Automatic classification of folk narrative genres

Dong Nguyen[1], Dolf Trieschnigg[1], Theo Meder[2], Mariët Theune[1]
[1]University of Twente, Enschede, The Netherlands
[2]Meertens Institute, Amsterdam, The Netherlands
{d.nguyen,d.trieschnigg}@utwente.nl
theo.meder@meertens.knaw.nl,m.theune@utwente.nl

Abstract

Folk narratives are a valuable resource for humanities and social science researchers. This paper focuses on automatically recognizing folk narrative genres, such as urban legends, fairy tales, jokes and riddles. We explore the effectiveness of lexical, structural, stylistic and domain specific features. We find that it is possible to obtain a good performance using only shallow features. As dataset for our experiments we used the Dutch Folktale database, containing narratives from the 16th century until now.

1 Introduction

Folk narratives are an integral part of cultural heritage and a valuable resource for historical and contemporary comparative folk narrative studies. They reflect moral values and beliefs, and identities of groups and individuals over time (Meder, 2010). In addition, folk narratives can be studied to understand variability in transmission of narratives over time.

Recently, much interest has arisen to increase the digitalization of folk narratives (e.g. Meder (2010), La Barre and Tilley (2012), Abello et al. (2012)). In addition, natural language processing methods have been applied to folk narrative data. For example, fairy tales are an interesting resource for sentiment analysis (e.g. Mohammad (2011), Alm et al. (2005)) and methods have been explored to identify similar fairy tales (Lobo and de Matos, 2010), jokes (Friedland and Allan, 2008) and urban legends (Grundkiewicz and Gralinski, 2011).

Folk narratives span a wide range of genres and in this paper we present work on identifying these genres. We automatically classify folk narratives as *legend, saint's legend, fairy tale, urban legend, personal narrative, riddle, situation puzzle, joke* or *song*. Being able to automatically classify these genres will improve accessibility of narratives (e.g. filtering search results by genre) and test to what extent these genres are distinguishable from each other. Most of the genres are not well defined, and researchers currently use crude heuristics or intuition to assign the genres.

Text genre classification is a well-studied problem and good performance has been obtained using surface cues (Kessler et al., 1997). Effective features include bag of words, POS patterns, text statistics (Finn and Kushmerick, 2006), and character n-grams (Kanaris and Stamatatos (2007), Sharoff et al. (2010)).

Finn and Kushmerick (2006) argued that genre classifiers should be reusable across multiple topics. A classifier for folk narrative genres should also be reusable across multiple topics, or in particular across story types[1]. For example, a narrative such as *Red Riding Hood* should not be classified as a fairy tale because it matches a story type in the training set, but because it has characteristics of a fairy tale in general. This allows us to distinguish between particular genres, instead of just recognizing variants of the same story. In addition, this is desirable, since variants of a story type such as *Red Riding Hood* can appear in other genres as well, such as jokes and riddles.

[1]Stories are classified under the same type when they have similar plots.

Most of the research on genre classification focused on classification of text and web genres. To the best of our knowledge, we are the first to automatically classify folk narrative genres. Our dataset contains folk narratives ranging from the 16th century until now. We first give an overview of the dataset and the folk narrative genres. We then describe the experiments, discuss the results and suggest future work.

2 Dataset

2.1 Overview

Our dataset is a large collection of folk narratives collected in the Netherlands[2]. Although the collection contains many narratives in dialect, we restrict our focus to the narratives in standard Dutch, resulting in a total of 14,963 narratives. The narratives span a large time frame, with most of them from the 19th, 20th and 21th century, but some even dating from the 16th century as can be seen in Table 1[3]. Each narrative has been manually annotated with metadata, such as genre, named entities, keywords and a summary.

2.2 Narrative genres

Folk narrative genres vary between cultures. Legend, myth and folktales are major genres that are present in many cultures. Bascom (1965) proposed a formal definition of these genres, based on belief, time, place, attitude and principal characters of the narratives. In this work, we restrict our attention to genres that are applicable to the Dutch folk narratives. The selected genres are described below and based on how annotators assign narratives to genres in the Dutch Folktale database.

