On Conceptual Indexing for Data Summarization

Henrik Bulskov[1] Troels Andreasen[1]

1. Department of Computer Science, Roskilde University,
P.O. Box 260, DK-4000 Roskilde, Denmark Email: bulskov@ruc.dk, troels@ruc.dk

Abstract— A summary is a comprehensive description that grasps the essence of a subject. A text, a collection of text documents, a query answer can be summarized by simple means such as an automatically generated list of the most frequent words or ”advanced” by a meaningful textual description of the subject. In between these two extremes are summaries by means of selected concepts exploiting background knowledge providing selected key concepts. We address in this paper an approach where conceptual summaries are provided through a conceptualization as given by an ontology. The idea is to restrict a background ontology to the set of concepts that appears in the text to be summarized and thereby provide a structure, a so-called instantiated ontology, that is specific to the domain of the text and can be used to condense to a summary not only quantitatively but also conceptually covers the subject of the text.

Keywords— conceptual clustering, conceptual descriptions, conceptual summaries, ontologies.

1 Introduction

The purpose of a summary is to provide a simplification to highlight the major points from the subject, e.g. a text or a set of texts such as a query answer. The aim is to provide a summary that grasps the essense of the subject.

Most common are summaries as those provided manually by readers or authors as a result of intellectual interpretation. Summaries can however also be provided automatically. One approach, in the Question Answering style, such as this is investigated in for instance the DUC and TREC conferences (see for instance [6],[5],[7]), is to provide a full natural language generation based summary construction while a less ambiguous, in the same tradition, is rather to perform a sentence selection from the text to be summarized.

In the other end the most simple approach is to select a reasonable short list of words among the most frequent and/or the most characteristic words from the set of words found in the text to be summarized. So rather than a coherent text the summary is a simple set of items.

Summaries in the approach presented here are also sets of items, but involves improvements over the simple set of words approach in two respects. First, we go beyond the level of keywords and aim to provide conceptual descriptions from concepts identified and extracted from the text. Second, we involve background knowledge in the form of an ontology. Strictly these two aspects are closely related – to use the conceptualization in the ontology we need means to map from words and phrases in the text to concepts in the ontology.

In this paper we present two different directions to conceptual summaries based on a background ontology. In both cases an ontology plays a key role as reference for the conceptualization. The general idea is from a world knowledge ontology to form a so-called “instantiated ontology” by restricting to a set of instantiated concepts.

First, we consider conceptual clustering over the instantiated concepts based on a semantic similarity measure and second we present an approach based on probabilities.

Below we first introduce to the the ontology notion, then we discuss extraction of conceptual descriptions, and finally we describe the various approaches to conceptual summaries.

2 Representing background knowledge – Ontology

Background knowledge is knowledge that complements the primary target data (the text or text collection / database) that is subject of the summarization with information that is essential to the understanding of this. Background knowledge can take different forms varying from simple lists of words to formal representations. In this context, however, our goal is conceptual summaries provided as sets of words or concepts so background knowledge to support this can range from unstructured lists of words to ontologies.

A simple list of words can be applied as a filter, mapping from a text to the subset of the word list that appears in the text. Such a controlled list of keywords or a vocabulary of topics can by obvious means be improved to capture also morphology by stemming or inflection patterns. Taxonomies, partonomies, semantic networks and ontologies are structures that potentially contribute also to knowledge-based summarization. Our main focus here is on ontologies ordered around taxonomic relationship. Rather than the common description logic based approach we choose here a simpler concept algebraic approach to ontologies.

One important rationale for this is that our goal here is not ontological reasoning in general but rather extraction of sets of mapped concepts and manipulation of such sets (e.g. contraction).

ISBN: 978-989-95079-6-8

2.1 *An algebraic approach to ontologies*

Given a basis taxonomy that situates a set of atomic term concepts $\mathcal{A}$ in a multiple inheritance hierarchy. Based on this we define a generative ontology by generalization of the hierarchy to a lattice an by introducing a (lattice-algebraic) concept language (description language) that defines an extended set of well-formed concepts, including both atomic and compound term concepts.

The concept language used here, ONTOLOG[9], has as basic elements concepts and binary relations between concepts. The algebra introduces two closed operations *sum* and *product* on concept expressions φ and ψ, where $(\varphi + \psi)$ denotes the concept being either φ or ψ and $(\varphi \times \psi)$ denotes the concept being φ and ψ (also called *join* and *meet* respectively).

Relationships r are introduced algebraically by means of a binary operator (:), known as the Peirce product $(r : \varphi)$, which combines a relation r with an expression φ. The Peirce product is used as a factor in conceptual products, as in $x \times (r : y)$, which can be rewritten to form the feature structure $x[r : y]$, where $[r : y]$ is an *attribution* of the concept x. Thus we can form compound concepts by attribution.

