of semantic distance. Contrastively, we observe a positive correlation between our baseline technique (*cf*) and that distance measure, although this correlation is on the weaker end of the scale. This indicates that *cf* more often ranked closely positioned terms to our concepts higher, than it ranked terms situated more distantly from our concepts in MeSH.

When we normalized the Leacock-Chodorow measure, we obtained positive correlation value for all techniques and they came to conform more to each other. (Table 3)

Feature Selection Technique	Rank Correlation
<i>cf</i>	0,35
<i>odds</i>	0,20
<i>chi</i>	0,19
<i>cc</i>	0,16

**Table 3:** Rank Correlations averaged over 2317 concepts. Paths via root node given a value of zero.

## 6 Discussion

We have evaluated and compared four feature selection techniques on the task of detecting significant terms for concepts in the medical domain. Our results suggest that all techniques behave similarly in respect to ranking significant terms. Both the Wilcoxon rank-sum test and precision at *n* gave a clear indication of this. Although we evaluated each feature selection technique on nine different ranking positions, it probably makes more sense to do it only on ranking positions 5—20. We can imagine a controlled vocabulary editor getting a

list of suggested terms to add to the terminology. In such a scenario, it is likely that the editor is only interested in verifying the relevance of 2—15 terms. Failing to notice significant terms appearing later in the list should be a minor concern.

However, we observed noticeable differences between the techniques when we compared their ordered set of terms with the semantic relatedness values of these terms. Results showed that the simplest technique (*cf*) conform to the conceptual relations among terms in MeSH the most, while the more sophisticated techniques tended to rank terms in contradictory order. We are aware that these results can be different if we choose some other semantic similarity metric. However, to the best of our knowledge, evaluating feature selection techniques using semantic similarity measures has never been tested and we consider semantic relatedness measures as interesting alternatives to the other evaluation metrics and they should provide us with some additional information regarding the behavior of feature selection techniques. In the future, we intend to investigate the justifications of semantic similarity measures and the role these measures can have in our setting.

What our study boils down to is that of determining whether the task we appoint to feature selection techniques in this setting is different from, similar or even identical to the task these techniques are intended to solve in text categorization. At this point, we cannot provide a straightforward answer to that question. It is reasonable to argue that the tasks are similar if we employ these techniques in some (semi-) automated scenario, where it is an absolute necessity that top ranking terms have high discriminating power. However, if these techniques are only part of, say, some editing tool where trained professionals can judge the outcomes, then we might want to consider the tasks as different.

## References

Alexander Budanitsky and Graeme Hirst. 2004. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics* 32(1):13—47.

Christiane D. Fellbaum. 1998. *WordNet, an electronic lexical database*. MIT Press.

David Freedman, Robert Pisani, Roger Purves, and Ani Adhikari. 1991. *Statistics*. Second edition. Norton. New York.

Luigi Galavotti, Fabrizio Sebastiani, and Maria Simi. 2000. Experiments on the use of feature selection and negative evidence in automated text categorization. *Proceedings of ECDL-00, 4th European Conference Research and Advanced Technology for Digital Libraries*.

William Hersh. 1994. Ohsumed: An interactive retrieval evaluation and new large test collection for research. *Proceedings of the 17th Annual Intl. ACM SIGIR Conference on R&D in Information Retrieval*.

Claudia Leacock and Martin Chodorow. 1998. Combining Local Context and WordNet Similarity for Word Sense Identification. *WordNet: An Electronic Lexical Database*. C. Fellbaum, MIT Press: 265—283

Dunja Mladenic. 1998. Feature Subset Selection in Text-Learning. *European Conference on Machine Learning*.

Hwee T. Ng, Wei B. Goh and Kok L. Low. 1997. Feature selection, perceptron learning, and a usability case study for text categorization. *Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval*.

Fabrizio Sebastiani. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.* 34(1): 1—47.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. *Proceedings of ICML-97, 14th International Conference on Machine Learning*.

# Towards Explicit Semantic Features using Thresholded Independent Component Analysis

Jaakko J. Väyrynen and Timo Honkela and Lasse Lindqvist

Adaptive Informatics Research Centre

Helsinki University of Technology

P.O.Box 5400, FIN-02015 TKK, Finland

{jjvayryn, tho, llindqvi}@cis.hut.fi

## Abstract

Latent semantic analysis (LSA) can be used to create an implicit semantic vectorial representation for words. Independent component analysis (ICA) can be derived as an extension to LSA that rotates the latent semantic space so that it becomes explicit, that is, the features correspond more with those resulting from human cognitive activity. This enables nonlinear filtering of the features, such as hard thresholding that creates a sparse word representation where only a subset of the features is required to represent each word successfully. We demonstrate this with semantic multiple choice vocabulary tests. The experiments are conducted in English, Finnish and Swedish.

## 1 Introduction

Latent semantic analysis (LSA) (Landauer and Dumais, 1997) is a very popular method for extracting information from text corpora. The mathematical method behind LSA is singular value decomposition (SVD) (Deerwester et al., 1990), which removes second order correlations from data and can be used to reduce dimension. LSA has been shown to produce reasonably low-dimensional latent semantic spaces that can handle various tasks, such as vocabulary tests and essay grading, at human level (Landauer and Dumais, 1997). The found latent components, however, are implicit and cannot be understood by humans.

Independent component analysis (ICA) (Comon, 1994; Hyvärinen et al., 2001) is a method for re-

moving higher order correlations from data. It can be seen as whitening followed by a rotation, where whitening can be produced with SVD. Independent component analysis can thus be seen as an extension of LSA. The rotation should find components that are statistically independent of each other and that we think are meaningful. In case the components are not truly independent, ICA should find “interesting” components similar to projection pursuit.

ICA has been demonstrated to produce unsupervised structures that well-align with that resulting from human cognitive activity in text, images, social networks and musical features (Hansen et al., 2005). We will show that the components found by the ICA method can be further processed by simple nonlinear methods, such as thresholding, that give rise to a sparse feature representation of words. An analogical approach can be found from the analysis of natural images, where a soft thresholding of sparse coding is seen as a denoising operator (Oja et al., 1999). The ICA can be, e.g., used to detect topics in document collections (Isbell and Viola, 1999; Bingham et al., 2001). Earlier we have shown that the ICA results into meaningful word features (Honkela and Hyvärinen, 2004; Honkela et al., 2004) and that these features correspond to a reasonable extent with syntactic categorizations created through human linguistic analysis (Väyrynen et al., 2004).

In this paper, we present experimental results that show how the ICA method produces explicit semantic features instead of the implicit features created by the LSA method. We show through practical experiments that this approach exceeds the capacity of the LSA method.

## 2 Data

We have a collection of texts as our source of natural language for English, Finnish and Swedish. Our unsupervised learning methods are singular value decomposition and independent components analysis. The semantic representations learned with the methods are applied to multiple choice vocabulary tasks that measure how well the emergent word representations capture semantics.

### 2.1 Europarl Corpus

The Europarl corpus (Koehn, 2005) contains texts from the Proceedings of the European Parliament in 11 languages. We concentrated in English, Finnish and Swedish in our experiments. XML tags and special characters were removed from the texts and uppercase characters were replaced with respective lowercase ones. The English text had 26 million tokens (word forms in running text) and 83 thousand types (unique word forms). The Finnish text had 19 million tokens and 480 thousand types. The Swedish text had 24 million tokens and 240 thousand types.

### 2.2 Gutenberg Corpus

A more general example of a natural text is a collection of 4966 free English e-books that were extracted from the Project Gutenberg website<sup>1</sup>. The texts were pruned to exclude poems and the e-book headers and footers were removed. The texts were then concatenated into a single file, special characters were removed, numbers were replaced with a special token and uppercase characters were replaced with respective lowercase ones. The final corpus had 319 million tokens and 1.41 million types.

### 2.3 Vocabulary Test Sets

Semantic word representations can be evaluated with multiple choice vocabulary tests that measure some semantic concept, such as synonymity. In a multiple choice test, the task is to select the correct word from a list of alternatives when given a stem word or a cue word.

For the English language, there exists free electronic resources that can be used to conduct such tests. For many other languages of interest, however, such resources may not be directly available.

<sup>1</sup><http://www.gutenberg.org>

We briefly introduce one famous but small and two large semantic resources for English, as well as one for many European languages.

Performance of the compared methods is measured with precision: the ratio of correct answers to the number of questions in the test set. The higher the precision is, the better the method has captured the type of semantics the questions cover. The vocabulary and the test questions were chosen so that recall was 100 percent. Especially this means that only single word terms were considered for test questions.

