eingereicht bei 'Soft Methods in Probability and Statistics' SMPS - 2002

Statistical profiles of words for ontology enrichment

Andreas Faatz, Cornelia Seeberg, Ralf Steinmetz

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

1. Introduction

The following paper focuses on the semi-automatic enrichment of ontologies based on statistical information.
An ontology is a structured network of concepts from an knowledge domain and interconnects the concepts by semantic relations and inheritance [Stumme, Mädche]. Ontologies give a formal representation and conceptualisation of a knowledge domain. For a given ontology we find propositions automatically, which could extend the ontology by new concepts. This means we

- use a text corpus
- detect a set of candidate concepts from the corpus
- finally select a subset of those candidate concepts by ranking their similarity to concepts already
- existing in the given ontology.

The final selection ends up in possible new concepts for the ontology to be proposed to a (human) ontology engineer.
The concepts in the ontology have one or more descriptors, which are words or phrases from natural language. On the other hand, the extractable information from large text corpora are words or phrases. For our technique this means that we develop a method of finding suitable definitions for the semantic similarity and dissimilarity of words. Throughout the paper we treat the ontological concepts and their descriptors as the same objects.
The paper is organised in the following way:

- we give the definitions in need
- we formally explain the enrichment algorithm
- we focus on properties of the algorithm which extend and systematically treat the linguistic properties we take under consideration
- we point out related work from the area of word clustering
- finally we summarise our results and open research issues

2.1 Definitions

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 $\sigma : R \longrightarrow \wp(CXC)$ is a function [Stumme/Mädche].
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.

A *distance measure* on Ω is a function d: $(C \text{ X } C) \longrightarrow [0,1]$. 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
2) $d(x,y)=1$, if there exists a relational path between the concepts x and y and $d(x,y)=0$, if there does not exist a relational path between the concepts x and y.
3) [Resnik] defined criteria for distance measures in thesauri, which can be applied to the restriction of an ontology $\Omega := (C, \text{is\_a}, R, \sigma)$ to the pair $(C, \text{is\_a})$.
[Faatz] showed, that there is an infinite number of distance measures fulfilling more restrictive characteristics than 1) and 2).

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 occuring in ζ. Note that ζ might be extended 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.
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.

2.2 The basic optimisation for ontology enrichment

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 $k = \{k_1, \dot{} , k_n\}$ 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 [Ku] and was applied successfully to statistical language modelling and predicition problems [CoTo], [Da]. The Kullback-Leibler $D(x,y)$ divergence for two words x, y is defined as

$$D(x, y) = \sum_{w} P(w|x) \log \frac{P(w|x)}{P(w|y)} \qquad (1.1)$$

In the basic version of the Kullback-Leibler divergence, which is expressed by formula (1.1), w is a linguistic property and $P(w|x)$ ist the probability of this property being fulfilled for the word x. In the sum indicated by formula (1.1), w ranges over all liguistic properties one includes in a corpus analysis. In our case the frequencies of observing the linguistic properties are denoted by $M(C, \rho, \zeta)$. For our purposes we change (1.1) in such a way, that k weighs the influence of each property w

$$D_k(x, y) = \sum_{w} k(w) P(w|x) \log \frac{P(w|x)}{P(w|y)} \qquad (1.2)$$

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

$$P(w|x_i) = \sum_{l=1}^{|\rho|} \left[\frac{m_{il}}{\sum_{n=1}^{|\rho|} m_{in}} \right] \qquad (2)$$

Let us clarify the notation of formula (2):
x_i denotes the i-th concept from C. Correspondingly in (2) the m_{il} are the matrix entries in $M(C, \rho, \zeta)$ in the row expressing the linguistic properties of x_i. With this notation $k(x_i) = k_i$ holds. In that sense, we will be able to determine an optimal $k = \{k_1, \ , k_n\}$.
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

$$\min_{k} \sum_{i=1}^{|C|} \sum_{i=1}^{|C|} (d(x_i, x_j) - D_k(x_i, x_j))^2 \tag{3}$$

where $k = \{k_1, \ , \ , k_n\}$ 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 (3) 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. Discussion of the algorithm - extending the properties from ρ

The application of the algorithm needs a rule set ρ as one of the parameters given. The following section focuses on a technique which systematically selects and constructs ρ We will use the fact, that - for several distance measures d - our experiments showed, that the optimisation (3) ended up in supressing most of the rules from ρ setting the corresponding k_i to zero.

Applying the optimisation to derive a configuration k we chose a ρ in the following way: for each concept from C (Ω given) we collect the collocators occuring in the same sentence at a distance of at most five words in ζ, create a list of all these collocators and finally check for each $c \in C$, how often it was collocated in the same sentence, but at maximum distance of five words within ζ.

We choose this particular ρ, as for the German language, with which we carry out our experiments. The property is a standard configuration of the German online corpus analysis tools [COSMAS] and [Wortschatz].

We obtained two general observations, which characterise all of our ongoing enrichment experiments and which imply interesting further developments of the algorithm, because they point to a compression of the property set ρ while applying the algorithm. The vast majority of the $k_i \in k$ are zero (in our first experiments about 90% of the $k_i \in k$) and there are many minor influences of nonzero k_i, if we also consider the fact, that that similarity must exceed a threshold T for a candidate concept to become a proposition.

