Usability of Expressive Description Logics – A Case Study in UMLS

R. Cornet MSc, A. Abu-Hanna PhD
Dept. Medical Informatics, Academic Medical Center, University of Amsterdam, The Netherlands

Abstract
Research in (medical) terminological knowledge representation is showing an increased interest in the family of Description Logics (DLs), as they allow for automatic reasoning. This interest is driven by an increase in demands on the quality of and reasoning ability with medical terminological knowledge. Recent advances in Computer Science have demonstrated the computational decidability and empirical tractability of quite expressive DLs. The question arises whether this expressivity is usable and useful. This paper motivates and describes an exploratory study to address this question by examining the surplus value of individual DL constructors based on an investigation of UMLS terms. Our study indicates that the disjunction and negation operators comprise very valuable extensions to current DLs. The impact of formalization depends on the involved semantic type; "Injury and Poisoning" is one of the semantic types in which a large portion of concepts will benefit from the extension.

Introduction
Terminological knowledge, such as definition of concepts, can be used as a basis to describe and code patient information -e.g. by using terms and codes corresponding to concepts-, facilitate aggregation of patient groups, and enable automated reasoning. Reasoning is an important task for supporting advanced querying of patient information, but also for automatic classification of concepts and for maintaining the consistency of the terminological knowledge itself. Reasoning however requires a formal, logic-based, representation of concepts and their relationships. Most contemporary medical terminological systems lack formal representation of concepts, if they use explicit concepts at all, although SNOMED RT/CT[1] and GALEN[2] are prominent examples of systems that use formal Description Logics (DLs) for concept representation augmented by free-text terms that designate them.
Currently, these systems are based on DLs that are computationally efficient but offer limited expressivity for formal specification of the meaning of concepts. For example, the concept designated by the term "Hemorrhoids" can be modeled as: varicosis located in rectal veins. This latter representation supports automated classification of the concept as a disease of the rectal veins, and as a varicosis. However "Hemorrhoids without complication", is indistinguishable from "Hemorrhoids" because negation (in this case: absence of complication) cannot be formally modeled in the simple DLs. Hence, automatic classification is hampered due to inexpressivity, gradually leading to inconsistencies in a model when the knowledge base becomes very large.
Expressivity comes however at a computational cost. The full expressivity of first-order logic leads to undecidability (meaning that no computer can find optimal solutions in any reasonable amount of time). Description logics, characterized by the set of operators allowed, strike a balance between the formal rigor of first-order logic and expressivity to allow for decidability. Recent advances in the field of Computer Science have identified decidable DLs that are more expressive than those currently used in medicine, raising the issue of their usability and usefulness.
In this paper we investigate the semantic types and the extent to which medical terminological systems may benefit from the use of more expressive DLs. To this end we have gathered grammatical constructions that potentially indicate an implicit meaning that could be made explicit by using a DL operator. For example, the constructions "with" and "and" indicate conjunction of concepts. As a heuristic measure of the potential usability of various constructors, we have determined the incidence of these constructions in the Unified Medical Language System (UMLS) Metathesaurus (release 2002)[3]. In addition to the overall incidence, we have also determined the incidence per semantic type, in order to determine if the usability of concept constructors is domain dependent.

Description Logics
Modeling essentially involves definition of concepts and relations (roles) between concepts. Description Logics provide a means of concept and role definition with explicit and agreed-upon semantics, as opposed to Frames, where semantics often depend on interpretation. DLs are characterized by the constructors (e.g. “and”, “not”) they allow for representing concepts and roles. SNOMED RT and CT use a DL that allows for conjunction, existential quantification, and the top-concept. GALEN uses GRAIL, that extends the DL of SNOMED with role hierarchies, inverse roles, role chaining, and transitive roles. Hence, GRAIL and SNOMED use the same concept constructors, and differ in that GRAIL allows for additional role constructors.

