Ontology-Based Incremental Annotation of Characters in Folktales

Thierry Declerck
DFKI GmbH
Stuhlsatzenhausweg, 3
66123 Saarbrücken, Germany
declerck@dfki.de

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

Hans-Ulrich Krieger
DFKI GmbH
Stuhlsatzenhausweg, 3
66123 Saarbrücken, Germany
krieger@dfki.de

Abstract

We present on-going work on the automated ontology-based detection and recognition of characters in folktales, restricting ourselves for the time being to the analysis of referential nominal phrases occurring in such texts. Focus of the presently reported work was to investigate the interaction between an ontology and linguistic analysis of indefinite and indefinite nominal phrase for both the incremental annotation of characters in folktales text, including some inference based co-reference resolution, and the incremental population of the ontology. This in depth study was done at this early stage using only a very small textual base, but the demonstrated feasibility and the promising results of our small-scale experiment are encouraging us to deploy the strategy on a larger text base, covering more linguistic phenomena in a multilingual fashion.

1 Introduction

In this submission we present on-going work dealing with the automatic annotation of characters in folktales. Focus of the investigation lies in using an iterative approach that combines an incremental ontology population and an incremental linguistic annotation. Our starting point is given by an in-house developed ontology, which is having as its core the description of family relations, but also some typical elements of folktales, like supernatural entities, etc.

The use of ontologies in the field of folktales is not new, but to our knowledge no attempt has been done so far to use ontologies in combination with natural language processing for automatizing the annotation of folktales, or for automatically populating a knowledge base of characters of folktales. The work by (Peinado et al., 2004) is dealing in first line with the Proppian functions that character can play and is also geared towards generation of interactive stories. (Zoelllner-Weber, 2008) is much closer to our aim, and in fact the author is proposing lines of research we want to implement in the near future, but her work is explicitly not dealing with the automation of annotation, and she is also not concerned with linguistic annotation in particular, but with general TEI annotation[1] of text structures.

At the present stage of development, we restricted ourselves to investigate the role indefinite and definite nominal phrases (NPs) can play for the detection of characters and their storage as instances of ontology classes. This decision is echoing well-established investigations on one possible function of indefinite NPs, namely to introduce a new referent in a discourse (see among others (von Heusinger, 2000)), whereas indefinite NPs can be used in the subsequent text to refer back to the introduced referential entities. This fact has also been acknowledged in the field of folktales and narratives and (Herman, 2000), for example, stressed the importance of analyzing sequences of referring expressions for achieving a more complete and accurate view on the role of participants in narratives.

Discourse models resulting from the sequence of referring expressions can thus support the comprehension of narratives (see (Herman, 2000),p. 962). Agreeing with this study, we further think that the automated analysis of referential expres-

[1] TEI stands for "Text Encoding Initiative, see www.tei-c.org

Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 30–34,
Avignon, France, 24 April 2012.

sions in folktales, delivering essential elements for the character models used in the interpretation of narratives, can be of help in the automated analysis of the whole folktale, and more generally for the automated analysis of narratives.

While (Herman, 2000) treats the role of anaphora used in transcripts of ghost stories, we deal (for the time being) only with the relation between characters introduced by indefinite NPs and their subsequent enunciation by definite NPs.

In the next sections we present the main components of the current version of our system. We discuss also the results of a first evaluation study, and conclude with indication on future work.

2 The Ontology

As mentioned above, our starting point is an ontology, developed at our lab. This ontology will be made publicly available, after merging it with further ontological elements relevant to the field of narratives, as those are for example described in (Zoelllner-Weber, 2008), and associating its classes and relations with elements relevant for the linguistic and semantic annotation of folktales, as described for example in (Scheidel and Declerck, 2010).

The class hierarchy and the associated relations (or properties) are equipped with natural language labels and comments. The labels are available in four languages: Bulgarian, English, German and Russian. But our work is dealing for the time being only with English.

