Evaluation of Extraction Techniques for Ontology Excerpts*

Jieying Chen[1], Michel Ludwig[2,3], Yue Ma[2], and Dirk Walther[2,3]

[1] Jilin University, College of Computer Science and Technology, China
chenjy12@mails.jlu.edu.cn
[2] TU Dresden, Theoretical Computer Science, Germany
[3] Center for Advancing Electronics Dresden, Germany
{michel, mayue, dirk}@tcs.inf.tu-dresden.de

Abstract. We introduce the notion of an ontology *excerpt* as being a fixed-size subset of an ontology that preserves as much knowledge as possible about the terms in a given vocabulary as described in the ontology. We consider different extraction techniques for ontology excerpts based on methods from Information Retrieval. To evaluate these techniques, we measure the degree of incompleteness of the resulting excerpts using the notion of logical difference. We provide an experimental evaluation of the extraction techniques by applying them on the biomedical ontology SNOMED CT.

1 Introduction

Ontologies based on Description Logics (DL) [2] have become a well-established paradigm for representing knowledge. An increasing number of ontologies is being developed, e.g., the CPO, FMA, GALEN, and SNOMED CT ontologies in the biomedical domain, which are made available in repositories such as NCBO Bioportal.[1] Ensuring efficient access to the knowledge contained in such repositories is thus becoming an important concern.

We consider the problem of ontology selection. Suppose that we are in the presence of a repository containing a number of ontologies that represent knowledge in a certain domain of expertise. The task that we consider is for a human user to efficiently select an ontology from the repository that is most "relevant" for a given vocabulary of interest, called *signature*. Such a decision task could be supported, for example, by providing summaries of the respective ontologies w.r.t. a signature Σ. As ontologies can become quite large, it is important, however, to limit the size of summaries to effectively support the decision process, in particular, when human users are involved.

The notion of modules provides a way to extract a subset of an ontology that exactly captures the knowledge of the ontology regarding a signature. However,

* Partially supported by German Research Foundation (DFG) via the Research Unit FOR 1513 (Project B1) and the Cluster of Excellence 'cfAED'.

[1] http://bioportal.bioontology.org/

a module notion typically allows for little *control* over the number of axioms that are included in a module. Depending on the signature of interest, even minimal modules can be as large as the entire ontology, essentially defeating the purpose of module extraction. To influence the size of a module, our only option is to adapt the signature for which the module is extracted. Generally, we have that the smaller the signature, the smaller are the modules of an ontology. But no strict upper bound on the module size can be guaranteed in this way. Given a relatively large signature of interest, obtaining small modules requires selecting smaller subsets of the signature. However, finding a suitably small subset of the initial signature may not be feasible. On the one hand, it may not be clear which symbols can be removed, and on the other hand, an exhaustive search for the "right" signature involves a combinatorial blowup.

In this paper, we introduce the notion of an *ontology excerpt* as a fixed-size subset of an ontology. For the purpose of selecting the most relevant ontology in a repository, one can first extract excerpts of all the ontologies in the repository. Based on such excerpts the user should be able to make an informed decision for selecting the most relevant ontology. Obviously, when the user is interested in terms contained in a signature Σ, the excerpts should capture as much knowledge of the terms from Σ as possible. We propose to use excerpt extraction techniques w.r.t. a signature from the area of Information Retrieval, i.e. a research area which is generally concerned with developing techniques to extract the "most relevant" documents to a query from large data sources.

To evaluate the quality of ontology excerpts, we employ a semantics-based measure, called *Gain*, based on Logical Difference [7] to quantify how much semantic meaning is preserved in an excerpt w.r.t. the original ontology.

The size of the excerpts can be made dependent, for instance, on the number of considered ontologies and the similarity of their content to comply with the time constraints associated with the decision process. If the user can invest more time, excerpts may be larger. Similarly, if the content of the ontologies should be similar, larger excerpts could help distinguish them more easily. Nevertheless the size should be chosen such that the resulting excerpts can still be easily understood by the user.

The paper is organized as follows. After some preliminaries, we introduce the notion of ontology excerpts in Section 3. Extraction techniques for excerpts are proposed in Section 4, and we conclude with a practical evaluation of the techniques in Section 5.

2 Preliminaries

We briefly recall basic notions related to the description logic $\mathcal{EL}$ [1], modularity of ontologies [6, 8] and the logical difference between terminologies [7, 9].

2.1 The Description Logic $\mathcal{EL}$

Let $\mathsf{N_C}$ and $\mathsf{N_R}$ be mutually disjoint (countably infinite) sets of concept names and role names. In the following we use A, B, X, Y, Z to denote concept names, and r, s stand for role names. The set of *$\mathcal{EL}$-concepts* C and the set of *$\mathcal{EL}$-inclusions* α are built according to the following grammar rules:

$$C ::= \top \mid A \mid C \sqcap C \mid \exists r.C$$
$$\alpha ::= C \sqsubseteq C \mid C \equiv C \mid r \sqsubseteq s$$

where $A \in \mathsf{N_C}$ and $r, s \in \mathsf{N_R}$. An *$\mathcal{EL}$-ontology* $\mathcal{O}$ is a finite set of $\mathcal{EL}$-inclusions, which are also referred to as *axioms*.

