Visualizing the Content of a Children's Story in a Virtual World: Lessons Learned

Quynh Ngoc Thi Do[1], **Steven Bethard**[2], **Marie-Francine Moens**[1]
[1]Katholieke Universiteit Leuven, Belgium
[2]University of Arizona, United States
quynhngocthi.do@cs.kuleuven.be
bethard@email.arizona.edu
sien.moens@cs.kuleuven.be

Abstract

We present the problem of "bringing text to life" via 3D interactive storytelling, where natural language processing (NLP) techniques are used to transform narrative text into events in a virtual world that the user can interact with. This is a challenging problem, which requires deep understanding of the semantics of a story and the ability to ground those semantic elements to the actors and events of the 3D world's graphical engine. We show how this problem has motivated interesting extensions to some classic NLP tasks, identify some of the key lessons learned from the work so far, and propose some future research directions.

1 Introduction

Our primary goal is to take as input natural language text, such as children's stories, translate the text into formal knowledge that represents the actions, actors, plots, and surrounding world, and render this formal representation as virtual 3D worlds via a graphical engine. We believe that translating text to another modality (in our case, a visual modality) is a good test case for evaluating language understanding systems.

We have developed an initial approach to this text-to-virtual-world translation problem based on a probabilistic graphical model that maps text and its semantic annotations (generated by more traditional NLP modules, like semantic role labelers or coreference resolvers) to the knowledge representation of the graphical engine, which is defined in predicate logic. In the process, we discovered several failings of traditional NLP systems when faced with this task:

Semantic Role Labeling We observed that current state-of-the-art semantic role labeling (SRL) systems perform poorly on children's stories, failing to recognize many of the expressed argument roles. Much of this is due to the domain mismatch between the available training data (primarily newswire) and our evaluation (stories for 3D visualization).

To address this, we introduced a technique based on recurrent neural networks for automatically generating additional training data that was similar to the target domain (Do et al., 2014; Do et al., 2015b). For each selected word (predicate, argument head word) from the source domain, a list of replacement words from the target domain which we believe can occur at the same position as the selected word, are generated by using a recurrent neural network (RNN) language model (Mikolov et al., 2010). In addition, linguistic resources such as part of speech tags, WordNet (Miller, 1995), and VerbNet (Schuler, 2005), are used as filters to select the best replacement words.

We primarily targeted improving the results of the four circumstance roles AM-LOC, AM-TMP, AM-MNR and AM-DIR, which are important for semantic frame understanding but not well recognized by standard SRL systems. New training examples were generated specifically for the four selected roles. In an experiment with the out-of-domain setting of the CoNLL 2009 shared task and the SRL system of (Björkelund et al., 2009), training the semantic role labeller on the expanded training data outperforms the

Proceedings of EMNLP 2016 Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods, pages 39–42,
Austin, TX, November 5, 2016.

model trained on the original training data by +3.36%, +2.77%, +2.84% and +14% F1 over the roles AM-LOC, AM-TMP, AM-MNR and AM-DIR respectively (Do et al., 2015b), but we still need linguistic resources to filter the words obtained by the language model. In an experiment where the same model was again trained on CoNLL 2009 training data, but the RNN training included a collection of 252 children stories (mostly fairy tales), we obtained F1 gains of +9.19,% +7.67%, +17.92% and +7.84% respectively over the four selected roles AM-LOC, AM-TMP, AM-MNR and AM-DIR, when testing on the story "The Day Tuk Became a Hunter" (Ronald and Carol, 1967) (Do et al., 2014).

Coreference Resolution We observed that current state-of-the-art coreference resolution systems are ignorant of some constraints that are important in storytelling. For example, a character is often first presented as an indefinite noun phrase (such as "a woman"), then later as a definite noun phrase (such as "the woman"), but this change in definiteness often resulted on missed coreference links.

To address this, we replaced the inference of the Berkeley coreference resolution system (Durrett and Klein, 2013) with a global inference algorithm which incorporated narrative specific constraints through integer linear programming (Do et al., 2015a). Our formulation models three phenomena that are important for short narrative stories: local discourse coherence, which we model via centering theory constraints, speaker-listener relations, which we model via direct speech act constraints, and character-naming, which we model via definite noun phrase and exact match constraints. When testing on the UMIREC[1] and N2[2] corpora with the coreference resolution system of (Durrett and Klein, 2013) trained on OntoNotes[3], our inference substantially improves the original inference on the CoNLL 2011 AVG score by +5.42 (for UMIREC) and +5.22 (for N2) points when using gold mentions and by +1.15 (for UMIREC) and +2.36 (for N2) points when using predicted mentions. When testing on the story "The Day Tuk Became a Hunter" (Ronald and Carol, 1967), our inference outperforms the original inference by 4.46 points on the CoNLL 2011 AVG score[4].

