R. Baud, M. Fieschi, P. Le Beux, P. Ruch (eds.): *The New Navigators: From Professionals to Patients.* (Proceedings of Medical Informatics Europe 2003). Amsterdam: IOS Press, 2003, 391-396.

Using Cross-Lingual Information to Cope with Underspecification in Formal Ontologies

Werner Ceusters[a], Ignace Desimpel[a], Barry Smith[b], Stefan Schulz[c]

[a] *Language and Computing nv., Zonnegem, Belgium*

[b] *Institute for Formal Ontology and Medical Information Science, Leipzig, Germany*

[c] *Department of Medical Informatics, Freiburg University Hospital, Germany*

Abstract

Description logics and other formal devices are frequently used as means for preventing or detecting mistakes in ontologies. Some of these devices are also capable of inferring the existence of inter-concept relationships that have not been explicitly entered into an ontology. A prerequisite, however, is that this information can be derived from those formal definitions of concepts and relationships which are included within the ontology. In this paper, we present a novel algorithm that is able to suggest relationships among existing concepts in a formal ontology that are not *derivable from such formal definitions. The algorithm exploits cross-lingual information that is implicitly present in the collection of terms used in various languages to denote the concepts and relationships at issue. By using a specific experimental design, we are able to quantify the impact of cross-lingual information in coping with underspecification in formal ontologies.*

Keywords*:*

Formal ontology, multi-lingual ontologies, language driven quality control

1 Introduction

We use the term 'ontology' in what follows to refer to any theory or system that aims to describe, standardize or provide rigorous definitions for terminologies used in the medical domain. Formal methods in general, and logics such as description logics or F-logic in particular, have been used to improve the quality of ontologies in this sense [1]. However, several factors can reduce the beneficial effect of such methods, especially when they are used while building the sorts of huge ontologies characteristic of the medical domain. More relevant for this paper is the fact that, while these methods – which use mechanisms such as role-restrictions and axiomatisation – can do a good job in preventing certain sorts of *mistakes* in an ontology, they typically fail in identifying *missing information* and *underspecification*. Some, it is true, adhere to a *minimal ontological commitment* paradigm, arguing that an ontology should make as few claims as possible about the world being modeled [2]. On our view, however, the job of ontology is not the construction of simplified models; rather, an ontology should correspond to reality itself in a manner that

maximizes descriptive adequacy within the constraints of formal rigour and computational usefulness.

A truly challenging project is one in which a large number of fine-grained relationships is to be used by several ontology builders collaborating on one project. Consider the concept “traumatic joint hemorrhage”. Very few formal systems (if any) and even very few experienced medical ontologists are able to identify the underspecification that is involved in a definition such as: (hemorrhage)(which HasLocation joint)(which HasCause trauma). For the latter does not represent the fact that the trauma must have acted on the very same joint which suffered hemorrhage and that it must have occurred within a short period of time thereafter. Contrast this with the concept “post-traumatic headache”, where the condition can occur in the wake of traumas in which the head is not involved at all and after much longer time spans. Underspecification is a problem to watch out for particularly in realist ontologies that want to describe what is the case. This is because ontology builders, when trying to avoid over-generalisation, may overlook essential information. Finally there is the problem of establishing just what resources of an existing ontology should be used in defining new concepts when the ontology is updated in a multi-author editing environment.

In this paper, we describe an algorithm that uses informal cross-lingual information to detect underspecification in a very large, multi-authored medical ontology that is used for natural language understanding. We report on a quantitative analysis of its effectiveness in exploiting the existence of terms for single concepts in multiple languages as a means for discovering better formal descriptions.

2 Material and methods

2.1 Domain ontology

LinKBase® is a large-scale medical ontology developed by Language and Computing nv using the authoring environment LinKFactory® [3]. LinKBase® contains over one million language-independent medical and general-purpose concepts. These are associated with more than 4 million terms in several natural languages [4]. A *term* can consist of one or more *words*, which can function in their turn as terms for other concepts. Concepts are linked together into a semantic network structure in which some 450 different link types are used to express formal relationships. The latter are derived from formal-ontological theories of mereology and topology [5, 6], time and causality [7], and also from the specific requirements of semantics-driven natural language understanding [8, 9]. Link types form a multi-parented hierarchy in their own right. At the heart of this network is the formal subsumption relationship, but this covers in LinKBase® only some 15% of the total number of relationships involved. As such, LinKBase® has a much richer structure than do description-logic-based terminological ontologies in which the ontology builder is limited effectively to relationships such as: *is strictly narrower than* and *is strictly broader than*. LinKBase® is a living ontology in which data are changed at a rate of some 2000 to 4000 modifications a day and in such a way that concepts can be added even before they have been completely defined [10]. In addition, the set of available relationships has been periodically expanded to accommodate new demands and finer ambiguity resolution in ways which have necessitated thorough revision of its existing concept definitions.