The semantics is defined using interpretations $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, where the domain $\Delta^{\mathcal{I}}$ is a non-empty set, and $\cdot^{\mathcal{I}}$ is a function mapping each concept name A to a subset $A^{\mathcal{I}}$ of $\Delta^{\mathcal{I}}$ and every role name r to a binary relation $r^{\mathcal{I}}$ over $\Delta^{\mathcal{I}}$. The *extension* $C^{\mathcal{I}}$ of a possibly complex concept C is defined inductively as: $(\top)^{\mathcal{I}} := \Delta^{\mathcal{I}}$, $(C \sqcap D)^{\mathcal{I}} := C^{\mathcal{I}} \cap D^{\mathcal{I}}$, and $(\exists r.C)^{\mathcal{I}} := \{x \in \Delta^{\mathcal{I}} \mid \exists y \in C^{\mathcal{I}} : (x, y) \in r^{\mathcal{I}}\}$.

An interpretation $\mathcal{I}$ *satisfies* a concept C, an axiom $C \sqsubseteq D$, $C \equiv D$, or $r \sqsubseteq s$ if $C^{\mathcal{I}} \neq \emptyset$, $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$, $C^{\mathcal{I}} = D^{\mathcal{I}}$, or $r^{\mathcal{I}} \subseteq s^{\mathcal{I}}$, respectively. We write $\mathcal{I} \models \alpha$ if $\mathcal{I}$ satisfies the axiom α. Note that every $\mathcal{EL}$-concept and axiom is satisfiable, but a particular interpretation does not necessarily satisfy a concept/axiom. An interpretation $\mathcal{I}$ is a *model* of $\mathcal{O}$ if $\mathcal{I}$ satisfies all axioms in $\mathcal{O}$. An axiom α *follows* from an ontology $\mathcal{O}$, written $\mathcal{O} \models \varphi$, if for all models $\mathcal{I}$ of $\mathcal{O}$, we have that $\mathcal{I} \models \alpha$.

An *$\mathcal{EL}$-terminology* $\mathcal{O}$ is an $\mathcal{EL}$-ontology consisting of axioms α of the form $A \sqsubseteq C$, $A \equiv C$, or $r \sqsubseteq s$, where A is a concept name, C an $\mathcal{EL}$-concept and no concept name A occurs more than once on the left-hand side of an axiom. A terminology is said to be *acyclic* if it can be unfolded (i.e., the process of substituting concept names by the right-hand sides of their defining axioms terminates). We denote with $|\mathcal{O}|$ the number of axioms in the ontology $\mathcal{O}$. A signature Σ is a finite subset of $\mathsf{N_C} \cup \mathsf{N_R}$. For a syntactic object X, the signature $\mathsf{sig}(X)$ is the set of concept and role names occurring in X.

2.2 Modularity

A *module* $\mathcal{M}$ of an ontology $\mathcal{O}$ w.r.t. a signature Σ is a subset of $\mathcal{O}$ that preserves all entailments formulated in Σ: for all inclusions α with $\mathsf{sig}(\alpha) \subseteq \Sigma$, $\mathcal{M} \models \alpha$ if and only if $\mathcal{O} \models \alpha$. The interesting direction of the equation in this definition is from right to left; the other direction holds by monotonicity of description logics. We can equivalently define $\mathcal{M}$ with the condition that $\mathcal{O}$ is a Σ-conservative extension of $\mathcal{M}$. For practical purposes, we are generally interested in modules to be as small as possible. Note that minimal modules are not necessarily unique. Computing minimal modules (which is equivalent to deciding whether or not an ontology is a conservative extension of the module) is usually harder than standard reasoning in the underlying DL, and often it is even undecidable [5, 11]. For $\mathcal{EL}$-terminologies, however, minimal modules are unique and they can be computed in polynomial time [8].

Example 1. Let $\mathcal{O}$ consist of the following four axioms:

$$\begin{array}{llll} \alpha_1: & A \sqsubseteq B \sqcap \exists r.X & \alpha_2: & B \sqsubseteq A \\ \alpha_3: & X \equiv A \sqcap B & \alpha_4: & Y \equiv B \sqcap \exists r.(X \sqcap \exists s.A) \end{array}$$

and let $\Sigma_1 = \{A, B\}$. Then $\mathcal{M} = \{\alpha_1, \alpha_2, \alpha_3\}$ is the min. module of $\mathcal{O}$ w.r.t. Σ.

In this paper, due to the availability of tool support, we employ locality-based modules [6]. These are approximations of minimal modules and can be efficiently computed in polynomial time. We use the $\bot\top^*$-locality notion. We denote with $\mathsf{Mod}^*_{\mathcal{O}}(\Sigma)$ the $\bot\top^*$-local module of the ontology $\mathcal{O}$ w.r.t. the signature Σ. Note that locality-based modules are unique, and they contain any minimal module for a signature and possibly more axioms. However, it was shown that locality-based modules are not much larger than minimal modules on real ontologies [3].

2.3 Logical Concept Difference

We now recall basic notions related to the logical difference [7, 9] between two $\mathcal{EL}$-terminologies for $\mathcal{EL}$-inclusions as query language.