An example of class of the ontology, with its labels is given just below:

```
<owl:Class rdf:about="#BiolDaughter">
  <owl:equivalentClass>
    <owl:Class>
       <owl:intersectionOf rdf:parseType="Collection">
           <owl:Restriction>
               <owl:onProperty rdf:resource="#hasBiolParent"/>
               <owl:onClass rdf:resource="#BiolParent"/>
               <owl:minQualifiedCardinality
               rdf:datatype="&xsd;nonNegativeInteger">
                   1</owl:minQualifiedCardinality>
           </owl:Restriction>
           <owl:Restriction>
               <owl:onProperty rdf:resource="#hasGender"/>
               <owl:hasValue>f</owl:hasValue>
           </owl:Restriction>
       </owl:intersectionOf>
    </owl:Class>
  </owl:equivalentClass>
  <rdfs:subClassOf rdf:resource="#BiolChild"/>
  <rdfs:subClassOf rdf:resource="#Daughter"/>
  <rdfs:comment>The class of biological daughther is
     a subclass of biological child and of daughter.
     This class designates all biological daughters.
     Each member of this class has gender female
     and at least one biological parent.
  </rdfs:comment>
  <dc:language xml:lang="bg">
     &#1073;&#1080;&#1086;&#1083;&#1086;&#1075;&#1080;&#1095;
     &#1085;&#1072; &#1044;&#1098;&#1097;&#1077;&#1088;&#1103;
  </dc:language>
  <dc:language xml:lang="de">
     biologische Tochter
  </dc:language>
  <dc:language xml:lang="en">
     biological Daughter
  </dc:language>
  <dc:language xml:lang="ru">
      &#1073;&#1080;&#1086;&#1083;&#1086;&#1075;&#1080;
     &#1095;&#1077;&#1089;&#1082;&#1072;&#1103; &#1044;
     &#1086;&#1095;&#1100;
  </dc:language>
</owl:Class>
```

The ontology also encodes inference rules that allow establishing automatically family relations between entities (characters) that have been stored at the instance level, which is the process of ontology population resulting from detection in the text.

1. hasParent(?x, ?x1), hasParent(?x, ?x2), hasParent(?y, ?x1), hasParent(?y, ?x2), hasGender(?x, ”f”), notEqual(?x, ?y) → Sister(?x)
2. Daughter(?d) , Father(?f) , Son(?s) → hasBrother(?d, ?s), hasChild(?f, ?s), hasChild(?f, ?d), hasSister(?s, ?d)

The first rule is about class inference. It states that if two different individuals in the ontology (represented by the variables `x` and `y`) share the same parents (represented by the variables `x1` and `x2`) and if the gender of one of the two individuals is female, then this individual can be considered as an instance of the ontology class *Sister*.

According to the second rule, various relations (or properties) can be inferred from the fact that we have three individuals (represented by the variables `d`, `f` and `s`) that are instances of the classes

Daughter, *Father* and *Son* respectively. The first inferred relations state that the *Daughter* has the *Son* as her *Brother* and the *Son* reciprocally has the *Daughter* as his *Sister*. In addition, the *Father* is being assigned twice the HASCHILD property, once for the *Daughter* and second for the *Son*.

3 Processing Steps

We submit first a folktale, here the "Magic Swan Geese"[2], to a linguistic processing engine (the NooJ platform, see (Silberztein, 2003)), applying to the text nominal phrases recognition rules (including coordination), which are differentiated in being either indefinite or definite (we do not consider pronouns for the time being). All annotated NPs are indexed with ID numbers.

In the following step, our algorithm extracts from the whole annotated text the nominal heads of the indefinite NPs and compares them with the labels present in the ontology. In case a match can be established, the nominal head of this phrase is used for populating the corresponding ontology class as an individual and the text is annotated with the nominal head being a (potential) character, as can be seen in the annotation example below, where the reader can observe that the (potential) characters are also enumerated, this time on the base of the (unique) ID they get during the ontology population phase. In this step thus, all candidate characters in text are automatically marked-up with both linguistic and character information derived from the ontology.

```
<text>
There lived
<NPCOORD id="z_coord_ph1" Nb="p" HEAD1="man"
  HEAD2="woman" Type="and">
  <NP id="indef_ph1" SPEC="a" HEAD="man"
      Gender="m" Num="s">
  <CHAR id="ch1" TYPE="man" Gender="m"
   Num="s">
  an old man</CHAR>
  </NP>
  and
  <NP id="indef_ph2" SPEC="a" HEAD="woman"
      Gender="f" Num="s">
  <CHAR id="ch2" TYPE="woman" Gender="f"
   Num="s">
  an old woman</CHAR>
  </NP>
</NPCOORD>
;
they had
<NPCOORD id="z_coord_ph2" Nb="p" HEAD1=
"daughter" HEAD2="son" Type="and">
<NP id="indef_ph3" SPEC="a"
HEAD="daughter" Gender="f" Num="s">
<CHAR id="ch3" TYPE="daughter"
Gender="f" Num="s">
a daughter</CHAR>
</NP>
  and
  <NP id="indef_ph4" SPEC="a" HEAD="son"
   Gender="m" Num="s">
<CHAR id="ch4" TYPE="son" Gender="m"
 Num="s">
a little son</CHAR>
</NP>
</NPCOORD>
```