#### 2.3.1 TOEFL Synonyms

A famous test case for English is the synonym part of the TOEFL data set<sup>2</sup>. It was provided for us by the Institute of Cognitive Science, University of Colorado, Boulder. The task is to select the synonym for each stem word from four alternatives. An example question is shown below with the correct answer emphasized.

**figure:** list, *solve*, divide, express

LSA has been shown to get 64.4% correct for the TOEFL data set, which is statistically at the same level as for a large sample of applicants to US colleges from non-English speaking countries (Laudauer and Dumais, 1997). Even a precision level of 97.5% has been reached by combining several methods, including LSA and an online thesaurus (Turney et al., 2003).

However, the TOEFL test set has only 80 questions and comparison of the methods with only this test set is not sufficient. Also, the baseline precision with guessing from four alternatives is 25% and chance might play a big role in the results.

#### 2.3.2 Moby Synonyms and Related Words

The Moby Thesaurus II<sup>3</sup> of English words and phrases has more than 30 000 entries with 2.5 million synonyms and related terms. We generated multiple choice questions by selecting a stem from the Moby thesaurus, and combining one of the listed synonyms with a number of random words from our vocabulary as alternatives. This method allows

<sup>2</sup><http://www.ets.org>

<sup>3</sup><http://www.dcs.shef.ac.uk/research/ilash/Moby/>

us to have more questions and alternatives than the TOEFL data set, which makes the test more robust in terms of confidence intervals for precision. On the other hand, the generated questions are very likely to lack the finesse of the hand-crafted TOEFL questions and no human level performance is known. An example entry in the thesaurus is shown below.

**approve:** OK, accede to, accept, accord to, accredit, admire, adopt, affiliate, affirm, . . .

We generated 16 638 questions from the Moby thesaurus with 16 alternatives. At most one question was generated from each entry. The baseline precision is 6.25% with guessing from 16 alternatives. An example of a generated question is shown below.

**constitute:** *validate*, washington, wands, paper-based, convention, aérospatiale, vanhecke, indifference, kaklamanis, possess, criminalization, grouping, shari, reorganisations, diluents

### 2.3.3 Idiosyncratic Associations

The free association norms data set<sup>4</sup> from the University of South Florida contains idiosyncratic responses in English, that is, responses given only by one human subject, to more than five thousand cue words. On average, there are approximately 22.15 idiosyncratic responses per cue word with high variation. An example entry is shown below.

**early:** before, classes, frost, on time, prompt, sleepy, sun, tired, years

Similarly to the generated Moby questions, the idiosyncratic association data set was used to generate 4 582 multiple choice questions with 16 alternatives. An example of a generated question is shown below.

**corrupt:** *crook*, plaice, wfp, a5-0058, administrated, vega, 1871, a5-0325, h-0513, toolbox, compelling, 1947, crashing, vac, illating, indemnity

### 2.3.4 Eurovoc Thesaurus

The multilingual Eurovoc thesaurus<sup>5</sup> covers fields that are of importance for the activities of the European institutions. It is available in many European languages and contains different relationships

<sup>4</sup><http://w3.usf.edu/FreeAssociation/>

<sup>5</sup><http://europa.eu/eurovoc/>

between the terms in the thesaurus. Each field is divided into several microthesauri, e.g., the field “trade” contains seven microthesauri, including “tariff policy” and “consumption”. An excerpt of an English microthesaurus is shown below.

- political system

**RT** political science (3611)

**NT1** authoritarian regime

**NT1** change of political system

**RT** political reform (0431)

**RT** transition economy (1621)

**NT1** constitutional monarchy

**RT** parliament (0421)

We set the task to be identification of terms in the same microthesaurus. Related terms (RT) in other microthesauri were not included. For each pair of terms in a microthesaurus, one term was selected as a cue word and the other was mixed with a number of random words as alternatives. Only fields “finance”, “law”, “politics” and “trade” were included in these experiments. This procedure gave 2 312 questions for English, 1 848 for Finnish, and 7 564 for Swedish. An example of a generated question in English is shown below.

**republic:** *oligarchy*, alps, spits, seventy, greeks, progressivity, deflationary, endorsing, renowned, understated, cogently, miscalculations, 0306, range, heralding, lèse-majesté

## 3 Methods

It has been known already for some time that statistical analysis of the contexts in which a word appears in text can provide reasonable amount of information on the syntactic and semantic roles of the word (Ritter and Kohonen, 1989; Church and Hanks, 1990). A typical approach is to calculate a document-term matrix in which the rows correspond to the documents and the columns correspond to the terms. A column is filled with the number of occurrences of the particular term in each document. The similarity of use of any two terms is reflected by the relative similarity of the corresponding two columns in the document-term matrix. Instead of considering the whole documents as contexts, one can also

choose a sentence, a paragraph or some other contextual window. A related approach, that is taken here, is to calculate the number of co-occurrences of the particular term with a number of other terms in contextual windows around the instances of the analyzed term in the text. This produces a context-term matrix, where each context is defined using terms instead of documents.

### 3.1 Contextual Information

Contextual information is a standard way of filtering more dense data from running text. Frequencies of term occurrences, or co-occurrences, in different chunks of texts are typically calculated. The idea behind this is that relations of words manifest themselves by having related words occur in similar contexts, but not necessary together. Raw contextual data is too sparse for practical use and it has been shown that finding a more compact representation from the raw data can increase the information content by generalizing the data (Landauer and Dumais, 1997).

A context-term matrix  $\mathbf{X}$  was calculated using the Gutenberg corpus or one of the analyzed languages in the Europarl corpus. The rows in the matrix correspond to contexts and the columns represent the terms in the analyzed vocabulary. The context contained frequencies of the 1 000 most common word forms in a 21 word window centered around each occurrence of the analyzed terms. The terms included the 50 000 most common word forms.

The contextual information was encoded with a bag-of-words model and the matrix  $\mathbf{X}$  was of size  $1\,000 \times 50\,000$ . A separate matrix with its own vocabulary was calculated for each corpus and language.

The raw frequency information of the terms is typically modified using stop-word lists and term weighting, such as the tf-idf method that is suitable for document contexts. We did not use stop-word lists and frequency rank information was preserved by taking the logarithm of the frequencies increased by one.

### 3.2 Singular Value Decomposition

Singular value decomposition learns a latent structure for representing data. Input to singular value decomposition is a  $m \times n$  matrix  $\mathbf{X}$ . The SVD method

finds the decomposition  $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ , where  $\mathbf{U}$  is an  $m \times r$  matrix of left singular vectors from the standard eigenvectors of square symmetric matrix  $\mathbf{X}\mathbf{X}^T$ ,  $\mathbf{V}$  is an  $n \times r$  matrix of right singular vectors from the eigenvectors of  $\mathbf{X}^T\mathbf{X}$ ,  $\mathbf{D}$  is a diagonal  $r \times r$  matrix whose non-zero values are the square roots of the eigenvalues of  $\mathbf{X}\mathbf{X}^T$  or (equivalently)  $\mathbf{X}^T\mathbf{X}$ , and  $r = \min(n, m)$  is the rank of  $\mathbf{X}$ . A lossy dimension reduction to  $l \leq r$  components can be achieved by discarding small eigenvalues.

In SVD-based latent semantic analysis, the input matrix  $\mathbf{X}$  is a context-term matrix representing the weighted frequencies of terms in text passages or other contexts. The method can handle tens of thousands of terms and contexts. Dimension is typically lowered to a few hundred components, that reduces noise and generalizes the data by finding a latent semantic representation for words. Words and texts can be compared by their respective vectorial representations in the latent space.

### 3.3 Independent Component Analysis

Independent component analysis uses higher-order statistics compared to singular value decomposition that only removes second-order correlations. ICA finds a decomposition  $\mathbf{Z} = \mathbf{B}\mathbf{S}$  for a data matrix  $\mathbf{Z}$ , where  $\mathbf{B}$  is a mixing matrix of weights for the independent components in the rows of matrix  $\mathbf{S}$ . The task is usually to find a separating matrix  $\mathbf{W} = \mathbf{B}^{-1}$  that produces independent components  $\mathbf{S} = \mathbf{W}\mathbf{Z}$ .