Definition 1 (Logical Difference). *Let $\mathcal{O}_1$ and $\mathcal{O}_2$ be $\mathcal{EL}$-terminologies, and let Σ be a signature. The $\mathcal{EL}$-*concept inclusion difference *between $\mathcal{O}_1$ and $\mathcal{O}_2$ w.r.t. Σ is the set $\mathsf{Diff}_\Sigma(\mathcal{O}_1, \mathcal{O}_2)$ of all $\mathcal{EL}$-inclusions α of the form $C \sqsubseteq D$ for $\mathcal{EL}$-concepts C and D such that $\mathsf{sig}(\alpha) \subseteq \Sigma$, $\mathcal{O}_1 \models \alpha$, and $\mathcal{O}_2 \not\models \alpha$.*

In case two terminologies are logically different, the set $\mathsf{Diff}_\Sigma(\mathcal{O}_1, \mathcal{O}_2)$ consists of infinitely many concept inclusions. The *primitive witnesses theorems* from [7] allow us to consider only certain inclusions of a simpler syntactic form.

Theorem 1. *Let $\mathcal{O}_1$ and $\mathcal{O}_2$ be $\mathcal{EL}$-terminologies and let Σ be a signature. If $\alpha \in \mathsf{Diff}_\Sigma(\mathcal{O}_1, \mathcal{O}_2)$, then either $A \sqsubseteq C$ or $D \sqsubseteq A$ is a member of $\mathsf{Diff}_\Sigma(\mathcal{O}_1, \mathcal{O}_2)$, where $A \in \mathsf{sig}(\alpha)$ is a concept name, and C, D are $\mathcal{EL}$-concepts occurring in α.*

Definition 2 (Primitive Witnesses). *Let $\mathcal{O}_1$ and $\mathcal{O}_2$ be $\mathcal{EL}$-terminologies and let Σ be a signature. We say that $\mathcal{EL}$-*concept inclusion difference witnesses *in Σ w.r.t. $\mathcal{O}_1$ and $\mathcal{O}_2$ are concept names contained in Σ that occur on the left-hand side of inclusions of the form $A \sqsubseteq C$ in $\mathsf{Diff}_\Sigma(\mathcal{O}_1, \mathcal{O}_2)$ or on the right-hand side of inclusions of the form $D \sqsubseteq A$ in $\mathsf{Diff}_\Sigma(\mathcal{O}_1, \mathcal{O}_2)$. The set of all such witnesses will be denoted by $\mathsf{Wtn}_\Sigma(\mathcal{O}_1, \mathcal{O}_2)$.*

Observe that $\mathsf{Wtn}_\Sigma(\mathcal{O}_1, \mathcal{O}_2) \subseteq \Sigma$. Consequently, this set is finite and it can be seen as a succinct representation of the set $\mathsf{Diff}_\Sigma(\mathcal{O}_1, \mathcal{O}_2)$ in the sense that: $\mathsf{Diff}_\Sigma(\mathcal{O}_1, \mathcal{O}_2) = \emptyset$ iff $\mathsf{Wtn}_\Sigma(\mathcal{O}_1, \mathcal{O}_2) = \emptyset$ [7].

Example 2. Let $\mathcal{O}$ be defined as in Ex. 1. For $\Sigma = \{A, B\}$, $\mathsf{Wtn}_\Sigma(\mathcal{O}, \{\alpha_1, \alpha_2\}) = \mathsf{Diff}_\Sigma(\mathcal{O}, \{\alpha_1, \alpha_2\}) = \emptyset$ and $\mathsf{Wtn}_\Sigma(\mathcal{O}, \emptyset) = \Sigma$ as $A \sqsubseteq B, B \sqsubseteq A \in \mathsf{Diff}_\Sigma(\mathcal{O}, \emptyset)$. If $\Sigma = \{A, r\}$, we have that $\mathsf{Wtn}_\Sigma(\mathcal{O}, \mathcal{O} \setminus \{\alpha_1\}) = \{A\}$ as $A \sqsubseteq \exists r.\top \in \mathsf{Diff}_\Sigma(\mathcal{O}, \mathcal{O} \setminus \{\alpha_1\})$.

Algorithms for computing the witness sets and, thus, for deciding whether a logical difference w.r.t. a signature exists, have been implemented in the CEX2.5 tool.[2] Given two acyclic $\mathcal{EL}$-terminologies and a signature Σ as input, CEX2.5 computes and outputs the set $\mathsf{Wtn}(\mathcal{O}_1, \mathcal{O}_2)$ in a fully automatic way.

[2] CEX2.5 is available under an open-source license from `http://lat.inf.tu-dresden.de/~michel/software/cex2/`.

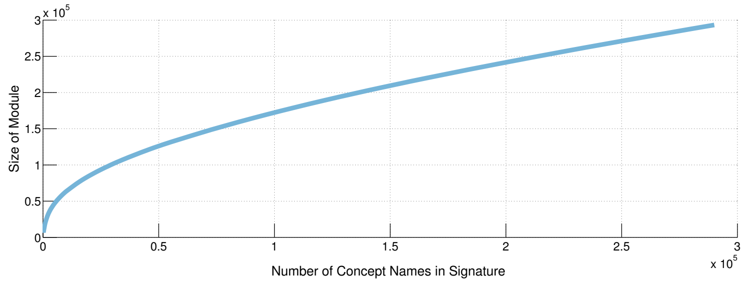

Fig. 1. Size of signature Σ vs. size of $\bot\top^*$-local module w.r.t. Σ of SNOMED CT

We note that a new approach for computing logical differences that can also handle large cyclic terminologies has recently been introduced [4].