Fairy tales are set in an unspecified time (e.g. the well-known *Once upon a time* ...) and place, and are believed not to be true. They often have a happy ending and contain magical elements. Most of the fairy tales in the collection are classified under the Aarne-Thompson-Uther classification system, which is widely used to classify and organize folk tales (Uther, 2004).

[2]Dutch Folktale database: http://www.verhalenbank.nl/.

[3]Although standard Dutch was not used before the 19th century, some narratives dating from before that time have been recorded in standard Dutch.

Time period	Frequency
- 1599	8
1600 - 1699	11
1700 - 1799	24
1800 - 1899	826
1900 - 1999	8331
2000 -	4609
Unknown	1154

Table 1: Spread of data over time periods

Legends are situated in a known place and time, and occur in the recent past. They were regarded as non-fiction by the narrator and the audience at the time they were narrated. Although the main characters are human, legends often contain supernatural elements such as witches or ghosts.

Saint's legends are narratives that are centered on a holy person or object. They are a popular genre in Catholic circles.

Urban legends are also referred to as contemporary legends, belief legends or FOAF (Friend Of A Friend) tales in literature. The narratives are legends situated in modern times and claimed by the narrator to have actually happened. They tell about hazardous or embarrassing situations. Many of the urban legends are classified in a type-index by Brunvand (1993).

Personal narratives are personal memories (not rooted in tradition) that happened to the narrator himself or were observed by him. Therefore, the stories are not necessarily told in the first person.

Riddles are usually short, consisting of a question and an answer. Many modern riddles function as a joke, while older riddles were more like puzzles to be solved.

Situation puzzles, also referred to as *kwispels* (Burger and Meder, 2006), are narrative riddle games and start with the mysterious outcome of a plot (e.g. *A man orders albatross in a restaurant, eats one bite, and kills himself*). The audience then needs to guess what led to this situation. The storyteller can only answer with 'yes' or 'no'.

Jokes are short stories for laughter. The collection contains contemporary jokes, but also older jokes that are part of the ATU index (Uther, 2004). Older jokes are often longer, and do not necessarily contain a punchline at the end.

Songs These are songs that are part of oral tradition (contrary to pop songs). Some of them have a story component, for example, ballads that tell the story of *Bluebeard*.

Genre	Train	Dev	Test
Situation puzzle	53	7	12
Saint's legend	166	60	88
Song	18	13	5
Joke	1863	719	1333
Pers. narr.	197	153	91
Riddle	755	347	507
Legend	2684	1005	1364
Fairy tale	431	142	187
Urban legend	2144	134	485
Total	8311	2580	4072

Table 2: Dataset

2.3 Statistics

We divide the dataset into a training, development and test set, see Table 2. The class distribution is highly skewed, with many instances of legends and jokes, and only a small number of songs. Each story type and genre pair (for example *Red Riding Hood* as a fairy tale) only occurs in one of the sets. As a result, the splits are not always even across the multiple genres.

3 Experimental setup

3.1 Learning algorithm

We use an SVM with a linear kernel and L2 regularization, using the liblinear (Fan et al., 2008) and scikit-learn libraries (Pedregosa et al., 2011). We use the method by Crammer and Singer (2002) for multiclass classification, which we found to perform better than one versus all on the development data.

3.2 Features

The variety of feature types in use is described below. The number of features is listed in Table 3. The frequency counts of the features are normalized.

I **Lexical features**. We explore *unigrams*, and *character n-grams* (all n-grams from length 2-5) including punctuation and spaces.

II **Stylistic and structural features**.

POS unigrams and bigrams (CGN[4] tagset) are extracted using the Frog tool (Van Den Bosch et al., 2007).

Punctuation. The number of punctuation characters such as *?* and *"*, normalized by total number of characters.

Whitespace. A feature counting the number of empty lines, normalized by total number of lines. Included to help detect songs.

Text statistics. Document length, average and standard deviation of sentence length, number of words per sentence and length of words.

III **Domain knowledge**. Legends are characterized by references to places, persons etc. We therefore consider the number of *automatically tagged named entities*. We use the Frog tool (Van Den Bosch et al., 2007) to count the number of references to persons, organizations, location, products, events and miscellaneous named entities. Each of them is represented as a separate feature.