If data  $\mathbf{Z}$  is white, it suffices to find a rotation that produces maximally independent components (Hyvärinen et al., 2001). The right singular values  $\mathbf{V}$  produced by SVD are uncorrelated and thus SVD can be seen as a direct preprocessing step to ICA, if the data  $\mathbf{X}$  has zero mean. This mathematical relation is showed in Figure 1. The ICA rotation should find components that are more interesting and structure the semantic space in a meaningful manner, as illustrated in Figure 2.

### 3.4 Thresholding

Thresholding is an example of a nonlinear filtering method. It forces a word representation to be more sparse by retaining only a subset of the features. For a successful usage of a such thresholded feature representation in a semantic task, it is necessary that features containing most of the semantic informa-

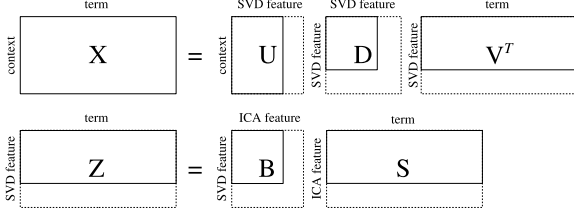


Figure 1: Mathematically, for zero-mean data  $\mathbf{X}$ , ICA can be represented as an extension of SVD, where the white SVD components  $\mathbf{Z} = \sqrt{n}\mathbf{V}^T$  for the  $n$  terms are generated by a rotation  $\mathbf{B}$  from the ICA components  $\mathbf{S}$ . SVD is approximated for a reduced dimension from the original dimension of the data matrix  $\mathbf{X}$ , marked here with the solid and dashed lines, respectively.

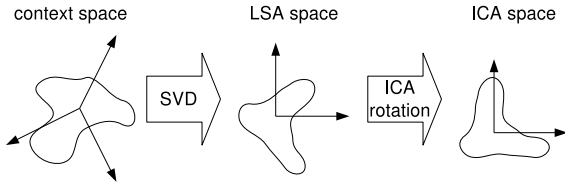
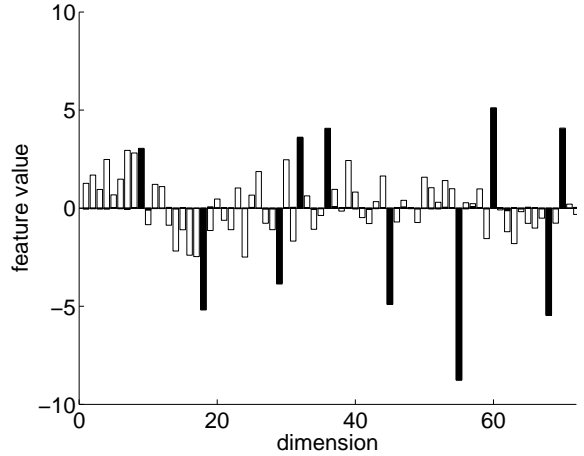


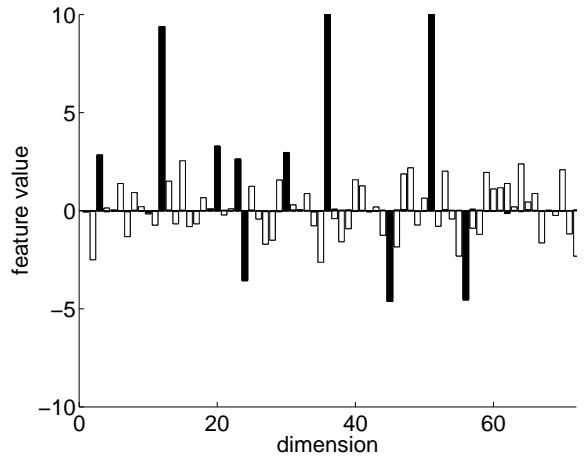
Figure 2: The distribution of terms in contexts can be approximated by a low-dimensional LSA space. ICA can be seen as an additional rotation of the latent space that finds interesting components.

tion are kept while less informative features are discarded. It is also important that the underlying representation models each word with as less features as possible, which is basically the definition of sparseness.

Our features produced by ICA and SVD have zero mean and have the same variance. For each term in our vocabulary, the features with the lowest absolute values can be considered inactive and be thresholded to zero value. Thus the remaining active features depend on the particular term. For comparison purposes, the same number of active features were kept for each word. An example of thresholded word features is shown in Figure 3. We compare thresholded ICA and thresholded SVD with different number of dimensions and show precisions of the representations for all values of the thresholding parameter. Results are also reported for standard SVD, that is also used for selecting the dimensionality for the thresholded versions.



(a) Feature vector for the word “election”.



(b) Feature vector for the word “candidate”.

Figure 3: ICA feature vectors for the word “election” (a) and “candidate” (b). The outlined bars show the original feature values and the filled bars show the thresholded values with ten active dimensions. Any comparison based on the dot product of the thresholded feature vectors depends only on the jointly active dimensions 36 and 45.

## 4 Results

Here we compare SVD and ICA as feature extraction methods by evaluating the emerging semantic word representations using multiple choice vocabulary tests in three languages. In order to show how ICA finds an explicit feature representation, we threshold the word features and show that ICA produces better results than SVD. In our experiments, the similarity of words was measured as the cosine

of the angle between the respective words vectors.

We have previously reported results for the English Gutenberg corpus and the Moby and idiosyncratic test sets (Väyrynen et al., 2007). The main results are reproduced in this paper. We present here additional results for representations learned from the English, Finnish and Swedish parts of the Europarl corpus. Suitable tests sets for the Europarl were generated from the multilingual Eurovoc thesaurus. The dimension for the thresholded versions of ICA and SVD was selected as approximately the dimension that produced the highest precision with the basic SVD method without thresholding. Some interesting results with other dimensions are also shown. In this section, the number of active components for each word, i.e., the level of thresholding, is varied and the precision of the thresholded representation is measured in a multiple choice vocabulary test. The ICA and SVD methods converge when no thresholding is done. The fewer active dimensions there are, the sparser the word representations are. If the sparse representation also succeeds in the tests measuring semantic content, the features are explicit in this sense.

The representation learned from the Gutenberg corpus was evaluated with the Moby test set and the idiosyncratic test set. The results indicate that thresholding with ICA outperforms standard SVD and that thresholding with SVD does not improve the results. The reproduced results are shown in Figure 4 and Figure 5.

Results for the TOEFL data set with the Gutenberg corpus (Väyrynen et al., 2007), are similar to the Eurovoc test with the Finnish part of the Europarl corpus, shown in Figure 6. In both cases the thresholded ICA and SVD have very similar performance. The hand-made questions in the TOEFL would make the semantics of the alternatives closer to each other, that would make the thresholding process more accurate as the word vectors would have more similar features. It is still unclear why this happens also with the Finnish Eurovoc test.

The English and Swedish word representations learned from the Europarl corpus behave more like the Gutenberg results. The Swedish result, shown in Figure 8, is a good example of how the thresholded ICA can maintain a high precision even when more than half of the features in each word are ig-

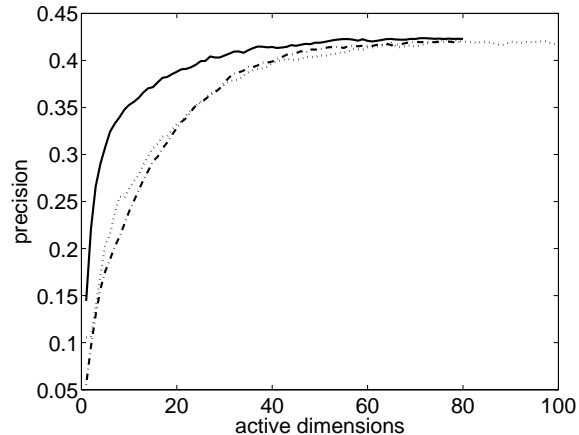


Figure 4: Precisions of the SVD (dotted), SVD with thresholding with 80 components (dashed) and ICA with thresholding with 80 components (solid) with the Moby data set w.r.t. the number of active components. The representations were learned from the Gutenberg corpus.

