Towards a WordNet based Classification of Actors in Folktales

Thierry Declerck[1,2], **Tyler Klement**[3], **Antonia Kostova**[3]

[1]DFKI GmbH, Language Technology Lab
Saarbrücken, Germany
[2]Austrian Centre for Digital Humanities (ACDH)
Vienna, Austria
[3] Saarland University, Dept. of Computational Linguistics
Saarbrücken, Germany
<declerck@dfki.de, klement.tyler@gmail.com, akostova@coli.uni-saarland.de>

Abstract

In the context of a student software project we are investigating the use of WordNet for improving the automatic detection and classification of actors (or characters) mentioned in folktales. Our starting point is the book "Classification of International Folktales", out of which we extract text segments that name the different actors involved in tales, taking advantage of patterns used by its author, Hans-Jörg Uther. We apply on those text segments functions that are implemented in the NLTK interface to WordNet in order to obtain lexical semantic information to enrich the original naming of characters proposed in the "Classification of International Folktales" and to support their translation in other languages.

1 Introduction

This short paper reports on the current state of a student software project aiming at supporting the automatized classification of folktales along the line of the classification proposed by Hans-Jörg Uther (2004). This classification scheme is considered as a central source for the analysis work of folklorists. It builds on former work by Antti Aarne (1961) and Stith Thompson (1977). In the following, we are using the acronym ATU for referring to (Uther, 2004): ATU standing for Aarne-Thompson-Uther.

We focus in the current work on the detection of common superclasses to the naming of the main actors (or characters) that are mentioned in the various types of folktales listed by Uther (2004). In doing this we are able to propose more generic classes of characters and an extended vocabulary, and so to link to other classification systems, like the Motif-Index of Folk-Literature proposed by Stith Thompson[1]. In general, we are aiming at a WordNet[2] based generation of lexical semantic relations for building a terminology network of actors/characters mentioned in folktales. Our work is anchored in the field of Digital Humanities (DH), where there is an increased interest in applying methods from Natural Language Processing (NLP) and Semantic Web (SW) technologies to literary work.

In the following sections we will present first the data we are dealing with and the transformations we applied on those for being able to use the NLTK interface to WordNet[3]. We describe then the functions of NLTK we are using and how we can benefit from those for building a more generic vocabulary and extending the basic terminology for classifying actors/characters in folktales.

Related work on this topic is presented in Declerck (2012), which is more focused on the use of Wiktionary for translation and also dealing rather with the formal representation of the terminology used in ATU.

2 The Data Source

We are taking the ATU classification scheme as our starting point. Just below we display the initial part of a *type* of folktale, which in ATU is marked using an integer, possibly followed by a letter. In this example we deal with type 2, which is included in the list of types "Wild Animal" (from type 1 to type 99), and more specifically within the list "The Clever Fox (Other Animal)" (from type 1 to type 69)[4].

[1]See the online version of the index: `http://www.ruthenia.ru/folklore/thompson/index.htm`.

[2]See (Fellbaum, 1998) and (Miller, 1995).

[3]NLTK is described in (Bird et al., 2009), with an updated online version: `http://www.nltk.org/book/`. At `http://www.nltk.org/howto/wordnet.html` the WordNet interface is described in details.

[4]See also `https://en.wikipedia.org/wiki/Aarne-Thompson_classification_systems`,

2 ***The Tail-Fisher.*** A bear (wolf) meets a fox who has caught a big load of fish. He asks him where he caught them, and the fox replies that he was fishing with his tail through a hole in the ice. He advises the bear to do likewise and the bear does. When the bear tries to pull his tail out of the ice (because men or dogs are attacking him), it is frozen in place. He runs away but leaves his tail behind [K1021]. Cf. Type 1891.
Combinations: This type is usually combined with episodes of one or more other types, esp. 1, 3, 4, 5, 8, 15, 41, 158, and 1910.

In this example, we can see the number of the type ("2"), its label ("The Tail-Fisher") and a text summarizing the typical motifs of this type of folktale. At the end of this "script", a link to a corresponding Thompson Motif-Index is provided ("[K1021]"). Finally, types are indicated, with which the current type is usually combined.

For us, a very interesting pattern in the description part of the type entry is "A bear (wolf)". This way (and also using more complex patterns), the author specifies variants of actors/characters that can play a role within a folktale type. We found this pattern interesting because our assumption is that in most of the cases only semantically related actors/characters can be mentioned in this text construct. And those pairs of variants give us a promising basis for trying to generate more generic terms from WordNet for classifying actors in folktales and so to support the linking of ATU to other classification schemes.

Our work consisted first in extracting from ATU the relevant text segments corresponding to such patterns and then to query WordNet in order to see if the characters named in such text segments are sharing relevant lexical semantic properties.

