In: Proceedings of the ESSLI 2001 Workshop on Semantic Knowledge Acquisation and Categorism. August, 2001

Conceptual enrichment of ontologies by means of a generic and configurable approach

Andreas Faatz (*, *) Stefan Hörmann*, Cornelia Seeberg*, Ralf Steinmetz(*, **)**

{afaatz, hoermann, seeberg, steinmetz}@kom.tu-darmstadt.de

*
Industrial process- and system-communications (KOM)
Technical University of Darmstadt
25 • 64283 Darmstadt

**
GMD IPSI
Integrated Publication and Information Systems Institute
Dolivostr. 15 • 64293 Darmstadt

***
intelligent views GmbH
Julius-Reiber-Str. 17 • 64283 Darmstadt

Abstract

In this paper we focus on systematic enrichment of ontologies by candidate concepts. To achieve this goal we compare semantic distance measures between concepts in an ontology with similarity and dissimilarity information introduced by statistical analysis of text corpora. Moreover we present methods, how the corpus extraction and the processing of the statistical information is also enhanced by a given ontology. Summing up these ontological methods we state the enrichment problem as an optimization problem. Its solution will yield two results: the best candidate concepts for the enrichment and the position of these candidates relative to the existing concepts of the ontology.

1. Introduction

Investigating the contemporary techniques of textual knowledge acquisition and formalisation, we identify two tendencies, an old one and a rather new one: on the one hand people preserve and interchange their knowledge creating natural language texts - a fundamental cultural technique of the civilized human being. On the other hand formal knowledge representation tries to find mappings of the real world or parts of the real world to a machine-readable and simplified model of the real world.

If we wish to share such a model of reality, it has to fulfil preliminary design principles [Gr]. The design principles have to hold while editing the knowledge representation. They prescribe formal rules, which are the basis of understanding, what the encoding of knowledge in the formal knowledge representation actually means.

The way we express our knowledge every day is completely different from formal knowledge engineering: we write it down in natural language. We observe a worldwide growth of the number of text documents, messages, textual archives and hypertextual informations, which are not formalised. All these ’uncontrolled’ archives of knowledge grow significantly faster than the formal knowledge representations in any domain - and use natural language.

The question arises, if there exists an alternative to the strict formalisation of knowledge - an alternative, which benefits from the high quantity of textual information. In this paper we examine this question concerning the following situation: imagine there already exists a domain ontology, that is the conceptualisation of a knowledge domain [Gr]. Additionally there are large collections of text documents available, among them documents, which could form a basis for enriching a given ontology The question of opening formal knowledge representations for knowledge expressed in natural language transforms to two rather concrete ones:

- how can we extract concepts from the large collection of documents, which *fit semantically* to the knowledge, i.e. to the concepts, represented by the ontology?

- how do we *group the concepts* identified in such a manner to the concepts in the ontology?

In this paper we envisage an approach, which tries answer these questions of ontological enrichment: we extract concepts from very large collections of texts and insert them as candidate concepts of an ontology.

2. Overview

In this paper we exploit the comparison of semantic distance measures between concepts in an ontology and similarity introduced by statistical information about the usage and especially the collocated usage of words in text corpora. From the comparison we develop candidate concepts, which can for instance be presented to a knowledge engineer as a possible enrichment of the ontology.

The paper is organised as follows: in section 3.1 we characterise the formal knowledge we want to enrich. We introduce the notion of ontologies and define heuristics for a computation of conceptual distance in 3.2. Afterwards we characterise the less formally structured resources of natural language we want to use for an enrichment of the ontology (4.1 and 4.2). We are in need of an identification and an isolation of concepts from the textual resources. Founding on statistical information in section 4.3 we explain a generic and configurable approach, which originally was applied to automatic ontology generation [BiNeCa]. Since our goal is ontology enrichment we undertake a further development of the approach which is twofold: we refer to the generic features (the so called rule-base) of the approach inserting attributes, which can only be defined with an ontology at hand.