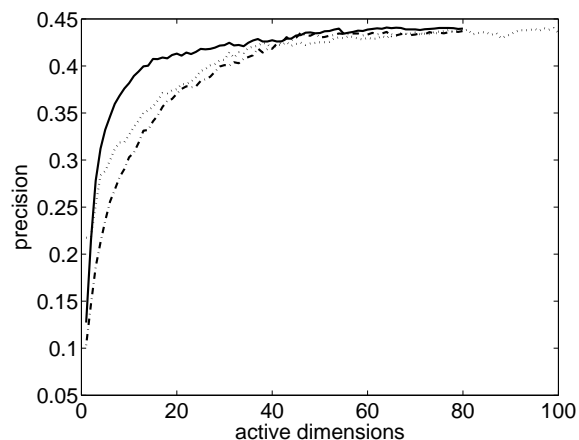


Figure 5: Precisions of the SVD (dotted), SVD with thresholding with 80 components (dashed) and ICA with thresholding with 80 components (solid) with the idiosyncratic association data set w.r.t. the number of active components. The representations were learned from the Gutenberg corpus.

nored. The English test with Europarl did not give equally clear results, but even here the thresholded ICA method does not worse than the standard SVD and outperforms the thresholded SVD method.

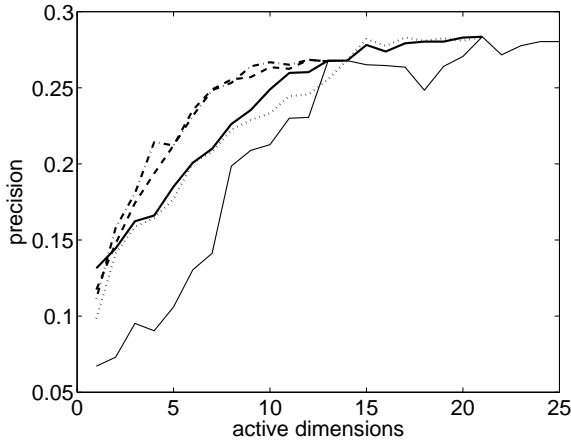


Figure 6: Precisions of the SVD (thin solid), SVD with thresholding with 21 components (dotted) and 13 components (dash dotted) and ICA with thresholding with 13 components (thick solid) and 13 components (dashed) with the Finnish Eurovoc test set w.r.t. the number of active components. The representations were learned from the Europarl corpus.

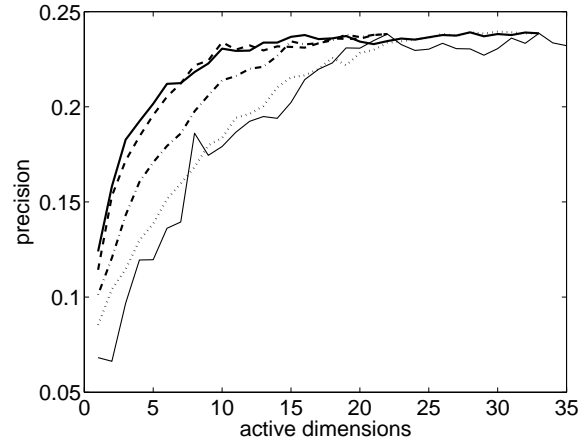


Figure 8: Precisions of the SVD (thin solid), SVD with thresholding with 33 components (dotted) and 22 components (dash dotted) and ICA with thresholding with 33 components (thick solid) and 22 components (dashed) with the Swedish Eurovoc test set w.r.t. the number of active components. The representations were learned from the Europarl corpus.

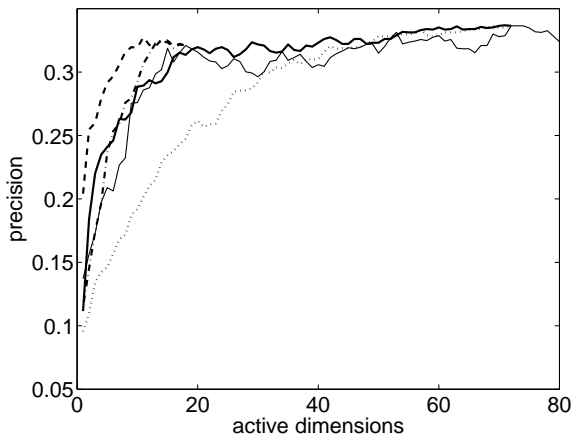


Figure 7: Precisions of the SVD (thin solid), SVD with thresholding with 72 components (dotted) and 18 components (dash dotted) and ICA with thresholding with 72 components (thick solid) and 18 components (dashed) with the English Eurovoc test set w.r.t. the number of active components. The representations were learned from the Europarl corpus.

## 5 Conclusions

In this paper, we showed how the explicit semantic features for words produced by independent compo-

nent analysis align more to cognitive components resulting from human activity. We applied a nonlinear filtering, thresholding, to the word vectors produced by ICA and SVD and studied these thresholded semantic representations in multiple choice vocabulary tests.

The results shown in this article indicate that it is possible to create automatically a sparse representation for words. Moreover, the emergent features in this representation seem to correspond with some linguistically relevant features. When the context is suitably selected for the ICA analysis, the emergent features mostly correspond to some semantic selection criteria. Traditionally, linguistic features have been determined manually. For instance, case grammar is a classical theory of grammatical analysis (Fillmore, 1968) that proposes to analyze sentences as constituted by the combination of a verb plus a set of deep cases, i.e., semantic roles. Numerous different theories and grammar formalisms exist that provide a variety of semantic or syntactic categories into which words need to be manually classified.

Statistical methods such as SVD and ICA are able to analyze context-term matrices to produce auto-

matically useful representations. ICA has the additional advantage, especially when combined with some additional processing steps reported in this article, over SVD (and thus LSA) that the resulting representation is explicit and sparse: each active component of the representation is meaningful as such. As the LSA method is already very popular, we assume that the additional advantages brought by this method will further strengthen the movement from a manual analysis to an automated analysis.

## References

- Ella Bingham, Ata Kabán, and Mark Girolami. 2001. Finding topics in dynamical text: application to chat line discussions. In *Poster Proceedings of the 10th International World Wide Web Conference (WWW'10)*, pages 198–199.
- Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.
- Pierre Comon. 1994. Independent Component Analysis, a new concept? *Signal Processing*, 36(3):287–314.
- Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.
- Charles J. Fillmore, 1968. *Universals in Linguistic Theory*, chapter The Case for Case, pages 1–88. Holt, Rinehart, and Winston, New York, USA.
- Lars Kai Hansen, Peter Ahrendt, and Jan Larsen. 2005. Towards cognitive component analysis. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, pages 148–153.
- Timo Honkela and Aapo Hyvärinen. 2004. Linguistic feature extraction using independent component analysis. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2004)*, pages 279–284.
- Timo Honkela, Aapo Hyvärinen, and Jaakko Väyrynen. 2004. Emergence of linguistic features: Independent component analysis of contexts. In *Proceedings of the 9th Neural Computation and Psychology Workshop (NCPW9): Modeling Language Cognition and Action*, pages 129–138.
- Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. 2001. *Independent Component Analysis*. John Wiley & Sons.
- Charles Lee Isbell, Jr. and Paul Viola. 1999. Restructuring sparse high dimensional data for effective retrieval. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS 1998)*, pages 480–486.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.
- Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.
- Erkki Oja, Aapo Hyvärinen, and Patrik Hoyer. 1999. Image feature extraction and denoising by sparse coding. *Pattern Analysis & Applications*, 2(2):104–110.
- Helge Ritter and Teuvo Kohonen. 1989. Self-organizing semantic maps. *Biological Cybernetics*, 61(4):241–254.
- Peter D. Turney, Michael L. Littman, Jeffrey Bigham, and Victor Schnayder. 2003. Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, pages 482–489.
- Jaakko J. Väyrynen, Timo Honkela, and Aapo Hyvärinen. 2004. Independent component analysis of word contexts and comparison with traditional categories. In *Proceedings of the 6th Nordic Signal Processing Symposium (NORSIG 2004)*, pages 300–303.
- Jaakko J. Väyrynen, Lasse Lindqvist, and Timo Honkela. 2007. Sparse distributed representations for words with thresholded independent component analysis. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2007)*. To appear.

# Representing achievements from Estonian transitive sentences

Anne Tamm

anne.tamm@unifi.it

University of Florence, Italy

Institute for the Estonian Language,

Tallinn, Estonia

## Abstract

The article opens the complex issue of representing and acquiring aspectual (event structural) information from surface syntactic elements scattered over several phrases in an Estonian sentence. The more detailed example contains the problems of “reading semantics off the syntax and morphology” in sentences containing Estonian transitive achievement verbs and their partitive objects. The article presents a representation in terms of a unification-based approach in Lexical Functional Grammar (LFG), where the aspectual features of verbs and case are modeled via unification at the syntactic level of functional structure.