In the field of Computer Science, ongoing research is scrutinizing the trade-off between expressivity of representations and complexity of algorithms. This has recently demonstrated that a DL called $\mathcal{SHIQ}$ is computationally decidable and empirically tractable[4]. This language extends GRAIL by disjunction, negation and qualified number restriction, but does not allow for role chaining. The DL community currently investigates another constructor: the epistemic operator, which makes it possible to define what is *known* about a concept.

The advantage of using an expressive DL lies in the possibility to better capture the semantics of a concept explicitly and formally, thereby reducing ambiguity. This is important, as non-ambiguity is a key requirement for terminological systems[5]. Another advantage of expressive DLs is the possibility of advanced inferencing, which can contribute to maintaining a consistent terminological system, and to enhanced possibilities for querying and aggregation.

Language and formal representation

In order to determine which constructors can contribute most to formalize concepts, we have first looked at the representation of various concept constructors in natural language. We focus on concept constructors, as role constructors are rarely explicitly represented in terminological systems (see the examples in Table 1). We have limited this study to the concept constructors that have a clear representation in natural language. For example, the distinction between universal and existential quantification is only rarely made in terms describing concepts in a terminological system.

Interpretation of "and" and "or"

A study on SNOMED 3.5 has demonstrated that the use of "and" is ambiguous[6]. In about 50% of the cases, "and" represented a logical and, the other 50% represented an (inclusive or exclusive) or. The semantics of "or" were almost evenly distributed between an "inclusive or" and an "exclusive or". The implication of their study on our results is twofold. Firstly, the actual number of logical conjunctions may be significantly lower than the incidence of "and", while the incidence of disjunctions may be underestimated, as a considerable number of the "and" terms will represent a logical disjunction. Secondly, to explicitly model an "exclusive or", the use of negation is required, hence the use of negation may be higher than estimated by the inventory we describe below.

Unrevealed semantics

Many terms will not reveal their semantics as concept constructors. Some concepts hold intrinsic definitions that are not expressed in English. For example, patients are diagnosed with "Rheumatoid Arthritis" if they comply with five out of seven criteria. This can be modeled using number restriction and disjunction, but this cannot be derived from the term "Rheumatoid Arthritis". Other examples are "bilateral" (i.e. "both left and right") and "unilateral" ("left or right, but not both"). This also illustrates that the study we

Table 1: Concept and role constructors with examples in natural language

	Name	Syntax	Natural Language Examples (taken from UMLS)
Concept constructors	Conjunction	$C \sqcap D$	Acute duodenal ulcer **with** hemorrhage **and** obstruction
	Universal quantification	$\forall R.C$	Mother with other multiple birth, **all** liveborn Progestogen **only** oral contraceptive
	Disjunction	$C \sqcup D$	Tremors **and/or** seizures Open treatment of sternoclavicular dislocation, acute **or** chronic
	Negation	$\neg C$	Fistula of intestine, **excluding** rectum and anus **Non**-venereal urethritis
	Existential quantification	$\exists R.C$	Measles **with intestinal complications**
	(Un)qualified number restriction	$(\geq nR.C)$ $(\leq nR.C)$	Fracture of **eight or more** ribs Uterus with **only one** functioning horn
	Concrete Domain		Good response to steroid therapy, dosage **15 mg/day**, **one week** Birth weight **999 g or less**
	Epistemic operator	$K C$	Fever of **Unknown** Origin
Role constructors	Role Hierarchy	$Q \sqsubseteq R$	**part_of** is a specialization of **physically_related_to**
	Role Chaining	$Q \circ R$	**abnormality_of** mitral valve, mitral valve **part_of** heart ➔ **abnormality_of** heart
	Inverse Roles	R^-	**part_of** is the inverse of **has_part**
	Transitive Roles	$R \in R_+$	Phalanx **part_of** Finger; Finger **part_of** Hand ➔ Phalanx **part_of** Hand

have performed probably underestimates the added value of the various constructors.

