Towards Extracting Ontology Excerpts

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

[1] College of Computer Science and Technology, Jilin University, China
chenjy12@mails.jlu.edu.cn

[2] Theoretical Computer Science, TU Dresden, Germany
{michel, dirk}@tcs.inf.tu-dresden.de

[3] Laboratoire de Recherche en Informatique, Université Paris-Sud, France
yue.ma@lri.fr

Abstract. In the presence of an ever growing amount of information, organizations and human users need to be able to focus on certain key pieces of information and to intentionally ignore all other possibly relevant parts. Knowledge about complex systems that is represented in ontologies yields collections of axioms that are too large for human users to browse, let alone to comprehend or reason about it. We introduce the notion of an ontology excerpt as being a fixed-size subset of an ontology, consisting of the most relevant axioms for a given set of terms. These axioms preserve as much as possible the knowledge about the considered terms described in the ontology. We consider different extraction techniques for ontology excerpts based on methods from the area of information retrieval. To evaluate these techniques, we propose to measure the degree of incompleteness of the resulting excerpts using the notion of logical difference.

1 Introduction

Ontologies based on Description Logics (DL) [2] have become a well-established paradigm used in the Web Ontology Language OWL [11] and by several biomedical ontologies like CPO, FMA, GALEN, SNOMED CT, etc. An increasing number of ontologies of large sizes have been developed and made available in repositories such as the NCBO Bioportal.[1] Ensuring efficient access to the knowledge contained in such ontologies has become an import concern.

The sheer size of some real-world ontologies is too large for human users to browse, let alone to comprehend or reason about it. Also, for automated reasoning systems these tasks could be challenging to accomplish within certain resource bounds. To facilitate the reuse of the knowledge contained in ontologies, module extraction [4] and approximate reasoning techniques [10], among others, have been suggested.

Partially supported by German Research Foundation (DFG) within the Cluster of Excellence ‘cfAED’.

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

A module $\mathcal{M}$ of an ontology $\mathcal{O}$ for a signature Σ, i.e. a set of concept and role names, is a subset of $\mathcal{O}$ that preserves the knowledge of the terms in Σ. The idea is that $\mathcal{M}$ can serve as a substitute for $\mathcal{O}$ regarding the terms in Σ. The smaller the module compared to the size of the ontology, the better it can be understood by a human user, and the more efficiently it can be distributed and reasoned with. Typically, entailment-based modularity notions are considered [4]. The meaning of the terms in Σ is preserved when $\mathcal{M}$ and $\mathcal{O}$ give the same answers to queries about the Σ-terms. However, this module notion allows for little *control* over the number of axioms that are included in a module. Even minimal modules can be as large as the entire ontology. To influence the size of a module, our only option is to adapt the signature for which the module is extracted and the query language underlying the module notion. Generally, we have that the smaller the signature and the weaker the expressivity of the query language, the smaller the modules of an ontology are. But no strict upper bound on the module size can be guaranteed this way.

In this paper, we introduce the notion of an *ontology excerpt* as a fixed-size subset of an ontology that captures as much as possible of the "meaning" of the terms in a given signature. Ontology excerpts facilitate comprehension by human users by aiding them to focus on a relatively small part of an ontology that is relevant for a considered signature.

To evaluate the quality of ontology excerpts, we define a semantics-based measure *Gain*, using Logical Difference [5], to quantify how much semantic meaning is preserved in an excerpt w.r.t. the original ontology. The logical difference is taken to be the set of queries relevant to an application domain that produce different answers when evaluated over ontologies that are to be compared. In this paper we are only interested in concept subsumption queries.

Using an exhaustive search to find the excerpts of an ontology that best preserve the semantic information w.r.t. the ontology is futile as it involves computing all (i.e. exponentially many) subsets of the ontology. We therefore want to investigate the feasibility of using, among others, excerpt extraction techniques stemming from the area of information retrieval (IR) [9], i.e. a research area which is generally concerned with developing techniques to extract the "most relevant" documents for a query from large data sources.