## 1 Credits

This article contains unpublished material and proposes a novel way of representing and acquiring the semantic content of predicates, more specifically, the event’s type on the basis of the verb classification and argument frame combined with other types of information such as object case. The analysis of the two classes of achievement verbs is novel and follows the theoretical grounds and the empirical generalizations in Tamm (2004) and the main points in Tamm (2007). It is an elaboration of Tamm (2005) on the Estonian verb classes in computational lexicography. Further main sources that form the theoretical basis of this account are Butt (2006), Butt and King (2005), Butt et al. (1997), Kiparsky (1998) and Nordlinger and Sadler (2004).

## 2 Introduction

The paper presents theoretical aspects of aspectual semantic content representation and the preliminaries of acquisition, that is, a sketch of a possible method and the tools of identifying the semantic content. The representational issues are addressed in the Lexical Functional Grammar (LFG) framework, which is equally accessible for formal generative as well as computational linguists, the more traditional functional as well as lexically oriented approaches.

Identifying the semantic content is understood as identifying the event types or the aspectual meaning in the text. The examples discussed are Estonian simple sentences, and the aspectual semantic event type of achievements (in the traditional Vendlerian classification). This seemingly narrow focus is due to many factors.

Firstly, simpler sentences are chosen since there are extremely few theories dealing with the topic in Estonian that are easily convertible into methods that could accomplish the identification of the aspectual meaning. Although aspect is obligatorily expressed in all Estonian sentences, there are no methods developed yet for capturing aspectual meaning.

Secondly, there are no works known to the author that would deal with the statistical extraction of the aspectual meaning in this language, but taking stock of at least some of the predictable future problems of statistical approaches is one of the goals of this article. Therefore, the article has chosen the data—achievement sentences—so that the information presented here could be valuable for both linguistic and statistical methods. A statistical

(surface) method can be predicted to give a reasonable solution to a high percentage of cases, but to regularly fail in cases where a combined method with a linguistic (deep) method can improve the results steadily. As an example from Hungarian speech synthesis and automatic stress assignment (Tamm and Olaszy 2005), the results of a combination with a linguistic, deep method and a more statistically oriented one show that the statistical method gives good results with little effort, but from a certain point it reaches its limits and the improvement process slows down. Improving the stress assignment algorithm with a more linguistically oriented method requires more effort, but allows for steady improvement of the error rate.

Thirdly, in general, the connections between the acquisition and representation of the volatile aspectual semantic content of the world's languages are a little studied topic. By volatile I mean the variety in the grammaticalization of the TAM categories across languages (and the terminologies used for describing the phenomena), the distribution of the morphemes and the division of labor between the purely lexical and the purely non-lexical; that is, between verbs, their internal argument NPs, PPs and a wide array of morphemes attaching to them and influencing the semantic content of the sentence they appear in. However, in the coming era of machine translation and other technical applications that build on the decoding and encoding of the semantic content, the bottlenecks of cross-linguistic TAM categories are better identified sooner than later, and preferably in a language with rich and explicit morphology, such as Estonian.

For several better studied, predominantly Indo-European languages (e.g., English, Dutch, Italian), the detection of argument frame, information about the verb's classification, the presence of certain adverbials or quantifiers, and the quantificational properties of the internal argument (the object) gives a fairly solid ground for the prediction of the compositional aspectual value of the sentence (cf. Verkuyl 1993). In the Slavic languages, the composition of the aspect is considerably reduced and the (morphologically complex) verb determines many key properties of the event (Filip 2001, Kiparsky 1998). However, the information about the morphological complexity of the verbs and the components of aspectual composition mentioned above is not sufficient for acquiring the semantic properties of the events in several Finnic lan-

guages. On the one hand, in languages such as Finnish, Livonian, Votic, Vepsan, or Estonian, the object case of a transitive verb is a far better indicator of the event type (this article defines it in terms of the opposition in boundedness as defined in Kiparsky 1998) than the morphological complexity or the quantificational properties of the internal argument. The objects of Estonian transitive verbs in active affirmative indicative clauses are marked with either the partitive or the total case; the latter is also known as the accusative in typological and theoretical approaches and as the morphological genitive or nominative. Roughly, partitive objects appear in unbounded sentences and total objects in bounded sentences. It is highly probable that this correlation between the event type and the semantically conditioned case (atelic-partitive, telic-total) will be detected by statistical methods. Do we need any additional semantic or syntactic information, then, if the event type is directly read off from the object's case?

I argue that we do, since on the other hand, the transitive verb's object case does not give sufficient ground for detecting the further details of the event's structure, such as its iterative, progressive etc nature of an event described in a sentence containing an achievement verb. In order to access further semantic content, an indication of the lexical aspectual verb class is necessary, and, as I will argue on the basis of the achievement verbs, providing the Vendler type is not sufficient. I propose to identify the semantic content in a combined way. Here I concentrate on the following intersection: the aspectual indications in the lexicon of the verb and the aspectual indications of the object cases. Here follow the possible components of the algorithm for sentence aspect.

1. The syntactic analyzer (by Kaili Müürisep, University of Tartu) that identifies the object and its case: partitive or non-partitive, total (see Appendix C for an analysis of the examples in this article).

2. The entries for the (aspectual) object case (as earlier theoretical works indicated in Section 1).

3. The aspectual lexicon of verbs (under construction at the Institute of Estonian Language, Tallinn).

4. LFG-type functional structures (for the unification of aspectual information from verbs and case).

This article has a rather specific focus: achievement (bounded event) meanings from transitive sentences with achievement verbs of the two types *leidma* ‘find’ and *võitma* (*kedagi*) ‘win (somebody)’ and with semantically conditioned partitive NP objects. The main difference between the two classes of achievement verbs is the following: the *leidma* ‘find’ type of achievement verbs is a so-called total object verb class and the *võitma* ‘win’ type is a partitive-object verb class (although the verb is polysemous and appears with total object with or without a resultative particle as well). This means that while the *find*-verbs typically appear with non-partitive (nominative or total) object case marking, the *win*-verbs in the given lexical achievement meaning appear with partitive object case only. However, in terms of the compatibility of verbs and object case, there is a marked difference between singular and plural (or mass) noun phrases. Namely, both *find* and *win* verb classes may appear with the partitive case in plural count (and mass singular) NP objects. In short, the *win* type is the problematic one, since the verb denotes an achievement, but the object case is not total as with event verbs, but partitive.

The problems and solutions are presented as follows: Section 3 introduces the basic facts about Estonian object case alternation and verbs. Section 4 presents the data that are the focus of the model. Section 5 addresses the verbs and Section 6 views the partitive objects. Section 7 illustrates the unification, and Section 8 is a conclusion.

### 3 Aspect and Object Case

Estonian clausal aspect is not entirely determined by the verb. Rather, the alternation of the partitive (1) and total-accusative (here, morphologically genitive) (2) object cases corresponds more closely to the aspectual oppositions as most clearly illustrated by the accomplishment verbs of creation.

- (1) *Mari kirjutas*  
 M.nom write.3.sg.past  
       *raamatut.*  
       book.part  
 ‘Mari was writing a/the book.’
- (2) *Mari kirjutas*  
 M.nom write.3.sg.past  
       *raamatu.*  
       book.gen  
 ‘Mari wrote a book.’

The example with the partitive object is aspectually unbounded (1); the example with the total object is aspectually maximally bounded (2). The telic *find* type of achievements resembles the accomplishments in object case matters, since it appears in maximally bounded sentences and with total objects as in (3).

- (3) *Mari leidis raamatu.*  
 M.nom find.past.3.sg book.gen  
 ‘Mary found a/some book.’

Unexpectedly, the telic verbs of the *win* type do not have aspectual case alternation. The article addresses this data in Section 4 and then proposes a way to understand and represent the partitive object case for achievement transitive verb classes of the types *find* and *win*.

### 4 Specific Data to Represent

As opposed to the type of event verbs such as *write* and *find*, event verbs of the *win* type do not have aspectual case alternation (4).

- (4) *Mari võitis Jürit.*  
 M. won George.part  
 ‘Mary won George.’