Phrases and Exclusions

For each DL constructor considered, we generated a set of phrases that may indicate the constructor's applicability to a concept term. Based on these phrases we then collected descriptions from the English preferred terms in the UMLS. Review of the results demonstrated the existence of terms that should be excluded, as a phrase was used with apparently different semantics than those of the constructor. For example, the phrase "with or without" is neither "and", nor "or", nor "not". A set of phrases and their exclusions is presented in Table 2.

Incidence of Phrases in the Metathesaurus

We have searched the UMLS Metathesaurus (release 2002) for terms containing the phrases and excluded those matching the exclusion criteria. The number of matching English preferred terms (and hence the number of concepts) are shown in Table 2. This results in a total number of concepts that have an English preferred term indicating the semantics of the various constructors. The percentages are based on the total of 776940 concepts in the Metathesaurus, each having exactly one English preferred term.

Phrases indicating conjunction, disjunction and negation turned out to occur the most frequently .

As mentioned above, the numbers are only a first rough estimate: projecting the results found by Menconça[6], there may be up to 15000 "and" phrases that represent logical disjunctions, and about 9000 "or" phrases that represent "exclusive or", and hence imply negation. Besides, the unrevealed semantics in concepts such as "Rheumatoid Arthritis" forms another source for underestimation.

Semantic type dependency

To study whether the incidence of constructors differs for various semantic types, the complying concepts are categorized along to their semantic types. The results of this categorization are presented in Table 3, showing the semantic types with the highest incidences of terms that indicate the use of various constructors.

Whereas the overall incidence of various constructors is at most 6% (see Table 2), there are semantic types that have much higher percentages of composed concepts, especially for conjunction, disjunction and (to a lesser degree) negation. The order of highest incidence per semantic type follows the same pattern that was found in the order of the overall incidence in Table 2: conjunction (29%, in "Biological function"),

Table 2: Constructors, and phrases indicating the semantics of the constructor. The numbers refer to English preferred terms, hence the number of different concepts found in the UMLS Metathesaurus, release 2002, containing 776940 concepts.

Constructor Name	Phrase	# concepts	Exclusion criteria	# excluded concepts	# remaining concepts
Conjunction	" and " " with "	30934 19950	" **with** mention of " " **with** or without " "between … **and** "	137 709 597	45759 (6 %)
Disjunction	" or " " and/or "	19352 1135	" **or** less" " **or** more" " with **or** without "	99 480 709	19280 (2 %)
Negation	" not " "^non" "[-]non[-]" "without" "exclude" "excluding" "exclusion"	3208 991 1709 4930 24 183 28	" **not** elsewhere" " **not** specified" " with or **without** " " **without** mention of "	515 106 709 1271	8548 (1 %)
Epistemic operator	"essential " "idiopath" "known" "primary"	81 334 534 1398			2328 (0.3 %)
(Un)qualified number restriction	" at least " " at most " " or less " " or more " "exactly " "less than" "more than"	53 1 99 480 0 314 422			1339 (0.2 %)

disjunction (22%, in "Phenomenon or Process"), negation (6%, in "Injury or Poisoning"), epistemic operator (6%, in "Physical Object", this actually involves only two concepts), and number restriction (1.5%, in "Injury or Poisoning"). Manual review of the 53 "conjunction" concepts in the semantic type "Biological function" demonstrates the ambiguity issue pointed out by Mendonça[6], and the need for formalization to reduce ambiguity, as all 53 concepts turn out to represent disjunction, e.g. "Functions and Abnormal Functions". As the semantic type "Injury or Poisoning" is in the "top 5" for "and", "or", "not", and "number restriction", this seems to be a good domain for further explorations.

Now that we have presented an indication of the usability of various types of constructors in various semantic types, we take a closer look at the usefulness of explicitly modeling disjunction and negation, as these have relatively high incidences, but are currently rarely used.