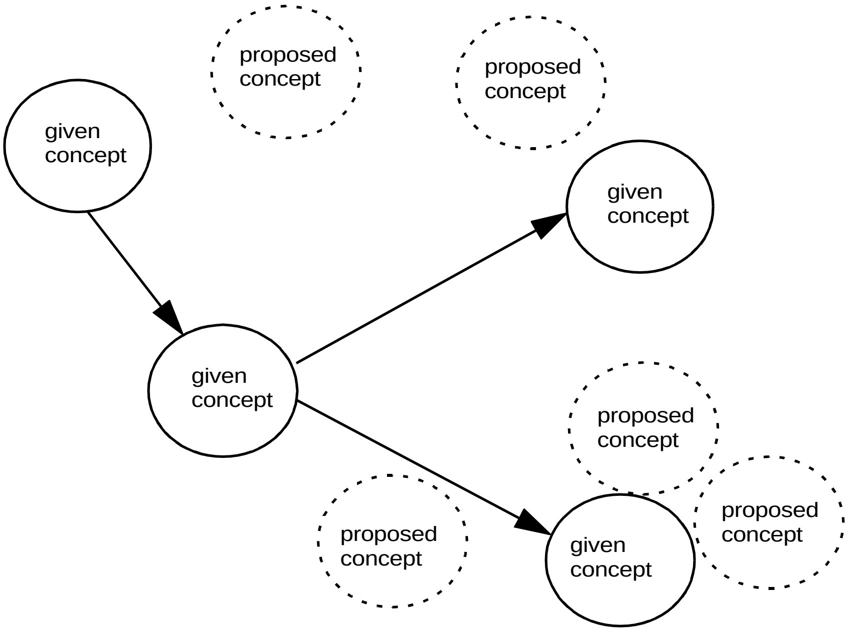

Figure 1

Secondly referring to the configurable features of the approach in section 4.4 we derive an optimization problem.

The solution of the optimization problem is able to fill a given ontology with proposals of new concepts, as shown in 4.5. The solution leads to the least possible numerical contradiction between ’similarities’ in our textual and ’similarities’ in our ontological world. Our vision is depicted in figure 1, where nodes indicating given concepts and arcs indicating ’is_a’ relations among the given concepts attract further nodes, which means further concepts.

In section 5 we present our conclusions and future work to be done.

3. Ontologies

3.1 Definition and characterisation

Throughout this paper Ω will always denote an ontology.

In the literature about knowledge representation there exist many different definitions of ontologies. One of the most common definitions is due to Gruber [Gr], who defines an ontology to be "*a conceptualisation of a knowledge domain*". Guarino [Gu] put efforts into an at least uniform characterisation of the components an ontology necessarily holds. Together with [Gu] we state, that an ontology has the following parts: *concepts*, which represent the things existing in our knowledge domain, *relations* connecting the concepts semantically and *axioms* as formal laws for the ontology. Throughout this paper we assume that a concept has a name existing in a dictionary. This assumption is not natural for arbitrary ontologies, because for example in description logics [Ho] one could define concepts like 'all countries with the same number of inhabitants as Malta'. Clearly there does not exist a name from a dictionary for this concept. Moreover we do not want to consider ontologies containing constructed names like for example the medical ontology GALEN does [OG]. GALEN distinguishes 'skin-as-organ' and 'skin-as-tissue', contrary to that we would refer to 'skin'. Throughout this paper we assume the identity of a concept and its name.

The second ingredient of an ontology is a set of relations. A relation *r* in our case is binary and establishes statements about concepts *x, y* : we write *r(x, y)*, if it is true, that a relation *r* holds between *x* and *y*. For example, if *r* is 'is a' it may connect the concepts 'salmonella' and 'bacterium' to a statement, which we interpret as 'a salmonella is a bacterium'.

The third part of an ontology, the axioms, are responsible for the supervision of concept creation and deletion and for relation creation and deletion [StM]. For example 'is a' almost ever is a transitive relation, i.e. if for concepts x, y, z the relations *r(x, y)* and *r(y, z)* hold, then also *r(x, z)* holds. As our approach will propose extensions of a given state of an ontology, we do not examine axioms in detail.

Note that our characterisation of an ontology by concepts, relations and axioms implies that we can visualise such an ontology as a directed graph $\Gamma(\Omega)$, with nodes corresponding to concepts and edges corresponding to relations.

Our ontological enrichment or extension identifies new concepts and groups them semantically to concepts from a given set of concepts from an existing ontology Ω: we translate the concepts of Ω and the relations among them to a semantic distance measure in the ontology. Technically this is based on relational paths in $\Gamma(\Omega)$ and on topological properties of $\Gamma(\Omega)$.