3 Ontology Excerpts

Ontologies appear to exhibit a strong dependency between the size of a signature Σ and the size of a module for the symbols in Σ. This dependency is a natural consequence of the structure of the ontology which, depending on the application at hand, we may not be able to afford due to resource restrictions. We are interested in gaining more control over the size of a module to be able to reuse the knowledge in an ontology in a scenario with limited resources.

Figure 1 illustrates the dependency between signature size and module size in the case of SNOMED CT consisting of 297 090 axioms, 297 079 concept names, and 62 role names. The coordinates of a point in Figure 1 are a pair (n, m) of numbers, where n corresponds to the number of terms in a signature Σ and m to the number of axioms in a subset $\mathcal{S}$ of the ontology $\mathcal{O}$. The curve connects over 30 data points, each of which represents the median value of the sizes of 500 $\bot\top^*$-local modules of SNOMED CT. Each module is extracted w.r.t. a signature consisting of n concept names, for n with $200 \leq n \leq 290\,000$, and 30 role names that are randomly selected from the signature of SNOMED CT. The special role name 'RoleGroup' is always selected. Note that we fixed the number of role names arbitrarily; similar results can be expected for different numbers of role names. The time needed to extract 500 modules ranges from about 30 min for small signatures (containing 200 concept names) to about 90 min for large signatures (containing 250 000 concept names).[3]

We can see from Figure 1 that the module sizes increase with the size of the input signature. For small signatures, the slope is steep, showing that the modules are relatively large compared to the signature size. However, with increasing signature sizes, the slope flattens. For different signatures of the same size, there are still variations in the sizes of the modules for these signatures. In

[3] The experiments were conducted on a PC equipped with an Intel Xeon E5-2640 CPU running at 2.50GHz and with 100GB of RAM. We used Debian GNU/Linux 7.3 as operating system, Java version 1.7.0 51 and OWLAPI version 3.4.8.

this experiment, the module sizes vary from 2 633 to 4 086 for signatures up to 100 000 concept names and 30 role names. For larger signatures the variation in module size reduces to 224 for signatures with 290 000 concept names and 30 role names, and converges to 0 as signature being expanded to the whole signature of SNOMED CT.

Let us consider the coordinates of a point in the chart in Figure 1 as a pair (n, m) of numbers, where n corresponds to the size of a signature Σ and m to the size of a subset $\mathcal{S}$ of the ontology $\mathcal{O}$. Note that, for a signature Σ' and a subset $\mathcal{S}' \subseteq \mathcal{O}$ that correspond to a point in the area above the curve for the ontology $\mathcal{O}$ in Figure 1, we may have that $\mathsf{Mod}_{\Sigma'}(\mathcal{O}) \subsetneq \mathcal{S}'$. In this case, $\mathcal{S}'$ likely contains axioms that do not contribute to the meaning of the symbols in Σ'. Therefore, we are mainly interested in the area below the curve for an ontology $\mathcal{O}$. Let (n, m) be a point in that area, and let $\mathcal{S} \subseteq \mathcal{O}$ and Σ be such that $|\Sigma| = n$ and $|\mathcal{S}| = m$. We have that $\mathcal{S}$ contains fewer axioms than the module $\mathsf{Mod}_{\Sigma}(\mathcal{O})$ (not considering the variation of module sizes), i.e. $|\mathcal{S}| \leq |\mathsf{Mod}_{\Sigma}(\mathcal{O})|$. Therefore, $\mathcal{S}$ is likely incomplete in capturing the meaning of the symbols in Σ. The trade-off for obtaining full control over the size of $\mathcal{S}$ is a certain degree of incompleteness of $\mathcal{S}$. This inspires the notion of *ontology excerpts* as introduced below.

Definition 3 (Ontology Excerpt). *Let $\mathcal{O}$ be an ontology and let $k > 0$ be a natural number. A k-*excerpt *of $\mathcal{O}$ is a subset $\mathcal{E} \subseteq \mathcal{O}$ consisting of k axioms, i.e. $|\mathcal{E}| = k$.*

An ontology excerpt is a subset of the ontology of a certain size. However, we are interested in those excerpts that preserve the meaning of the symbols in a signature of interest. To quantify the meaning of an excerpt, we need some metric μ. We assume that the lower the value of μ for an excerpt is, the more meaning is preserved by the excerpt. This is made precise as follows.

Definition 4 (Incompleteness Measure). *Let $\mathcal{O}$ be an ontology. An* incompleteness measure *μ is a function that maps every triple $(\mathcal{O}, \Sigma, \mathcal{E})$ consisting of an ontology $\mathcal{O}$, a signature Σ, and an excerpt $\mathcal{E} \subseteq \mathcal{O}$ to a non-negative natural number.*

In this paper we use as incompleteness measure μ the number $\mathsf{ldiff}(\mathcal{O}, \Sigma, \mathcal{E})$ of $\mathcal{EL}$-concept inclusion difference witnesses in Σ w.r.t. $\mathcal{O}$ and $\mathcal{E}$, which is defined formally as $\mathsf{ldiff}(\mathcal{O}, \Sigma, \mathcal{E}) = |\mathsf{Wtn}_{\Sigma}(\mathcal{O}, \mathcal{E})|$.

