[FSSt02] *Andreas Faatz, Cornelia Seeberg, Ralf Steinmetz;* **Ontology Enrichment with Texts from the WWW;** 13th European Conference on Machine Learning (ECML'02), Helsinki, August 2002, S. 20-34.

Ontology Enrichment with Texts from the WWW

Andreas Faatz, Ralf Steinmetz

KOM – Multimedia Communications Lab
Darmstadt University of Technology, Merckstr. 25, 64283 Darmstadt, Germany
{afaatz, rst}@kom.tu-darmstadt.de

Abstract. The following paper explains, how we can enrich an existing ontology by mining the WWW. The use of such an ontology may be manifold, for example as a component of information systems or multimedial repositories. The enrichment process is based on the comparison between statistical information of word usage in a large text collection, a so called text corpus, and the structure of the ontology itself. The text corpus will be constructed by using the vocabulary from the ontology and querying the WWW via Google.
We define similarity measures by optimising their parametrisation and examine the central properties of the enrichment approach - along with the presentation and evaluation of experimental results. Parametrisation of a similarity measure means assigning weights to each word collocation feature we first check in the text corpus and thereafter integrate into the representation of a word or a concept.

1 Introduction

Automatic thesaurus and ontology construction dates back from the last three decades [1]. Our approach is a further development of methods to construct the whole ontology automatically. In contrast to these approaches our algorithm can only be applied, if we enrich an existing ontology instead of fully constructing the ontology.
The following paper focuses on requirements for the semi-automatic enrichment of medical ontologies based on the statistical information of word usage. An ontology is a structured network of concepts from an knowledge domain and interconnects the concepts by semantic relations and inheritance. [2] gives a precise technical definition of an ontology, that we will refer to throughout this paper:

Definition 1: An *ontology* is a 4-tuple $\Omega := (C, \text{is\_a}, R, \sigma)$, where C is a set we call *concepts*, *is_a* is a partial order relation on C, R is a set of relation names and $\sigma : R \rightarrow \wp(CXC)$ is a function [2].
Throughout this paper we assume that a concept has a character string as a descriptor. This character string may be a word or a phrase.

For our purposes we will neglect R as well as σ and focus on *is_a* as the particular relation, which is responsible for superconcept-subconcept dependencies. For example *bacteria* is a superconcept of the concept *pathogenic bacteria*. Whenever we talk of 'relations' or 'relational paths' in the following sections, we refer to the *is_a* relation. We also define

Definition 2: We call the restriction of an ontology $\Omega := (C, \text{is\_a}, R, \sigma)$ to $(C, \text{is\_a})$ the *hierarchical backbone* of Ω.

Ontologies give a formal representation and conceptualisation of a knowledge domain, which is useful for the administration of large multimedial resource collections: if ontologies reflect an agreement of a group of experts and are rich enough in the sense of a sufficient number of concepts, ontologies are able to handle information exchange across the borders of one expert's vocabulary. For example, one could ask an ontology to return the names of all bacteria causing diarrhoea and in this way access domain knowledge without the barrier of finding out the names of the particular bacteria by reading texts from the domain of infectiology.
Naturally the construction of an ontology is hard and expensive, as one has to train domain experts in formal knowledge representation. This is the motivation behind the idea, that for a given ontology we focus on finding new concepts automatically. Those new concepts are propositions, which extend the given ontology. For this we

use a special text corpus derived from WWW search results
detect a set of candidate concepts from the corpus
finally select a subset of those candidate concepts ranking their similarity to concepts already existing in the given ontology.

The final selection ends up in new concepts for the ontology to be proposed to a (human) ontology engineer.
The concepts have one or more descriptors, which are words or phrases from natural language. This implies that we develop our method finding suitable definitions for the semantic similarity of words or ordered sets of words.
The paper is organised as follows: in section 2.1 we describe our approach informally, whereas in 2.2 we give a survey on formalisation, which is explained in detail in section 2.3. Section 3 deals with the experimental results on two very different kinds of text corpora and especially on mining propositions from Google search hits. Section 4 discusses related work. At last section 5 summarises and points to future work.