3.2 Distance measures

The distance between two concepts *x, y* indicates, how strong the semantic similarity between *x* and *y* is. Semantic similarity and the distance measure are correlated: the bigger the distance between *x* and *y*, the bigger the dissimilarity between *x* and *y*. Unlike the approaches of [Tv], which is based on hierarchical structures, and [Leng], which is based formal concept lattices, we do not define our similarity measures by the attributes or the extension of concepts. Nevertheless the visualisation by [Leng], which is based on the construction of paths corresponding to an attribute-driven similarity definition of a distance, has commonalities with our premise: semantic distances between concepts in an ontology Ω are defined by path lengths in the graph $\Gamma(\Omega)$ we introduced in the previous section.

To enhance a reasonable distance measure we add a restriction to our ontology Ω, namely some our relations have to be hierarchical, for example 'is a'- or 'subconcept of'-relations. The parts of $\Gamma(\Omega)$ remaining form a multihierarchy as shown in figure 2, where the plain lines visualise hierarchical and the interrupted lines express non-hierarchical relations. .

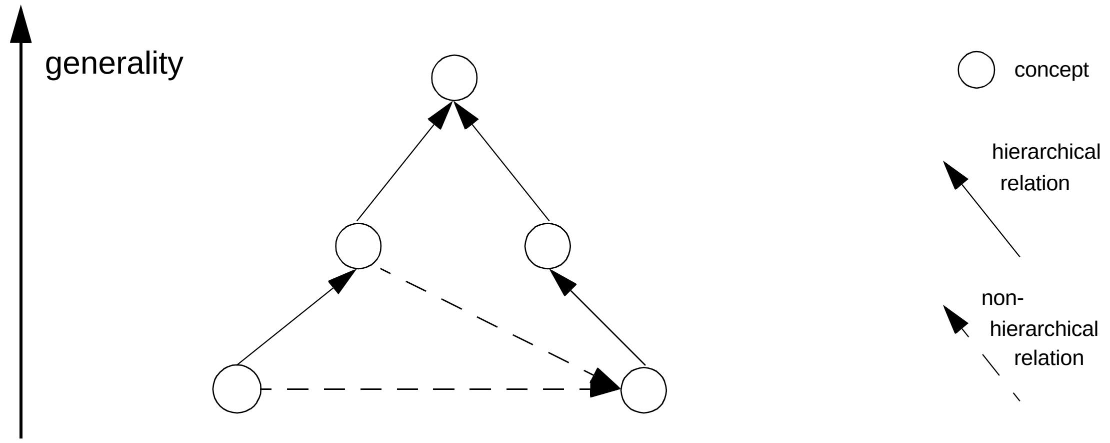

Figure 2

In figure 2 we also find an axis ('generality'), which expresses the fact, that the visualisation places the more general or abstract concepts at the top and the more special concepts to the bottom. We denote by $\Gamma'(\Omega)$ those parts of $\Gamma(\Omega)$, which are formed by the concepts (as nodes) and the hierarchical relations (as edges). Thus $\Gamma'(\Omega)$ would result from figure 2 omitting the interrupted lines. We will only use the hierarchical relations in the definition of semantic distance and claim, that the measure should fulfil four principles:

(i) longer paths in $\Gamma'(\Omega)$ are correlated with a higher semantic distance.
(ii) up-posting (generalisation) leads to a lower semantic distance than stepping down the hierarchy. Finding superconcepts is more fault tolerant than referring to senseless specialisations.
(iii) siblings increase the distance: this means the more subconcepts one concept *z* has, the higher is the distance between the subconcepts and also between *z* and its subconcepts. We posed this principle, because too many siblings often indicate missing abstraction levels in a hierarchy.
(iv) for isomorphic paths the distance between abstract or general concepts is higher than between special concepts at the bottom of the hierarchy, i.e. we judge 'the same' path on an abstract level to be vague in comparison to its pendant on a concrete level.

How do we achieve (i)-(iv) mathematically? An example distance design is given by the formula

(1) $$D(x,y) = \left[\frac{\alpha}{\gamma}\left(\frac{1}{\kappa}+\frac{1}{2\lambda}\right)\right]^{-2}$$

where the distance *D(x,y)* between two concepts *x, y* is computed with α denoting the average abstraction level of the concepts *x, y,* whereas γ denotes the average number of siblings of *x* and *y*. By κ we denote the steps upwards in the hierarchy while moving from *x* to *y* via the shortest path in $\Gamma'(\Omega)$, by λ the steps downwards in the hierarchy. Let us define exceptions for formula (1). If $x = y$ then set *D(x,y)=0,* if $\kappa = 0$ then set