2 Preliminaries

We briefly recall basic notions related to the description logic $\mathcal{ELH}$ [1], modularity of ontologies [4,6] and the logical difference between ontologies [5,7].

2.1 The Description Logic $\mathcal{ELH}$

Let $\mathsf{N_C}$ and $\mathsf{N_R}$ be mutually disjoint and 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 sets of *$\mathcal{ELH}$-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 $r, s \in \mathsf{N_R}$. $\mathcal{ELH}$-inclusions that are not of the form $r \sqsubseteq s$ are called *$\mathcal{EL}$-concept inclusions*. An *$\mathcal{ELH}$-ontology* $\mathcal{O}$ is a finite set of $\mathcal{ELH}$-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 is satisfiable. An interpretation $\mathcal{I}$ is a *model* of $\mathcal{O}$ iff $\mathcal{I}$ satisfies all axioms in $\mathcal{O}$. An axiom α *follows* from an ontology $\mathcal{O}$, written $\mathcal{O} \models \alpha$, iff for all models $\mathcal{I}$ of $\mathcal{O}$, we have that $\mathcal{I} \models \alpha$.

An *$\mathcal{ELH}$-terminology* $\mathcal{O}$ is an $\mathcal{ELH}$-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* iff it can be unfolded (i.e., the process of substituting concept names by the right-hand sides of their defining axioms terminates).

We denote the number of axioms in an ontology $\mathcal{O}$ with $|\mathcal{O}|$. 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 Logical Concept Difference

We now recall basic notions related to the logical difference [5,7] between two $\mathcal{EL}$-ontologies for $\mathcal{EL}$-inclusions as query language.

Definition 1 (Concept Inclusion Difference). *Let $\mathcal{O}_1$ and $\mathcal{O}_2$ be two $\mathcal{ELH}$-ontologies, 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 ontologies 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 [5] allow us to consider only certain inclusions of a simpler syntactic form.

Theorem 1. *Let $\mathcal{O}_1$ and $\mathcal{O}_2$ be $\mathcal{ELH}$-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 $\mathcal{ELH}$-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 the set $\mathsf{Wtn}_\Sigma(\mathcal{O}_1, \mathcal{O}_2)$ is finite as Σ is finite. Consequently, 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 $\mathsf{Wtn}_\Sigma(\mathcal{O}_1, \mathcal{O}_2) = \emptyset$ [5]. In the remainder of this paper, we use the size of the set $\mathsf{Wtn}_\Sigma(\mathcal{O}_1, \mathcal{O}_2)$ as a measure for the concept inclusion difference between $\mathcal{O}_1$ and $\mathcal{O}_2$ w.r.t. Σ. We leave investigating alternative measures which allow for a possibly more faithful representation of the logical difference for future work.

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

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

For $\Sigma = \{A, B\}$, it holds that $\mathsf{Wtn}_\Sigma(\mathcal{O}, \{\alpha_1, \alpha_2\}) = \mathsf{Diff}_\Sigma(\mathcal{O}, \{\alpha_1, \alpha_2\}) = \emptyset$ 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 hence 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 can compute and output the set $\mathsf{Wtn}(\mathcal{O}_1, \mathcal{O}_2)$ in a fully automatic way.

We still note that a new approach for computing logical differences that can also handle large cyclic terminologies has recently been introduced [3,8].

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. We are interested in gaining more control over the size of a module in order to be able to reuse the knowledge contained in an ontology in a scenario where resources are restricted in terms of cognitive ability in human users, and time and space available in technical systems.

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 (as much as possible) 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.

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

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 $\mathsf{Idiff}(\mathcal{O}, \Sigma, \mathcal{E})$ of $\mathcal{EL}$-concept inclusion difference witnesses in Σ w.r.t. $\mathcal{O}$ and $\mathcal{E}$, which is formally defined as $\mathsf{Idiff}(\mathcal{O}, \Sigma, \mathcal{E}) = |\mathsf{Wtn}_\Sigma(\mathcal{O}, \mathcal{E})|$. In the remainder of this paper we only consider this incompleteness measure. We leave investigating and comparing alternative notions of incompleteness measures for future work.