Given atomic concepts $\mathcal{A}$ and semantic relations $\mathcal{R}$, the set of well-formed terms $\mathcal{L}$ is:

$$\mathcal{L} = \{\mathcal{A}\} \cup \{x[r_1 : y_1, ..., r_n : y_n] | x \in \mathcal{A}, r_i \in \mathcal{R}, y_i \in \mathcal{L}\} \quad (1)$$

Compound concepts can thus have multiple as well as nested attributions. For instance with $\mathcal{R}$= $\{\text{WRT}, \text{CHR}, \text{CBY}, \text{TMP}, \text{LOC}, \ldots\}$[1] and $\mathcal{A}$ ={*entity, physical_entity, abstract_entity, location, town, cathedral, old*} we get:

$$\begin{aligned}&\mathcal{L} = \\ &\{entity, physical\_entity, abstract\_entity, \\ &location, town, cathedral, old, \\ &\ldots, \mathit{cathedral}[\text{LOC}: \mathit{town}, \text{CHR}: \mathit{old}], \\ &\mathit{cathedral}[\text{LOC}: \mathit{town}[\text{CHR}: \mathit{old}]], \ldots\}\end{aligned}$$

2.2 *Modelling Ontologies*

Obviously modelling ontologies from scratch is the best way to ensure that the result will be correct and consistent. However, for many applications the effort it takes is simply not at disposal and manual modeling have to be restricted to narrow and specific subdomains while the major part have to be derived from relevant sources. Sources that may contribute to modeling of ontologies may have various forms. A taxonomy is an obvious choice and it may be supplemented with, for instance, word and term lists as well as dictionaries for definition of vocabularies and for handling of morphology. Among the obviously useful resources are the semantic network WordNet [11] and the Unified Medical Language System (UMLS) [4] and unifies several other resources in the biomedical science area.

To go from a resource to an ontology is not necessarily straightforward, but if the goal is a generative ontology and the given resource is a taxonomy, one option is to proceed as follows. Given a taxonomy $\mathcal{T}$ over the set of atomic concepts $\mathcal{A}$ and a language $\mathcal{L}$, over $\mathcal{A}$ for a given set of relations $\mathcal{R}$, being derived as indicated in (1) above. Let $\hat{\mathcal{T}}$ be the transitive closure of $\mathcal{T}$. $\hat{\mathcal{T}}$ can be generalized to an inclusion relation "$\leq$" over all well-formed terms of the language $\mathcal{L}$ by the following

$$\begin{aligned}\text{“} \leq \text{“} = \;& \hat{\mathcal{T}} \\ & \cup \{\langle x[\ldots, r : z], y[\ldots]\rangle | \langle x[\ldots], y[\ldots]\rangle \in \hat{\mathcal{T}}\} \\ & \cup \{\langle x[\ldots, r : z], y[\ldots, r : z]\rangle | \langle x[\ldots], y[\ldots]\rangle \in \hat{\mathcal{T}}\} \\ & \cup \{\langle z[\ldots, r : x], z[\ldots, r : y]\rangle | \langle x, y\rangle \in \hat{\mathcal{T}}\}\end{aligned}$$

where repeated ... denote zero or more attributes of the form $r_i : w_i$.

The general ontology $\mathcal{O} = (\mathcal{L}, \leq, \mathcal{R})$ thus encompasses a set of well-formed expressions $\mathcal{L}$ derived in the concept language from a set of atomic concepts $\mathcal{A}$, an inclusion relation generalized from the taxonomy relation in $\mathcal{T}$, and a supplementary set of semantic relations $\mathcal{R}$. For $r \epsilon \mathcal{R}$, we obviously have $x[r : y] \leq x$, and that $x[r : y]$ is in relation r to y. Observe that $\mathcal{O}$ is generative and that $\mathcal{L}$ therefore is potentially infinite.

An example is given in figure 1 showing a segment of a generative ontology build with WordNet as resource.

3 Referencing the background knowledge – providing descriptions

As already indicated the approach involves surveying text through the ontology provided and delivering summaries on top of the conceptualization of the ontology. For this purpose we need to provide a description of the text to be summarized in terms of the concepts in the ontology. So words and/or phrases must be extracted from the text and mapped into the ontology. This is a knowledge extraction problem, and obviously such knowledge extraction can span from full deep natural language processing (NLP) to simplified shallow processing methods.

Here we will consider the latter due to the counterbalance between the need for a full interpretation and the computational complexity of getting it. A very simple solution would match words in text with labels of concepts in the ontology, hence make a many-to-many relation between words in text and labels in the ontology that just accepts the ambiguity of natural language. Improvements can easily be obtained through pattern based information extraction / text mining and through methods in natural language processing.