(1a) $$D(x,y) = \left[\frac{\alpha}{2\gamma\lambda}\right]^{-2}$$

and analogously, if $\lambda = 0$ set

(1b) $$D(x,y) = \left[\frac{\alpha}{\gamma\kappa}\right]^{-2}$$

Formula (1), (1a) and (1b) together fulfil (i)-(iv), because

(i) increasing path lengths κ and λ increase *D(x,y)*

(ii) $(2\lambda)^{-1}$ decreases faster than κ^{-1}

(iii) as γ increases also *D(x,y)* increases

(iv) we work with an abstraction factor α, which we define as 1 divided by the average number of steps along the shortest path in $\Gamma'(\Omega)$ from *x* and *y* to one of the top level concepts, which means a concept without any hierarchical relations posting up from it. The abstraction factor α is high on a high level of abstraction and low on a low level of abstraction.

Note that *D(x,y)* and *D(y,x)* do not have to be equal. The exponent -2 stretches the distances, such that high distances become relatively even higher.

For technical reasons, namely turning semantic distance into semantic similarity to compute the optimization formula, α should be modified by a multiplication with a positive real in such a way, that $0 \le D(x, y) \le 1$ for every pair of concepts *x* and *y*. A consequent modification of α would yield 1 for the maximal occurring distance *D(x,y)*.

With *n* concepts given, we obtain $n \times n$ distances, which will serve as the input for the comparison mechanism optimising the configuration of the statistic knowledge acquisition process, which we will describe in the next section.

4. Generic and configurable acquisition and enrichment

4.1 Overview of the approach

For the remainder of this paper we consider an ontology Ω to be given. In this section we refer to a generic approach to determine similarities between concepts. It is based on computing statistics of word usage in natural language. Large text corpora are the basis for such statistics. [BiNeCa] first described the design and the capabilities of such an approach and applied it to conceptual clustering without a given ontology. Consequently the major differences between [BiNeCa] and our approach are

- we focus on creating algorithms for ontological enrichment and not on automatic ontology generation.
- we control our textual input by the concepts from Ω. This is motivated by experiments outlined in [BiNeCa]: specialised and pruned corpora seem to be the ideal input to unbias the word usage statistics.
- we only will use the initial step of [BiNeCa] and define conceptual similarities. We will modify it because we want to determine conceptual similarities within and by the help of Ω.

We undertake the following steps: we list the concepts from the ontology Ω and query the text corpus. This extracts from the corpus all natural sentences with at least one concept from the ontology Ω (see 4.2). We fill a representation of the collocation phenomena in these extracted sentences (see 4.3). From this sample representation we define similarity measures between any pair of concepts: note that a concept may belong to the *known* concepts from Ω and the additional concepts from the extracted sentences, which were formerly *unknown*. This finally yields similarities between each pair of concepts. These similarities will be optimised (see 4.4 and 4.5) in such a way, that at least the *known* concepts from the ontology Ω have similarities, which do not contradict the distances from section 3.2.

4.2 Corpus split and extraction

Again an ontology Ω and a large text corpus are given. By $B_\Omega = \{b_1, , , , b_n\}$ we denote the set of concepts from Ω. In this first step we derive a corpus C_Ω, which contains all sentences from the given corpus with at least one $b_i \in B_\Omega$. The preparation of a corpus C_Ω is motivated by the work of [BiNeCa] on the one hand, who found evidence for the fact that specialised corpora with a restricted vocabulary tend to be a good basis for ontology learning. A partial restriction of the vocabulary and the domain in our case will be due to the set of concepts B_Ω. The work of [ChGr] shows, that a corpus, which is as general as the one we use (a newspaper corpus in both cases) may be split to several artificial corpora, that were also a good basis for ontology learning. [ChGr] called their method 'corpus split' because they grouped sentences from the corpus to thematic domains: a sentence was added to the sample of the thematic domain, if it contained significantly many words from one and only one thematic domain. The thematic domains were characterised by typical words occurring in domain (such as 'judge', 'crime', 'law' for the domain if justice) together with a value of importance: for the domain of justice judge(0.8), law(0.5), crime(0.6) for example would indicate, that the occurrence of 'judge' in a sentence is highly significant for grouping a sentence containing 'judge' to the domain of justice, i.e. more significant than the occurrence of 'law' in a sentence. The domain significance information was determined by a collocation network of the classical news categories in newspapers (politics, foreign affairs, arts, media, justice etc.).