2.1 Pre-Processing the ATU Catalogue

In order to be able to apply functions of the WordNet interface of NLTK to the ATU classification scheme, we first had to transform the original document into a punctuation separated text format, using for this a Python script. For the type 6, just to present another example of an ATU type, we have now the following text format:

```
6˜Animal Captor Persuaded to
Talk.˜ A fox (jackal, wolf)
catches a chicken (crow, bird,
hyena, sheep, etc. ) and is
about to eat it. The weak animal
asks a question and the fox
answers. Thus he releases the
prey and it escapes. ˜K561.1
```

With this new format, where the sign " ˜ " is used as the separator, it is very easy to write code that is specialized for dealing with parts of the ATU entries. For our work, we concentrate only on the third field of the " ˜ " separated input file. This way we avoid the "noise" that could be generated if considering the use of parentheses in the second field (the label of the type), like:

```
Torn-off Tails (previously The
Buried Tail).
```

which is used in the label of type 2A.

2.2 Pattern Extraction

On the basis of a manual analysis of the ATU entries, regular expressions for detecting the formulation of variants of actors/characters have been formulated and implemented in Python. Below we show some examples of extracted text segments, on the basis of the Python script:

- A master (supervisor)
- an ox is so big that it takes a bird a whole day (week, year)
- A sow (hare)
- A giant has sixty daughters (sons)
- a brook (sea)
- A man puts a pot with hot milk (chocolate)
- A man who has recently been married meets a friend (neighbor, stranger)
- A wolf (bee, wasp, fly)
- A suitor (suitors)

with more details given in the French or German corresponding pages.

- a flea (fly, mouse)
- a series of animals (hen, rooster, duck, goose, fox, pig)
- a person (animal)
- An ant (sparrow, hare)

As the reader can see, each text segment starts with an indefinite Nominal Phrase (NP) and ends with a closing parenthesis. This pattern is consistently used in ATU, and corresponds to our intuition that a referent in discourse is mostly introduced by an indefinite NP. For the first step of our investigation of the use of WordNet for generating more generic terms for the mentioned actors, we decided to concentrate on the simple sequence "A/An Noun (Noun)", like for example "A fox (wolf)".

2.2.1 Accessing WordNet with the NLTK Interface

NLTK provides for a rich set of functions for accessing WordNet. The first function we applied was the one searching for the least common hypernym for the two words used in the pattern "A/An Noun (Noun)". Some few results on such a search for all the synsets of the considered noun-pairs are displayed below for the purpose of exemplification, where we indicate the least common hypernym with the abbreviation LCH:

- Synset(man.n.01) & Synset(fox.n.05) => LCH(Synset(person.n.01))
- Synset(fox.n.01) & Synset(jackal.n.01) => LCH(Synset(canine.n.02))
- Synset(fox.n.01) & Synset(cat.n.01) => LCH(Synset(carnivore.n.01))
- Synset(raven.n.01) & Synset(crow.n.01) => LCH(Synset(corvine_bird.n.01))

It is for sure interesting to see that depending on the word they are associated with, synsets of "fox", for example, can be related to a different hypernym. In the case of "fox.n.05" and "man.n.01" sharing the hypernym "person.n.01", we have to check if this case should be filtered out, since the hypernym is too generic. We tested for this the NLTK function "path_similarity", which computes a measure on the basis of the respective length of the path needed for each synset to the shared LCH. For "man.n.01" and "fox.n.05" the function "path_similarity" gives '0.2', while for "fox.n.01" and "jackal.n.01" it gives '0,33'. We might have '0,33' as a threshold for accepting the selected hypernym as a relevant generalization of the words used in the patterns of ATU we are investigating. Or allowing also lower similarity measures, but filtering out the selected hypernym on the basis of the length of the path leading from it to the root node. The LCH "canine.n.02' has a much longer path to "entity" as does the LCH "person.n.01". Our first experiments seem to indicate that the longer the path of the hypernym to the root node, the more informative is the generalization proposed by querying WordNet for the least common hypernym.

Additionally to those two functions of the NLTK interface to WordNet, we make use of the possibility to extract from WordNet all the hyponyms of the involved synsets. This can offer an extended word base for searching in folktale texts for relevant actors/characters. While this assumption seems reasonable in certain cases, like for example for the synset "overlord.n.01" for which we can retrieve hyponyms like "feudal_lord", "seigneur' and "seignior", it is not clear if it is beneficial to retrieve all the scientific names listed as hyponyms of the synset "fox.n.01", like "Urocyon_cinereoargenteus" or "Vulpes_fulva". But in any case, the terminology basis of the words used in ATU can this way be extended.

Last but not least, we take advantage of the multilingual coverage of WordNet, using for this another function implemented in NLTK. As an example, for the following pairs mentioned in ATU, we get from WordNet the French equivalents:

- Synset(fox.n.01) & Synset(wolf.n.01) => ['renard'] & ['loup', 'louve']
- Synset(dragon.n.02) & Synset(monster.n.04) => ['dragon'] & ['démon', 'monstre', 'diable', 'Diable']
- Synset(enchantress.n.02) & Synset(sorceress.n.01) => ['sorcière'] & ['enchanteur', 'ensorceleur', 'sorcière']

As part of future work, we are considering those multilingual equivalents provided by WordNet as a starting point for providing for a multilingual extension of the ATU classification.

