Software Projects for developing Digital Humanities Resources

Thierry Declerck
DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg, 3
D-66123 Saarbrücken
delecork@dfki.de

Abstract

In this short paper we report on experiences gained from bachelor and master theses, and from a series of software projects conducted in cooperation with the Department of Computational Linguistics of the Saarland University. Those bachelor/master theses and software projects were dealing with the application of Natural Language Processing and Semantic Web technologies to the representation and analysis of folktales. Data, codes and results of the software projects have been made available in various repository management services, like GitLab, GitHub or Bitbucket. We think that it will be important to discuss the design of such openly accessible repositories in order to ensure their re-usability and further extensions across various educational institutions.

1 Introduction

In the past 3-4 years we proposed in cooperation with the Computational Linguistics (CL) department of the Saarland University a series of bachelor/master theses and software projects, which were dealing with various aspects related to the wider field of folktales and therefore introducing Digital Humanities (DH) topics to students trained primarily to learn and apply computational methods of language technologies.

Our diagnosis was that the approach building on software projects for introducing CL students, and some few students from other departments, to Digital Humanities topics has been very successful. It is also the case that some of the projects we conducted have gained the interest of a larger group of students (they can choose between different software projects). We describe in the following sections the types of approaches we followed and the results that the students generated and made available on various repository management services, like GitLab, GitHub or Bitbucket. The idea of having software projects as a platform followed the work done by two students in their master and bachelor theses, which were written in the context of their Research Assistant appointments within a larger national project. We describe briefly the results of all those endeavours in the following sections.

2 Annotations

In the context of cooperation between the past D-SPIN and AMICUS projects, a master thesis was written by the student Antonia Scheidel on the http://www.abitur-und-studium.de/Bilder/Jana-Ott-Christian-Eisenreich-Christian-Willms-Studenten-von-Thierry-Declerck-haben-ein-Programm-entwickelt-das-Maerchen-vorlesen-kann.aspx

or http://kulturellebildung.de/fa/user/Fachbereiche/Literatur_Sprache/Aktuelles/141121_PRESSE_Erzaehlen.pdf

We think that a main aspect of this success story lies in the fact that the students had to work together, building teams for working on modules and meeting for integrating the work done so far.

In all the 4 different software projects conducted until now, we could observe that the folktale topic was a driver calling for participation of a larger group of students (they can choose between different software projects). We describe in the following sections the types of approaches we followed and the results that the students generated and made available on various repository management services, like GitLab, GitHub or Bitbucket. The idea of having software projects as a platform followed the work done by two students in their master and bachelor theses, which were written in the context of their Research Assistant appointments within a larger national project. We describe briefly the results of all those endeavours in the following sections.

2 Annotations

In the context of cooperation between the past D-SPIN and AMICUS projects, a master thesis was written by the student Antonia Scheidel on the
annotation of fairy tales with Propp’s functions. Vladimir Propp “was a Soviet folklorist and scholar who analysed the basic plot components of Russian folk tales to identify their simplest irreducible narrative elements.” Those basic plot elements are called by Propp “functions” and he identified 31 such functions, like “Interdiction”, “Delivery” or “Rescue”, etc. Propp also introduced circa 150 sub-functions that are specialisations of the 31 top-level functions. Complementary to the functions, Propp identified 7 broad characters, like “the villain”, “the donor” or the “hero”. The “morphology of the tale” described by Vladimir Propp was based on a subset of the so-called Afanasyev collection of Russian Folktales.

Antonia Scheidel developed a new annotation scheme according to which fairy tales can be queried for texts, temporal structures, characters, dialogues, and Propp’s functions. The annotation scheme has been named APftML, standing for “Augmented Propp fairy tale Mark-up Language”. Antonia Scheidel’s work is documented in (Declerck and Scheidel, 2010) and (Declerck et al., 2011). Annotated fairytale textual data is important in that automated systems have a data set against which they can map their results (see, for example (Declerck et al., 2012) and (Koleva et al., 2012) for more details on the results of her work).

Once we had those resources, i.e an annotation framework for folktales, based in a first instance of the mark-up of Proppian functions, and an ontology framework in which characters playing a role in folktales are stored as instances of domain-specific classes, the idea was to extend those to a larger framework supporting DH application scenarios.

4 Approaches to Story Segmentation

In a first software project which was building on the top of the two resources mentioned in the previous sections, a division of work could be established between the four members of the project team. One task consisted in offering a meaningful segmentation of the tales. The approach for this consisted in automatically segmenting the tales along the lines of the dialogue structure. This had one motivation: to offer a base for the integration of a text-to-speech system supporting the “read aloud” of a tale, in which voices are associated to each contributors to the dialogues (and for sure one voice for the narrator). This application is described in more details in the next section.

The students worked in this project mainly on the English version of the “Froschkönig” tale (The Frog Prince). Following those new steps, the initial annotation format has been augmented with detailed dialogue descriptions. And the ontology has also been extended, including now a description of dialogues (questions, answers, monologues etc.), including the encodings of the participants.
and the dialogue turns. In the two most recent and currently still running software projects the students are implementing a strategy on additionally segmenting a tale by the locations in which events are occurring. There is an interesting correlation between the segmentation by dialogues and the one by locations, as in this kind of narratives the participants to a dialogue are often sharing a location.