2 Enrichment Approach

2.1 Overview

Similarity between words is a topic from the theory of word clustering algorithms and requires statistical information about the context, in which the words are used. Many approaches check collocation features of the words in large text corpora, such that a word is represented by a large vector. The vector has entries communicating, how often a collocation feature was fulfilled in the corpus. The vectors are sparse [3]. The notion of similarity definitions by vector representations normally does not assign a weight to every single dimension of the vectors. In this paper we argue, that this is possible by a soft method using the information already defined in the given ontology. The influence of the ontological structure on the word (-vector) similarities results in an optimisation

problem, which determines which dimension in the word (-vector) representation is influential for the similarity computation. The following definition should clarify, what a collocator is.

Definition 3: Let a word w be given. A *collocator* of w is a word, which occurs together with w due to a predefined rule in a text collection (text corpus).

Thus for example in the phrase '*Medical ontology enrichment in the k-med project*', '*enrichment*' and '*medical*' would be collocators for predefined rules like for instance 'maximal distance 5 words' or also 'occurrence in the same sentence'.
The way we include the ontological information from $\Omega := (C, \text{is\_a}, R, \sigma)$ may be guided by different heuristics on a numerical interpretation of is_a, R and σ. For example the abstraction level of a concept, the interconnection by relations, relational paths and their lengths or the local granularity of the modelling can establish distance measures on a given ontology Ω.
Our goal is a comparison of this distance measures to the information about collocators in a text corpus.

2.2 Enrichment as an optimisation problem

The core idea of our approach is computing enrichment rules, which do not contradict the distance information already given by the ontology we want to enrich.
We first have to state a basic assumption.

Assumption 1: There exists a consistent distance measure d expressing semantic distances between the concepts in Ω. The distance measure is based on the relational interconnections between the concepts in the hierarchical backbone of Ω.

By consistency we mean, that d underlies some characteristic heuristics: a long relational path between two concepts rises the distance between them, the abstraction level in the hierarchical backbone influences the distance measure d as well as the number of concepts, which are subconcepts to the same superconcept. Both abstraction and the number of siblings rise the distance. The distance measure d, which we will from now on denote by $d(x,y)$ for concepts x and y from Ω also differentiates between generalisation and specialisation in an ontology. We showed in [4] that such distance measures exist indeed. Thereby our notion of 'consistent' does not necessarily imply a good enrichment quality, but just means, that the above heuristics are fulfilled. The quality of the heuristics and the resulting distance measures have to be judged by the results of enrichment experiments.
We also assume a text corpus ζ to be given and determine the ordered set K of the n most frequent collocators which cooccur with at least two concepts from Ω. Let $v(x) \in R^n$ be a vector; the i-th entry of this vector $v(x)$ expresses, how often the descriptor of the concept $x \in \Omega$ was a collocator in ζ with the i-th element of K.

Let now f_k be a component-wise monotonic function of the dissimilarity D between two concepts x and y from Ω. For the dissimilarity D we postulate, that it must be computable from the vectors $v(x)$ with concepts x from Ω.
The parametrisation $k = (k_1, \ , \ , k_n)$ just weighs each collocation feature positively, it indicates, how strong the j-th dimension is involved in the dissimilarity computation. The optimisation process consequently fits the average of the $f_k\,(D\,(v(x), v(y)))$ for possible $k = (k_1, \ , \ , k_n)$ with each $k_i \geq 0$, to the distances $d(x,y)$ for each pair of concepts from Ω.
To sum it up briefly, the optimisation establishes a dissimilarity measure, which is as near as possible to the distance measure d in Ω. In the next section we present a formalisation of the algorithm. We decided to explain the details of this formalisation to make a repetition of experiments with the algorithm possible.

2.3 Formalisation of the algorithm

A *distance measure* on Ω is a function d: $(C \times C) \rightarrow [0,1]$. Examples of distance measures are:

1) $d(x,y) = e^s$, where e is Euler's constant and s denotes the number of steps along the shortest relational path between the concepts x and y. This distance definition corresponds to the heuristics, that long relational paths rise the distance between given concepts.
2) $d(x,y)=1$, if there exists a relation between the concepts x and y and $d(x,y)=0$, if there does not exist a relation between the concepts x and y. This definition has only a reasonable application to ontologies, if the transitivity of the is_a-Relation and the concatenation of different relations from R is clearly stated in the a of axioms of Ω.
3) [5] defined criteria for similarity measures in thesauri, which in turn can be applied to distances in the hierarchical backbone of the ontology $\Omega := (C, \text{is\_a}, R, \sigma)$.
[4] showed, that there is an infinite number of distance measures on the hierarchical backbone of an ontology fulfilling more restrictive characteristics than 1) and 2). For further details we have to refer to [4].