Definition 5 (Best k-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*

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

Example 2 (Ex. 1 contd.). The values $\mathsf{Idiff}(\mathcal{O}, \Sigma, \mathcal{E})$ for all 2-excerpts $\mathcal{E}$ of $\mathcal{O}$ are given in the second row of the table below.

$\{\alpha_1, \alpha_2\}$	$\{\alpha_1, \alpha_3\}$	$\{\alpha_1, \alpha_4\}$	$\{\alpha_2, \alpha_3\}$	$\{\alpha_2, \alpha_4\}$	$\{\alpha_3, \alpha_4\}$
0	2	2	2	2	2

One can thus see that $\{\alpha_1, \alpha_2\}$ is the best 2-excerpt of $\mathcal{O}$ w.r.t. Σ under Idiff.

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 Idiff-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 $\binom{|\mathcal{O}|}{k}$ subsets of size k would have to be enumerated. In the next section, we give introduce two excerpt extraction techniques and evaluate them subsequently.

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 [9]: 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 and of a signature to a real number. We can then define a ranking of axioms w.r.t. Σ that is 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 $\{ \alpha \in \mathcal{O} \mid |\{ \beta \in \mathcal{O} \mid s(\beta, \Sigma) > s(\alpha, \Sigma) \}| \leq k \}$, named a* similarity based excerpt*.*

A k-excerpt 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 such k-excerpts of $\mathcal{O}$ for Σ under s 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 (Ex. 2 contd.). Let α_1, α_2, α_3, α_4 be four axioms defined as in Example 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 (Ex. 2 contd.). Let $\mathcal{O}$ be the ontology defined as in Example 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 $A \prec B \prec X \prec Y \prec r \prec s$. Then we obtain the following signature vector for Σ and axiom vectors for each axiom of $\mathcal{O}$:

$$\overrightarrow{\Sigma} = [1,1,0,0,1,0] \qquad \overrightarrow{\alpha_1} = [1,1,1,0,1,0] \qquad \overrightarrow{\alpha_2} = [1,1,0,0,0,0]$$
$$\overrightarrow{\alpha_3} = [1,1,1,0,0,0] \qquad \overrightarrow{\alpha_4} = [1,1,1,1,1,1]$$

Then we can define the distance of an axiom and a set of signature by the distances measures 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 follows:*

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

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

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

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

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 first evaluation of the proposed excerpt extraction techniques. To this end, we implemented the previously introduced excerpt extraction methods, and we compared them on the following real-world biomedical ontologies with the help of a normalized evaluation metric based on ldiff.

We consider four prominent biomedical ontologies: SNOMED CT (SM) from IHTSDO[3] (first release of 2012), MESH[4], NCBI[5] and NCI[6] (version 10.02d). Table 1 presents the metrics of these ontologies, including the number of logical axioms as well as the number of concept names and role names.

[3] http://www.ihtsdo.org/snomed-ct/

[4] http://bioportal.bioontology.org/ontologies/MESH

[5] http://bioportal.bioontology.org/ontologies/NCBITAXON

[6] http://evs.nci.nih.gov/ftp1/NCI_Thesaurus/

	SM	MESH	NCBI	NCI	SM-f	MESH-f	NCBI-f
Nr. of logical axioms	291156	403210	847755	75239	50034	49991	51879
Nr. of concepts	291145	286380	847760	76708	50520	50888	82778
Nr. of roles	62	0	0	124	62	0	0

Table 1: Metrics of the Considered Ontologies

5.1 Experimental Setup

In our experiments, for the four considered biomedical ontologies SNOMED CT, MESH, NCBI, and NCI, we first removed non-$\mathcal{EL}$ axioms from them to be able to use the CEX2.5 tool to compute ldiff values. Note that, however, the proposed extraction techniques can operate on ontologies formulated in any DL. To speed up the experiments, we then selected fragments of SM, MESH, and NCBI, which will be denoted using a '-f' suffix as given in Table 1.