In the next step, the inference rules of the ontology are applied to the candidate characters (the individuals stored so far in the knowledge base). In the particular tale we are analysing the class associated with the potential character `ch2` (*Woman*) can be equated with the class *Mother* and its associated string in the label (`mother`), so that all occurrences of the two strings in the text can be marked as referring to the same character.

Also some relations are established between the individuals by the application of the inference rules described in the ontology section above: Wife_of, Mother_Of, etc. (together with the strings listed in the labels). If the related strings are found in definite NPs in the text, the corresponding segment can then be annotated by

[2] http://en.wikipedia.org/wiki/The_Magic_Swan_Geese

our linguistic processing engine with the original character identifier. In the example below, the reader can see that on the base of the inference rules, the string `mother` in the definite NP (with ID DEF_PH6) is referred to `ch2` (see the first annotation example above for the first occurrence of `ch2`).

```
<NP id="def_ph6" SPEC="the" HEAD="mother"
  Gender="f" Num="s">
  the mother</NP>
 said: "Daughter, daughter, we are going
  to work; we shall bring you back
 <NP id="indef_ph7" SPEC="a" HEAD="bun"
  Gender="n" Num="s">
 <CHAR id="ch6" TYPE="bun" Gender="n"
  Num="s">a little bun</CHAR>
 </NP>
 , sew you <NP id="indef_ph8" SPEC="a"
 HEAD="dress" Gender="n" Num="s">
 <CHAR id="ch7" TYPE="dress"
  Gender="n" Num="s">
 a little dress</CHAR>
</NP>
```

The same remark is valid for the `ch3` (”a daughter” introduced in the first sentence of the tale). In the definite NP with ID 12, the string (`daughter`) is occurring in the context of a definite NP, and thus marked as referring to `ch3`. The string (`girl`) is occurring four times in the context of definite NPs (with IDs 18, 25, 56 and 60) and for all those 4 occurrences the inference driven mark-up of the nominal head with `ch3` turns out to be correct.

In this annotation example, the reader can also see that the heads of all indefinite NPs are first considered as potential characters. A preliminary filtering of such expression like `dress` is not possible, since in folktales, every object can be an actant. So for example in this tale, an `oven`, an `apple tree` or a `river of milk` are playing an important role, and are characters involved in specific actions. Our filtering is rather taking place in a post-processing phase: strings that get only once related to a potential character ID and which are not involved in an action are at the end discarded.

The next steps are dealing with finding other occurrences of the potential characters (within definite NPs), or to exclude candidates from the set.

4 Evaluation of the approach

In order to be able to evaluate our approach, even considering that we are working on a very small text base, we designed a first basic test data and annotated manually the folktale ”The magic swan geese” with linguistic and character annotation. The linguistic annotation is including coreferential information. In the longer term, we plan to compare our work applied to more folktales with a real gold standard, the UMIREC corpus (http://dspace.mit.edu/handle/1721.1/57507)

Our evaluation study shows results in terms of correct detection of tale characters in comparison with the manually annotated data. Eight of the real characters were correctly classified by the tool. Three of the instances are actually characters but they were not detected. One candidate is not a character according to the manually annotated data, but the system classified it as character. Seven entities were correctly detected as non characters. On this small basis, we calculated the accuracy of the tool, which is 79%. We also computed the precision, the recall and the F-measure. The precision amounts to 88%; the recall to 73%; and the value of the balanced F-measure is 80%. So these metrics confirm what the accuracy has been already expressing: the results are encouraging.