A *text corpus* ζ is a collection of text documents written in exactly one natural language. We assume ζ to be electronically available. From a text corpus we define a set of words or phrases to be the candidate concepts. A *proposition* for the ontological enrichment is a word or a phrase from ζ, which is used similarly to the concepts from the given ontology. Candidates are to be predefined, for example as all nouns occurring in ζ. Note that ζ might also be extended by additional text material. This may happen during or after the application of the enrichment algorithm.
A *rule set* ρ is a finite set of linguistic properties, each of which can be tested in terms of its fulfilment frequency in the text corpus. In our case we will always consider collocation properties for the rule set ρ.
The entries m_{ij} of a *representation matrix* $M(C, \rho, \zeta)$ list, how often the j-th property from ρ was fulfilled in ζ for the descriptor the i-th concept from C.
The enrichment algorithm processes information available from ζ and Ω. It computes an optimal solution for the problem of fitting the distance information among the concepts expressed by Ω and the dissimilarity information between words or phrases to be extracted from the word usage statistics considering ζ.

Let us assume a given $M(C, \rho, \zeta)$. We search for a set $k = \{k_1, \ldots, k_n\}$ of non-negative reals with $|k| = |\rho|$, which will be called *configuration* of the rule set ρ. Each k_i corresponds to a rule ρ_i.
The configuration k decides about the quantities of dissimilarity we derive from $M(C, \rho, \zeta)$.
The *Kullback-Leibler divergence* generally measures the dissimilarity between two probability mass functions [6] and was applied successfully to statistical language modelling and prediction problems [7]. The Kullback-Leibler $D(x,y)$ divergence for two words x, y is defined as

$$D(x, y) = \sum_w P(w|x) \log \frac{P(w|x)}{P(w|y)} \quad (1)$$

In the basic version of the Kullback-Leibler divergence, which is expressed by formula (1), w is a linguistic property and $P(w|x)$ is the probability of this property being fulfilled for the word x. In the sum indicated by formula (1), w ranges over all linguistic properties one includes in a corpus analysis. In our case the frequencies of observing the collocation properties are denoted by $M(C, \rho, \zeta)$. We change (1) in such a way, that k weighs the influence of each property w

$$D_k(x, y) = \sum_w k(w) P(w|x) \log \frac{P(w|x)}{P(w|y)} \quad (2)$$

with $k(w) \in k$ in our case
Considering our representation matrix notation $M(C, \rho, \zeta)$ we obtain

$$P(w|x_i) = \sum_{l=1}^{|\rho|} \left[\frac{m_{il}}{\sum_{n=1}^{|\rho|} m_{in}} \right] \quad (3)$$

Let us clarify the notation of formula (3):
x_i denotes the i-th concept from C. Correspondingly in (3) the m_{il} are the matrix entries in $M(C, \rho, \zeta)$ in the row expressing the collocation properties of x_i. With this notation $k(x_i) = k_i$ holds. In that sense, we will be able to determine an optimal $k = \{k_1, \ldots, k_n\}$.
Taking the distances from the ontology Ω as an input, which should be approximated by the $D_k(x, y)$ as well as possible, the question of finding an optimal configuration k reduces to the question:
which configuration k minimises the average squared error expressed by the differences

$$(d(x, y) - D_k(x, y))^2 \; ?$$

Finally we present a formulation of this question in terms of a quadratic optimisation formula. Searching for an optimal k means searching for a minimum of the following fitness expression:

$$\min_k \sum_{i=1}^{|C|} \sum_{j=1}^{|C|} (d(x_i, x_j) - D_k(x_i, x_j))^2 \quad (4)$$