Both achievement verbs appear in bounded sentences, but they differ in their cumulativity in tests. The type of boundedness of these verbs is, consequently, different. The sentences with the *win* type achievements are semantically diverse, cumulative and not divisive (see Kiparsky, 1998) and the *find* type are diverse, not cumulative and not divisive. Diverse, cumulative and not divisive boundedness (4) is further referred to as minimal boundedness. Diverse, not cumulative and not divisive boundedness (3) is referred to as maximal boundedness. A predicate is not divisive if the proper parts of the event described by the predicate are not in the denotation of the predicate. This definition classifies the predicates *võitma* ‘win’ and *leidma* ‘find’ as non-divisive, since a part of a finding or a winning event cannot be always qualified as finding or winning. A predicate is cumulative if the sum of the events that are in the denotation of the predicate is in the denotation of the predicate (understood as temporally adjacent). This definition classifies the predicate *leidma* ‘find’ as non-cumulative, since another event denoted by that predicate cannot be qualified as finding the object referent, it can only be qualified as performing acts of ‘finding’ it again. On the contrary, despite its clearly similar

eventive character, *võitma* ‘win’ is lexically cumulative, since two events that qualify as winning (somebody) can still be qualified as winning him or her in this lexical meaning (‘more’, as in another game, not necessarily ‘again’ or ‘once more’ as with the *find* achievement). Both verbs allow objects and form bounded sentences with partitive case marked objects, but the restriction is that of mass or plural, as in (5) and (6).

- (5) Mari leidis raamatuid.  
 M.nom find.past.3.sg  
 book.part.pl  
 ‘Mary found a/some book.’
- (6) Mari võitis poisse.  
 M. won boys.part.pl  
 ‘Mary won the boys.’

Testing captures the elusive eventive nature of Estonian sentences with the *leidma* type verbs “telic” verbs and partitive marked subjects or objects that are mass or plural NPs, as in *leidis raamatuid* ‘found books’. The predicate is not maximally bounded by the definition applied here, since it is cumulative; finding more books is still finding books, so the sum of the events that are in the denotation of the predicate *leidis raamatuid* is also in the denotation of the respective predicates. In my terminology, the sentences with partitive plural objects are minimally bounded (as the win verbs) in case of both achievement verbs.

The aspect of the sentence is needed for MT and also TTS applications need the exact object case, to mention some. For speech production it is important how to pronounce the case of objects if the object NP is a numeral, such as in the following real life example from the internet *Helen Mirren solvas Elizabeth II hingepõhjani* ‘HM offended Elizabeth II deeply’. As these verbs may appear with total objects as well, a purely statistical MT method for Estonian-English may wrongly translate the sentences with partitive objects to Progressive English verbs (*was winning*, *was finding*).

The boundedness feature that corresponds to the observed aspectual distinctions is a scalar (gradable) feature as proposed in Tamm (2004). The scale is formed from zero boundness (1) via minimal boundedness (4), (5), (6) to maximal boundedness (2), (3). The question is how to represent the cause of the following effect: the minimal boundedness of the sentence is in some cases due to the semantically conditioned partitive object, as in (5) and (6) and, in other cases, due to the verb, as

in (4). The article follows an analysis of Tamm (2004): the object case alternation is an aspectual semantic and functional syntactic phenomenon. Accordingly, the special focus is on the modeling of verbs and the aspectual object case at the syntactic level of the functional structure.

## 5 Achievement Verbs’ Entries

The main puzzle to solve in a model representing aspect, verbs, and case is the lexical minimal boundedness of the *win* type achievements, the relative aspectual freedom (boundability) of the *find* type achievements and the semantic effect of semantically conditioned partitive objects. The sentences with partitive plural objects are minimally bounded in case of both achievement verbs. Indications about the boundedness and boundability belong to the functional specifications in the entries and in the respective terminal node of the constituent structure (Tamm, 2004). The functional specifications that are associated with different nodes constrain the functional structures.

### 5.1 Bounded Achievement Verbs: *win*

The intuition to capture with these verbs is that their lexical grammatical properties prevent them from appearing in bounded sentences with total objects. If a verb is lexically minimally bounded, then its boundedness feature must be fully specified in the lexical entry. These specifications have the form of defining equations as in the verb entry of *võitma* ‘win’ (7).

- (7) *võitma*, V:  
 $(\uparrow \text{PRED}) = \text{‘WIN} < (\uparrow \text{SUBJ}), (\uparrow \text{OBJ}) > \text{’}$   
 $(\uparrow \text{B}) = \text{MIN}$

The information containing the functional specifications in the entry in (7) are mapped from the constituent structure (c-structure) to functional structure (f-structure) as illustrated in Figure 1.



Figure 1. *võitma* ‘win’ at the f-structure.

In this case, boundedness is specified in the lexical entry of the verb, and clausal aspect is determined by the verb. As the result of the mapping from constituent structure to functional structure, the f-

structure is constrained to contain the specified boundedness feature, that is, an attribute with a “fixed” value (Figure 1). Having a fully specified feature (a defining equation) as part of its lexical entry, such as  $(\uparrow B) = \text{MIN}$ , captures that the verb’s boundedness is lexicalized, that *win* is an inherently bounded verb (lexical sense of the verb). Since clausal aspect is modeled in terms of the unification of boundedness features at the f-structure, the failure in unification explains the restrictions on case marking patterns in the model, where case contributes different values. The effect of the constraint is that the verb is not boundable by further elements in syntax, and the range of aspectual case marking possibilities available for the verb is restricted.

## 5.2 The boundable achievement verbs: *find*

If the verb is boundable, that is, aspectually free, then its boundedness feature must be partially specified in the lexical entry. This specification has the form of an existential constraint as in the verb entry of *leidma* ‘win’ (8).

$$(8) \textit{leidma}, V: \\ (\uparrow \text{PRED}) = \langle \text{WIN} \langle (\uparrow \text{SUBJ}), (\uparrow \text{OBJ}) \rangle \rangle \\ (\uparrow B)$$

Having an existential constraint  $(\uparrow B)$  means that the attribute B must be present in the f-structure feature matrix that corresponds to the verb in c-structure. As clausal aspect is modeled in terms of the unification of boundedness features in the functional structure, the possibility of the unification with features with different values explains the wider range of case marking patterns. The information containing the functional specifications in the entry in (8) are mapped from the c-structure to f-structure as illustrated in Figure 2.

$$\left[ \begin{array}{l} \text{PRED 'FIND' } \langle \text{SUBJ, OBJ} \rangle \\ B \end{array} \right]$$

Figure 2. *leidma* ‘find’ at the f-structure.

The next question is: given the incomplete f-structure, how will the values be obtained? In this model, the “underspecified” features become fully specified by the features of case-marked objects. Before discussing the verbs’ contribution to the

sentence in their interaction with case-marked objects, the features associated with the partitive case marker are presented.

## 6 Partitive

In their aspectual behavior, the partitive objects that do not have a restriction on the semantics of the object NP (example 1, 4) are different from those that have (example 5, 6). As this article concentrates on the new results in the representation of the semantic partitive, the semantically unrestricted partitive is regarded as a default (Tamm, 2004).

### 6.1 The Default Partitive

In several previous sources, the semantically unrestricted Finnish partitive is regarded as the default case. Estonian can be regarded similar in this respect, and the default is captured in an annotation at a c-structure node (9).

$$(9) X' \rightarrow X^0 \text{ XP} \\ ((\downarrow \text{CASE}) = \text{PART})$$

On the other hand, the entry for the partitive case (10) encodes only the constraint that the sentence is not maximally bounded. As a default, the boundedness feature has the value “0”, for unbounded.

$$(10) \text{PARTITIVE: } (\uparrow \text{CASE}) = \text{PART} \\ ((\text{OBJ } \uparrow) B) = \neg \text{MAX} \\ ((\text{OBJ } \uparrow) B) = 0$$

The indication  $(\text{OBJ } \uparrow)$  is the inside-out function application (Nordlinger and Sadler 2004). The association between the nominal and its grammatical function is established by virtue of the case marker attached to it. I leave the semantic constraints that constrain the mapping between the f-structure and semantic structure aside. A defining equation captures the constraint on boundedness on the corresponding f-structures (Figure 3).

$$\left[ \begin{array}{l} \text{OBJ} \\ \left[ \text{CASE PART} \right] \end{array} \right]$$

Figure 3. The partitive case at the f-structure.

The result is that the partitive NPs specify the information in the f-structure feature matrix as in Figure 3.

## 6.2 The Semantic Partitive

Singular mass noun and plural count noun partitive objects are specified as follows (adding the tentative possible semantic restrictions in prose for descriptive clarity) in (11).