Definition 5 (Best Excerpt). *Let $\mathcal{O}$ be an ontology, let Σ be a signature, and let $k > 0$ be a natural number. Additionally, let μ be an incompleteness measure. A* best k-excerpt *of $\mathcal{O}$ w.r.t. Σ under μ is a k-excerpt $\mathcal{E}$ of $\mathcal{O}$ such that*

$$\mu(\mathcal{O}, \Sigma, \mathcal{E}) = \min\{\, \mu(\mathcal{O}, \Sigma, \mathcal{E}') \mid \mathcal{E}' \textit{ is a } k\textit{-excerpt of } \mathcal{O} \,\}.$$

To preserve the largest possible amount of semantic information in a k-excerpt, it would be preferable to extract k-excerpts that have the lowest ldiff-value among all the subsets of size k. However, it is difficult in general to compute all such excerpts in an exhaustive way as all the $\binom{|\mathcal{O}|}{k}$ many subsets of size k would have to be enumerated.

4 Extraction Techniques

In this section, we introduce two different k-excerpt extraction approaches. One is based on the simple intuition that axioms comprising more elements from Σ should be preferred to be included in an excerpt for Σ. The other approach is inspired by ideas from the area of Information Retrieval (IR): we view each axiom in $\mathcal{O}$ as a document, and the input signature Σ as the set of keywords from a query. The top-k retrieved documents for the given keywords then correspond to a k-excerpt. These two approaches share a common methodology in the sense that they define a "similarity" between each axiom w.r.t. a given signature such that selecting the k axioms closest to the given signature results in a k-excerpt. We make this idea more precise in the following definition.

Definition 6. *Let $\mathcal{O}$ be an ontology and let $\Sigma \subseteq \mathsf{sig}(\mathcal{O})$. Additionally, let s be a function that maps every pair (α, Σ) consisting of an $\mathcal{EL}$-axiom α together with a signature Σ to a real number. We define the* ranking $\rhd$ of axioms w.r.t. Σ induced by s *as follows: $\alpha \rhd \beta$ if, and only if, $s(\alpha, \Sigma) > s(\beta, \Sigma)$. Given an integer k, we define a* k-excerpt of an ontology $\mathcal{O}$ for a signature Σ under s *as the set $\{ \alpha \in \mathcal{O} \mid |\{ \beta \in \mathcal{O} \mid s(\beta, \Sigma) > s(\alpha, \Sigma) \}| \leq k \}$.*

A k-excerpt hence consists of those axioms α in $\mathcal{O}$ for which there are at most $k-1$ axioms β in $\mathcal{O}$ that precede α w.r.t. $\rhd$. Note that such a definition leaves the possibility that k-excerpts can contain more than k axioms due to an equivalent distance of several axioms w.r.t. Σ. In real-world applications there would exist different remedies to such a situation. Since we aim to compare different excerpt extraction techniques in this paper, we choose to apply a random cut whenever there are more than k axioms contained in a k-excerpt.

4.1 Common Signature based k-Excerpts

A naïve extraction method for k-excerpts w.r.t. a signature Σ simply consists in a random selection of k axioms from the considered ontology. As a first improvement of the random selection, it is possible to guide the selection of the axioms by considering the number of concept and role names shared by an axiom and Σ, defined formally as follows:

Definition 7. *Given an axiom α and a signature Σ, the COM-similarity between α and Σ is defined as $s_{com}(\alpha, \Sigma) = |\mathsf{sig}(\alpha)\ \cap\ \mathsf{sig}(\Sigma)|$.*

Example 3. Let α_1, α_2, α_3, α_4 be four axioms defined as in Ex.1 and let $\Sigma = \{A, B, r\}$. Then we have $s_{com}(\alpha_1, \Sigma) = 3$, $s_{com}(\alpha_2, \Sigma) = 2$, $s_{com}(\alpha_3, \Sigma) = 2$, and $s_{com}(\alpha_4, \Sigma) = 3$. Therefore, the ranking of the axioms will be: $\alpha_1, \alpha_4 \rhd \alpha_2, \alpha_3$. The first and the last axiom are ranked higher than the other two, but no preference between α_1 and α_4, or between α_2 and α_3, exists.

4.2 Information Retrieval based *k*-Excerpts

In IR vector representations of documents and queries are a fundamental tool to model problems, based on which different retrieval strategies can be applied. We first define the vector representation for axioms and signatures.

In the remainder, we assume that every ontology $\mathcal{O}$ is associated with a strict total order $\prec$ on the elements of $\mathsf{sig}(\mathcal{O})$. Whenever we want to access the i-th signature element of $\mathcal{O}$ we refer to the i-element w.r.t. the assumed order $\prec$, starting from the smallest element. For a signature $\Sigma \subseteq \mathsf{sig}(\mathcal{O})$ or axiom $\alpha \in \mathcal{O}$, we can define the signature vector of Σ and the axiom vector of α as follows:

Definition 8 (Signature and Axiom Vector). *For a signature $\Sigma \subseteq \mathsf{sig}(\mathcal{O})$, the signature vector of Σ, written $\overrightarrow{\Sigma} = [v_1, v_2, \cdots]$, is a vector of length $|\mathsf{sig}(\mathcal{O})|$ such that $v_i = 1$ if the i-th element of $\mathsf{sig}(\mathcal{O})$ appears in Σ, otherwise $v_i = 0$. Similarly, for an axiom $\alpha \in \mathcal{O}$ we define $\overrightarrow{\alpha} = \overrightarrow{\mathsf{sig}(\alpha)}$.*

