An Ontology-Based Iterative Text Processing Strategy for Detecting and Recognizing Characters in Folktales

Thierry Declerck, Nikolina Koleva, Hans-Ulrich Krieger

DFKI GmbH, Stuhlsatzenhausweg, 3
66123 Saarbrücken, Germany
declerck@dfki.de, Nikolina.Koleva@dfki.de, krieger@dfki.de

Abstract We describe on-going work on the combined use of ontologies and natural language processing for the detection and incremental recognition of relevant character in folktales. Our approach implements various iterations of ontology-based textual analysis of a folktale that allow to first detect potential characters, to consolidate this knowledge and finally to propose a reference resolution for all the mentioning of all characters in the tale.

1. Introduction

Detecting and marking consistently through a folktale the participants that are playing a role in the story can help a lot in assigning in an automatic fashion the typical functions to characters, as those are for example described by (Propp, 1968), and we equally expect that the Proppian narrative functions can also be better automatically detected and marked-up in text, if an accurate recognition of the main participants in the story has been performed beforehand. (Lendvai et al., 2010) addresses the issue of semi-automatically assigning Proppian characters and action types to text segments mainly on the base of linguistic analysis.

In this poster/demo article, we describe a complementary approach, which relies first on a knowledge base, in the form of an ontology formalizing family relationships, which is getting populated by iterative applications of the ontology components to a linguistically annotated tale, whereas different natural language expressions referring to an unique character are marked in the iteratively updated knowledge base using the OWL[1] "sameAs" property[2]. We developed a detailed family ontology[3], which is for the time being embedded in a small folktale ontology that describes the world of the "Magical Swan Geese" tale[4]. The class hierarchy of this ontology is displayed in Figure 1, in the Appendix.

The current focus on the family ontology is guided by the fact that family relationships play a central role in many tales. Modeling other participants in tales is much more difficult, since their behavior very often do not correspond to the "normal" entities (so for example a speaking "river of milk", which acts as a "helper" in the tale). Nevertheless our approach allows detecting also such entities as characters.

2. Knowledge-based Reference Resolution

A problematic issue in processing folktales is the detection and corresponding annotation of co-referring expressions in text. Folktales are particular in this respect, since people (or characters) are relatively rarely mentioned by name, but are prevalently introduced by their function ("the King"), family status ("the father") or their mere existence ("there lived a woman"). This phenomenon, together with very vague contextual spatio-temporal descriptions in text, makes the recognition of co-referring expressions on the basis of mere linguistic features quite cumbersome. This is a reason why we developed the family ontology, in order to support knowledge-based reference resolution of entities detected in text. On the

basis of this semantic resource one can store as a specific individual in the knowledge base each entity of the tales that has been associated with a particular biological or family status.

For the purpose of the knowledge-based reference resolution, we equipped the class hierarchy with a set of inference rules, which are acting in a complementary fashion to the Protégé built-in Pellet reasoner[5]. The rules ensure that instances of the classes **Man** and **Woman** are encoded as instances of the class **Parents,** and therefore are identical to instances of the classes **Father** and **Mother** in case enough evidence about the marital or biological status is given by the text. Different instances of the class **Children** are at the same time instances of the class **Siblings**, using similar heuristics as for the class **Parents**, so that family relationships extracted from text can be completed by the inference rules, and made available for the incremental analysis of the text.

Every class and relation encoded in the ontology is associated with a label in natural language (in four languages: English, German, Russian and Bulgarian)[6]. The labels serve as the interface between the ontology and the text processing system, which is in our case the NooJ platform (Siberztein, 2003).

3. A first Iteration of Text Processing: Annotation of Nominal Phrases

We process with NooJ the whole tale and mark especially all nominal phrases (NPs), being simple ("The mother"), coordinated ("a old man and a old woman") or recursive ("a river of milk flowing in banks of pudding") NPs[7].

Our textual analysis is further specifying if an NP is indefinite or definite on the basis of the determiners used ("a woman" – indefinite – vs "the mother" – definite –), at least for languages using this kind of determiners, like English, German, etc.[8] A specific property of indefinite nominal phrases as introducing discourse referents has been widely discussed in the field of computational semantics and (von Heusinger, 2000) is giving a good overview of the past discussions. In the special case of tales (Herman, 2000) is providing for examples supporting this view on indefinite nominal phrases, relating them to the introduction of characters of tales. Our actual work with NooJ is implementing some of the views described in the work of Herman. The first step of our iterative approach to text analysis is resulting in the linguistic annotation of the folktale in terms of indefinite and definite NPs.

4. A second Iteration: Storing core Elements of indefinite NPs as candidate Characters in the Knowledge Base

The next iteration is dealing then with the application of the knowledge base to the indefinite NPs in the text. The main elements of the indefinite NPs – the nouns – are extracted and compared with the labels of the classes in the ontology. So the noun "daughter" within an indefinite NP in the tale is matching the label of the class **Daughter** of the family ontology. As a consequence, this noun is stored in the knowledge base as a potential character of the tale and gets the ID "ch3" (since before this the program has identified "man" and "wife" as the first potential characters occurring in the text), marking it as an individual of the class **Daughter.** This procedure is applied to all indefinite NPs occurring in the tale.