IV **Meta data**. We explore the added value of the manually annotated metadata: *keywords, named entities* and *summary*. Features were created by using the normalized frequencies of their tokens. We also added a feature for the manually annotated *date* (year) the story was written or told. For stories of which the date is unknown, we used the average date.

Feature type	# Features
Unigrams	16,902
Char. n-grams	128,864
Punctuation	8
Text statistics	6
POS patterns	154
Whitespace	1
Named entities	6
META - Keywords	4674
META - Named entities	803
META - Summary	5154
META - Date	1

Table 3: Number of features

3.3 Evaluation

We evaluate the methods using precision, recall and F_1-measure. Since the class distribution is highly skewed, we focus mainly on the macro average that averages across the scores of the individual classes.

[4]Corpus Gesproken Nederlands (Spoken Dutch Corpus), http://lands.let.kun.nl/cgn/ehome.htm

	Precision	Recall	F_1
Unigrams	0.551	0.482	0.500
Char. n-grams	0.646	0.578	0.594

Table 4: Baselines (Macro average)

	Precision	Recall	F_1
All	0.660	0.591	0.608
-Unigrams	0.646	0.582	0.597
-Char. n-grams	0.577	0.511	0.531
-Punctuation	0.659	0.591	0.607
-Text statistics	0.659	0.589	0.606
-POS patterns	0.659	0.590	0.607
-Whitespace	0.660	0.591	0.608
-Domain knowl.	0.660	0.591	0.607

Table 5: Ablation studies, without meta data (Macro average)

4 Results

The penalty term C of the error term for SVM was set using the development set. All results reported are on the test set. We first report two baselines in Table 4, using only unigrams and character based n-grams. Results show that the character based n-grams are highly effective. This is probably because they are able to capture punctuation and endings of words, and are more robust to spelling mistakes and (historical) variations.

Next, ablation studies were performed to analyze the effectiveness of the different feature types, by leaving that particular feature type out. We experimented with and without metadata. The results without metadata are reported in Table 5. We find that only character n-grams contribute highly to the performance. Removing the other feature types almost has no effect on the performance. However, without character n-grams, the other features do have an added value to the unigram baseline (an increase in macro average of 0.50 to 0.53 in F_1 score). One should note that some errors might be introduced due to mistakes by the automatic taggers for the POS tokens and the domain knowledge features (named entities), causing these features to be less effective.

The results with all features including the metadata are reported in Table 6. We find that when using all features the F_1 score increases from 0.61 to 0.62. The ablation studies suggest that especially the keywords, summary and date are effective. However, overall, we find that only using character n-grams already gives a good performance. We therefore believe they are a valuable alternative against more sophisticated features.

	Precision	Recall	F_1
All	0.676	0.600	0.621
-META - Keywords	0.659	0.595	0.611
-META - Named Entities	0.682	0.599	0.623
-META - Summary	0.664	0.596	0.614
-META - Date	0.674	0.592	0.614

Table 6: Ablation studies all features (Macro average)

	Precision	Recall	F_1
Sit. puzzle	0.70	0.58	0.64
Saint's legend	0.81	0.40	0.53
Song	0.00	0.00	0.00
Joke	0.93	0.71	0.81
Pers. narr.	0.69	0.52	0.59
Riddle	0.86	0.83	0.84
Legend	0.82	0.92	0.86
Fairy tale	0.69	0.53	0.60
Urban legend	0.59	0.91	0.71

Table 7: Results per genre

We also find that including the date of a narrative as a feature leads to an increase from 0.614 to 0.621. This feature is effective since some genres (such as urban legends) only occur in certain time periods. Noise due to the many documents (1154 of 14963) for which the date is not known, could have affected the effectiveness of the feature.

In Table 7 the results per genre are listed for the run including all features (with metadata). The best performing genres are jokes, riddles and legends. We find that songs are always incorrectly classified, probably due to the small number of training examples. Personal narratives are also a difficult category. These narratives can be about any topic, and they do not have a standard structure. Fairy tales are often misclassified as legends. Some of the fairy tales do not contain many magical elements, and therefore look very similar to legends. In addition, the texts in our dataset are sometimes interleaved with comments that can include geographical locations, confusing the model even more.