Looking at the errors made by the tool, we know that it does not consider the characters that are mentioned only one time. In our text, *a hedgehog* occurs only once. However, the human intuition is that it is a character and differs from the phrases *a bun* and *a dress*, which have just descriptive function. In a next version of the tool, it will be checked if the head of an indefinite NP, which is present only once in the text, is having an active semantic role, like Agent. In this case, it can be considered as a character.

Another problem of our actual approach is that we do not consider yet the possessive phrases and pronominal expressions. Precise analysis of these anaphoric expressions will improve the approach

in augmenting the number of occurrences of candidate characters. We also expect the availability of related instances in the knowledge base to help in resolving pronominal co-reference phenomena.

The applied method does not detect one of the main characters in the sample text namely the *swan-geese*. The *swan-geese* are introduced in the discourse only via a definite noun phrase. If there are some equivalent phrases, for example occurring in the title of the tale, they can be annotated as character by the tool. An additional problem we have, is the fact that our NP grammar has analyzed the words *swan* and *geese* as separate nouns and not as a compound noun. So that the linguistic analysis for English compounds has to be improved.

5 Conclusion and future work

Our in depth investigation of the interaction of an ontology and language processing tools for the detection of folktale characters and their use for incrementally populating an ontology seems to be promising, and it has allowed for example to associate a unique character ID to occurrences of different nominal heads, on the base of their inferred semantic identity. A possible result of our work would lie in the constitution of larger database containing characters of narratives extracted automatically from text.

We plan to tackle the processing of pronominal and possessive expressions for completing the co-reference task. We plan also to extend our work to other languages, and we already started to do this for anaother folktale in German, in which much more complex family relationships are involved (the German version of the tale "Father Frost"). But more challenging will be to deal with languages, which do not know have the difference between indefinite and definite NPs.

Acknowledgments

The work presented 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

Marc Cavazza and David Pizzi. 2006. Narratology for interactive storytelling: A critical introduction. In *TIDSE*, pages 72–83.

Maja Hadzic, Pornpit Wongthongtham, Tharam Dillon, Elizabeth Chang, Maja Hadzic, Pornpit Wongthongtham, Tharam Dillon, and Elizabeth Chang. 2009. Introduction to ontology. In *Ontology-Based Multi-Agent Systems*, volume 219 of *Studies in Computational Intelligence*, pages 37–60. Springer Berlin / Heidelberg.

Knut Hartmann, Sandra Hartmann, and Matthias Feustel. 2005. Motif definition and classification to structure non-linear plots and to control the narrative flow in interactive dramas. In *International Conference on Virtual Storytelling*, pages 158–167.

David Herman. 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.

Harry R. Lewis and Christos H. Papadimitriou. 1998. *Elements of the theory of computation*. Prentice-Hall.

Deborah L. McGuinness and Frank van Harmelen. 10 February 2004. OWL Web Ontology Language Overview. W3C Recommendation.

Federico Peinado, Pablo Gervás, and Belén Díaz-Agudo. 2004. A description logic ontology for fairy tale generation. In *Language Resources for Linguistic Creativity Workshop, 4th LREC Conference*, pages 56–61.

Vladimir IA. Propp, American Folklore Society., and Indiana University. 1968. *Morphology of the folktale / by V. Propp ; first edition translated by Laurence Scott ; with an introduction by Svatava Pirkova-Jakobson*. University of Texas Press, Austin :, 2nd ed. / revised and edited with a preface by louis a. wagner ; new introduction by alan dundes. edition.

Antonia Scheidel and Thierry Declerck. 2010. Apftml - augmented proppian fairy tale markup language. In Sándor Darányi and Piroska Lendvai, editors, *First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts: Poster session. International Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts (AMICUS-10), located at Supporting the Digital Humanities conference 2010 (SDH-2010), October 21, Vienna, Austria*. Szeged University, Szeged, Hungary, 10.

Max Silberztein. 2003. Nooj manual. available for download at: www.nooj4nlp.net.

Klaus von Heusinger. 2000. The reference of indefinites. In K. von Heusinger and U. Egli, editors, *Reference and Anaphoric Relations*, pages 247–265. Kluwer.

Amelie Zoelllner-Weber. 2008. *Noctua literaria - A Computer-Aided Approach for the Formal Description of Literary Characters Using an Ontology*. Ph.D. thesis, University of Bielefeld, Bielefeld, Germany, may.