Having corrected some of the more serious failures of NLP systems on stories, we turn to the problem of mapping the semantic analysis of these NLP systems to the knowledge representation of the graphical engine. Our initial approach is implemented as a probabilistic graphical model, where the input is a sentence and its (probabilistic) semantic and coreference annotations, and the output is a set of logical predicate-argument structures. Each structure represents an action and the parameters of that action (e.g., person/object performing the action, location of the action). The domain is bounded by a finite set of actions, actors and objects, representable by the graphical environment. In our implementation, decoding of the model is done through an efficient formulation of a genetic algorithm that exploits conditional independence (Alazzam and Lewis III, 2013) and improves parallel scalability.

In an evaluation on three stories ("The Day Tuk Became a Hunter" (Ronald and Carol, 1967), "The Bear and the Travellers"[5], and "The First Tears"[6]), this model achieved F1 scores of 81% on recognizing the correct graphical engine actions, and above 60% on recognizing the correct action parameters (Ludwig et al., Under review). Example scenes generated by the MUSE software are shown in Figure 1, and a web-based demonstration can be accessed at `http://roshi.cs.kuleuven.be/muse_demon/`.

2 Lessons learned

Studying the problem of translating natural language narratives to 3D interactive stories has been instructive about the capabilities of current natural processing for language understanding and the battles that still have to be fought. The truthful rendering of language content in a virtual world acts as a testbed for

[1] `http://dspace.mit.edu/handle/1721.1/57507`

[2] `http://dspace.mit.edu/handle/1721.1/85893`

[3] `https://catalog.ldc.upenn.edu/LDC2011T03`

[4] We only evaluate the entities that are available in our virtual domain such as tuk, father, mother, bear, sister, igloo, sled, etc.

[5] `http://fairytalesoftheworld.com/quick-reads/the-bear-and-the-travellers/`

[6] `http://americanfolklore.net/folklore/2010/09/the\_first\_tears.html`

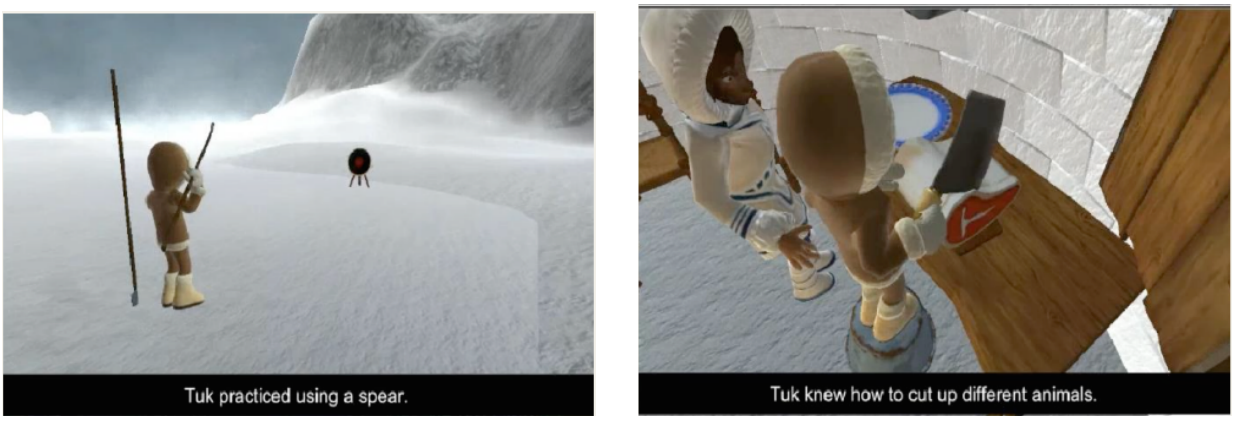

Figure 1: MUSE-generated scenes from "The Day Tuk Became a Hunter" (Ludwig et al., Under review).

natural language understanding, making this multimodal translation a real-life evaluation task.

On the positive side, some NLP tasks such as semantic role labeling and coreference resolution proved to be useful for instantiating the correct action frames in the virtual world with their correct actors. However, some NLP tasks that we imagined would be important turned out not to be. For example, temporal relation recognition was not very important, since children's stories have simpler timelines, and since the constraints of the actions in the 3D interactive storytelling representation could sometimes exclude the inappropriate interpretations. Moreover, across all of the NLP tasks, we saw significant drops in performance when applied to narratives like children's stories. While we introduced solutions to some of these problems, much work remains to be done to achieve easy and effective transfer of the learning models to other target texts.