Our approach to extract sentences from a corpus works from the point of view Ω - and especially B_Ω - instead of a collocation network. Another difference to [ChGr] is, that we do not categorise the extracted sentences to get specialised text corpora. We rely on the powers of the distance measures for Ω defined in 3.2 to implicitly group the concepts from C_Ω. This will be done in the further steps.

Although the ontology enrichment approach we will show in the remainder of this paper is a technique, which may be applied to other languages, our methodology will be applied to German. For German the IDS ('Institut für deutsche Sprache', i.e. German Language Institute) located at Mannheim/Germany offers online access to very large newspaper corpora [IDS]. The general online corpus at the IDS contains several volumes of more than ten newspapers respectively. Moreover we find the online query system COSMAS I [IDS] supporting extraction queries like the one described at the beginning of 4.2. and in addition to this systematic stemming. Consequently for the German language we find an ideal environment for the establishment of the corpus C_Ω

4.3 The representation matrix

The next basic step is filling a representation matrix $M(C_\Omega, \rho)$ for a finite rule set ρ and the corpus C_Ω defined in the previous section. We firstly explain, what a matrix entry means and then we explain how we install a rule set ρ. Let us state, that C_Ω by definition contains some of the concepts B_Ω and further nouns, which we aggregate to a set $B(C_\Omega)$. B_Ω are *known* concepts from the ontology Ω, whereas $B(C_\Omega)$ are formerly *unknown* concepts.

We define $B_\Omega \cap B(C_\Omega)$ to be empty, $B(C_\Omega)$ consists of the concepts (i.e. nouns) *additionally* found in C_Ω and is the set of the potential candidates for the enrichment.

Each row in the representation matrix includes the information concerning the properties of exactly one concept from the unified set of concepts $B_\Omega \cup B(C_\Omega)$. More precisely reading the i-th row of $M(C_\Omega, \rho)$ verifies, if the rules in ρ are fulfilled for the i-th concept of $B_\Omega \cup B(C_\Omega)$. That means the j-th entry in the i-th row indicates, if the j-th rule is fulfilled or how often it is fulfilled in C_Ω. The rules ρ used in [BiNeCa] were of a syntactic type. The deployment of a syntactic parser found concepts, which appeared in at least one sentence from the corpus in a subject-predicate-object relationship with another concept. We give an example of a syntactical rule:

> Take the sentence '*The bakerman is baking bread*' as an example sentence occurring in the corpus. The rule ' '*bread* ' occurs as an object of '*bakerman*' ' would be fulfilled, as the sentence exists, and in the row for '*bread* ' there would be a positive entry at the j-th position, if j is the number of the rule 'concept x occurs as an object in sentence with *bakerman* as a subject'.

As we do not have syntactic parsers for German at hand we define a rule set ρ, which reflects another usage information. The genericity of the approach comes into play here: technically every rule set ρ reflecting usage of concepts in a corpus may design the columns of our matrix $M(C_\Omega, \rho)$.

The query system, which [IDS] offers, is able to return the words, which occurred at a maximal distance δ_W in a sentence. δ_W just counts, how many words one has to pass in a sentence from one word to another. At this point we remark, that the notion of δ_W should not be mixed up with the semantic distances from section 3.2. In our example sentence '*The bakerman is baking bread*' we obtain the values for δ_W listed in table 1

Table 1:

word	word	δ_W
bakerman	*bread*	3
bakerman	*baking*	2
baking	*bread*	1
bread	*bread*	0

For the rule set ρ we chose a fixed δ_W and constitute the j-th rule dependent from the ontology Ω and its set of concepts $B_\Omega = \{b_1, , , , b_n\}$. We define the j-th rule ρ_j to be fulfilled for a concept $x \in B_\Omega \cup B(C_\Omega)$, if x occurs at a predefined maximal distance δ_W to the concept $b_j \in B_\Omega$ in a sentence from C_Ω.

Once we organised this particular rule set ρ we declare the entries m_{ij} of the matrix $M(C_\Omega, \rho)$ as follows:

(2) $m_{ij} = 0$, if the *i*-th concept in $B_\Omega \cup B(C_\Omega)$ does not fulfil the *j*-th rule ρ_j in ρ

(3) $m_{ij} = 1$, if the *i*-th concept in $B_\Omega \cup B(C_\Omega)$ fulfils the *j*-th rule ρ_j in ρ

For a more refined evaluation of the fact *how often* a rule ρ_j was fulfilled we could modify m_{ij} from (3) and count the number of occurrences of rule ρ_j for the *i*-th concept.