The fact that we observe many zeros in the solution k is related to the sparse structure of the optimisation problem (3). But even if we cannot predict the exact structure of $M(C, \rho, \zeta)$ beyond sparsity, the sparsity of the data belonging to candidate concepts additionally leads to properties with minor influence. Although we admit, that this observation needs a further strict systematic fundament, we use it as a working hypothesis.

Our first experiments also point to a fact, which we already expected intuitionally: if a concept becomes a proposition and its similarity was only determined by exactly one feature ρ_i from ρ (leaving $k_j = 0$ for $i \neq j$) we detected a higher risk of bad propositions in the sense of a semantic mismatch or an overgeneral proposition.

Another complication may arise, if we extend the initial corpus while applying the algorithm. Such a strategy is useful, if we start with a small specialised corpus and a few concepts in Ω. In that case the initial corpus may contain not enough information, consequently information must be added by including new texts in the initial corpus [25 von BiNeCa]. Only in that sense it is true, that specialised corpora perform well in domain dependent conceptual clustering or ontology enrichment problems like [BiNeCa] stated. But extending a corpus goes along with introducing a fulfilment of properties, which we did not observe in the initial corpus [Da], which gives a bias to our computation of important properties via the optimal configuration k.

For a systematic treatment of the difficulties we discussed in this section - arbitrary choice of ρ, bias with propositions guided by exactly one linguistic property, bias after extending corpus - we give a prospect on a stepwise application of the algorithm: after applying the algorithm once we make up a new property set $\tilde{\rho}$ keeping the properties, which turned out to be influential, and adding new properties.

An example for new properties is a larger context, in which a collocation may occur. As long our data remain sparse, several extensions of the linguistic property set can result in a rich representation making the selection of our the modified $\tilde{\rho}$ less arbitrary. In the special case of word-cooccurence in contexts the extension of ρ by a stepwise application of the enrichment algorithm becomes systematic, if we start with narrow contexts (as the distance of five words we used in the first experiments) and broaden the contexts monotonically in every step.

4. Related work

Similarity between words is a topic from the theory of word clustering algorithms and requires statistical information about the contexts, 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, which has entries communicating, how often a collocation feature was fulfilled in the corpus. The vectors are sparse [Sahlgren].

The notion of similarity definitions by vector representations normally does not weight every single dimension of the vectors. In the paper we stated 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 gives an answer to the question which dimension in the word (-vector) representation is influential for the similarity computation.

Automatic thesaurus and ontology construction dates back from the 1970s [Spark Jones]. Our approach is a further development of methods, which try to construct the whole ontology. The soft method of introducing heuristics for the ontological information given can only be applied, if we enrich an existing ontology instead of fully constructing the ontology. Besides the question about the heuristics guiding the optimisation procedures described above, a number of other interesting questions

arises along with the approach. Among them are the following: how do we construct a suitable corpus to learn from, which linguistic preprocessing is necessary or helpful, do we need absolute, relative or probabilistic vector entries, how does the approach scale for larger ontologies.
The question of evaluating the results is interesting for related areas such as Kohonen maps for documents and word clustering algorithms [Lagus].

5. Summary and future work

We presented an algorithm, which returns propositions for the enrichment of an ontology. The algorithm selects from a set of linguistic properties regarding the information encoded in the ontology, for wich we wish an enrichment: at this point, our soft method is based on a modified Kullback-Leibler divergence for each single given enrichment problem.
We also gave a prospect on a more systematic setup of the algorithm, which has to undergo genuine evaluation to overcome its merely constructive status. The area of evluation methods for the algorithm together with further experiments will be the focus of our future work on the subject.

References

[COSMAS]: the Cosmas corpus querying service, http://corpora.ids-mannheim.de/~cosmas/

[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

[RoJu]: Roland D. and Jurafsky D.: *How Verb Subcategorisation Frequencies Are Affected By Corpus Choice*, Proceedings of the International Conference on Computational Linguistics (COLING'98), 1998

[Da]: 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

[Ku]: Kullback, S.: *Information Theory and Statistics*. John Wiley and Sons, New York, 1959.

[Lagus]: Lagus, K.: *Studying similarities in term usage with self-organizing maps*. Proceedings of NordTerm 2001, 13-16 June, Tuusula, Finland. pp. 34-45. ,2001

[Resnik]: 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, vol. **11**, 1999

[Sahlgren]: 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

[Spark Jones]: Spark-Jones K.: *Readings in Information Retrieval,* Morgan Kaufmann, 1997

[Stumme, Mädche]: Stumme, G., Mädche A.:
FCA-Merge: A Bottom-Up Approach for Merging Ontologies JCAI '01 - Proceedings of the 17th International Joint Conference on Artificial Intelligence, Seattle, USA, August, 1-6, 2001, San Francisco/CA: Morgan Kaufmann, 2001

[Wortschatz]: the Wortschatz corpus querying service, http://wortschatz.uni-leipzig.de/