First, a heuristic part of speech tagging can be performed on the text, and provided that word classes are assigned to the concepts given in the ontology this enables a word class based disambiguation.

Second, a stemming or, provided lexical information is available, a transformation to a standardized inflectional form can significantly improve the matching.

Third, given part of speech tagged input, simple syntactic natural language grammars can be used to chunk words together forming utterances or phrases [3], that

[1] for *with respect to, characterized by, caused by, temporal, location*, respectively.

ISBN: 978-989-95079-6-8

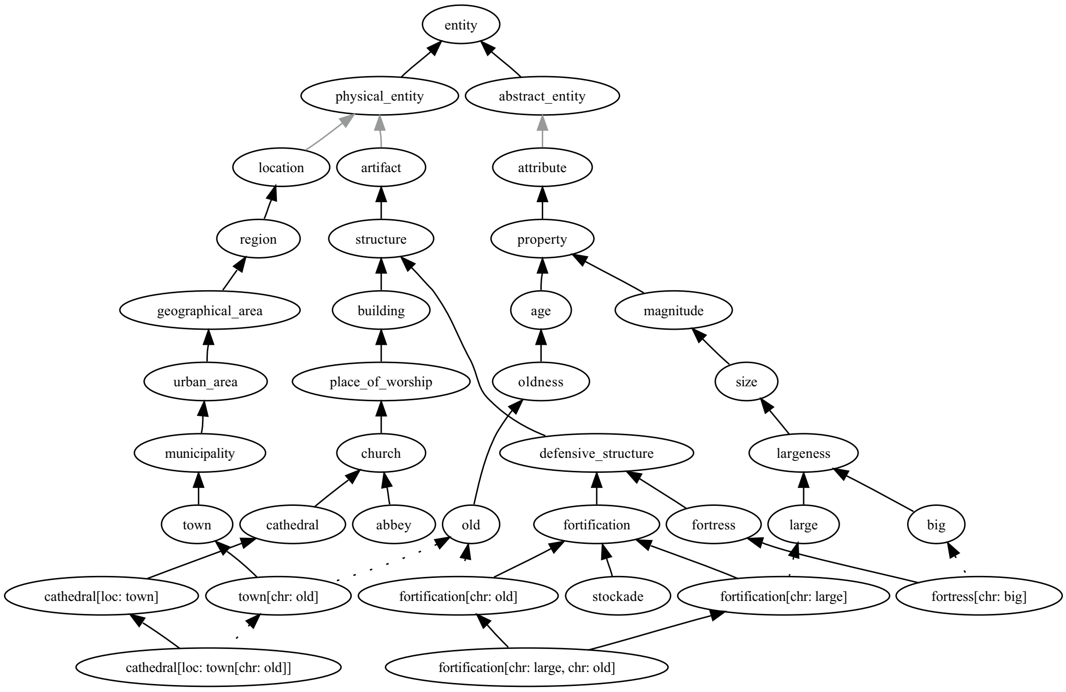

Figure 1: A segment of an ontology based on Wordnet. Does also correspond to an instantiated ontology for the set of instantiated concepts $\{$*cathedral*$[\textsc{loc}\colon$ *town*$[\textsc{chr}\colon$ *old*$]]$, abbey, *fortification*$[\textsc{chr}\colon$ *large*$, \textsc{chr}\colon$ *old*$]$, stockade, *fortress*$[\textsc{chr}\colon$ *big*$]\}$

.

can be used as the basis for matching against compound concepts in the ontology. Obviously the matching of chunks from the text and concepts in the ontology is in principle the same complex NLP problem over again, but the chunks identified will often correspond to meaningful concepts and therefore lead to a more refined and better result of the matching and, in addition, allow for a simple pattern-based approaches. We refer to [18] and [1] for more refined approaches. Here we will only cover a simple pattern-based approach.

Finally, some kind of word sense disambiguation [20] can be introduced in order to narrow down the possible readings of words, hence ideally mapping words of phrases to exactly one concept in the ontology.

A very simple approach along these lines is the following. Given a part-of-speech tagged and NP-chunked input a grammar for interpretation of the chunks is the following:

$$\begin{aligned} \text{Head} &::= \text{N} \\ \text{NP} &::= \text{A* N* Head} \mid \text{NP P NP} \end{aligned} \tag{2}$$

where A, N and P as placeholders for adjective, noun and preposition respectively. A very course-grained mapping strategy on top of this interpretation can be formed using the following transition rules, where premodifying adjectives relates to the head through *characterized by* (CHR) while premodifying composite nouns and prepositions both relates through *with respect to* (WRT):