Initially, annotators of the Dutch Folktale database were allowed to assign multiple genres. From this, we observe that many narratives were classified under multiple genres (these narratives were excluded from the dataset). This is evidence that for some narratives it is hard to assign a single genre, making it unclear what optimal performance can be achieved.

5 Conclusion

Folk narratives are a valuable resource for historical comparative folk narrative studies. In this paper we presented experiments on classifying folk narrative genres. The goal was to automatically classify narratives, dating from the 16th century until now, as *legend, saint's legend, fairy tale, urban legend, personal narrative, riddle, situation puzzle, joke* or *song*. Character n-grams were found to be the most effective of all features. We therefore plan to explore historical texts in non standard Dutch, since character n-grams can be easily extracted from them. We also intend to explore features that help detect difficult genres and generalize across specific stories. For example, features that can detect humorous components.

6 Acknowledgements

This research was supported by the Folktales as Classifiable Texts (FACT) project, part of the CATCH programme funded by the Netherlands Organisation for Scientific Research (NWO).

References

J. Abello, P. Broadwell, and T. R. Tangherlini. 2012. Computational folkloristics. *Communications of the ACM*, 55(7):60–70, July.

C. O. Alm, D. Roth, and R. Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of HLT/EMNLP*, pages 579–586.

W. Bascom. 1965. The Forms of Folklore: Prose Narratives. *The Journal of American Folklore*, 78(307).

J. H. Brunvand. 1993. *The Baby Train and Other Lusty Urban Legends*. W. W. Norton & Company.

P. Burger and T. Meder. 2006. -A rope breaks. A bell chimes. A man dies.- The kwispel: a neglected international narrative riddle genre. In *Toplore. Stories and Songs*, pages 28–38.

K. Crammer and Y. Singer. 2002. On the learnability and design of output codes for multiclass problems. *Mach. Learn.*, 47(2-3):201–233, May.

R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. 2008. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9(6/1/2008):1871–1874.

A. Finn and N. Kushmerick. 2006. Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57(11):1506–1518.

L. Friedland and J. Allan. 2008. Joke retrieval: recognizing the same joke told differently. In *Proceedings of CIKM*, pages 883–892.

R. Grundkiewicz and F. Gralinski. 2011. How to Distinguish a Kidney Theft from a Death Car? Experiments in Clustering Urban-Legend Texts. In *Proceedings of the Workshop on Information Extraction and Knowledge Acquisition*.

I. Kanaris and E. Stamatatos. 2007. Webpage Genre Identification Using Variable-Length Character n-Grams. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Volume 02*, ICTAI '07, pages 3–10, Washington, DC, USA. IEEE Computer Society.

B. Kessler, G. Nunberg, and H. Schuetze. 1997. Automatic Detection of Text Genre. In *Proceedings of the 35th ACL/8th EACL*.

K. A. La Barre and C. L. Tilley. 2012. The elusive tale: leveraging the study of information seeking and knowledge organization to improve access to and discovery of folktales. *Journal of the American Society for Information Science and Technology*, 63(4):687–701.

P. V. Lobo and D. M. de Matos. 2010. Fairy Tale Corpus Organization Using Latent Semantic Mapping and an Item-to-item Top-n Recommendation Algorithm. In *Proceedings of LREC*.

T. Meder. 2010. From a Dutch Folktale Database towards an International Folktale Database. *Fabula*, 51(1-2):6–22.

S. Mohammad. 2011. From once upon a time to happily ever after: tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, pages 105–114.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830.

S. Sharoff, Z. Wu, and K. Markert. 2010. The Web Library of Babel: evaluating genre collections. In *Proceedings of LREC*.

H.-J. Uther. 2004. *The Types of International Folktales: A Classification and Bibliography Based on the System of Antti Aarne and Stith Thompson. Vols 1-3.* Suomalainen Tiedeakatemia, Helsinki.

A. Van Den Bosch, B. Busser, S. Canisius, and W. Daelemans. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Computational Linguistics in the Netherlands Selected Papers from the Seventeenth CLIN Meeting*, pages 99–114. OTS.