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

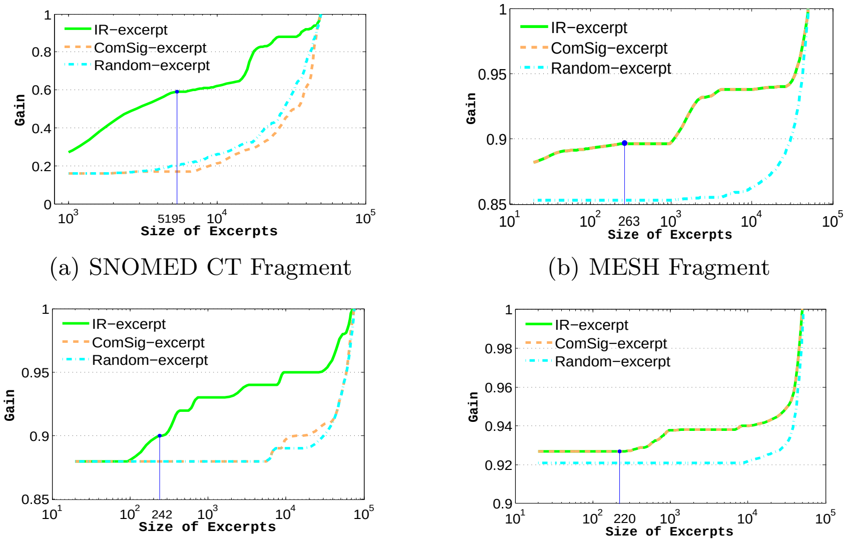
$$G_{\mathcal{O}}(\mathcal{E}, \Sigma) = 1 - \frac{\mathsf{ldiff}(\mathcal{O}, \Sigma, \mathcal{E})}{|\Sigma \cap \mathsf{sig}(\mathcal{O}) \cap \mathsf{N_C}|}.$$

That is, 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 is, the more semantic information is preserved by $\mathcal{E}$.

5.2 Results

The four charts in Figure 1 report on the results for the different excerpt extraction techniques on the considered ontologies. The values along the x-axis in each chart represent the parameter k, i.e. the excerpt size, whereas the Gain value of the corresponding k-excerpts is shown along the y-axis. The excerpts were generated for each ontology w.r.t. one randomly generated signature, containing 100 concept names and 30–50 role names in the case of SM and NCI, and 1 000 concept names and no role names for the remaining two ontologies. The vertical line in each chart represents the size of the locality-based module for the signatures.

From the charts 1(a)–1(d) one can see that the Gain values for IR-based excerpts are higher than or equal to the values for other excerpt extraction strategies. In the case of the NCBI and MESH ontologies, one can observe that the ComSig- and IR-based excerpts result in the same Gain values. Indeed, these two strategies yield the same axiom ranking if the signature of all the axioms contains the same number of signature elements, which is the case for NCBI and MESH (each axiom is of the form $A \sqsubseteq B$ for concept names A and B). In all, we



(a) SNOMED CT Fragment

(b) MESH Fragment

(c) NCI

(d) NCBI Fragment

Fig. 1: Gain-Measure for k-Excerpts of Various Ontologies

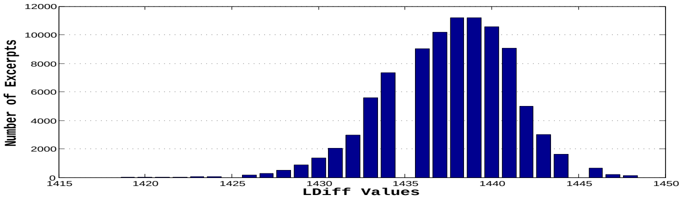


Fig. 2: Distribution of 93170 Random Excerpts over Respective ldiff-Values

can conclude that the excerpts produced by the IR-technique consistently have higher Gain values than excerpts obtain by using the other two methods on the tested ontologies and signatures.