$$\begin{aligned} &\mathrm{A}_1, \ldots, \mathrm{A}_n \mathrm{N}_1, \ldots, \mathrm{N}_m \mathrm{Head} \mapsto \\ &\quad \mathrm{Head}[\textsc{chr}\colon \mathrm{A}_1, \ldots, \textsc{chr}\colon \mathrm{A}_n, \textsc{wrt}\colon \mathrm{N}_1, \ldots, \textsc{wrt}\colon \mathrm{N}_m] \\ &\mathrm{NP}\ (\mathrm{P\ NP})_1, \ldots, (\mathrm{P\ NP})_n \mapsto \\ &\quad \mathrm{NP}[\,\textsc{wrt}\colon \mathrm{NP}_1, \ldots, \textsc{wrt}\colon \mathrm{NP}_n] \end{aligned} \tag{3}$$

To test this approach we consider the the Metathesaurus in the Unified Medical Language System (UMLS) [13] as resource and build a generative ontology from this. For part of speech tagging and phrase chunking we use the MetaMap application [2].

Consider the following utterance[2] as an example:

> [...] *the plasma patterns of estrogen and progesterone under gonadotropic stimulation simulating early pregnancy* [...]

The first part of the analysis leads to part of speech tagging and phrase recognition as shown in table 1.

By applying the grammar (2) this can be transformed into the following three noun phrases:

plasma/N patterns/N of/P estrogen/N

progesterone/N under/P gonadotropic/A stimulation/N

early/A pregnancy/N

and by using the transition rules (3) we can produce the following compound expressions:

[2]This utterance is from a small 50K abstracts fraction of MEDLINE [14] having both *Homones* and *Reproduction* as major topic keywords.

ISBN: 978-989-95079-6-8

Table 1: Part of speech tagging and phrase recognition

Phrase	Type	Word	POS
Noun Phrase	det	the	det
	mod	plasma	noun
	head	patterns	noun
Preposition	prep	of	prep
	head	estrogen	noun
	conj	and	conj
Noun Phrase	head	progesterone	noun
Preposition	prep	under	prep
	mod	gonadotropic	adj
	head	stimulation	noun
	verb	simulating	verb
Noun Phrase	mod	early	adj
	head	pregnancy	noun

patterns[WRT: plasma, WRT: estrogen]

progesterone[WRT:stimulation[CHR:gonadotropic]]

pregnancy[CHR:early]

3.1 *Instantiated Ontology*

The description $d_O(T)$ of a text T given the ontology O comprise a set of concepts in O and as indicated the purpose here is to summarize based on relations in the ontology. Now given the set of concepts (the description) $d_O(T)$ an obviously relevant subontology is a subontology that covers all elements of $d_O(T)$. Such a subontology can be consider an instantiation of the text T (or the set of concepts $d_O(T)$).

Given an ontology $\mathcal{O} = (\mathcal{L}, \leq, \mathcal{R})$ and a set of concepts C we define the instantiated ontology $\mathcal{O}_C = (\mathcal{L}_C, \leq_C, \mathcal{R})$ as a restriction of $\mathcal{O}$ to cover only the concepts in C, that is, C and every concept from $\mathcal{L}$ that subsumes concepts in C or attributes for concepts in C. $\mathcal{L}_C$ can be considered an "upper expansion" of C in $\mathcal{O}$. More specifically, with C^+ being C extended with every concept related by attribution from a concept in C:

$$\begin{aligned} \mathcal{L}_C &= C \cup \{x | y \in C^+, x \in \mathcal{L}, y \leq x\} \\ \text{“} \leq_C \text{“} &= \{(x, y) | x, y \in \mathcal{L}_C, x \leq y\} \end{aligned} \tag{4}$$

Thus $\mathcal{O}_C$ is not generative. "$\leq_C$" may be represented by a minimal set "$\leq'_C$"$\subseteq$"$\leq_C$" such that "$\leq_C$" is derivable from "$\leq'_C$" by means of transitivity of "$\leq$" and monotonicity of attribution:

$$\begin{aligned} transitivity: x &\leq y, y \leq z \Rightarrow x \leq z \\ monotonicity: x &\leq y \Rightarrow z[r: x] \leq z[r: y] \end{aligned}$$

Figure 1 shows an example of an instantiated ontology. The general ontology is based on (and includes) WordNet and the ontology shown is "instantiated" wrt. the following set of concepts:

$$C = \{\textit{cathedral}[\text{LOC}: \textit{town}[\text{CHR}: \textit{old}]], abbey, \\ \textit{fortification}[\text{CHR}: \textit{large}, \text{CHR}: \textit{old}], stockade, \\ \textit{fortress}[\text{CHR}: \textit{big}]\}$$

4 Data summarization through background knowledge

The general idea here is to exploit background knowledge through conceptual summaries, that are to provide a means to survey textual data, for instance a query result. A set of concepts from the background knowledge is first identified in the text and then contracted into a smaller set of, in principle, most representable concepts.