(11) SEMANTIC PARTITIVE:

(↑CASE) = PART

(OBJ↑)

((OBJ↑)B) = MIN

the referent is homogeneous

The semantically conditioned partitive maps to the information in the f-structure feature matrix as illustrated in Figure 4, specifying the constraint on the aspectual minimal boundedness of the sentence.

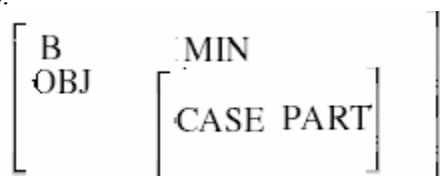


Figure 4. The semantic partitive case at the f-structure.

My representation is syntactic, but the semantic nature of several restrictions—as in (11)—points to the necessity of more explicit semantic structure, which ideally interacts with the morphosyntactic and lexical analysis as well as pragmatic discourse structures.

## 7 Feature Unification

Fully lexically inflected words enter the LFG c(onstituent)-structure terminal nodes. The lexical entries for the verbs *find* (12) and *win* (14) and the semantically conditioned partitive and the object are represented as in (13) and (15).

(12)

*leidis*, V:

(↑PRED) =

'find<(↑SUBJ), (↑OBJ)>'

(↑TNS) = PAST

(↑PERS) = 3

(↑NUM) = SG

(↑B)

(13)

*raamatuid*, N:

(↑PRED) = 'BOOK'

(↑CASE) = PART

(↑NUM) = PL

(OBJ↑)B) = MIN

(OBJ↑)

(14)

*võitis*, V:

(↑PRED) =

'win<(↑SUBJ), (↑OBJ)>'

(↑TNS) = PAST

(↑PERS) = 3

(↑NUM) = SG

(↑B) = MIN

(15)

*poisse*, N:

(↑PRED) = 'BOY'

(↑CASE) = PART

(↑NUM) = PL

((OBJ↑)B) = MIN

(OBJ↑)

The possible unifications determine the possible aspectual semantics for sentences in the unification-based approach of LFG. The aspectual features of verbs and case are unified at the functional structure. The lexical entries in the computational lexicon for transitive verbs are provided with valued or unvalued aspectual features in the lexicon. The *win*-verbs fully determine the sentential aspect, and the aspectual feature is valued in the functional specifications of the lexical entry of the verb; this is realized in the form of defining equations. If the aspect of the verb is variable, as with the *find*-verbs, the entry's functional specifications have the form of existential constraints. The partitive case is the default complement case and the case of the objects of unbounded predicates; mass and plural partitive NPs, however, can be optionally telizers or bounders (but not of the maximal type, though). The general well-formedness conditions on functional structures secure the sensitivity of aspectual case to verb classification.

The following two figures illustrate the unification of the aspectual information that has the form of constraints associated with the verb and object entries. Figure 5 corresponds to sentence (5) and Figure 6 corresponds to sentence (6).

PRED	'FIND <SUBJ, OBJ>'
B	MIN
TNS	PAST
NUM	SG
PERS	3
SUBJ	[ PRED 'MARI' CASE NOM ]
OBJ	[ PRED 'BOOK' CASE PART NUM PL ]

Figure 5. The unification at the f-structure: *leidma* 'win' and the semantically conditioned partitive.

The minimal boundedness feature is contributed by the constraint associated by the object (see (13)) in the feature matrix in Figure 5, which corresponds to sentence (5) with the verb *find*. Appendix A provides the constituent structures for sentence (5) for an illustration.

In the feature matrix of Figure 6, which corresponds to sentence (5) with the verb *win*, the minimal boundedness feature is contributed by the constraint associated with the verb (see (14)). Appendix B provides the constituent structures for sentence (5).

PRED	'WIN <SUBJ, OBJ>'
B	MIN
TNS	PAST
NUM	SG
PERS	3
SUBJ	[ PRED 'MARI' CASE NOM ]
OBJ	[ PRED 'BOY' CASE PART NUM PL ]

Figure 6. The unification at the f-structure: *võitma* 'win' and the semantically conditioned partitive.

## 8 Summary

This article proposes a way to represent and some preliminaries of how to acquire event structural semantic content from Estonian transitive sentences. The parser identifies the finite verb, its object and its object's case; the lexicon contains

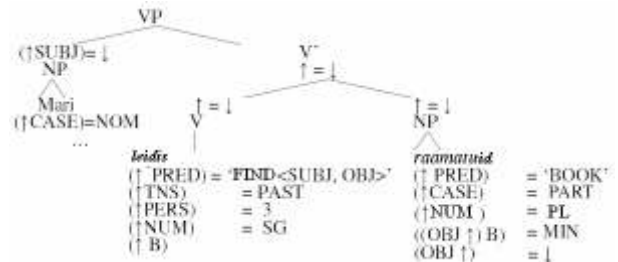
entries with aspectual information attached to verbs and case morphemes. The article presents a unification based model of two classes of Estonian transitive achievement verbs and partitive objects, including those that have semantic restrictions on the NP. This classification accommodates the systematic compatibility of verb classes with clausal aspectual object case marking patterns.

The article applies the Lexical Functional Grammar (LFG) formalism and its methodology. Clausal aspect is understood in terms of boundedness and represented with a clausal boundedness feature. Clausal boundedness is encoded in the form of features at the LFG's syntactic level of f(unctional)-structures. A clause or a sentence is maximally bounded if it describes an event with a definite, maximal endpoint. A clause or a sentence is minimally bounded if it describes an event with an endpoint that is not maximal.

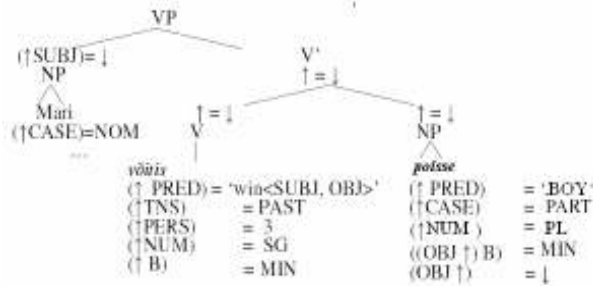
Achievement verbs fall into two aspectual classes, distinguished from each other according to the pattern of the attributes and values in the functional specifications of the verbs' lexical entries. The lexical entries for the *find* type achievements are provided with underspecified boundedness features in the proposed LFG lexicon. The lexical entries for the *win* type achievements are provided with specified boundedness feature MIN (for minimally bounded). The lexical entries for the case markers specify constraints as well. The semantically conditioned partitive is provided with specified boundedness feature MIN.

The aspect-related attributes and values encode constraints on the f-structure, which interfaces with semantics.

### Appendix A. Minimally bounded sentence (5). The constituent structure of the Estonian achievement verb *find* and a partitive plural object.



**Appendix B. Minimally bounded sentence (6). The constituent structure of the Estonian achievement verb *win* and a partitive plural object.**



**Appendix C. Illustration of the relevant parts of the output of Mürisep's syntactic analyzer based on Mürisep (2000), sentences (5) and (6).**

```

Mari
Mari+0 // _S_ prop sg nom #cap // **CLB @SUBJ
mari+0 // _S_ com sg nom #cap // **CLB @SUBJ
mari+0 // _S_ com sg nom #cap // **CLB @SUBJ
võitis
võit+is // _V_ main indic impf ps3 sg ps af #FinV #NGP-P
#In // @+FMV
poisse
poiss+e // _S_ com pl part // @OBJ
poi+sse // _S_ com sg ill // @ADVL
...
leidis
leid+is // _V_ main indic impf ps3 sg ps af #FinV #NGP-P
// @+FMV
raamatuid
raamat+id // _S_ com pl part // @OBJ

```

**Acknowledgements.**

I acknowledge the useful discussions at the LFG 2006 conference and at the Research Institute for Linguistics, Budapest.

**References**

Miriam Butt. 2006. *Theories of Case*. Cambridge University Press, Cambridge.

Miriam Butt and Tracy Holloway King. 2005. The Status of Case. In: Veneeta Dayal and Anoop Mahajan (eds.). *Clause Structure in South Asian Languages*. Springer Verlag, Berlin.

Miriam Butt, Mary Dalrymple and Anette Frank. 1997. An Architecture for Linking Theory in LFG. In: Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG 97 Conference*. CSLI Publications, Stanford.

Hana Filip. 2000. The Quantization Puzzle. In: James Pustejovsky and Carol Tenny (eds.). *Events as grammatical objects, from the combined perspectives of lexical semantics, logical semantics and syntax*. CSLI Press, Stanford: 3-60.