5 Emotions Detection and Text-to-Speech Modules

One student had the task to implement a program able to detect emotions. For this the original annotation scheme has been extended, supporting the mark-up of 6 basic emotions (fear, grief, joy, etc.), which are also encoded in the ontology. The automatic processing of the text (based in this case on the NLTK package\footnote{NLTK stands for “Natural Language Toolkit” and is written in Python, including a lot of corpus processing and statistical libraries. See \url{http://www.nltk.org/} for more details.}) was then marking the emotion detected in one sentence, on the base of an emotion lexicon build from annotated examples that served as a seed that was completed by consulting the WordNet\footnote{See \url{http://www.nltk.org/howto/wordnet.html} for more details.} module implemented in NLTK\footnote{See \url{https://www.w3.org/RDF/} for more details.}.

A major extension of the past work in this software project was that synthetic voices also play a role. Once a character has been recognized, for example the princess (in the fairy tale “Frog King”) additional features are coded (for example age, gender, emotion, etc.). Then a previously defined synthetic voice is automatically added to the character. And when the text is processed by the system, the story can be “told” by the voices. If there is no detected character in a dialogue situation, it is assumed that the narrator is the speaker and the reader the receiver. A demo can be heard in the corresponding Bitbucket repository\footnote{The data, algorithms and results of the projects are stored in \url{https://bitbucket.org/ceisen/apftml2repo} for more details.}. In this software project we made use of the “Mary” Text-To-Speech System\footnote{See \url{http://mary.dfki.de/} for more details.}. The overall results of the projects are described also in \cite{Eisenreich2014}.

6 Iterative Ontology Developments

A major extension of the past work in this software project was that synthetic voices also play a role. Once a character has been recognized, for example the princess (in the fairy tale “Frog King”) additional features are coded (for example age, gender, emotion, etc.). Then a previously defined synthetic voice is automatically added to the character. And when the text is processed by the system, the story can be “told” by the voices. If there is no detected character in a dialogue situation, it is assumed that the narrator is the speaker and the reader the receiver. A demo can be heard in the corresponding Bitbucket repository. In this software project we made use of the “Mary” Text-To-Speech System. The overall results of the projects are described also in \cite{Eisenreich2014}.

We described in sections\footnote{See \url{https://sites.ualberta.ca/~urban/Projects/English/Motif_Index.htm}.} and\footnote{See \url{https://en.wikipedia.org/wiki/Motif_(folkloristics)}.} how the original ontology has been enriched with additional features. In a second software project, work was dedicated in the ontologisation of classical knowledge – indexation and classification – resources in the field of folklore. We were considering in this software project two such resources: The “Motif-index of folk-literature”\footnote{See \url{https://sites.ualberta.ca/~urban/Projects/English/Motif_Index.htm}.} (Thompson, 1955–1958) and the “Types of International Folktales”\footnote{See \url{https://en.wikipedia.org/wiki/Motif_(folkloristics)}.} (Uther, 2004). The first resource, which we abbreviate as TMI, is available as an on-line resource.\footnote{See \url{http://mary.dfki.de/} for more details.} A folk tale motif can be defined as a “repeated story element, e.g., a character, an object, an action, or an event that can be found in several stories.”\footnote{See \url{https://www.w3.org/TR/owl-semantics/} for more details.} In TMI all motifs are organized in a tree structure, so that each motif has a more abstract class that describes a span of subordinated motifs. One motif entry consists of a motif-id, motif name, motif description (optional), and references to literature where it occurs.

The second resource builds on former work by Antti Aarne (Aarne, 1961) and Stith Thompson. This classification system was extended by Hans-Jörg Uther\footnote{See \url{https://en.wikipedia.org/wiki/Motif_(folkloristics)}.}, and in the following we are using the acronym ATU for referring to this resource. A folk tale motif can be described as a main story line that can be found in several cultures. The parts of this story line can refer to specific story elements also known as motifs. A folk tale type is therefore a bigger unit than a motif.

Our approach consisted in extracting from those knowledge resources, which are stored in different formats, classification relevant information and to re-organize them in two interrelated ontologies, using for this the W3C standards OWL\footnote{See \url{https://www.w3.org/TR/owl-semantics/} for more details.} RDF(s)\footnote{See \url{http://www.w3.org/TR/rdf-schema/} for more details.} and RDF\footnote{See \url{https://www.w3.org/RDF/} for more details.}.

The integrated ontology resulting from the software project, also after curation done in the context of an internship at DFKI, contains 46,950 motifs for the TMI domain and 2802 elements for the ATU domain, most of them interrelated by corresponding properties. Results of this software project are available in a
An application of this new integrated ontology for the classification of characters in folktales has been presented in (Declerck et al., 2016) and more recent developments related to this integrated ontology are described in (Declerck et al., 2017).

7 Conclusion

We did report on specific teaching activities in the field of the representation and processing of folktales by students (mainly) in the field of computational linguistics. The specificity of the experiences we are reporting is that those activities took place in the context of software projects or internships, thus with a focus on practical implementation and development works. We noticed that this kind of team work, or also compact work done in the context of an internship, is delivering a very large amount of resources that are potentially very relevant for being reused in other type of teaching activities. Maybe also a coordinated action between universities and other educational institutions toward the organization of such software projects could be an idea to discuss and implement. Last but not least, many of the results presented in this short paper have been submitted to and accepted at relevant workshops and conferences, bringing the students thus also closer to this type of academic achievements.

References