This can be seen as one direction in a more general conceptual querying approach where queries can be posed, and answers be presented, by means of conceptual abstractions. For a general discussion on other means, except from conceptual summaries, of conceptual querying, where also a dedicated language constructs for this purpose is presented we refer to [8]. Here we discuss summaries only.

In the approach to summarization described here we assume an ontology to guide the summarization and, for the text to be summarized, an initial extraction of concepts as described in the previous section. Thus, we can assume an initial set of concepts C and we a facing a challenge to provide a smaller set of representative concepts covering C, that is, an appropriate summary that grasps what´s most characteristic about C. For computation of the summary we restrict to the subontology $\mathcal{O}_C = (\mathcal{L}_C, \leq_C, \mathcal{R})$ corresponding to the instantiated ontology for C.

We introduce two directions for deriving summaries below: one based on clustering of the input concept set C and the other realized as a probability-based restriction of the input ontology $\mathcal{O}_C$. Towards the end we discuss possibilities for combining these approaches.

4.1 *Similarity Clustering*

Given a similarity measure summaries can be derived from a clustering of concepts applying this measure. Obviously, if the measure is derived from an ontology, and thereby do reflect this, then so will the clustering. We will below assume an ontology-based similarity measure sim. A simple example of such a measure can be derived from the path lenght in the ontology graph (Rada' *Shortest Path Length* [12]). More refined approaches are *Information Content* [16] and *Weighted Shared Nodes* [17].

4.1.1 *A hierarchical similarity-based approach*

With a given path-length dependent similarity measure derived from the ontology a lub-centered, agglomerative, hierarchical clustering can be performed as follows.

Initially each "cluster" corresponds to an individual element of the set to be summarized. At each particular stage the two clusters which are most similar are joined together. This is the principle of conventional hierarchical clustering. However rather than replacing the two joined clusters with their union as in the conventional approach they are replaced by their lub. Thus given a set of concepts $C = \{c_1, \ldots, c_n\}$ summarizers can be derived as follows.

ALGORITHM – Hierarchical clustering summary
INPUT: Set of concepts $C = \{c_1, \ldots, c_n\}$

ISBN: 978-989-95079-6-8

OUTPUT: Generalizing description $\delta(C)$ for C.

1) Let the instantiated ontology for C be $\mathcal{O}_C = (\mathcal{L}_C, \leq_C \mathcal{R})$
2) Let $T = \{\langle x, y\rangle | sim(x, y) = max_{z,w\in C}(sim(z, w))\}$
3) Let $U = min\left(\{u | u \in \mathcal{L}_C \wedge \exists x, y \in \mathcal{L}_C : x < u \wedge y < u\}\right)$
4) $L = \{x | \langle x, y\rangle \in T \vee \langle y, x\rangle \in T\}$
5) set $\delta(C) = C \cup U/L$

As was also the case with the connectivity clustering, to obtain an appropriate description of C we might have to apply δ several times and at some point m we have that $\delta^m(C) = Top$.

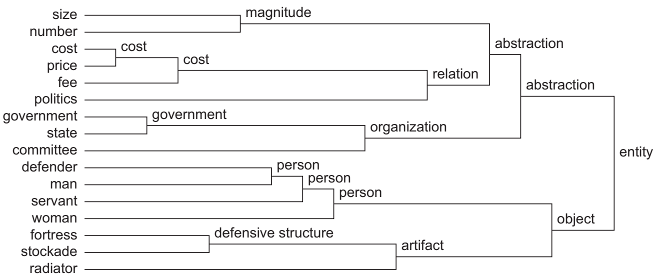

Figure 2: An illustration of the hierarchical clustering summary. The merging of two clusters are shown with their lub.

Figure 2 illustrates the application of δ a total of 15 times to the set of concepts from the previous example, where we might have (depending on the exact similarity values) for instance:

C = {*number, size, committee, government, state, defender, man, servant, woman, bribe, cost, price, fee, fortification, fortress, stockade*}

$\delta^1(C)$ = {*number, size, committee, government, state, defender, man, servant, woman, bribe,* ***cost****, fee, fortification, fortress, stockade*}

$\delta^2(C)$ = {*number, size, committee,* ***government****, defender, man, servant, woman, bribe, cost, price, fee, fortification, fortress, stockade*}

etc.

Thus summaries are generated iteratively and at each step the two closest concepts are clustered and the result is replaced by the corresponding lub.