Paul Kiparsky. 1998. Partitive Case and Aspect. In: Miriam Butt and Willem Geuder (eds.) *The Projection of Arguments*. CSLI Publications, Stanford: 265 – 307.

Kaili Mürisep, 2000. *Eesti keele arvutigrammatika: süntaks*. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu.

Rachel Nordlinger and Louisa Sadler. 2004. Tense Beyond the Verb: Encoding Clausal Tense/Aspect/Mood on Nominal Dependents. *Natural Language and Linguistic Theory* 22: 597–641.

Anne Tamm. 2005. Boundedness features in a computational lexicon: Estonian transitive verbs in LFG. In: Kiefer, Ferenc, Júlia Pajzs (eds.) *Proceedings of COMPLEX 2005*, Budapest.

Anne Tamm. 2007. Estonian transitive verbs and object case. In: Butt, Miriam and Tracy Holloway King (eds.) *Proceedings of the LFG06 Conference, Universität Konstanz*. CSLI Publications, Stanford: 484-504.

Anne Tamm. 2004. *Relations between Estonian verbs, aspect, and case*. Doctoral dissertation, ELTE, Theoretical Linguistics Program, Budapest.

Anne Tamm and Gábor Olaszky. 2005. Kísérlet automatizált szövegelemzési módszerek kialakítására a szóhangsúlyok meghatározásához. *III. Magyar Számítógépes Nyelvészeti Konferencia NSzNy 2005. 2005. december 8-9. Szeged*.

Henk Verkuyl. 1993. *Theory of Aspectuality. The Interaction between Temporal and Atemporal Structure*. CSIL 64. Cambridge University Press.

## Demonstrations

# Ontological-Semantic Internet Search

Christian F. Hempelmann<sup>1</sup>, Victor Raskin<sup>1,2</sup>, Riza C. Berkan<sup>1</sup>, and Katrina E. Triezenberg<sup>1,2</sup>

<sup>1</sup>hakia Inc.  
New York, NY 10006  
{chempelmann, rberkan, vraskin}@hakia.com

<sup>2</sup>Purdue University  
West Lafayette, IN 47907  
{vraskin, kattriez}@purdue.edu

An ontological semantic (OntoSem) system represents the meaning of input text, not by trying to reduce it to numbers or insufficient first-order logic, but by instantiating and relating property-rich concepts of events, objects, and relations, thus emulating the mental processes of a human. Our current inventory (as of 3/26/2007) includes the following resources:

- a 6,724-concept language-independent ontology,
- several ontology-based lexicons, including a 47,025-entry English lexicon with 77,156 senses, and a several smaller lexicons for other languages,
- onomastica, dictionaries of proper names for several languages; the current one with 19,352 entries and a total of 24,328 senses,
- a text-meaning-representation (TMR) language, an ontology-based knowledge representation language for natural language meaning,
- a fact repository, containing the growing number of implemented TMRs,
- a preprocessor analyzing pre-semantic (ecological, morphological, and syntactic) information,
- an analyzer (ontological parser) transforming text into TMRs, and
- a generator translating TMRs into text, data, and potentially images.

Users have become conditioned over time to deal with BOW-based search engines, and tweak their queries by reducing natural-languages questions to only nouns and verbs, inserting Boolean operators and punctuation commands, and making multiple searches for synonyms. OntoSem makes it possible for the naïve user (as well as the experienced one) to achieve optimum search results with a single search. For example, suppose that a user with a pounding headache wants to know what remedies are available and appropriate. A BOW query might be “aspirin headache,” or “cure headache,” and neither would produce all of the desired results. Our OntoSem search engine, on the other hand, takes the natural language query *does aspirin cure headaches?* and automatically expand upon the query to produce a thorough search. “Aspirin” would trigger a search not just for the word *aspirin*, but rather for all words linked to its ontology concept, and words linked to that concept’s parent and child concepts—not only “aspirin” but “acetylsalicylic acid” and all of its known brand names, as well as generic words and brand names of conceptually similar drugs. The same would be done for *cure*, bringing up search results for other similar words

such as *treat* and *relieve*, and for *headache*, looking up results for specific types of headaches (child concepts of headache), as well as other similar painful conditions (parent concepts of headache).

The overall goal of our symbiotic effort, uniting bags of words (BOW) and statistical approximations where useful, with OntoSem, has been as follows:

- to optimize the output(s) of OntoSem systems for the implementation of Internet search,
- to differentiate between the use of full-fledged OntoSem resources for offline operations in Internet search, namely crawling, semantic parsing, and storage and retrieval of the parsing results, and
- to develop unprecedented battery of quick, cheap incremental enhancements to each current phase of the search engine development which dynamically and asymptotically move the product to the optimal meaning representation at runtime.

Besides developing what is hoped to be a successful next-generation search engine, which will raise the users’ expectations significantly to one-click-brings-all, the OntoSem approach to NLP brings forth an essentially new discipline of Meaning Processing, as opposed to the BOW-and-statistics NLP. In the latter area, dominated for a variety of academic and sociological reasons, by non-linguists and non-semanticists, the approach emphasizes the significance of the meaning resources underlying human understanding of language and the commitment to developing them. The current NLP approach, on the other hand, is still trying to get at the meaning without penetrating the semantic substance of language, while trying to use ready-made (and, rarely, to create) word and frequency lists, WordNets, OWL formalisms, and other resources that are simple to acquire and used for that very reason. From the point of view of OntoSem, these attempts to get at the meaning without doing semantics look like the *perpetuum mobile* project, probably highly desirable but not realistic and, most certainly, not accurate enough to be acceptable to the human user, as they can only capture what is regular in language. It must be noted, however, that this attitude is highly contaminated by our decisive commitment to the representationalist, AI-type position.

## References

<http://www.hakia.com>  
<http://www.ontologisemantics.com>

# Infomat

## A Vector Space Visualization Tool

Magnus Rosell  
KTH CSC  
100 44 Stockholm  
Sweden  
rosell@csc.kth.se

Infomat is a vector space visualization tool aimed at Information Retrieval (IR) and text clustering in particular. However, it could be used in many areas of language technology, as well as in other fields when information can be stored in a matrix (Infomat – information matrix).

As an example we give an IR matrix where rows represent texts and columns words, see Figure 1. In each matrix element the  $tf*idf$  weight for the word in the text is stored. Similarity between texts is calculated with the cosine measure.

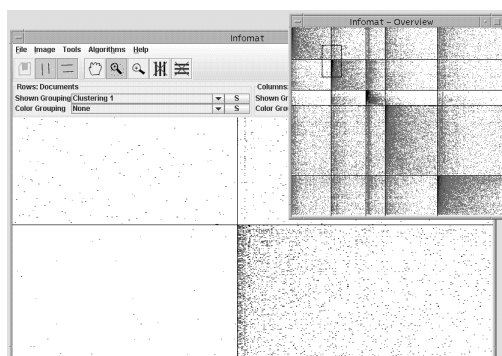


Figure 1: Infomat. The upper right window is the overview window, displaying the whole matrix. The small rectangle in it indicates the part that is displayed in the main window. Using K-Means 2500 Swedish newspaper articles have been clustered to five clusters along the rows. The columns represent 5663 words, clustered to five clusters relative to the row (article) clusters. Hence the diagonal pattern in the overview.

In Infomat the matrix is presented as a scatter plot, where the opacity of each pixel is propor-

tional to the weight of the corresponding matrix element(s). When the mouse pointer is placed over a pixel textual information about its content is presented.

When the matrix is larger than the picture it is compressed and each pixel presents the average value of the corresponding matrix elements. This can be justified when the objects of adjacent rows and columns are related (have high similarity). One way of obtaining such relatedness is through clustering. Infomat provides basic clustering functionality.

Infomat has many functions. Among other things it is possible to zoom in and out of the matrix. To visualize several groupings (clusterings, categorizations, etc.) at the same time it is possible to color objects belonging to different groups in different colors, both in rows and columns.

Many existing IR visualization methods calculate the similarity between all objects and project this relationship down to two or three dimensions (Baeza-Yates and Ribeiro-Neto, 1999). Such methods do not usually give much information as to why objects are deemed similar. In Infomat similarity between adjacent rows and columns appear as patterns, reflecting the distributional definition of similarity.

Infomat is developed in Java and uses an xml-format for reading and writing matrixes. It is freely available<sup>1</sup> together with more information.

## References

R. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley.

<sup>1</sup><http://www.csc.kth.se/tcs/humanlang/>