2.2 Algorithm design

LinKBase’s TermModelling algorithm uses conceptual and linguistic information to seek out missing relationships. Input is in the form of terms in a given language. The algorithm then works by attempting to find concepts which enjoy the closest (where possible an

exact) match to these terms. To achieve this, the algorithm makes use not only of terms stored but also of generated linguistic variations and of the ontological descriptions of the corresponding concepts, whether or not these are complete. In the following paragraphs we describe the algorithm in its simplest form (which is to say: without the optimizations that had to be implemented for efficient searches over a huge ontology such as LinKBase®). We first describe the existing Find-Relation-Via-Path (FRVP) algorithm that takes concepts as input. We then explain the mechanisms by which an extended algorithm decides what concepts to present as input to FRVP on the basis of input terms.

2.2.1 Base algorithm

All concepts in LinKBase® are represented in a directed graph, the links (L1, …) representing subsumption (*isa*) and other associative relations between the concepts. FRVP uses these links to build paths through the graph starting from each given input concept (c1, …, c4) and concluding where paths intersect. At this stage, all path intersections are found, both partial (such as x1, x2, and x4) and complete (such as x3 and x5) (Figure 1).

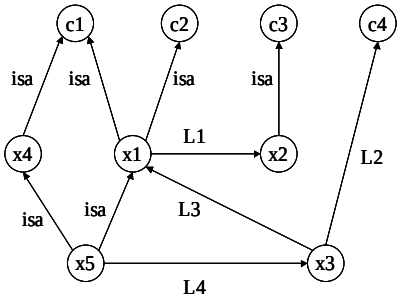

Figure 1: Find relation via path algorithm

In order to assess whether x3 is a better solution than x5, an edge-based *cost* calculation is performed. The smaller the cost related to a path, the better the solution. As such, it is easy to verify that x5 provides a closer match to the input concepts than x3. The algorithm is implemented in such a way that, when no complete intersections can be found, partial results are proposed.

The basic TermModelling algorithm is a naïve variant of the FRVP algorithm in the sense that search starts not from concepts but from terms. Given a search term T1 made out of words W1, ..., Wn, the simplest way to find the needed concepts would be to find all the concepts that have as term any substring composed of W1, … Wn. Note that for polysemous words it is already possible to find more concepts than the given number of words. This is shown in figure 2. The picture shows a search term T1 consisting of two words W1 and W2, where W1 is triply polysemous in a way which yields three distinct LinKBase® concepts. To adjust for this problem, the FRVP algorithm was modified to find the intersections of the paths between groups of one or more concepts, S1 to Sn, called sections. This modification can be viewed as if we would be applying the FRVP algorithm to each of the possible combinations of concepts associated with given words. For the example in the figure we would need to apply the FRVP algorithm three times to find the complete intersections for the concepts (c1, c4), (c2, c4) and (c3, c4). The ranking of the possible solutions is still the same. In figure 2, the complete intersections are x6, x7, x3 and x8. The complete intersection x6 will be best ranked, since all other complete intersections are reached by using incoming links from x6, regardless of the type of links involved.

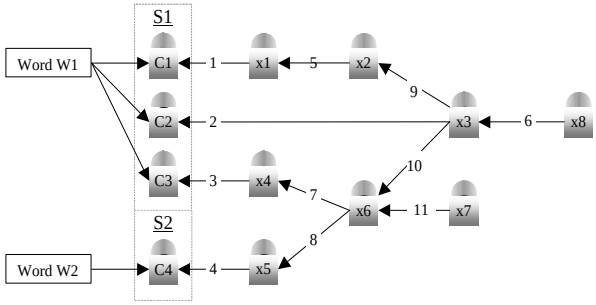

Figure 2 : Starting search from terms

2.2.2 Extended algorithm

Several extensions were required to make the TermModelling algorithm find all solutions present in the ontology. Most of them are implementations of ideas described in [8].

Problems are posed by terms containing words that are not themselves associated with LinKBase® concepts and by the verbal overspecifications of concepts involved in terms such as “dorsal back pain”, or “knee joint arthropathy”. To accommodate for these problems, the algorithm was modified in such a way that it also picks up concepts associated with terms containing only a subset of the words from the query term. The FRVP algorithm is then used to find complete intersections using this larger set of concepts. As such, the ontology structure is used as a means to validate whether the given input term makes sense at all. If words are combined that do not make sense, then intersecting paths would not be found, or at least the cost of the paths would be very long.