4.1.2 Simple least upper bound-based approach

The principle of replacing clusters by their least upper bound can by applied on top of in principle any clustering approach. A straightforward similarity based approach is simply to apply a crisp clustering to the set of concepts $C = \{c_1, \ldots, c_n\}$ leading to $\{C_1, \ldots, C_k\}$ and then provide the set of lub's $\{\widehat{c}_1, \ldots, \widehat{c}_k\} = \{lub(C_1), \ldots, lub(C_k)\}$ for the division of C as summary. However to take into account also the importance of clusters in terms of their sizes the summary can be modified by the support of the generalizing concepts, $support(x, C)$, that for a given concept specifies the fraction of elements from the set C covered:

$$support(x, C) = \frac{|\{y | y \in C, y \leq x\}|}{|C|} \tag{5}$$

leading to a fuzzyfied (weighted) summary, based on the division (crisp clustering) of C into $\{C_1, \ldots, C_k\}$:

$$\Sigma_i support\left(lub(C_i), C\right)/lub(C_i) \tag{6}$$

To illustrate this lub-based approach consider table 2. Five groups are given that are derived as clusters of synsets in wordnet [3]

Table 2: A set of clusters and their least upper bounds from WordNet.

cluster	lub
{number, size}	magnitude
{committee, government, state}	organization
{defender, man, servant, woman}	person
{bribe, cost, fee, price}	cost
{fortification, fortress, stockade}	defensive structure

From these clusters the fuzzyfied summary {*.13/magnitude + .19/organization + .25/person + .25/cost + .19/defensive structure*} can be generated.

We may expect a pattern similar to hierarchical clustering in derivation of summaries in an approach based on similarity when the similarity measure reflects simple shortest path in the ontology.

This approach to summarization can be generalized using a fuzzyfied notion in place of the least upper bound as candidate representative. The generalization reduces the sensitivity against noise in the groups resulting from the initial clustering. This approach is described in [8].

4.2 A probability-based approach

We consider above a summary of textual input based on the concepts that appears in the text and how these are related in a background ontology. In the hierarchical approach candidate summarizers are chosen regardless of their coverage of the input, while the lub-based approached is introduced with a support that measures the degree to which all occurring input concepts are summarized. An obvious extension in this direction is to also consider frequencies of terms in the input text and thereby measure the probability of encountering an instance of a concept in the text.

Probabilities provides a means for selecting summarizers without taking other measures into account and thus allows for a straightforward approach as follows.

Given a set of concepts $C = \{c_1, \ldots, c_n\}$ and let $\mathcal{O}_C = (\mathcal{L}_C, \leq_C, \mathcal{R})$ be the instantiated ontology. Let further $child(c)$ denote the set of immediate children and $parent(c)$ the set of immediate parents for any concept $c \in \mathcal{L}_C$. The principle is to accumulate the frequencies to more general concepts but only so that a child c contributes with $\frac{1}{|parent(c)|}$ to each parent. A summary can be derived as follows.

ALGORITHM – Probability-based summary

INPUT: Set of concepts $C = \{c_1, \ldots, c_n\}$, their relative frequencies $C = \{freq(c_1), \ldots, freq(c_n)\}$ and a threshold α

[3] The first four are set of clusters and their least upper bounds in where $C_1, \ldots, C_4$ are from SEMCOR and C_5 is from the example ontology in figure 1.