where $k = \{k_1, \ldots, k_n\}$ and $k_l \geq 0$ for all $k_l \in k$. Note that we minimise over the set of all configurations, that means over all possible k. We now explain, which words phrases are propositions for the ontological enrichment.
Once we optimised formula (4) we obtain the configuration in need to compute all the distance measures between all the concepts from Ω and the candidates. We apply an enrichment step starting with the optimal similarity measures $D_k(x, y)$.
We only take into concern the $D_k(x, y)$ with $x \in C$ and a candidate y. If such a distance between a formerly known concept (i.e. its descriptor) and a candidate (i.e. a word from the corpus) formerly unknown to Ω is lower than a predefined threshold, y proposition to enrich Ω. A suitable threshold can for example be defined from the average of the distances $d(x,y)$ where $x \sim y$ holds for some $\sim \in R$.
Additionally the $D_k(x, y)$ with $x \in C$ and a candidate y carry even more information, namely an optimal *placement* of the candidate concepts. The candidate concepts and the concepts from Ω can be presented together, if a candidate turns out to be a proposition. This simplifies the knowledge engineer's understanding of how the candidate concepts evolved and in which semantic area of Ω they might belong.

3 Experimental results

3.1 Basic input: ontology and corpus

The ontology Ω we enriched in our experiments is a modelling carried out by a medical expert during the first phase of the k-med project. K-med is an abbreviation for 'knowledge based multimedia medical education'. This project tries to collect multimedial medical learning resources and for the sake of reuse describe the educational resources by applying a metadata scheme and a common medical ontology [8].
The ontology contains the most abstract concept *disease OR symptom* along with the subconcepts *measles*, *German measles*, *diarrhoea*, *intestinal infection*. *Diarrhoea* itself has the subconcepts *aqueous diarrhoea* and *sanguinary diarrhoea*. The ontology may be viewed as an incomplete test ontology for several reasons: it only uses hierarchical relations, the superconcept and subconcept relations and also the knowledge domain are not fully clarified (such that a construction like *'disease OR symptom'* with a logical *OR* becomes necessary) and at least one additional abstraction level (a concept like '*infections*') could make the modelling clearer. In fact, this ontology is just a part of a larger ontology under construction. It is based on the subjects, which have to be tought during the first semesters of medical education in Germany.
To sum it up, we enriched this intuitively modelled ontology in the sense of [9], without deductive or inductive rules or axioms. From our point of view this enrichment of incomplete ontologies is, along with the extension of thesauri and catalogues, the main

application area of our approach. Furthermore following [9], such a rather informal ontology represents a situation, where machine learning techniques should support the knowledge engineer.

The size of the ontology was kept small for the sake of a rapid application and evaluation. For the experiments presented in the remainder of this paper, it was easier to reduce the possible interdependencies by referring to this ontology chunk containing seven concepts.

As larger ontologies can be segmented to smaller ones - for instance to speed up the computation - we consider the enrichment of the ontological chunk a good starting point for experiments.

The computation of the represenation matrix and the derivation of the optimisation problem was carried out by an implementation of our own. For the sake of a possible later connection of this component to other existing ontology tools and the linguistic workbench TATOE [10] we used Smalltalk as the language to implement the algorithm. The quadratic optimisation (4) itself was carried out by the ampl-solver [11].

For our first experiments, we used a general, but very large (about 28.700.000 sentences) newspaper corpus available at [12]. The corpus ζ can be queried by on-line query tools, which also provide stemming and lemmatisation techniques. All queries are collocation queries determining, whether two words were used in ζ at a distance of predefined size. Although these experiments produced bad enrichment results from the medical expert's point of view, very important meta-properties of the algorithm, as compression of the rule set and a stability of the algorithm were found.

A second experiment was based on the search results of the web search engine Google [13]. We passed each descriptor of a concept to Google [13] and converted the documents belonging to the 10 search hits with the highest ranking into a text corpus. In contrast to the newspaper corpus this corpus is more specialised, consisting of 70 documents with 135.166 words and 15.570 sentences. Because of a restriction of the concordancer freeware in use (Wconcord, Darmstadt University of Technology) we did not apply stemming and lemmatisation to the specialised corpus we gained from the WWW. In our Smalltalk implementation we included a stop list consisting of auxiliary verbs, conjunctions, personal pronomina and prepositions.

For the rule set ρ we always tested, how often a collocation at maximum distance five tokens, but in the same sentence took place.

All of our enrichment results can be found in table 1. In the column at the very left the reader will find the concept from Ω, in the middle column the concepts from the general newspaper corpus proposed to the concept from Ω. At the right we find the propositions to Ω from the special corpus.