Actually all the m_{ij} from for technical reasons should be normalised to obtain $m_{ij} \in [0, 1]$. We achieve this by viewing all m_{ij} as a statistic sample, applying statistic standardisation and cutting off all values less than zero.

For the development of our approach we only focus on three characteristics of the m_{ij}:

(i) if a rule is not fulfilled for a concept, (2) applies and the entry equals zero.

(ii) if a rule is fulfilled for a concept, $m_{ij} \in (0, 1]$ must hold.

(iii) if we do not rely on the naive strategy (3), we should define m_{ij} positively correlated with the frequency of this rule appearing in the corpus C_Ω.

We now give an example of a representation matrix constructed according to (2) and (3). Note that the size of the example is not realistic. Indeed we have to deal with large corpora, so we just demonstrate the context of section 4. Let the set of known concepts be $B_\Omega = \{bread, butter, food\}$. Let C_Ω consist of two sentences:

The bakerman is baking bread.
She had bread and butter for breakfast.

The set of additionally retrieved concepts from the corpus C_Ω is $B(C_\Omega) = \{breakfast, bakerman\}$. With a rule set ρ founded on the word distance $\delta_W = 4$ in a sentence and the naive entry strategy (3) we obtain a matrix $M(C_\Omega"", \rho)$ like the inner cells we visualise in table 2:

Table 2:

	ρ_{bread}	ρ_{butter}	ρ_{food}	$\rho_{breakfast}$	$\rho_{bakerman}$
bread	1	1	0	1	1
butter	1	1	0	1	0
food	0	0	0	0	0
breakfast	1	1	0	1	0
bakerman	1	0	0	0	1

The matrix is not at all typical, regularly we expect sparse matrices in our approach. The next section will derive similarity measures from matrices $M(C_\Omega, \rho)$ by comparison of the respective rows.

4.4 Configurations, similarities and optimal configurations

We follow [BiNeCa] and assign a weight to each rule in our set of rules ρ. While [BiNeCa] assigned these weights with a focus on avoiding overgeneral clustering, a problem obscuring many approaches in automatic thesaurus and ontology construction, we will use the weights as a configuration, which bring into line the two semantic informations we compute: the distances implicitly given by $\Gamma'(\Omega)$ (see 3.2) and the similarities we define now.

Let us assume a given representation matrix $M(C_{\Omega}, \rho)$. A set $k = \{k_1, , k_n\}$ of positive (or zero) reals with $|k| = |\rho|$ will be called configuration of the rule set ρ. The configuration *k* decides about the similarities we derive from $M(C_{\Omega}, \rho)$ by the following definition. We define conceptual similarity *S(x,y)* between two concepts *x,* the *i*-th concept from $B_{\Omega} \cup B(C_{\Omega})$ and *y,* the *j*-th concept from $B_{\Omega} \cup B(C_{\Omega})$ to be

$$S(x, y) = \sum_{l=1}^{|\rho|} \frac{k_l}{|\rho|} min[(m_{il}, m_{jl})] \tag{4}$$

Thus a fixed configuration is needed to compute *S(x,y)*. The question arises, how we should chose such a fixed configuration *k*. Remember section 3.2: we already own - for the concepts B_{Ω} from the ontology Ω - a similarity, we just have to interpret semantic distance *D(x,y)* as the contrary of semantic similarity. Because we required for our distance measure $D(x, y) \in [0, 1]$ the transformation of such distance measures to similarity measures, $S_{\Omega}(x, y)$ can be achieved easily:

$$S_{\Omega}(x, y) = 1 - D(x, y) \tag{5}$$

Taking the distances the ontology Ω as an input, which approximately should be supported by the *S(x,y)* the question of finding an optimal configuration *k* reduces to the question:

> what configuration might minimise the average (squared) error expressed by the (squared) differences $(S(x, y) - S_{\Omega}(x, y))^2$?

Finally we present a formulation of this question in terms of a quadratic optimization formula. Searching for an optimal *k* means searching for a minimum of the expression

$$\sum_{i=1}^{N} \sum_{i=1}^{N} \left(1 - D(x_i, x_j) - \left(\sum_{l=1}^{|\rho|} \frac{k_l}{|\rho|} min[(m_{il}, m_{jl})]\right)\right) \tag{6}$$

with respect to $k = \{k_1, , , k_n\}$ and $k_l \geq 0$ for all $k_l \in k$. Let us clarify the notation of formula (6): we let *i* and *j* run from 1 to $|B_{\Omega}| = N$. This means, that in formula (6) x_i denotes the *i*-th concept from B_{Ω}. Correspondingly in (6) the m_{il} are the matrix entries in $M(C_{\Omega}, \rho)$ in the row of x_i, whereas the m_{jl} are the matrix entries of $M(C_{\Omega}, \rho)$ in the row of x_j.