Example 4. Let $\mathcal{O}$ be the ontology defined as in Ex. 1, and let $\Sigma = \{A, B, r\}$. We assume the strict total order $\prec \subseteq \mathsf{sig}(\mathcal{O}) \times \mathsf{sig}(\mathcal{O})$ given by $A \prec B \prec X \prec Y \prec r \prec s$. We obtain the following signature vector for Σ and axiom vectors for each axiom of $\mathcal{O}$:

$$\begin{array}{lll} \overrightarrow{\Sigma} = [1,1,0,0,1,0] & \overrightarrow{\alpha_1} = [1,1,1,0,1,0] & \overrightarrow{\alpha_2} = [1,1,0,0,0,0] \\ & \overrightarrow{\alpha_3} = [1,1,1,0,0,0] & \overrightarrow{\alpha_4} = [1,1,1,1,1,1] \end{array}$$

We can then define the distance between an axiom and a signature by a distance measure between the axiom and signature vectors. A first measure is the cosine value, resulting in the COS-k-module.

Definition 9 (COS-distance between Axiom and Signature). *Given an axiom α and a signature set Σ, the COS-distance between α and Σ is defined as*

$$d_{cos}(\alpha, \Sigma) = \cos(\overrightarrow{\alpha}, \overrightarrow{\Sigma}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}},$$

where $\overrightarrow{\alpha} = [x_1, x_2, ..., x_n]$ and $\overrightarrow{\Sigma} = [y_1, y_2, ..., y_n]$.

Example 5. Let $\mathcal{O}$ be the ontology as defined in Ex. 1, let $\prec$ be the total order on $\mathsf{sig}(\mathcal{O})$ as defined in Ex. 4, and let $\Sigma = \{A, B, r\}$. Then we have that:

$$\begin{array}{ll} d_{cos}(\alpha_1, \Sigma) = 3/(\sqrt{4}\sqrt{3}) \approx 0.8660 & d_{cos}(\alpha_2, \Sigma) = 2/(\sqrt{2}\sqrt{3}) \approx 0.8164 \\ d_{cos}(\alpha_3, \Sigma) = 2/(\sqrt{3}\sqrt{3}) \approx 0.6667 & d_{cos}(\alpha_4, \Sigma) = 3/(\sqrt{6}\sqrt{3}) \approx 0.707 \end{array}$$

Therefore, the ranking of the axioms will be $\alpha_1 \rhd \alpha_2 \rhd \alpha_4 \rhd \alpha_3$.

5 Evaluation

In this section, we present a detailed evaluation of the proposed excerpt extraction techniques. To this end, we implemented the excerpt extraction methods,

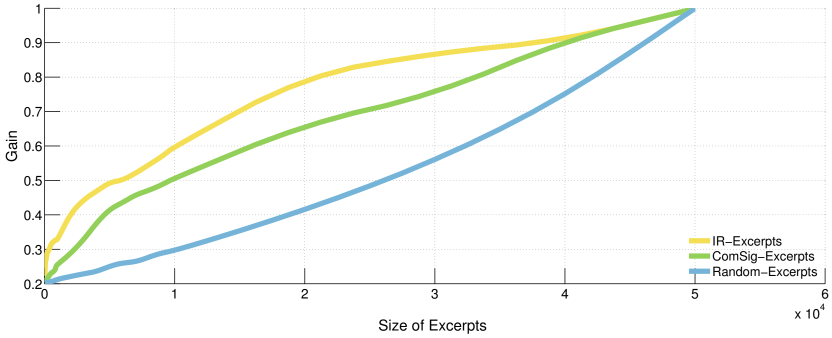

Fig. 2. Gain values of k-excerpts ($0 \leq k \leq 50\,034$) of the SNOMED CT fragment for different excerpt extraction techniques

and we compare them on SNOMED CT using a normalized evaluation metric based on ldiff. To speed up the experiments, we use a module of SNOMED CT for a randomly generated signature. The considered fragment consists of 50 034 axioms and it contains 50 587 concept and role names. We employ the CEX2.5 system as evaluation tool for the logical difference, which can currently process acyclic terminologies only. However, the proposed extraction techniques work for ontologies formulated in any DL.

As baseline, we use a *random choice* strategy which randomly selects k axioms from the input ontology to extract a k-excerpt. To estimate the quality of excerpts $\mathcal{E}$, we make use of the following metric, named Gain (G):

$$G_{\mathcal{O}}(\mathcal{E}, \Sigma) = 1 - \frac{\mathsf{ldiff}(\mathcal{O}, \Sigma, \mathcal{E})}{|\Sigma \cap \mathsf{sig}(\mathcal{O}) \cap \mathsf{N_C}|},$$

Note that Gain is inverse to ldiff normalized by the total number of possible witness concept names. Intuitively, the higher the Gain value of an excerpt $\mathcal{E}$ for a signature Σ, the more semantic information is preserved by $\mathcal{E}$.