In both experiments a candidate became a proposition, if we computed a dissimilarity below 0.5. The reason for the choice of this threshold was, that a path of length 2 in the hierarchical backbone of the test ontology led to a distance average of 0.5. All experiments were carried out in German, so we show translations here.

Table 1: enrichment results

concept from Ω	general corpus (28.700.000 sentences)	special (www) corpus (15.500 sentences)
disease OR symptom	*loss* *illness* *infections body* *leg* *wound* *animals* *vaccination* *fever* *combat*	
intestinal infection		*medical doctor* *cause*
diarrhoea	*ailment* *epidemic* *cough* *fever* *vaccination* *infections*	*vomit* *stomach ache* *nausea* *fever* *medical doctor*
measles		
German measles		
sanguinary diarrhoea		*vomit* *stomach ache* *nausea* *fever* *can*
aqueous diarrhoea		*vomit* *stomach ache* *nausea* *fever* *can*

With '*stomach ache*' and '*medical doctor*' in German we did not propose word groups but composita ('*Bauchschmerzen*' and '*Arzt*' in German).

3.2 Results for general corpora

We refer to the enrichment results depicted in table 1. As the rule set ρ we used collocation in the same sentence at maximal distance of five tokens in ζ.
Although we find some concepts like '*infection*' which is a missing abstraction level in the ontology, the enrichment results from general corpora are poor. We retrieve overly general propositions like 'body' and even flaws like the proposition of 'wound', 'animal' or 'leg'. These flaws occured especially with candidates, which only had one property from ρ.
Another problem occurs with special concepts - as expected, even a very large newspaper corpus does not contain enough information to get propositions for the subconcepts of diarrhoea.
Although we faced these problems, the experiments for general corpora were worthwhile, because we identified two considerable meta-properties of the approach: *stability* and *compression*.

a) *Stability*

Before dealing with the core of the enrichment - the proposal of concepts - we tested the inner stability of the approach and pruned $M(C, \rho, \zeta)$, which was a square matrix of size 102, in two different ways. If $M(C, \rho, \zeta)$ contained not enough information, then different pruning strategies would bare the danger of destroying the enrichment process, which means leading to inconsistent optimal configurations or enrichment results.

Table 2: stability

first strrategy	second strategy
suffer	*suffered*
diseases	*fever*
illness	*measles*
	hepatitis
die	*died*
pregnancy	*pregnancy*
percent	*percent*
stomach	

The first pruning strategy was keeping only the ten largest entries per row, resulting in a 102 X 34 matrix $M(C, \rho_{first}, \zeta)$. The second pruning strategy only kept the entries $c_{ij} \in M(C, \rho_{second}, \zeta)$ with $c_{ij}>10$, resulting in a 102 X 63 matrix. With both matrices we set up the optimization procedure, solved expression (4) respectively and derived two optimal configurations k_{first} and k_{second}. The collocators belonging to rules with nonzero weights are listed in table 2. Both strategies mean collocation at maximal distance five words with a descriptor from the respective column of table 2.
Let us comment the collocators remaining from the two pruning strategies and the optimization. In table 2 we listed the collocators in such a way, that we immedialy see the relation between the two sets. Some of the collocators do not differ at all, some of them are only different in terms of the grammatical context they stem from (for example '*suffer*' and '*suffered*'), some of them obviously carry a semantic relation ('*diseases*' and '*illness*' on the one hand and their more specialised pendants '*feaver*', '*measles*' and '*hepatitis*' on the other hand). The only collocator without any direct relative in the other set is '*stomach*', so we state, that the analysis of the resulting collocator sets of nonzero weight does not show any inner contradiction in the approach, the representation matrix in our case seems to be stable and even carrying redundant information.

b) *Compression of* ρ

With both pruning strategies only a few properties ρ_i from ρ achived a corresponding weight k_i from k, with $k_i > 0$. This compression of the rule set also occured for different definitions of the ontological distance d. We assume, that this compression is closely related to the sparsity of our representation matrix and to the structure of the optimisation formula (4). The reason for this assumption are further experiments with artificially and randomly generated matrices, which we used as pseudo-representation matrices with sparsity structures similar to the ones of $M(C, \rho_{first}, \zeta)$ and $M(C, \rho_{second}, \zeta)$. As these experiments ended in a similar compression, we will search for a proper mathematical reason why this reduction of influential features with nonzero weight takes place.