To better understand the distribution of Gain values that we have observed for the ontology MESH (cf. Chart 1(c)), we performed an experiment in which we randomly extracted excerpts and computed their ldiff-value w.r.t. a considered signature containing 5 000 concept names and no role names such that the corresponding locality-based module contained 1610 axioms. To limit the search space, we selected a subset of MESH containing 2 491 axioms, from which we randomly extracted 93 170 many k-excerpts, for $k = 100$. Indeed, for an ontology of that size and a k-value of 100, there exist around 6.2×10^{180} k-excerpts, which renders an exhaustive search through all the excerpts impossible. The results that we obtained are summarized in Table 2. The total number of possible

Nr. of Excerpts	ldiff-Value Intervals					
	[1 419, 1 420]	[1 421, 1 425]	[1 426, 1 430]	[1 431, 1 435]	[1 436, 1 440]	[1 441, 1 448]
6.2×10^{180}	2.14×10^{-5}	0.16	3.45	19.26	56.03	21.10

Table 2: Percentage of k-Excerpts Falling into Various ldiff-Value Intervals

k	Nr. of Excerpts	ldiff-Value Intervals							
		[0, 2]	[3, 5]	[6, 8]	[9, 11]	[12, 14]	[15, 17]	[18, 20]	[21, 23]
1	19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
2	171	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
3	969	0.00	0.00	0.00	0.00	0.00	0.00	0.10	99.90
4	3 876	0.00	0.00	0.00	0.00	0.00	0.00	0.52	99.48
5	11 628	0.00	0.00	0.00	0.00	0.00	0.02	1.52	98.46
6	27 132	0.00	0.00	0.00	0.00	0.00	0.10	3.46	96.44
7	50 388	0.00	0.00	0.00	0.00	0.03	0.34	6.67	92.97
8	75 582	0.00	0.00	0.00	0.00	0.10	0.87	11.45	87.58
9	92 378	0.00	0.00	0.00	0.00	0.31	1.94	17.89	79.87
10	92 378	0.00	0.00	0.00	0.00	0.78	3.96	25.58	69.68
11	75 582	0.00	0.00	0.00	0.00	1.78	7.63	33.31	57.28
12	50 388	0.00	0.00	0.00	0.05	3.83	13.79	38.80	43.52
13	27 132	0.00	0.00	0.00	0.35	8.10	22.56	39.18	29.81
14	11 628	0.00	0.00	0.01	1.94	16.10	31.49	32.68	17.78
15	3 876	0.00	0.00	0.70	7.22	27.73	34.73	20.92	8.69
16	969	0.00	0.00	4.75	19.09	37.36	26.73	8.98	3.10
17	171	0.00	2.34	18.71	34.50	31.58	10.53	1.75	0.58
18	19	0.00	36.84	31.58	26.32	5.26	0.00	0.00	0.00
19	1	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3: Percentage of k-Excerpts Falling into Various ldiff-Value Intervals

k-excerpts is given in the first column, and the ldiff-values that we observed, except for the value 1 416, were regrouped into several intervals that are shown in the 6 right-most columns of the table. The percentage of k-excerpts whose ldiff-value fell into the respective intervals is shown in the second row of these columns.

Figure 2 shows the distribution of the Gain value over the 93 170 excerpts of the MESH fragment, i.e. each bar in the chart shows the number of excerpts that have the ldiff-value shown on the x-axis that is associated with the bar (no excerpts having an ldiff-value of 1435 or 1445 were found). We note that the excerpt extracted using the IR-technique had a Gain value (lowest ldiff-value of 1 416) that was higher than the values of all the random excerpts we extracted.

Judging from the experimental results that we obtained so far, one could draw the conclusion that excerpts produced by the IR-technique appear to result in high Gain values (i.e. low ldiff-values) in general. To test this hypothesis, we conducted another experiment in which we limited the size of the ontology in such a way that an exhaustive enumeration of all its excerpts is feasible.