Fig. 2 reports on the results for the different excerpt extraction techniques on the considered ontology. The excerpts were generated w.r.t. ten randomly generated signatures, containing 100 concept names and 30 role names. The values along the x-axis represent the parameter k, i.e. the excerpt size, whereas the average Gain value of the corresponding k-excerpts over the 10 signatures is shown along the y-axis. From the chart, one can see that the Gain values for IR-based excerpts (yellow) are higher than or equal to the values for other excerpt extraction strategies, and that the Gain values for common signature-based excerpts (green) lie above the values for the random choice strategy (blue). The gain value reaches 1 in Fig. 2 only for values of k approaching the size of the input ontology $\mathcal{O}$. However, extracting excerpts of size larger than the size of the module $\mathsf{Mod}^*_{\mathcal{O}}(\Sigma)$ of $\mathcal{O}$ for a signature Σ is of limited use as selecting the axioms in the module is sufficient to obtain a Gain value of 1.

We note that the IR-based excerpt extraction technique does not guarantee to select axioms that belong to the module $\mathsf{Mod}^*_{\mathcal{O}}(\Sigma)$. The average Gain values in Fig. 2 can be improved upon by ensuring that only axioms from $\mathsf{Mod}^*_{\mathcal{O}}(\Sigma)$ are

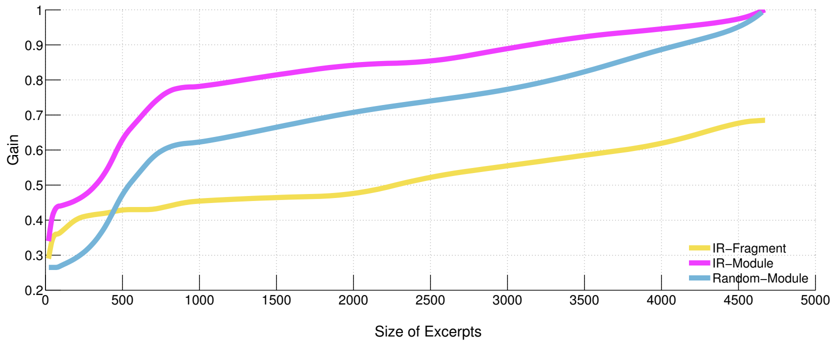

Fig. 3. Gain values of k-excerpts ($0 \leq k \leq 4673$) of the SNOMED CT fragment and one of its modules for IR-based and random excerpt extraction techniques

included in the excerpts as it is shown in Fig. 3. The curves in Fig. 3 show the difference between the Gain values for the IR-excerpts extracted from the entire SNOMED CT fragment (yellow) and those from the module of this fragment for the input signature (pink). We can see that the quality of IR-excerpts in Fig. 3 is improved significantly when using the module as an input, reaching 1 when k is close to the size of the module (i.e. 4 673) already. As baseline reference for selecting axioms from the module, the Gain values of the random choice strategy are also given (blue). Note that selecting about 500 axioms at random from the module already yields an excerpt of better quality than the IR-based excerpt of the same size of the entire ontology. One can see that using a precomputed module can improve the quality of the extracted excerpts. Excerpt extraction can thus benefit from state-of-the-art module computation tools as well.

However, for obtaining best excerpts (cf. Definition 4), we cannot solely base excerpt extraction techniques on algorithms for computing modules. We call $\mathcal{M}$ a *module-based k-excerpt for a signature Σ of an ontology $\mathcal{O}$* if there is a subsignature $\Sigma' \subseteq \Sigma$ such that $\mathcal{M} = \mathsf{Mod}^*_{\mathcal{O}}(\Sigma')$ and $|\mathcal{M}| = k$. Note that each such Σ' determines an excerpt size k with $k = |\mathsf{Mod}^*_{\mathcal{O}}(\Sigma')|$. The results depicted in Fig. 4 show that module-based excerpts (orange) are not of better quality than IR-based excerpts (pink). The values along the x-axis denote the size of subsignature Σ'. The values along the y-axis denote the average Gain values over the excerpts for ten randomly selected subsignatures Σ' of each size. The Gain values for the module-based excerpts are the values $G_{\mathcal{O}}(\mathcal{M}, \Sigma)$ with $\mathcal{M} = \mathsf{Mod}^*_{\mathcal{O}}(\Sigma')$ for every Σ', whereas $G_{\mathcal{O}}(\mathcal{E}, \Sigma)$ are the Gain values for the IR-based k-excerpts $\mathcal{E}$ for $k = |\mathsf{Mod}^*_{\mathcal{O}}(\Sigma')|$.

As a comparison, we include in Fig. 4 the average Gain values for the IR-based excerpt w.r.t. the entire SNOMED CT fragment (yellow). For small sizes of subsignatures Σ' up to 32 concept names and 30 role names, the IR-based excerpt of size $|\mathsf{Mod}^*_{\mathcal{O}}(\Sigma')|$ performs better than the corresponding module-based excerpt, even though the IR-based excerpt is not restricted to $\mathsf{Mod}^*_{\mathcal{O}}(\Sigma)$.