3.3 Results for a special corpus based on Google hits

As we mentioned, we passed each descriptor of a concept to a web search engine and converted the first 10 hits of the Google search result into text files, removing the HTML-specific tagging. The results can be found in table 1.
As the rule set ρ we used again collocation at maximal distance of five words in ζ. For pruning reasons from our representation matrix we kept only properties from ρ, which were fulfilled for at least two concepts from the small test ontology. This resulted in a representation matrix with 292 columns. This means, that - for the special corpus - we initially found significantly more rules than with the common corpus, but after the solution of (4) we obtained 12 properties from ρ with a nonzero weight. These were distance five in the same sentence with *chronic, infection, because of, seldom, diarrhoea, vaccinate, pneunomia, virus*.
The choice of the candidates for the enrichment was driven by the observation from the previous experiments: propositons with only one nonzero property induced flaws to the enrichment. Consequently we accepted candidates, which at least fulfilled two of the 12 remaining properties from ρ.

The main results of the experiment with the special corpus crafted from web search hits are

- enrichment of the special concepts *sanguinary diarrhoea* and *aqueous diarrhoea*
- identification of a group of symptoms (*vomit, stomach ache, nausea, fever)* as propositions
- lack of propositions for *measles* and *German measles*

The flaws in this enrichment are *can* which should actually be a member of the stoplist and *medical doctor* which is at least too overgeneral.

3.4 Discussion of the results

As a second general observation we state, that the main flaws identified during the evaluation in this section come from candidates, which share only one feature with a concept from Ω. We conclude, that a possible way to handle this may be an additional tuning of the definition of d. A potential technique for this is generating artificial usage profiles, which do not at all reflect real words, and searching for a measure, which properly discriminates between real words and randomly constructed feature sets.
There exist subjective and objective ways of evaluating the results. Roughly speaking the objective evaluation methods base on reference ontologies, the subjective evaluations are based on expert interviews [14].
The question posed in the subjective evaluation was:
Consider the ontology Ω and the table of propositions from table 1 to be given. Which strategy performs better? Which aspects of the enrichment results are positive, which ones are negative?
The subjective evaluation clearly showed, that the enrichment with the special corpus performs better, as there are less flaws like *wounds* or *leg* and also less overgeneralisations like *illness*. In addition to this, the results of the specialised enrichment are easier to perceive, as there are not too many propositions and the good propositions were more precise.
The fact that our specialised enrichment with a web-based corpus performs better is not as trivial as it seems. For instance, the symptom group *vomit, stomach ache, nausea, fever* was also on the candidate list for the general corpus experiments - but the collocation information was insufficient, although the corpus was very large.
Comparing the experiments, possible causes for lacking propositions for *measles* and *German measles* may be found in the structure of out test ontology. It contains more information about diarrhetic diseases. At least the special text corpus must be balanced, a preprocessing step we will include in future experiments.
Other good candidates (like '*virus*') did not become propositions in the special corpus experiment, as we did not use stemming and lemmatisation. This introduces all inflections of verbs to the candidate set. Instead of this, the inflections should be unified to one candidate.
The main goal of a series of evaluations should be finding a correlation between subjective and objective measures. If such a correlation exists, even imprecise objective evaluation measures can give answers to the quality of parameter tuning or candidate strategies. The reason for this is, that the objective evaluation measures have to correlate ordinally but not cardinally with the subjective ones.
Objective evaluation measures we are developing are guided by the notion of precision in document retrieval [1]. Naturally we need a larger series of experiments to detect a possible correlation between objective and subjective evaluation measures.