ISBN: 978-989-95079-6-8

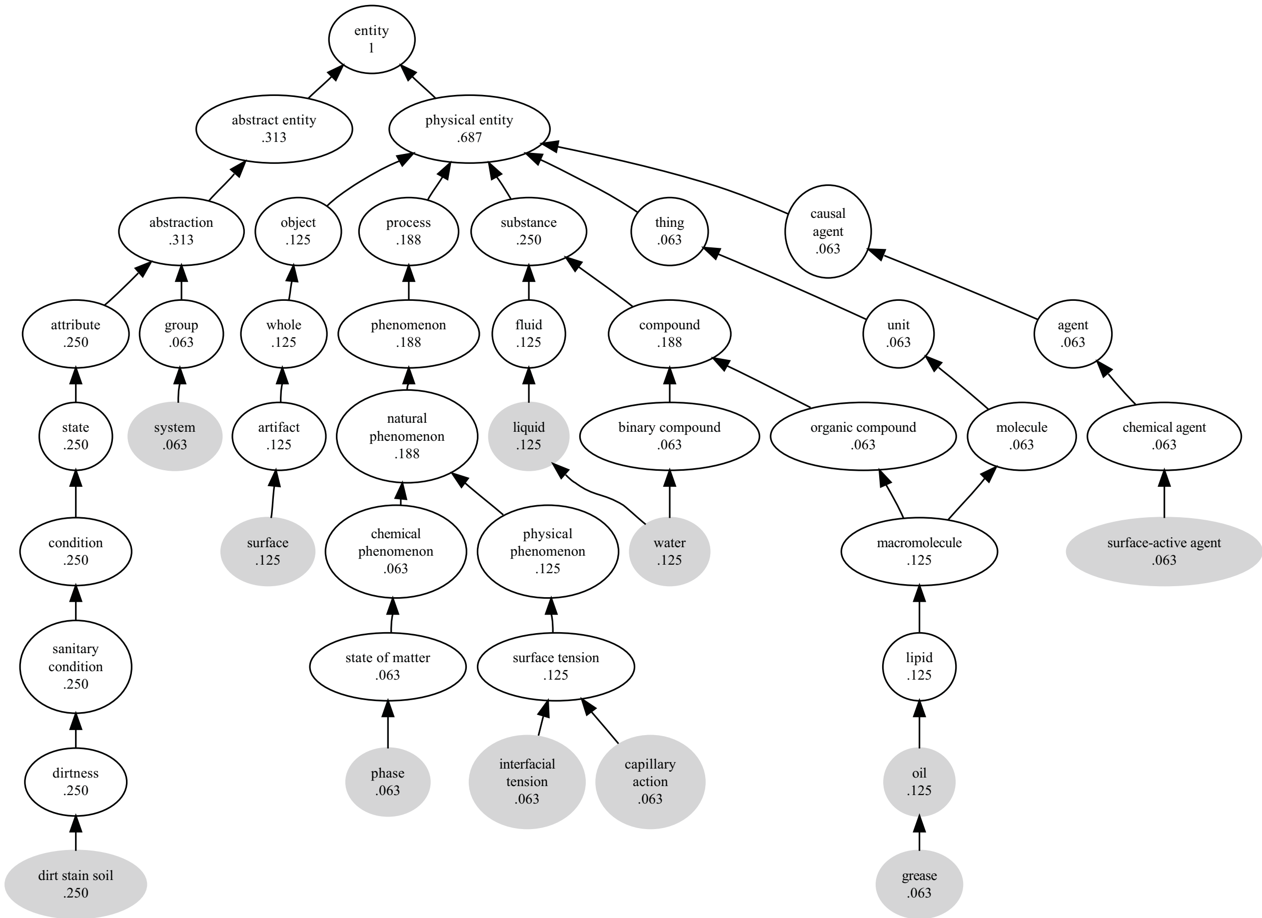

Figure 3: An instantiated ontology based on a paragraph from SEMCOR

OUTPUT: Generalizing description $D(C, \alpha)$ for C.

1) Let the instantiated ontology for C be $\mathcal{O}_C = (\mathcal{L}_C, \leq_C, \mathcal{R})$
2) Accumulate the frequencies in correspondence with the ontology such that $\forall c \in \mathcal{L}_C \backslash C : freq(c) = \sum_{c' \in child(c)} \frac{1}{|parent(c')|} freq(c')$
3) Let $N = |C|$ and $p(c) = freq(c)/N$ be the probability of encountering c.
4) Let $\mathcal{O}'_C = (\mathcal{L}'_C, \leq'_C, \mathcal{R}')$ be the restriction of $\mathcal{O}_C$ to the concepts that appear in $\{c \in \mathcal{L}_C | p(c) \geq \alpha\}$
5) Set the α-level summary of C to the most specific concepts appearing in $\mathcal{O}'_C$, that is $D(C, \alpha) = \{c \in \mathcal{L}'_C |\ \nexists c' \in \mathcal{L}'_C : c' < c\}$

As an example consider figure 3 that shows an instantiated ontology derived from WordNet for the following paragraph found in SEMCOR[10][4]:

> *Greases*, *stains*, and miscellaneous *soils* are usually sorbed onto the *soiled surface*. In most cases, these *soils* are taken up as *liquids* through *capillary action*. In an essentially static *system*, an *oil* cannot be replaced by *water* on a *surface* unless the *interfacial tensions* of the *water phase* are reduced by a *surface-active agent*.

[4]SEMCOR is a subset of the documents in the Brown corpus which has the advantage of being semantically tagged with senses from WordNet

Words in italics indicate the initial set of concepts[5]. Among the recognized concepts most appear only once, while the frequency of *surface* and *water* is two and the frequency of *soils* is 4 (includes *stains*). We have N =16 and C={*Greases*, *stains*, *soils*, *soiled*, *surface*, *soils*, *liquids*, *capillary action*, *system*, *oil*, *water*, *surface*, *interfacial tensions*, *water*, *phase*, *surface-active agent*} and thus get for instance

> $D(C, 0.1)$={*dirt stain soil*, *surface*, *surface tension*, *water*, *oil*}
>
> $D(C, 0.15)$={*dirt stain*, *soil*, *natural phenomenon*, *compound*}

5 Concluding remarks

In this paper we have considered how to use ontologies to provide data summaries with a special focus on textual data. Such summaries can be used in a querying approach where concepts describing documents, rather than documents directly, are retrieved as query answer. The summaries presented are conceptual due to fact that they exploit concepts from the text to be summarized and ontology-based because these concepts are drawn from a reference ontology.