In our experiments so far, IR-based excerpts performed better than other extraction techniques. However, the question arises whether the IR-based ex-

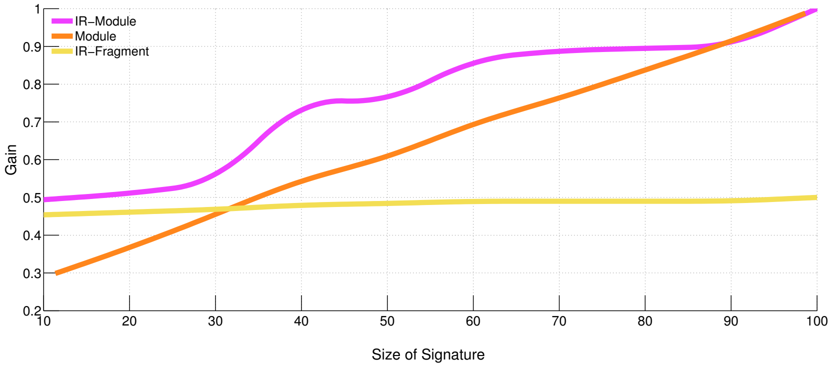

Fig. 4. Gain values of IR- and module-based excerpts of a SNOMED CT module and Gain values of IR-based excerpts of the SNOMED CT fragment

traction technique yields best excerpts. As the following example shows, this is not the case.

Example 6. Let $\mathcal{O}$ consist of the three axioms $\alpha_1 = A_1 \sqsubseteq B_1 \sqcap \exists r.X$, $\alpha_2 = A_3 \sqsubseteq A_2 \sqcap B_3$, $\alpha_3 = A_2 \sqsubseteq B_2$, and let $\Sigma = \mathsf{sig}(\mathcal{O})$. Then the ldiff-values for all 1- and 2-excerpts of $\mathcal{O}$ are respectively shown in the left- and right-hand side of the table below:

	$\{\alpha_1\}$	$\{\alpha_2\}$	$\{\alpha_3\}$	$\{\alpha_1, \alpha_2\}$	$\{\alpha_1, \alpha_3\}$	$\{\alpha_2, \alpha_3\}$
ldiff	4	5	6	3	4	2

The COS-distance between each of the three axioms α_i and Σ is as follows (using an implicit order on the signature elements): $d_{cos}(\alpha_1, \Sigma) \approx 0.707$, $d_{cos}(\alpha_2, \Sigma) \approx 0.612$, $d_{cos}(\alpha_3, \Sigma) = 0.5$. Thus, we obtain the following IR-ranking for the axioms: $\alpha_1 \rhd \alpha_2 \rhd \alpha_3$. Although the best 1-excerpt is $\{\alpha_1\}$, the best 2-excerpt is given by $\{\alpha_2, \alpha_3\}$ without having the highest ranked axiom α_1.

We conjecture that the size parameter k has to be an input parameter to any algorithm that aims at extracting best excerpts for a given signature.

6 Conclusion

We introduced the notion of ontology excerpts that can serve as a summary of an input ontology w.r.t. a signature of interest. We presented several strategies for excerpt extraction and we evaluated them based on how well the resulting excerpts capture the knowledge about the input signature. The extraction strategy based on IR-techniques clearly outperformed the others in our experiments.

We leave finding an algorithm for computing best excerpts as future work, for which we want to investigate the use of simulation-based techniques that are capable of identifying logical differences [4, 10].

References

1. Baader, F., Brandt, S., Lutz, C.: Pushing the $\mathcal{EL}$ envelope. In: Proceedings of IJCAI-05. pp. 364–369. Morgan-Kaufmann Publishers (2005)
2. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The description logic handbook: theory, implementation, and applications. Cambridge University Press (2007)
3. Del Vescovo, C., Klinov, P., Parsia, B., Sattler, U., Schneider, T., Tsarkov, D.: Syntactic vs. semantic locality: How good is a cheap approximation? In: Proceedings of WoMO'12. CEUR Workshop Proceedings, vol. 875. CEUR-WS.org (2012)
4. Ecke, A., Ludwig, M., Walther, D.: The concept difference for $\mathcal{EL}$-terminologies using hypergraphs. In: Proceedings of DChanges'13. CEUR Workshop Proceedings, vol. 1008. CEUR-WS.org (2013)
5. Ghilardi, S., Lutz, C., Wolter, F.: Did i damage my ontology? a case for conservative extensions in description logics. In: Proceedings of KR2006. pp. 187–197. AAAI Press (2006)
6. Grau, B.C., Horrocks, I., Kazakov, Y., Sattler, U.: Modular reuse of ontologies: theory and practice. JAIR 31, 273–318 (2008)
7. Konev, B., Ludwig, M., Walther, D., Wolter, F.: The logical difference for the lightweight description logic $\mathcal{EL}$. JAIR 44, 633–708 (2012)
8. Konev, B., Lutz, C., Walther, D., Wolter, F.: Semantic modularity and module extraction in description logics. In: Proceedings of ECAI'08. Frontiers in Artificial Intelligence and Applications, vol. 178, pp. 55–59. IOS Press (2008)
9. Konev, B., Walther, D., Wolter, F.: The logical difference problem for description logic terminologies. In: Proceedings of IJCAR'08. LNCS, vol. 5195, pp. 259–274. Springer (2008)
10. Ludwig, M., Walther, D.: The logical difference for $\mathcal{ELH}^r$-terminologies using hypergraphs. In: Proceedings of ECAI 2014 (2014)
11. Lutz, C., Wolter, F.: Deciding inseparability and conservative extensions in the description logic $\mathcal{EL}$. Journal of Symbolic Computation 45(2), 194–228 (Feb 2010)