4 Related work

Two main branches of automated ontology construction by natural language processing in general and checking collocators in our special case may be identified: those, which base the similarity of concepts or their descriptors on syntactic criteria or collocation directly (we will refer to those as first type) and those, which take statistic samples of the features of a concept or its descriptor. For example the first type declares concepts or their descriptors as similar, when they often occur together in one sentence. The works of [15]and [16] are examples of this type of enrichment.
The second type declares words as similar, if they are used in an similar context. For example, if a word w is used with a word v in the same sentence very often, and also a word u is used with a word v in the same sentence very often, then w and u would be similar according to the assumptions of the second type approaches, even if w and u never appear in the same sentence all over the text corpus. Note the significant difference between the approaches: the first type would state a high similarity between w and v, also between u and v. Representatives of the second type are the works of [17], [18], [19] and also [20]. The latter ones use syntactic parsing instead of pure collocation information. [20] moreover does not assume identity between concepts and their descriptors as we did in this paper.
If we try to group our approach among the first or second type approaches, we can clearly state that the way we define word similarities and dissimilarities is due to the second type as we pointed out in the previous section. Nevertheless the way we restrict our view to certain candidate concepts and their respective features also refers to the first type: a candidate concept is a concept, of which we determine similarities or dissimilarities to all ontological concepts. As for almost natural computational restrictions we are not able to compute the dissimilarities for all the words from a corpus, we must restrict our observations. Our candidate strategies in section 3 explain possible restrictions in detail. They are influenced by ideas of the first type.
Although - except [21] - none of the related works mentioned here directly focuses on ontology enrichment from the WWW, all these works have in common, that their automated construction features can be extended to our enrichment goals of proposition identification and placement.
[15] used pure collocation information for gaining new concepts, but also focuses on qualitative issues of the collocations to derive statements about relations and the behaviour of the relations.
[16] focused on the so-called salient words, which are able to disambiguate word meaning very well. In a way [16] is also a extending special form of ontology, namely a thesaurus. In contrast to our approach his focus is on disambiguation, which was further

developed by [21], but with web resources. In our approach a concept can be proposed to two different concepts, but may also disambiguate, depending on our ontological distance measures. Especially the identification of the symptom group *vomit, stomach ache, nausea, fever* propositions for several concepts from Ω would be simply impossible with the approaches of [16] or [21], as they found on discriminating descriptors. They achieve this by a test which detects the mostly diverging contexts of concepts of a given ontology (or thesaurus).
[19] designed the Mo'K workbench for word clustering and building hierarchies from the clusters in a second step. We also implemented a feature based environment, and our goal is even more specialised than word clustering. However we share the opinion with [19], that collecting many features and assigning weights to the features is an excellent basis for similarity definitions. Our implementation is in Smalltalk, whereas Mo'K uses C. Smalltalk remains a possible language, as we end up in very condensed representations with a few features of nonzero weight (compare section 3.2). Generally spoken, in contrast to our work (which systematically computes an optimal configuration k for a rule set ρ) [19] do not explicitly offer a strategy for choosing the weights.
An approach related to [19], but more founded on collocation networks and determining artificially specialised corpora can be found in [18] and a comparison of the performance of specialised and common corpora in word clustering can be found in [Asium].
Finally [17] experimented with word similarities expressed by an unsupervised neural network algorithm, the Kohonen map [23], but also for clustering, not for enrichment goals. Our evaluation methods result from the enrichment goal and could also be a basis for an evaluation of the Kohonen maps in word clustering, a task desired by [22].

5 Conclusion and future work

We presented a method for ontology enrichment and applied it to a medical ontology chunk. Our evaluation shows, that the strategy to derive propositions from a special corpus seems to end up in clearer and more error prone conceptual propositions to a domain expert. The experiments have to be repeated with other specialised corpora from the web, the major task for future work. Instead of Google one could refer to the hits of a medical search engine like Medivista.
From all our experiments we identified very important meta-properties of the algorithm. These are a possible basis for future extension of the algorithm: a more systematic treatment of the initial rule set ρ by gradually extending the word distances can be achieved, if we are able to keep the compression property of the algorithm.
From our point of view the integration of the presented algorithm in a Delphi method [9] for k-med-like projects or the evolving semantic web is very promising. In the context of the task of the k-med ontology, automatically identifying a whole group of symptoms is especially helpful for managing documents for case-based medical education.
A number of other interesting questions comes along with the approach. Among them are the following ones: how do we construct and balance a suitable corpus to learn from, which linguistic preprocessing is necessary or helpful, how does the approach scale for larger ontologies. The latter question is again closely related to our observations: the optimisation problems end up in an identification of a few relevant collocation features and the representation matrix can stand a pruning preprocessing.
Also the question of evaluating the results is interesting for related areas such as Kohonen maps for documents and word clustering algorithms [17].