As the time consumed for solving the quadratic optimization problem (6) depends on the number of rules, we mention at this point, that in case of large rule sets ρ these ρ can be reduced in two ways:

- if a rule is not at all or only once fulfilled for the concepts B_{Ω}, it does not influence *S(x,y)* for $x, y \in B_{\Omega}$ andshould be skipped.
- if two rules are strongly positively correlated, for example in case of synonyms constituting the rules, one could merge these rules to one rule.

Furthermore, we suggest a careful approach towards the number of concepts from Ω, that is on $|B_{\Omega}|$. All steps presented also work with any subset of B_{Ω}.

We now get to our conclusion and explain, which concepts are candidates for the ontological enrichment.

4.5 Ontological enrichment

Once we optimised formula (6) with respect to $k = \{k_1, , , k_n\}$ we obtain the configuration in need to compute all the similarity measures *S(x,y)* between all the concepts from $B_{\Omega} \cup B(C_{\Omega})$. We propose to apply an enrichment step starting with the optimal similarity measures *S(x,y)*.

Only take into concern the *S(x,y)* with $x \in B_{\Omega}$ and $y \in B(C_{\Omega})$. If such a distance between a formerly known concept and a formerly unknown concept exceeds a threshold, for example the average of the similarities $S_{\Omega}(x, y)$, it is a candidate concept, which should be communicated to a knowledge engineer creating or maintaining the ontology Ω.

Additionally the *S(x,y)* with $x \in B_{\Omega}$ and $y \in B(C_{\Omega})$ carry even more information, namely an optimal placement of the candidate concepts. The candidate concepts and the concepts from Ω can be presented together, a fact that simplifies the knowledge engineer's understanding of how the candidate concepts evolved, i.e. in which semantic area of Ω they might belong. We sketch how this can take place in a visualisation. Remember figure 2, which presented $\Gamma'(\Omega)$, the hierarchical parts of Ω as a graph. Figure 3 is an example of a visualisation of the enrichment process.

.

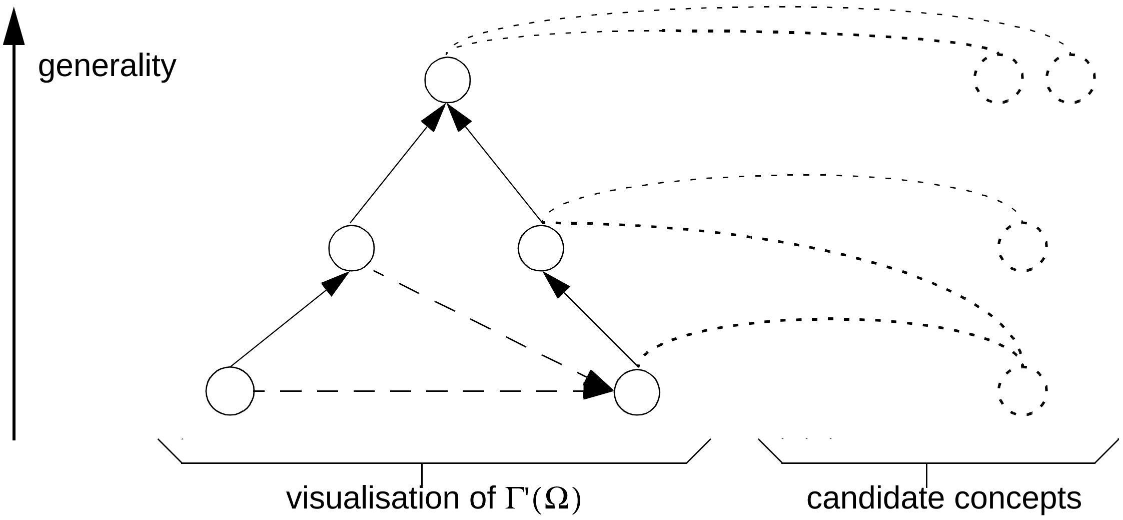

Figure 3

Figure 3 shows the visualisation of the known concepts on the left. On the right the candidate concepts were inserted. These candidate concepts have a similarity exceeding the threshold *T* we mentioned at the beginning of this subsection. Moreover, the dotted lines between candidate concepts and known concepts show, *which one* of the known concepts and *which one* of the candidate concepts have a similarity exceeding the threshold *T*. To sum it up we can visualise our ontological enrichment by drawing $\Gamma'(\Omega)$, adding all $y \in B(C_{\Omega})$, for which a $x \in B_{\Omega}$ with $S(x,y) > T$ exists. Finally we draw a line for each pair x, y with $S(x,y) > T$. Thus a candidate concept can refer to one or many existing concepts, and two or more candidate concepts can refer to one existing concept.