4. A third Iteration: Applying Inference Rules to the stored candidate Characters

We apply then the inference rules described above in Section 2 to the candidate characters stored in the knowledge base. Just to give a simple example: "ch3" ("daughter") is being

automatically encoded in the ontology as an instance of the classes **Girl** and **Sister**, while the relationships to the brother is also automatically inferred. These inferences can be draw also du to the fact that after the first iteration, it appeared that the tale is mentioning only one young female person and only one young person. This iteration offers thus also a kind of consolidation of the results of the preceding ontology population procedure.

5. A fourth Iteration: Merging the stored Characters with the core Elements of definite NPs

In Figure 2 in the Appendix, the reader can see that our approach manage to map the "ch3" (resulting from the indefinite NP "their daughter") with occurrences of the string "girl" and "sister" occurring in definite NPs elsewhere in the text. This step is for sure benefiting from the results of the application of the inference rules described in Section 4. We apply further a filtering procedure: candidate characters that are mentioned only once in the text (not being matched to the content of definite NPs, for example, or not being involved as agent in an event) are deleted from the knowledge base. On this basis we can eliminate the string "a handkerchief" from the list of potential characters (as an indefinite NPs), but we can keep the string "an apple tree" and consolidate the core element "apple tree" as a character of the tale, since it occurs also in the context of a definite NP, and it is involved in an agentive action (speaking).

5. Conclusion

We demonstrate the potential benefits of the combined use of an ontology, inference rules and textual analysis for identifying characters in the relatively small (and closed) world of a folktale. While first results of our on-going work are promising, we still have to apply the approach to more tales, in other languages, and to evaluate our approach. We plan to use for this purpose the UMIREC Corpus[9]

Notes

[1] OWL (Web Ontology Language) is a formal representation language, which is nowadays widely used for describing ontologies. OWL is a W3C standard. See http://www.w3.org/2001/sw/BestPractices/Tutorials for introduction material.

[2] The "owl:sameAs" property used in OWL knowledge bases offers a formal way for stating that various expressions (represented by various URIs) are referring to an identical person (or entity).

[3] The ontology has been encoded using the Protégé editor: http://protege.stanford.edu

[4] An online English version of this Russian tale is available at: http://www.fdi.ucm.es/profesor/fpeinado/projects/kiids/apps/protopropp/swan-geese.html. We are currently extending the coverage of the ontology, behind its family relationships component, integrating for example elements belonging to the description of characters of narratives, as those are in detail described in (Zöllner-Weber, 2008).

[5] We use this reasoner since it is incorporated in the Protégé ontology editor, which we selected for designing our ontology. More details about Pellet are given at: http://clarkparsia.com/pellet/

[6] The availability of the ontology class labels in 4 languages is reflecting our additional aim in providing for multilingual ontology resources for ontology-based text mining or information extraction procedures applied to folktales in various languages. In this our effort is complementary to two use cases (on the financial domain and eGovernment) defined in the European R&D project "Monnet" (Multilingual Ontologies for Networked Ontologies, see www.monnet-project.eu).

[7] It is important to note that in the actual version of the system only referential expressions are considered. In a next version, the reference resolution work will be extended to all kind of pronominal expressions.

[8] For languages not having determiners in their set of categories, like Russian, we are investigating the detection of other linguistic features that can mark indefiniteness, as those are described for example in (Geist, 2008).

[9] See http://dspace.mit.edu/handle/1721.1/57507.

Acknowledgements

The work reported in this paper has been partly supported by the R&D project "Monnet", which is co-funded by the European Union under Grant No. 248458.

References

Geist, L. (2008). Specificity as referential anchoring: evidence from Russian. In: *Proceedings of SuB12*, Oslo: ILOS 2008, 151-164.
Herman, D. (2000). *Pragmatic constraints on narrative processing: Actants and anaphora resolution in a corpus of North Carolina ghost stories*. Journal of Pragmatics, 32(7):959–1001.

Lendvai, P., Váradi, T., Darányi, S., Declerck, T. 2010, Assignment of Character and Action Types in Folk Tales. In: *Proceedings of the NooJ 2010 Conference*.

McCrae, J., Aguado-de-Cea, L., Buitelaar, P., Cimiano, P., Declerck, T., Gomez-Perez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. (2012). *Interchanging lexical resources on the Semantic Web*. Journal on Language Resources and Evaluation (in Press).

Propp, V.J. (1968). *Morphology of the folktale*. University of Texas Press: Austin.

Silberztein, M. (2003). *Nooj Manual*. http://www.nooj4nlp.net.