We performed an exhaustive computation of all the k-excerpts, with $1 \leq k \leq 19$, together with the ldiff-values of a fragment $\mathcal{O}_f$ of SNOMED CT that contains 19 axioms, using $\Sigma = \mathsf{sig}(\mathcal{O}_f)$ as signature. For every $1 \leq k \leq 19$ we also computed the excerpt returned by the IR method for Σ. The results that we obtained are shown in Table 3. The first column indicates the value of k and the total number of possible k-excerpts is given in the second column. The 24 ldiff-values that we observed were then regrouped into 8 intervals of three elements, and the percentage of k-excerpts whose ldiff-value fell into the respective intervals is shown in the last 8 columns. The interval that contained the ldiff-value for the excerpt computed by the IR-method is indicated using a background coloured in gray. One can see that in none of the cases for $k < 19$, the k-excerpt obtained using the IR-based technique had the lowest ldiff-value. In other words, the IR-based technique fails to extract the best excerpt for $k < 19$.

The previous experiment has thus established that our hypothesis was wrong, i.e. the IR-based technique cannot guarantee to find the best excerpts in every case. Moreover, we can derive an ever stronger conclusion using the following example.

Example 6. Let $\mathcal{O}$ consist of the following three axioms:

$$\alpha_1 : A_1 \sqsubseteq B_1 \sqcap \exists r.X, \qquad \alpha_2 : A_3 \sqsubseteq A_2 \sqcap B_3, \qquad \alpha_3 : A_2 \sqsubseteq B_2$$

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.

As the example shows, an extraction technique that is based on assigning a unique and static (i.e. independent of the excerpt size k) ranking to all the axioms contained in the input ontology cannot be used to extract the best k-excerpts for every value of k. 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 have introduced the notion of ontology excerpts as a fixed-size subset of an input ontology w.r.t. a signature of interest. We have 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 involving large ontologies. However, this work is a first application of IR-techniques to excerpt extraction. A more extensive evaluation is needed to investigate the advantages of IR-techniques.

We also showed, however, that a static axiom ranking technique (assigning unique rankings) cannot be used in general to obtain best excerpts for every excerpt size. 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 [3,8].

References

1. F. Baader, S. Brandt, and C. Lutz. Pushing the $\mathcal{EL}$ envelope. In *Proceedings of IJCAI'05*. Morgan-Kaufmann Publishers, 2005.
2. F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The description logic handbook: theory, implementation, and applications*. Cambridge University Press, New York, NY, USA, 2007.
3. A. Ecke, M. Ludwig, and D. Walther. The concept difference for $\mathcal{EL}$-terminologies using hypergraphs. In *Proceedings of DChanges'13*. CEUR-WS.org, 2013.
4. B. C. Grau, I. Horrocks, Y. Kazakov, and U. Sattler. Modular reuse of ontologies: theory and practice. *JAIR*, 31:273–318, 2008.
5. B. Konev, M. Ludwig, D. Walther, and F. Wolter. The logical difference for the lightweight description logic $\mathcal{EL}$. *JAIR*, 44:633–708, 2012.
6. B. Konev, C. Lutz, D. Walther, and F. Wolter. Semantic modularity and module extraction in description logics. In *Proceedings of ECAI'08*, volume 178 of *Frontiers in Artificial Intelligence and Applications*, pages 55–59. IOS Press, 2008.
7. B. Konev, D. Walther, and F. Wolter. The logical difference problem for description logic terminologies. In *Proceedings of IJCAR'08*, pages 259–274. Springer, 2008.
8. M. Ludwig and D. Walther. The logical difference for $\mathcal{ELH}^r$-terminologies using hypergraphs. In *Proceedings of ECAI'14*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, pages 555–560. IOS Press, 2014.
9. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
10. S. Rudolph, T. Tserendorj, and P. Hitzler. What is approximate reasoning? In *Proceedings of RR'08*, pages 150–164, Berlin, Heidelberg, 2008. Springer-Verlag.
11. W3C OWL Working Group. *OWL 2 Web Ontology Language: Document Overview*, 2009. http://www.w3.org/TR/owl2-overview/.