5. Conclusions and future work

In this paper we presented an approach of ontological enrichment. The approach leads to grouping new concepts identified in a large text corpus to concepts known in an ontology Ω. The way to achieve this goal is influenced by the given ontology Ω three times:

(i) we derived a text corpus C_{Ω}, which is particularly characteristic for Ω.

(ii) we created rules to determine, to which extend the usage of a concept in C_{Ω} was similar to the usage of another concept in C_{Ω}. The rule set ρ consisted of rules implied by the ontology Ω.

This step was possible because the approach has generic features principally allowing the use of many different rule sets.

(iii) during an optimization process we grouped candidate concepts to concepts from Ω. The optimization process was due to attaching a weight to each rule. The configuration of these weights was transformed into an optimization problem with semantic distances also derived from Ω.

By (i) to (iii) we extended an approach originally stemming from semi-automatic ontology engineering and conceptual clustering [BiNeCa].

The work presented in this paper should be continued with experiments. An evaluation of the experimental results from our point of view should investigate the following questions:

- do different distance measures fulfilling the requirements 3.2 (i)-(iv) yield very different enrichment results?
- what is a sensible choice for the threshold *T*? To answer this question, we must investigate, how many concepts are proposed as candidate concepts with different thresholds *T*.
- does the quality of the candidate concepts satisfy the needs of an ontology engineer?

In our opinion the latter question is crucial. Roughly speaking, there are two ways of finding answers to this evaluation question. We could take parts of existing large ontologies and see, if the parts left out evolve through our enrichment process. On the other hand, we could ask a knowledge engineer to build a (small) ontology and comment the candidate concepts found afterwards. This evaluation should include answering the questions, which concepts did the engineer expected to be added, if they did appear in the enrichment results, if the enrichment results contained unsuspected high-quality concepts and on the contrary bad candidates.

Also future work should experiment either with more complex concept extraction from the texts (like attributed nouns) or extending the candidates due to high similarity to another candidate concept.

From a more abstract point of view, future work also includes applications of the approach to ontology merging and mapping, i.e. to the question, what happens if we take more than one ontology Ω as basic input. Finally, the question arises if the approach presented here benefits automatic ontology generation, i.e. if there are also optimization problems to be stated and solved in this research area.

References

[BiNeCa] Bisson, G. and Nedellec, C. and L. Cañamero: “Designing clustering methods for ontology building - The Mo'K workbench”, *Proceedings of the Ontology Learning ECAI-2000 Workshop,* August 2000

[ChGr] G. de Chalendar and B. Grau: “SVETLAN' or how to Classify Words using their Context”, *proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2000,* Juan-les-Pins, France, October 2000

[Gr] T.R. Gruber: “Towards principles for the design of ontologies used for knowledge sharing”, International Journal of Human-Computer Studies, **43**, 1995

[Gu] N. Guarino: “Understanding, Building and Using Ontologies. A Commentary to "Using Explicit Ontologies in KBS Development"“, *International Journal of Human and Computer Studies*, **46**, 1997

[Ho] I. Horrocks: “Using an expressive description logic: FaCT or fiction?” , *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth International Conference (KR'98),* San Francisco, California, June 1998.

[IDS] Institut für Deutsche Sprache: *www.ids-mannheim.de*

[Leng] K. Lengnink: “Formalisierung vom Ähnlichkeit aus Sicht der formalen Begriffsanalyse”, *PhD Thesis, Technical University of Darmstadt, Shaker Verlag*, 1996

[OG] Open GALEN: *www.opengalen.org*

[StM] Steffen Staab, Alexander Mädche: Axioms are objects, too - Ontology Engineering Beyond the Modeling of Concepts and Relations *Proceedings of the ECAI 2000 Workshop on Ontologies and Problem-Solving Methods* Berlin, 2000

[Tv] A. Tversky: “Features of similarity”, *Psychological Review* **84**, 1977