3 An Ontology for ATU

In order to store all the results of the work described above, including the multilingual correspondences of the English terminology used in ATU, we decided to go for the creation of an ontology of ATU, a step which is also aiming at supporting the linking of this classification scheme to other approaches in the field. The ontology was generated automatically from the transformed ATU input data described in section 2.1., and encoded in the OWL and RDF(s) representation languages[5]. ATU not being a hierarchical classification, we decided to have only one class in the ontology, and to encode each type of ATU as an instance of this class. As a result, we have 2221 instances. The main class is displayed just below, using the Turtle syntax [6] for its representation:

```
:ATU
  rdf:type owl:Class ;
  rdfs:comment
  "\"Ontology Version of ATU\""@en ;
  rdfs:label "\"The Types of International
  Folktales Aarne-Thompson-Uther\""@en ;
  rdfs:subClassOf owl:Thing ;
.
```

An instance of this class, for example for the type 101, has the following syntax:

```
<http://www.semanticweb.org/tonka/
   ontologies/2015/5/tmi-atu-ontology#101>

 rdf:type :ATU ;

 linkToTMI <http://www.semanticweb.org/
   tonka/ontologies/2015/5/
 tmi-atu-ontology#K231.1.3> ;

 rdfs:comment "\"Type 101 of ATU\""@en ;

 rdfs:isDefinedBy "The Old Dog as Rescuer
   of the Child (Sheep). A farmer plans
 to kill his faithful old dog because
 it cannot work anymore. The wolf makes
 a plan to save the dog: The latter is to
 rescue the farmer's child from the wolf.
 The plan succeeds and the dog's life is
 spared. The wolf in return wants to
 steal the farmer's sheep. The dog
 refuses to help and loses the wolf's
 friendship . "@en ;

 rdfs:label "\"The Old Dog as Rescuer
   of the Child (Sheep)\""@en ;
.
```

The reader can see in this extensive example that each instance of the ATU class is named in the first line of the code by an Unique Resource Identifier (URI). The property "rdf:type" indicates that the object named by the URI is an instance of the class "ATU". The last element of the code, introduced by "rdfs:label", stores the original label in English ("en"). We will use this property "rdfs:label" to encode the multilingual correspondences. We encode the original description of the type as a value to the property "rdfs:isDefinedBy".

The property "linkToTMI" is the way we go for linking ATU types to Motifs listed in the Motif-Index of Folk-Literature (which we abbreviate with TMI). This linking is still in a preliminary stage, since we first have to finalize the corresponding TMI ontology, and also check the validity of the linking to TMI we extracted from the ATU book. This kind of linking is the one we will use for interconnecting all types of classification schemes used for folktales (and maybe also for other literary genres). We will add a property for including relevant hypernyms (and possibly hyponyms) extracted from WordNet to the current labels, contributing this way to the semantic enrichment of the original classification.

4 Conclusion and future Work

We presented work done in the context of a running student software project consisting in accessing WordNet for providing for lexical semantic information that can be used for enriching an existing classification scheme of folktales with additional terms gained from the extraction of relevant hypernyms (and to a certain extent from hyponyms) of words naming characters playing a central roles in folktales. The aim is to generate a WordNet based network of terms for the folktale domain.

As future work, an investigation will be performed in order to determine the optimal length of the path between a Lowest Common Hypernym (LCH) and the root node of WordNet as the filtering process for excluding irrelevant and noise introducing LCHs. We will also perform an evaluation of the extracted LCHs against a manually annotated set of ATU entries. And we will compare the French equivalents of the synsets proposed by WordNet with the French terms used in the French Wikipedia page for the AT. Additionally, we plan to compare our WordNet based approach as the basis for the linking between ATU and TMI to the machine learning approach to such a linking described in (Ofex et al., 2013).

[5] See `http://www.w3.org/2001/sw/wiki/OWL` and `http://www.w3.org/TR/rdf-schema/`.

[6] See http://www.w3.org/TR/turtle/ for more details.

Acknowledgments

Work done by the Saarland University has been supported by the PHEME FP7 project (grant No. 611233). We would like to thank Alyssa Price, Saarland University, for providing for the manual analysis of the patterns occurring in ATU. Our gratitude goes also to the two anonymous reviewers for the very helpful comments on the previous version of this short paper.

References

Antti Aarne. 1961. *The Types of the Folktale: A Classification and Bibliography*. The Finnish Academy of Science and Letters, Helsinki.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA.

Thierry Declerck, Karlheinz Mrth, Piroska Lendvai. 2012. Accessing and Standardizing Wiktionary Lexical Entries for the Translation of Labels in Cultural Heritage Taxonomies. *Proceedings of the Eight International Conference on Language Re-sources and Evaluation (LREC'12)*. Istanbul, Turkey.

Christiane Fellbaum (ed). 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, Vol. 38, No. 11: 39-41.

Thierry Declerck, Karlheinz Mrth, Piroska Lendvai. 2013. Linking Motif Sequences to Tale Types by Machine Learning. *Proceedings of the 2013 Workshop on Computational Models of Narrative*, 166-182. Dagstuhl, Germany

Stith Thompson. 1977. *The Folktale*. University of California Press, Berkeley.

Hans J. Uther. 2004. *The Types of the Folktale: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson*. The Finnish Academy of Science and Letters, Helsinki.