Thompson, S. (1955). *Motif-index of folk-literature: A classification of narrative elements in folktales, ballads, myths, fables, medieval romances, exempla, fabliaux, jest-books, and local legends. Revised and enlarged edition*. Bloomington, Indiana University Press, 1955-58

Tuffield, M. M., Millard, D. E. and Shadbolt, N. R. (2006) Ontological Approaches to Modelling Narrative. In: *2nd AKT DTA Symposium*, January 2006, AKT, Aberdeen University.

Uther, H.-J. (2004). *The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson*. FF Communications no. 284–286. Helsinki: Suomalainen Tiedeakatemia, 2004

Zöllner-Weber, A. 2008. *Noctua literaria : a computer-aided approach for the formal description of literary characters using an ontology*. Ph.D. Thesis, Bielefeld University.

Appendix 1

Figure Fehler! Nur Hauptdokument **Screen Shot of the Ontology**

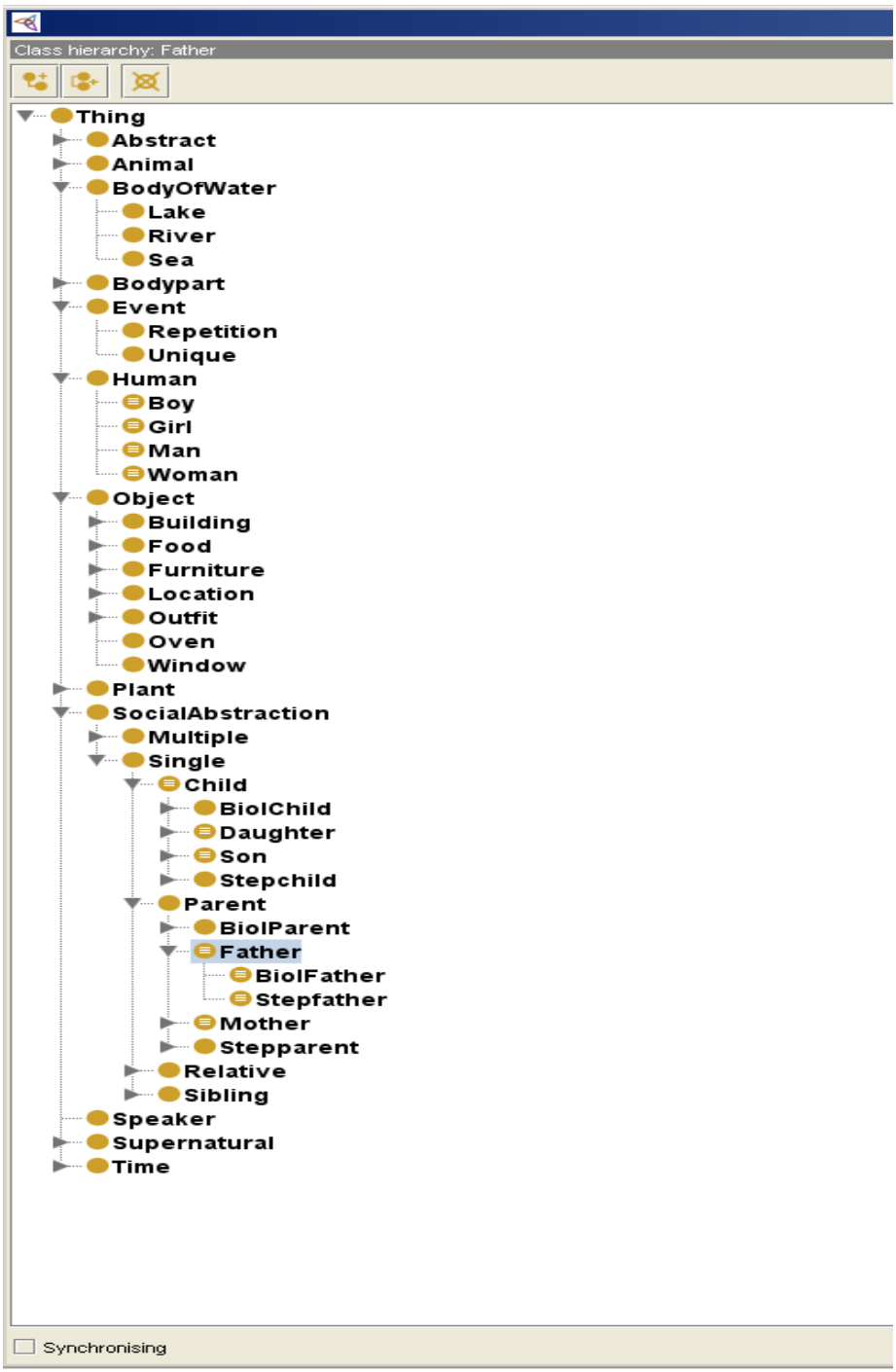

Figure 2: The knowledge base after round4: „ch3“ (Daughter) has been associated (Same individuals) with occurences of „girl“ and „sister“, as those has been identified in the phrases numbered 12, 21, 35, 85 and 91 in ur NooJ XML annotation.