Advantages of modeling disjunction

Making semantics explicit and hence reducing the ambiguity in the semantics of "and" and "or" can help the automated inferencing to improve the consistency of a terminological system. An example from the Metathesaurus: Clinical Terms Version 3 defines "open nephrostomy" and "open pyelostomy" as siblings of "open nephrostomy or pyelostomy", whereas they formally are descendants. An expressive DL could resolve this inconsistency.

Another advantage that holds especially for systems that support post-coordination of concepts and terms is that disjunction allows more detailed representation of uncertainty. Concepts such as "Inflammation caused by virus or bacterium" can only be modeled using disjunction. Without the use of disjunction, such concepts should be modeled at a more general level (e.g. "Inflammation caused by microorganism"), in which case the information is lost that the inflammation was not caused by a fungus. Likewise, more expressive queries can be built to aggregate concepts: "all inflammations that are caused by a virus or by a bacterium".

Advantages of modeling negation

Although there is much less ambiguity in terms that indicate negation than there is for "and" and "or" phrases, modeling negation will still contribute to more explicit semantics and hence reasoning. Without negation, no formal distinction is possible between absence of a phenomenon (e.g. "without infection") and the absence of mention of a phenomenon, e.g. between formal representation of "blister of ear" and "blister of ear without infection". Another valuable contribution of negation is that it makes it possible to express that concepts are disjoint. For example, "Virus", "Bacterium", and "Fungus" are all defined as "Micro-organism", but there is no means for expressing that any microorganism is

Table 3: Semantic types, their total number of concepts, and the incidence of concept constructors, ordered by descending overall incidence. The overall top-10 and the 5 highest percentages for each constructor are printed bold, the maximum in italic. For each constructor the 5 semantic types with the highest percentage are presented, as well as the 10 semantic types with the highest overall percentage of terms indicating the use of any of the constructors.

rank	TUI	Semantic Type	#	conjunc	disjunc	negat	epistem	Nr.restrict	Total %
1	T037	Injury or Poisoning	30926	**25 %**	**9 %**	***6 %***	0.1 %	***1.5%***	***35 %***
2	T065	Educational Activity	1994	**28 %**	8 %	1 %	0.3 %	0.1%	**35 %**
3	T067	Phenomenon or Process	924	10 %	***22 %***	2 %	-	-	**31 %**
4	T058	Health Care Activity	11003	**22 %**	**10 %**	2. %	0.6 %	**0.6%**	**30 %**
5	T038	Biologic Function	185	***29 %***	-	-	-	-	**29 %**
6	T061	Therap. or Prev. Proc.	63962	19 %	**9 %**	2 %	**1 %**	**0.7%**	**27 %**
7	T048	Mental or Behav. Dysf.	4729	19 %	6 %	2 %	1 %	**-**	**26 %**
8	T060	Diagnostic Procedure	10458	18 %	7 %	**3 %**	0.2 %	**0.9%**	**24 %**
9	T185	Classification	995	**22 %**	2 %	1 %	0.2 %	-	**23 %**
10	T068	Hum.-caused Phen. or Process	520	16 %	8 %	2 %	**-**	**-**	**23 %**
11	T190	Anatomical Abnormality	1995	13 %	7 %	2 %	**2 %**	-	21 %
12	T020	Acquired Abnormality	2930	14 %	5 %	**3 %**	0.5 %	0.0%	20 %
13	T047	Disease or Syndrome	48286	14 %	4 %	3 %	**2 %**	0.1%	20 %
16	T001	Organism	94	5 %	3 %	**3 %**	-	-	16 %
17	T072	Physical Object	34	6 %	3 %	-	***6 %***	-	15 %
18	T191	Neoplastic Process	12297	9 %	2 %	2 %	**2 %**	-	14 %
20	T066	Machine Activity	106	5 %	**9 %**	-	-	-	14 %
27	T057	Occupational Activity	730	8 %	2 %	2 %	-	**0.5%**	11 %
31	T033	Finding	43658	4 %	2 %	**3 %**	0.3 %	0.2%	10 %

either a virus, or a bacterium, or a fungus, but never a combination of those. Likewise, the concept "Hepatitis virus, non-A, non-B" can only be explicitly modeled using negation. If this type of knowledge is made explicit, automated inferencing can help checking and maintaining the consistency of a system, and automated classification. These are crucial issues as the size and complexity of terminological systems are continuously increasing. For post-coordination and querying, the advantages of allowing for negation are comparable to the advantages of modeling disjunction.