References

1. Spark-Jones K.: *Readings in Information Retrieval*, Morgan Kaufmann, 1997
2. Stumme, G., Mädche A.: *FCA-Merge: A Bottom-Up Approach for Merging Ontologies*, Proceedings of the 17th International Joint Conference on Artificial Intelligence, Seattle, USA, August, 1-6, 2001, San Francisco/CA: Morgan Kaufmann, 2001
3. Sahlgren,M.: *Vector-Based Semantic Analysis: Representing Word Meanings Based on Random Labels*, Proceedings of the ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation, Helsinki, Finland, 2001
4. Faatz, A., Hoermann, S., Seeberg, C., Steinmetz, R.: *Conceptual Enrichment of Ontologies by means of a generic and configurable approach*, workshop notes of the ESSLLI 2001-workshop on semantic knowledge acquisition and categorisation, Helsinki, Finland, August 2001
5. Resnik, P.: *Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language*, Journal of Artificial Intelligence Research, **11**, 1999
6. Kullback, S. and Leibler, R.A.: *On Information and Sufficiency*, Annals of MAthematical Statistics **22**, 1951
7. Dagan I., Perreira F., Lee L.: *Similarity-based Estimation of Word Co-occurence Probabilities*, Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, ACL'94, New Mexico State University, June 1994
8. Stefan Hörmann and Ralf Steinmetz: *Creating courses with learning object metadata*, Multimedia Systems, Springer, Berlin/Heidelberg, to appear 2002
9. Clyde W. Holsapple and K.D. Joshi: *A collaborative approach to ontology design*, Communications of the ACM, Vol. **45**, 2, February 2002
10. Alexa, M.: *Text type analysis with TATOE*. Storrer, A. & B. Harriehausen (eds.) (1998): Hypermedia für Lexikon und Grammatik. Gunter Narr Verlag, Tübingen.
[Asium]: D. Faure and C. Nedellec: *ASIUM: Learning subcategorization frames and restrictions of selection*, in Y. Kodratoff, editor, 10th Conference on Machine Learning (ECML 98) -- Workshop on Text Mining, Chemnitz, Germany, April 1998
11. ampl optimisation software, *http://www.ampl.com*
12. Institut für deutsche Sprache, *www.ids-mannheim.de*
13. the Google search engine, *www.google.de*
14. Andreas Faatz: *Enrichment Evaluation*, technical report TR-AF-01-02 at Darmstadt University of Technology
15. Heyer, G.; Läuter, M.;Quasthoff, U.; Wittig, Th.; Wolff, Chr.: *Learning Relations using Collocations*, Proceedings of the IJCAI Workshop on Ontology Learning, Seattle, USA, August 2001

16. David Yarowsky: *Word-Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora*, Proceedings of COLING-92, Nantes, France, 1992
17. Lagus, K.: *Studying similarities in term usage with self-organizing maps.* Proceedings of NordTerm 2001, 13-16 June, Tuusula, Finland. pp. 34-45. ,2001
18. Gaël de Chalendar and Brigitte 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, pages 203-216 Rose Dieng and Olivier Corby (Eds.), Springer, 2000,
19. Bisson, G. and Nédellec, C. and Cañamero L.: *Designing clustering methods for ontology building - The Mo'K workbench*, in Staab, S. and Maedche, A. and Nedellec C., editors, Ontology Learning ECAI-2000 Workshop, Berlin, August 2000
20. Hahn author Hahn, U. and Schnattinger}, *Ontology Engineering via Text Understanding*, IFIP'98 --- Proceedings of the 15th World Computer Congress, Vienna/Budapest, 1998
21. Eneko Aguirre, Mikel Lersundi: *Extracción de relaciones léxico-semánticas a partir de palabras derivadas usando patrones de definición*. Procesamiento del Lenguaje Natural 27: 165-172 (2001)
22. Lagus, K.: *Text Mining with the WEBSOM*. Acta Polytechnica Scandinavica, Mathematics and Computing Series no. **110**, Espoo, Finland 2000
23. Teuvo Kohonen, *Self Organising Maps*, 3rd Edition, Springer Series in Information Sciences, Vol. **30**, Springer, Berlin