Automated Narrative Information Extraction Using Non-Linear Pipelines

Josep Valls-Vargas
Drexel University
Philadelphia, Pennsylvania, USA
josep.vallsvargas@drexel.edu

Abstract

Our research focuses on the problem of automatically acquiring structured narrative information from natural language. We have focused on character extraction and narrative role identification from a corpus of Slavic folktales. To address natural language processing (NLP) issues in this particular domain we have explored alternatives to linear pipelined architectures for information extraction, specifically the idea of feedback loops that allow feeding information produced by later modules of the pipeline back to earlier modules. We propose the use of domain knowledge to improve core NLP tasks and the overall performance of our system.

1 Introduction and Background

Computational narrative systems, especially story generation systems, require their input hand-authored in some structured knowledge representation formalism [Zhu and Ontanón, 2010], a notoriously time-consuming task. To address this well-known "authorial bottleneck" problem we have been exploring natural language processing (NLP) techniques to automatically acquire structured narrative information from natural language. We have focused on extracting characters (i.e. sentient beings) and identifying their narrative roles based on their prototypical interactions and properties using a text corpus of translated and annotated Slavic folktales [Finlayson, 2012]. This particular domain poses additional problems to natural language processing as off-the-shelf NLP systems underperform in tasks such as coreference resolution and verb argument identification due to the complex rhetoric present in our corpus. On the other hand, narrative regularities exhibited by these folktales were the subject of interest in early research in the field of narratology and were the basis to the development of Propp's narrative theory [Propp, 1973]

In this general area, we identified one particular problem: most systems for narrative information extraction and language understanding rely on linear pipelined architectures [Clarke *et al.*, 2012] where the output of one module is the input to the next one. However, non-linear pipelined alternatives can improve the performance of certain tasks with respect to linear architectures [Roth and Yih, 2004; Marciniak and Strube, 2005].

2 Contributions

To address the general problem of acquiring structured narrative information from natural language, our contributions include:

- **Identifying narrative roles from characters**. We explored approaches to bridge narrative domain knowledge (Propp's narrative theory) and NLP. In this work we introduce the idea of representing the "sphere of action" of a character role (their prototypical actions in Propp's narrative theory) as a matrix encoding interactions between differentroles. For this work we used an annotated dataset to compute a matrix from a story and compare it against a reference matrix using the Wordnet hierarchy to find similarities. [Valls-Vargas *et al.*, 2013].
- **Extracting characters and their narrative roles from unannotated folktales**. We presented a framework to automatically extract referring expressions from unannotated text and an instance-based learning approach to classify the extracted mentions as characters (sentient beings) and non-characters. Our contributions include an analysis of how different feature sets perform for this task and the definition of a novel similarity measure (a continuous variant of the Jaccard distance) used for instance retrieval [Valls-Vargas *et al.*, 2014a]. Building on our previous contribution we extended the character extraction process with our "sphere of action" representation to identify narrative roles for extracted characters [Valls-Vargas *et al.*, 2014b].
- **Evaluation of information extraction pipelines**. We developed a methodology for the empirical study and evaluation of information extraction pipelines. Our methodology focuses on the study of the sources of error and how the error propagates through different modules of an information extraction pipeline. We applied this methodology to an empirical study of our narrative information extraction pipeline (under review).
- **Using feedback loops in information extraction pipelines**. We explored the idea of introducing feedback loops in information extraction pipelines. We applied the idea to our narrative information extraction pipeline and used the identified character roles (the final output) to inform the coreference resolution task and improving the overall performance. [Valls-Vargas *et al.*, 2015].

- **Voz**. For our experimental evaluation I have developed *Voz*, a narrative information extraction system that combines off-the-shelf natural language processing toolkits (e.g., Stanford CoreNLP, ClearNLP), common sense knowledge (e.g., WordNet, ConceptNet) and domain knowledge (Propp's narrative theory). We have been using a corpus of Slavic folktales collected and annotated by Mark A. Finlayson [2012].

