Use of OWL 2 to Facilitate a Biomedical Knowledge Base Extracted from the GENIA Corpus

Rafal Rak, Lukasz Kurgan, and Marek Reformat

Department of Electrical and Computer Engineering, University of Alberta
{rrak,lkurgan,reform}@ece.ualberta.ca

Abstract. The annotation of the GENIA corpus, a set of biomedical articles, targets the classification of biological entities based on their association with a domain-tailored taxonomy of categories. By incorporating information extraction process on the corpus we have developed a knowledge base (KB) that includes a more comprehensive taxonomy of categories, relationships between biological entities, and a hierarchy of relationships. We present our experiences in exploring the expressiveness of OWL to accommodate our KB. OWL proves to be sufficient to accommodate the extracted knowledge, however, it lacks expressive power when generalization of knowledge is needed. To this end we endorse recent endeavors in extending OWL with fuzzy description logic.

1 Introduction

The GENIA corpus [1] consists of a set of 2000 *annotated* abstracts fetched from the MEDLINE database using a query that concerns "transcription factors in human blood cells". The corpus annotation includes, but is not limited to, sentence boundaries, *biological entity* boundaries, and their association with biological categories. Biological entities (multi-word expressions that carry some biologically significant meaning) are assign one of 36 distinct categories. These categories together with additional 12 concepts that generalize them constitute the GENIA ontology. Since its development the corpus and the ontology have been extensively used by researchers in biological entity recognition, ontology creation and population, binary relation extraction, and query processing [2–5].

We propose an extension to the original GENIA ontology that not only makes the original ontology more comprehensive for reasoners but also accounts for additional information embedded in the GENIA corpus. We are particularly interested in possibilities of encoding this new knowledge in OWL, an ontology language that is becoming increasingly popular in both academic and commercial sectors. The proposed extension to the original ontology involves 1) enriching conceptually the structure of the original taxonomy of categories, 2) asserting the category membership of biological entities, 3) introducing binary relationships between biological entities, 4) building the hierarchy of relationships, and 5) connecting the ontology to an external, well-developed source of knowledge.

We employed an information extraction process on the corpus in order to evince the above mentioned features based on which the ontology is built. The process produced a set of biological entities, relationships between them, and their lexical decomposition. We also associated the biological entities with their descriptions using the UMLS Metathesaurus, a large vocabulary database that contains information about biomedical and health related concepts.

The first version of the extended GENIA ontology in OWL 1 has already been published [6]. Here we report on an ongoing work on the second version, which encompasses a more thorough knowledge base and is encoded in OWL 2.

2 Ontology Construction

The original GENIA taxonomy consists of only the declaration of classes and axioms about subclasses. The taxonomy "conceals" potentially useful pieces of information, such as: 1) there is a few *default* terminal classes that serve as placeholders for instances that do not belong to any of these classes' siblings, 2) only terminal classes have instances, and 3) an instance belongs to a single class.

The fact that the biological entities can directly belong to the terminal classes only can be embedded in the ontology by introducing *covering axioms*, such that $C \equiv D_1 \sqcup \ldots \sqcup D_n$, where $D_1, \ldots, D_n$ are subclasses of class C, i.e., if an individual is a member of class C it must also be a member of at least one of C's subclasses. We further narrow the "at least one" expression to "exactly one" by declaring that the sibling classes are disjoint, i.e., $D_1 \sqcap \ldots \sqcap D_n \sqsubseteq \bot$. The disjointness of classes is assumed by the fact that each distinct biological entity that appears in the corpus is annotated to one and only one class. The last axiom also addresses the issue of the default classes.

Asserting the membership of individuals (biological entities) is straightforward. Each annotated (and preprocessed) biological entity is a member of a class indicated by the annotation.

Due to the OWL entity naming constraints the names of individuals are encoded (but still fully understandable by humans). For the sake of clarity each individual additionally carries a *label* property that contains the original form of the biological entity.

We also introduce the `hasCUI` property that links an individual with UMLS Metathesaurus through its Concept Unique Identifiers. Following nested tags in the corpus we add additional properties, `stemsFrom` and `isRootFor`, that hold between individuals, the name one of which is lexically composed from the other.

The identification of acronyms (during the information extraction process) leads to another fact that can be stated about two individuals, one of which is an acronym of the other, namely that the two are the same.

The TBox of our ontology is extended by declaring a set of object properties that will be used to assert verb relationships between biological entities in the corpus. These relationships are fed by the relationship extraction process. The extracted triples in the form of (*subject*, *verb expression*, *object*) are used to enrich both the TBox and the ABox, by (1) deriving a hierarchical structure of object properties based on the syntax of *verb expressions*, and (2) asserting a relationship between the *subject* and the *object*, described by the *verb expression.*

The hierarchy of verb expressions is built by looking for expressions that have the same verb but different prepositions. We make an exception to this rule whenever the preposition *by* is encountered, which suggests that a verb expression with this preposition is in inverse relation to the verb alone.