[5]Notice that due to the use of SEMCOR there is no compound concepts in the initial set here.

ISBN: 978-989-95079-6-8

We have presented three summary principles. Two based on similarity and clustering and the third on probabilities derived from frequencies in the text to be summarized. Obviously a "meaningful" clustering may lead to good summaries if characterizing subsuming concepts can be found in the ontology, but as indicated also counting occurrences, rather than only recognizing presence, of concepts may contribute to encircling essential concepts. Thus a next step should be to develop a combination that draws on similarity to ensure for instance that close summarizers are joined and on frequencies to ensure that derived summarizers are important.

For a summarization principle to work in practise users clearly need some kind of guidance on how many times to iterate or how to set a threshold. This brings up the important question on evaluation of summaries. Initial considerations on the quality of summaries can be found in [19] but the issue is also an obvious direction for further work in continuation of what has been described here.

References

[1] P. Anker Jensen, J. Fischer Nilsson Ontology-based Semantics for Prepositions, in Syntax and Semantics of Prepositions, P. Paint-Dizier (ed.), Text, Speech and Language Technology, Vol. 29, Springer, 2006.

[2] S. R. Aronson Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, Proc AMIA Symp, pp 17-21, 2001.

[3] S. Abney Partial parsing via finite-state cascades, in: Proceedings of the ESSLLI´96 Robust Parsing Workshop, 1996.

[4] Olivier Bodenreider The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32, D267-D270. (2004)

[5] Udo Hahn, Inderjeet Mani The Challenges of Automatic Summarization Computer November 2000

[6] Gabor Melli, YangWang, Yudong Liu, Mehdi M. Kashani, Zhongmin Shi, Baohua Gu, Anoop Sarkar and Fred Popowich: Description of SQUASH, the SFU Question Answering Summary Handler for the DUC-2005 Summarization Task. Proceedings of DUC-2005, Vancouver, Canada pp 103-110 (205).

[7] Zhongmin Shi, Gabor Melli, Yang Wang, Yudong Liu, Baohua Gu, Mehdi M. Kashani, Anoop Sarkar and Fred Popowich Question Answering Summarization of Multiple Biomedical Documents Lecture Notes in Computer Science Springer Berlin / Heidelberg Volume 4509/2007 Advances in Artificial Intelligence pp 284-295 (2007)

[8] Andreasen, T.,Bulskov, H.: Conceptual Querying Through Ontologies. Fuzzy Sets and Systems, Elsevier, to appear (2008).

[9] Nilsson, J.F.: A logico-algebraic framework for ontologies – ONTOLOG. In Jensen, P.A., Skadhauge, P., eds.: First International OntoQuery Workshop, University of Southern Denmark (2001)

[10] Miller, G.A., Chodorow, M., Landes, S., Leacock, C., and Thomas R.G.: Using a semantic concordance for sense identification. In *Proc. of the ARPA Human Language Technology Workshop*, pages 240-243, 1994.

[11] Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM **38**(11) (1995) 39–41

[12] Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics **19**(1) (January 1989) 17–30

[13] Unified Medical Language System U.S. National Library of Medicine http://www.nlm.nih.gov/research/umls/

[14] Medical Literature Analysis and Retrieval System Online U.S. National Library of Medicine http://www.ncbi.nlm.nih.gov/pubmed/

[15] Bulskov, H., Knappe, R., Andreasen, T.: On measuring similarity for conceptual querying. In: FQAS '02: Proceedings of the 5th International Conference on Flexible Query Answering Systems, Springer-Verlag (2002) 100–111

[16] Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language (1999)

[17] Andreasen, T., Knappe, R., Bulskov, H.: Domain-specific similarity and retrieval. Proceedings IFSA 2005, Tsinghua University Press (2005) 496–502

[18] Andreasen, T. and Jensen, P. Anker and Nilsson, J. Fischer and Paggio, P. and Pedersen, B. Sandford and Thomsen, H. Erdman: Content-based text querying with ontological descriptors,Data Knowledge Engineering ,vol 48(2), pp 199–219, Elsevier Science Publishers

[19] Yager, R.R., Petry, F.E.: A Multicriteria Approach to Data Summarization Using Concept Hierarchies IEEE Trans. on Fuzzy Sys. vol 14 no 6 (2006)

[20] Zhou, Xiaohua, Han, Hyoil Survey of word sense disambiguation approaches 18th FLAIRS Conference (2005)

ISBN: 978-989-95079-6-8