A second extension involves the implementation of a language-specific term generator based on inflection-, derivation-, and clause-generation rules. Again, overgeneration could be tamed by checking whether such constructed combinations of words qualify as terms for an existing concept in LinKBase®.

Most important for our purposes here is the third extension, which generates larger sections for a given word by checking the ontology also for translations and/or possible synonyms of the word and its generated words in other languages. Suppose for example that there is a concept with which the terms “pulmonary infarction” and “lung infarction” are associated, but that “pulmonary” is not a known synonym for “lung”. The extended algorithm helps out by finding an association between the term “pulmonary embolism” with the concept for which the term “lung embolism” exists. In the cross-language version of this extension the concept annotated with the English term “lung embolism” is annotated in French with the term “embolie pulmonaire”. When there is a concept annotated in French with the annotations “infarction pulmonaire” and “infarctus du poumon” (but in this case without any English annotation) and when “lung” and “poumon” are terms in English and French respectively for the same concept, then the algorithm will also find the correct concept for the term “pulmonary embolism”. This method frees us from using additional external systems such as EuroWordNet (which has a very poor medical coverage) or the UMLS (which has a minimal coverage of languages other than English).

2.3 Experiment design

To quantify the effect of using cross-linguistic information in concept search, we ran the following experiment. For six languages, we randomly selected 100 terms from LinKBase®, all of them associated with concepts for which explicit conceptual information is lacking. The languages selected were German, Spanish, Italian, French, Dutch and English for which respectively 51,238, 60,308, 80,986, 98,218, 310,197, and 1,093,607 unique terms were present in LinKBase® at the time of the experiment. In addition, output of the Morphosaurus® system was used which extracts meaningful subwords of a text, and replaces them with language-independent identifiers, the so-called MIDs [11]. In this experiment, the MIDs for a number of compound German terms were considered as forming a separate (seventh) language. The advantage of mapping compound German words to MIDs lies in the capacity of Morphosaurus® to extract content-bearing fragments from complex nouns that are not expected to be covered by conventional lexicons. MIDs behave as additional words within a term. The hypothesis was that these morphosemantic components would result in a larger number of possible matches. 41,037 MID-terms were present in the experiment.

We ran 7 tests, for each of which a separate base language was chosen. Thus for the Dutch test we applied the TermModelling algorithm first in a restricted mode, allowing it to use only Dutch together with conceptual information already existing in LinKBase®. We then reprocessed the set of 100 Dutch terms allowing in addition German to be used, then adding

Spanish, and so on. The order of the languages consecutively added was dictated by the increasing number of terms available for the languages, as it could be expected that the more terms available, the greater the amount of implicit information available to be used. As an exception, the MID-language was always added last.

For quantification purposes we used the cost function as described above. Since the TermModelling algorithm guarantees that when using additional linguistic information it can never be the case that fewer concepts are found than before (modulo the elimination of redundant subsumers), and since a lower cost related to a term-concept match reflects a closer relationship, the *gain in cost* after applying additional linguistic information is a good measure for how much implicit information could be used.

3 Results

	German		Spanish		Italian		French		Dutch		English		MIDs	
	R	A	R	A	R	A	R	A	R	A	R	A	R	A
German			0,00	0,00	0,00	0,00	3,39	0,44	0,00	0,00	0,00	0,00	0,00	0,00
Spanish	0,00	0,00			84,95	0,14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Italian	19,59	2,73	94,50	26,98			0,00	0,00	0,00	0,00	9,93	0,04	0,00	0,00
French	19,59	2,80	0,66	0,26	0,00	0,00			0,00	0,00	0,00	0,00	0,00	0,00
Dutch	40,21	5,92	4,74	1,86	0,00	0,00	88,66	11,50			90,07	0,41	0,00	0,00
English	20,62	3,23	0,10	0,04	15,05	0,02	4,36	0,64	100,00	4,60			0,00	0,00
MIDs	0,00	0,00	0,00	0,00	0,00	0,00	3,59	0,53	0,00	0,00	0,00	0,00		
	100,00	14,67	100,00	29,13	100,00	0,16	100,00	13,11	100,00	4,60	100,00	0,45	0,00	0,00

Table 1: Relative (R) and absolute (A) cost gain percentage using linguistic information.