Discussion

Modeling a terminological system is a time and knowledge intensive effort. Admittedly the use of an expressive DL will require additional effort but the gain is found in the possibility of using automated inferencing, which will support maintaining consistency of the terminological system and automatic classification of concepts leading to trustworthy terminological systems. Although the results are only indicative, this study provides insight in the potential usability of various concept constructors, and the semantic types where the highest effect is expected.

Our study aims to contribute to the formalization of terminological systems, and is complementary to studies on the conversion of portions of the UMLS Metathesaurus to Description Logics[7], and studies on the representation of part-whole relationships in DL-based systems[8]. One issue that needs further study is the use of universal and existential quantification. We have left these constructors out of our study, because of their poor representation in natural language. However, to further strengthen the basis for formalization of terminological systems by means of an expressive representation, and to improve reasoning, these constructors will play an important role.

One of the next steps will be the actual formalization of (parts of) terminological systems, for which we have indicated appropriate semantic types. Inference engines that support reasoning with expressive Description Logics are essential for using these DLs. The Computer Science community has responded to this need by implementing such engines[9,10].

Conclusion

This study of terms in the UMLS shows that use of an expressive Description Logic can contribute to a more complete formal representation of concepts. A DL with conjunction, disjunction and negation will provide a large benefit, as these constructors are used relatively often. A number of semantic types, such as "Injury or Poisoning", have been identified, in which a considerable fraction of the concepts can be formally described using one or more of these constructors.

An important advantage of using conjunction and disjunction is the reduction of ambiguity of concepts that are described by terms containing "and" or "or".

Negation will not only contribute to explicitly express exclusion or absence, but also to state that concepts are disjoint (hence no overlap between such concepts is possible).

Modeling based on an expressive DL will be more complicated, but advantages are gained in consistency checking and automated classification, which are essential for development and maintenance of terminological systems.

Acknowledgement

This research is supported by the Dutch Organization for Scientific Research (NWO).

References

1. Spackman, K. A., K. E. Campbell, et al. (1997). "SNOMED RT: a reference terminology for health care." Proc AMIA Annu Fall Symp: 640-4.
2. Rector, A., S. Bechhofer, et al. (1997). "The GRAIL Concept Modelling Language for Medical Terminology." AIM 9: 139-171.
3. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. JAMIA. 1998;5(1):1-11.
4. I. Horrocks, U. Sattler, and S. Tobies. Practical Reasoning for Expressive Description Logics. In H. Ganzinger, editor, In: 6th International Conference on Logic for Programming and Automated Reasoning (LPAR'99), Lecture Notes in Artificial Intelligence 1705, 1999. Springer Verlag.
5. Cimino, J. J. (1998). "Desiderata for controlled medical vocabularies in the twenty-first century." Methods of Information in Medicine 37(4-5): 394-403.
6. Mendonça, E. A., J. J. Cimino, et al. (1998). "Reproducibility of interpreting "and" and "or" in terminology systems." Proc AMIA Symp: 790-4.
7. Haarslev, V. and R. Möller (2000). High Performance Reasoning with Very Large Knowledge Bases. International Workshop in Description Logics 2000 (DL2000), Aachen, Germany.
8. Schulz, S. and U. Hahn (2001). "Medical knowledge reengineering-converting major portions of the UMLS into a terminological knowledge base." Int J Med Inf 64(2-3): 207-21.
9. RACER, http://kogs-www.informatik.uni-hamburg.de/~race/, last visited July 12th, 2002
10. FaCT, http://www.cs.man.ac.uk/~horrocks/FaCT/ last visited July 12th, 2002