3 Research Plan

In my research I intend to further explore the idea of non-linear pipelined architectures for narrative information extraction. Specifically I am interested in the following areas:

- **Using narrative information to improve core NLP tasks**. I would like to explore how narrative information, both automatically extracted information and encoded domain knowledge, can be used to inform core NLP tasks. I am particularly interested in tasks related to verb argument identification and semantic role labeling. I expect to be able to exploit narrative regularities and prototypical interactions between different classes of entities (character roles and non-character entities in a taxonomy such as Chatman's [1980]).
- **Generalizing methodologies for building and evaluating non-linear pipelines**. I would like to formalize methodologies for building non-linear information extraction pipelines and adding non-linear features to existing linear pipelines. I plan on also generalizing our current methodology for linear pipelines to evaluate complex non-linear pipelines in order to better understand how some error can be mitigated using global inference or feedback loops.
- **Improving and generalizing narrative information extraction**. I plan on adding modules that extract additional and higher level narrative information such as Proppian functions and affect relationships between characters. I am also considering generalizing our approach to other narrative domains and narrative theories such as Campbell's monomyth [Campbell, 2008].
- **Voz**. Building on the work outlined in this section, I would like to culminate my research work by connecting *Voz* to a digital entertainment system such as *Game Forge* for generating game worlds [Hartsook *et al.*, 2011] or *Comme il Faut* [McCoy *et al.*, 2011] for authoring social models.

References

[Campbell, 2008] Joseph Campbell. *The hero with a thousand faces*, volume 17. New World Library, 2008.

[Chatman, 1980] Seymour Benjamin Chatman. *Story and discourse: Narrative structure in fiction and film*. Cornell University Press, 1980.

[Clarke *et al.*, 2012] James Clarke, Vivek Srikumar, Mark Sammons, and Dan Roth. An nlp curator (or: How i learned to stop worrying and love nlp pipelines). In *LREC*, pages 3276–3283, 2012.

[Finlayson, 2012] Mark A. Finlayson. *Learning narrative structure from annotated folktales*. PhD thesis, Massachusetts Institute of Technology, 2012.

[Hartsook *et al.*, 2011] Ken Hartsook, Alexander Zook, Sauvik Das, and Mark O. Riedl. Toward supporting stories with procedurally generated game worlds. In *Proceedings of the 2011 IEEE Conference on Computational Intelligence in Games*, pages 297–304. Ieee, August 2011.

[Marciniak and Strube, 2005] Tomasz Marciniak and Michael Strube. Beyond the pipeline: Discrete optimization in nlp. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 136–143. Association for Computational Linguistics, 2005.

[McCoy *et al.*, 2011] Joshua McCoy, Mike Treanor, Ben Samuel, Noah Wardrip-Fruin, and Michael Mateas. Comme il faut: A system for authoring playable social models. In *Proceedings of the Artificial Intelligence and Interactive Digital Entertainment Conference*, 2011.

[Propp, 1973] Vladimir Propp. *Morphology of the Folktale*. University of Texas Press, 1973.

[Roth and Yih, 2004] Dan Roth and Wen-tau Yih. A linear programming formulation for global inference in natural language tasks. Technical report, DTIC Document, 2004.

[Valls-Vargas *et al.*, 2013] Josep Valls-Vargas, Santiago Ontañón, and Jichen Zhu. Toward character role assignment for natural language stories. In *Proceedings of the Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*, pages 101–104, 2013.

[Valls-Vargas *et al.*, 2014a] Josep Valls-Vargas, Santiago Ontañón, and Jichen Zhu. Toward automatic character identification in unannotated narrative text. In *Proceedings of the Seventh Workshop in Intelligent Narrative Technologies*, 2014.

[Valls-Vargas *et al.*, 2014b] Josep Valls-Vargas, Jichen Zhu, and Santiago Ontañón. Toward automatic role identification in unannotated folk tales. In *Proceedings of the Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2014.

[Valls-Vargas *et al.*, 2015] Josep Valls-Vargas, Santiago Ontañón, and Jichen Zhu. Narrative hermeneutic circle: Improving character role identification from natural language text via feedback loops. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, page 2517, 2015.

[Zhu and Ontanón, 2010] Jichen Zhu and Santiago Ontanón. Story representation in analogy-based story generation in riu. In *Proceedings of the 2010 IEEE Conference on Computational Intelligence in Games*, pages 435–442. IEEE, 2010.