Table 1 shows the results for all test runs. Each row represents the absolute and relative cost gain percentages obtained when allowing the TermModelling algorithm to use the terms of the language specified in the first column in addition to those preceding it when processing the terms in the base language written in the first row. As an example, allowing the Term Modelling algorithm to use Italian terms in addition to German terms when searching the ontology for concepts related to Spanish terms, leads to an absolute cost gain of 26,98%, which is 94,50% of the total cost gain using all languages when processing Spanish.

4 Discussion

Table 1 shows that the number of terms available in a specific language is an important factor for successful application: to have a gain, there must be more terms in the contributing language than in the base language. Searches in English, the language for which the most terms are available, can only be improved by 0,45%.

The results seem to indicate that the winner takes nearly all: the first language able to result in a gain always becomes the biggest contributor. German is an exception for a reason that needs further investigation. Also the behavior of the MIDs is surprising, since they do not contribute to any gain, except a very small one for French. Probably it was a bad design choice to apply them last, since they are outnumbered by the other languages, and logically ought to have been added to the algorithm before any other language. This view is further supported by the observation that a search using MIDs cannot be improved by any other language either. On the other hand, it is difficult to consider the MIDs as a language, since the version used did not allow us to give formal semantics to an individual MID, specifically not in the realist sense which is the main paradigm employed by LinKBase®.

The TermModelling algorithm has been implemented in LinKFactory® in such a way that it helps ontology builders to discover underspecifications. Take the concept annotated in

English with the term “atrium septum defect” and in French with “communication auriculaire” as an example. When the cost under a French only paradigm for finding that concept drops when allowing the algorithm to use English also, then that can only be explained by the fact that either the implicit information in “atrium septum defect” or the implicit information in “communication interauriculaire” is not adequately represented in the system. In other words: the system is not formally aware that a hole in the septum leads to a shunt from one atrium to the other. When on the other hand the cost increases, this indicates that adding linguistic information results in false positives. In this case, the system is not aware of the possibility that a specific term might have an additional meaning.

For this last reason, it is our feeling that the figures above are only a lower bound for the actual cost gain since possible false positives are not excluded from the calculations above while they are associated with a high cost. This is a point that must be further investigated.

5 Conclusion

We have shown that there is an objectively measurable value to exploiting implicit linguistic-semantic information present in multi-lingual annotations of concepts in resolving the problem of formal underspecification in ontologies. Hence, multilingual annotations are an additional means for quality assurance in ontologies, adding a dimension that cannot be covered by description logics only.

6 References

[1] Dieng R, Corby O (eds) Knowledge Engineering and Knowledge Management. Methods, Models and Tools. 12th International Conference, EKAW 2000, Juan-les-Pins, France. Springer Verlag, 2000.

[2] Gruber T. Report KSL 93-04, Knowledge Systems Laboratory, Stanford, 1993.

[3] Ceusters W, Martens P, Dhaen C, Terzic B, *LinkFactory: an Advanced Formal Ontology Management System*. Interactive Tools for Knowledge Capture Workshop, KCAP-2001, October 20, 2001, Victoria B.C., Canada (http://sern.ucalgary.ca/ksi/K-CAP/K-CAP2001/).

[4] Montyne F, *The importance of formal ontologies: a case study in occupational health.* OES-SEO2001 International Workshop on Open Enterprise Solutions: Systems, Experiences, and Organizations, Rome, 14-15 September 2001 (http://cersi.luiss.it/oesseo2001/papers/28.pdf).

[5] Smith B, *Mereotopology: a theory of parts and boundaries*, Data and Knowledge Engineering 20 (1996), 287-301.

[6] Smith B, Varzi AC, *Fiat and Bona Fide Boundaries*, in Proc. COSIT-97, Springer-Verlag 1997, 103-119.

[7] Buekens F, Ceusters W, De Moor G, *The Explanatory Role of Events in Causal and Temporal Reasoning in Medicine*, Met Inform Med 1993, 32: 274 - 278.

[8] Ceusters W, Buekens F, De Moor G, Waagmeester A, *The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition.* Met Inform Med 1998; 37(4/5):327-33.

[9] Bateman JA. *Ontology construction and natural language*. In Proc. International Workshop on Formal Ontology. Padua, Italy, 1993, 83-93.

[10] Flett A, Casella dos Santos M, Ceusters W. *Some Ontology Engineering Processes and their Supporting Technologies*, in: Gomez-Perez A, Benjamins VR (eds.) Ontologies and the Semantic Web, EKAW2002, Springer 2002, 154-165.

[11] Schulz S, Hahn U: *Morpheme-based, cross-lingual indexing for medical document retrieval.* International Journal of Medical Informatics, 2000; 58-59: 87-99