The classes in our ontology mostly serve as "containers" for individuals, i.e., we cannot *directly* state anything about how they are related other than what was already discussed about their hierarchy. However, some information about classes can be inferred from the relationships between individuals, which is explicitly stated in the ontology by introducing additional axioms using *existential* and/or *universal quantifications.* By using the universal quantification construct we state that for each class C_D: $C_D \sqsubseteq \forall R.(C_{R1} \sqcup \ldots \sqcup C_{Rn})$, where $C_{Ri}, \ldots, C_{Rn}$ is a set of *filler* classes. This set of classes is obtained directly from all the relationship triples with the object property R which have an individual of class C_D on the left-hand side of the triple. The set of fillers, i.e., right-hand side classes, is obtained by looking up the membership of the right-hand side individuals. In order to enforce the *existence* of a relationship, the last axiom can be replaced with $C_D \sqsubseteq (\exists R.C_{R1} \sqcup \ldots \sqcup \exists R.C_{Rn})$.

The two types of quantifications can be combine to create a *closure axiom* of the form $C_D \sqsubseteq (\exists R.C_{R1} \sqcup \ldots \sqcup \exists R.C_{Rn}) \sqcap \forall R.(C_{R1} \sqcup \ldots \sqcup C_{Rn})$, or simply $C_D \sqsubseteq \exists R.\top \sqcap \forall R.(C_{R1} \sqcup \ldots \sqcup C_{Rn})$. The closure axioms suppress the *open world assumption* reasoning in OWL, i.e., they "close" the knowledge base to only what is known and derived from the corpus.

3 Discussion

In the previous section we tried to infer some information about relationships between classes. We achieved that by subsuming the classes with a set of quantifications, which resulted in extensive sets of axioms for each class. For some applications, however, such superfluous knowledge may actually be harmful. For example, there has been a number of attempts (e.g., [3, 5, 7], to name just a few focused exclusively on the GENIA corpus) to *generalize* knowledge extracted from text corpora. The atomic pieces of knowledge such as relationships between individuals are generalized to state certain facts about the relationships between classes. In fact, the end product usually yields a set of classes and relationships between them only, neglecting the individuals.

The process of generalizing relationships is usually based on assigning some *confidence*, with which a particular relationship holds between two classes, based on an individual membership distribution. By knowing the confidence of a relationship, useful information can be inferred about (1) the significance of such a relationship, and (2) the probability of an event that a pair of individuals will participate in this relationship.

Although the confidence of relationships can be calculated indirectly by an application that works on an OWL ontology, this type of information cannot be explicitly stated in OWL (in neither of its versions). OWL 2 provides the expressiveness of the $\mathcal{SROIQ}$ DL language, and as such, does not allow for any kind of uncertainty in individuals' membership, i.e., an individual (a pair of individuals) either belongs to a concept (a relationship) or not. This uncertainty can be introduced to the language by the fuzzy logic extension, originally proposed by [8]. The extension affects Boolean operators and quantifiers whose range is changed from the two-value set $\{0, 1\}$ to the interval $[0, 1]$.

The importance of fuzzy description logic has already been recognized. Major contributions in both defining the problem and proposing the OWL syntax of the fuzzy DL extension include [9, 10] and more recently (including new features of OWL 2) [11]. A fuzzy DL reasoner is also available as described in [12].

The capability to embed uncertainty in asserting the membership of individuals would be useful to evince some kind of generalization and, to a certain degree, assess a knowledge base.

We would like to thank Dr. Inge Christiaens for biomedical consultation.

References

1. Kim, J.D., Ohta, T., Tateisi, Y., ichi Tsujii, J.: GENIA corpus–a semantically annotated corpus for bio-textmining. ISMB **19** (2003) 180–182
2. Zhou, G.: Recognizing names in biomedical texts using mutual information independence model and SVM plus sigmoid. Int. J. Med. Inform. **75**(6) (2006) 456–467
3. Cimiano, P., Hartung, M., Ratsch, E.: Finding the appropriate generalization level for binary relations extracted from the GENIA corpus. In: Proc. of the Int. Conf. on Language Resources and Evaluation (LREC). (May 2006) 161–169
4. Rinaldi, F., Schneider, G., Kaljurand, K., Hess, M., Andronis, C., Konstandi, O., Persidis, A.: Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. Artif. Intell. Med. **39**(2) (2007) 127–136
5. Abulaish, M., Dey, L.: Biological relation extraction and query answering from medline abstracts using ontology-based text mining. Data Knowl. Eng. **61**(2) (2007) 228–262
6. Rak, R., Kurgan, L., Reformat, M.: xGENIA: A comprehensive OWL ontology based on the GENIA corpus. Bioinformation **1**(9) (2007) 360–362
7. Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J., Rojas, I.: Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In: Proc. 19th Int. Joint Conf. on AI, Edinburgh, Scotland (July 2005) 659–664
8. Yen, J.: Generalizing term subsumption languages to fuzzy logic. In: Proc. 12th Int. Joint Conf. on AI, Sydney, Australia (August 1991) 472–477
9. Gao, M., Liu, C.: Extending OWL by fuzzy description logic. In: Proc. 17th IEEE Int. Conf. on Tools with AI, Washington, DC (2005) 562–567
10. Straccia, U.: Towards a fuzzy description logic for the semantic web (preliminary report). In: Proc. of The Semantic Web: Research and Applications, Second European Semantic Web Conference, Heraklion, Greece (2005) 167–181
11. Stoilos, G., Stamou, G.: Extending fuzzy description logics for the semantic web. In: Proc. Int. Workshop of OWL: Experiences and Directions, Innsbruck, Austria (August 2007)
12. Bobillo, F., Straccia, U.: fuzzydl: An expressive fuzzy description logic reasoner. In: Proc. Int. Conf. on Fuzzy Systems, Hong Kong (June 2008) 923–930