Given the almost complete lack of training data for translating children's stories to the representations needed by the graphical engine (we only used two quite unrelated annotated stories to optimize the inference in the Bayesian network when our system parsed a test story), we had to rely on a pipelined approach. In this way we could exploit the knowledge obtained by the semantic role labeler and coreference resolver, which were trained on other annotated texts and adapted by the novel methods described above. The Bayesian framework of the probabilistic graphical model allows it to realize the most plausible mapping or translation to a knowledge representation given the provided evidences obtained from the features in a sentence and a previous sentence, the (probabilistic) outcome of the semantic role labeler and the (probabilistic) outcome of the coreference resolver, and to model dependencies between the variables of the network. This Bayesian framework for evidence combination makes it possible to recover from errors made by the semantic role labeler or coreference resolver.

Our most striking finding was that the text leaves a large part of its content implicit, but this content is actually needed for a truthful rendering of the text in the virtual world. For instance, often the action words used in the story were more abstract than the actions defined by the graphical engine (e.g., "take care" in reference to a knife, where actually "sharpen" was meant). Sometimes using word similarities based on embeddings (Mikolov et al., 2013) helped in such cases, but more often the meaning of such abstract words depends on the specific previous discourse, which is not captured by general embeddings. Adapting the embeddings, which are trained on a large corpus to the specific discourse context as is done in (Deschacht et al., 2012) is advisable. Moreover, certain content was not mentioned in the text, but a human could infer. For example, given "Tuk and his father tied the last things on the sled and then set off," a human would infer that the two people most likely sat down on the sled. Such knowledge is important for rendering a believable 3D interactive story, but can hardly be inferred from text data, instead needing grounding in the real world, perhaps captured by other modalities, such as images.

Another problem we encountered was scalability.

If the goal is to allow users to bring any text "to life", then all of the parsing and translation to the 3D world needs to happen online. Although the computational complexity when the machine comprehends the story is reduced by limiting the possible actions and actors (e.g., characters, objects) to the ones mentioned in the story and the ones inferred, parsing of the story is still slow. But even with the genetic algorithm inspired parallel processing we introduced, our graphical model is still too slow to operate in an online environment. Instead of considering parallel processing, it would be interesting to give priority to the most likely interpretation based on event language models (Do et al., Under review).

Finally, while working closely with researchers in 3D interactive storytelling, we learned that there is little consistency across designers of graphical worlds on the structure of basic actions, actors, objects, etc. Thus a model that has been trained to translate stories that take place in one digital world will produce invalid representations for other digital worlds. Generalizing across different digital worlds is a challenging but interesting future direction. Proposing standards could make a major impact in this field, and in addition could promote cross-modal translation between language and graphical content. We witness an increasing interest in easy to program languages for robotics that operate in virtual worlds (e.g., Mindstorms, ROBOTC) and in formalizing knowledge of virtual words by ontologies (Mezati et al., 2015). Although valuable, such approaches translate yet to another human made language. If we really want to test language understanding in a real-life setting, translation to perceptual images and video might be more suited, but more difficult to realize unless we find a way of composing realistic images and video out of primitive visual patterns.

Acknowledgments

This work was funded by the EU ICT FP7 FET project "Machine Understanding for interactive StorytElling" (MUSE). We thank the anonymous reviewers for their valuable comments.

References

Azmi Alazzam and Harold W. Lewis III. 2013. A new optimization algorithm for combinatorial problems. *International Journal of Advanced Research in Artificial Intelligence, IJARAI*, 2(5).

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 43–48, Stroudsburg, PA, USA. ACL.

Koen Deschacht, Jan De Belder, and Marie-Francine Moens. 2012. The latent words language model. *Computer Speech and Language*, 26(5):384–409, October.

Quynh Ngoc Thi Do, Steven Bethard, and Marie-Francine Moens. 2014. Text mining for open domain semi-supervised semantic role labeling. In *DMNLP@ PKDD/ECML*, pages 33–48.

Quynh Ngoc Thi Do, Steven Bethard, and Marie-Francine Moens. 2015a. Adapting coreference resolution for narrative processing. In *Proceedings of EMNLP 2015*.

Quynh Ngoc Thi Do, Steven Bethard, and Marie-Francine Moens. 2015b. Domain adaptation in semantic role labeling using a neural language model and linguistic resources. *IEEE/ACM Transactions on Audio, Speech & Language Processing*, 23(11):1812–1823.

Quynh Ngoc Thi Do, Steven Bethard, and Marie-Francine Moens. Recurrent neural semantic frame language model. Under review.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of EMNLP 2013*.

Oswaldo Ludwig, Do Quynh Thi Ngoc, Smith Cameron, Cavazza Marc, and Moens Marie-Francine. Translating written stories into virtual reality. Under review.

Messaoud Mezati, Foudil Cherif, Cdric Sanza, and Vronique Gaildrat. 2015. An ontology for semantic modelling of virtual world. *International Journal of Artificial Intelligence & Applications*, 6(1):65–74, janvier.

Tomas Mikolov, Martin Karafit, Lukas Burget, Jan Cernock, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, editors, *INTERSPEECH*, pages 1045–1048. ISCA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A. Miller. 1995. Wordnet: A lexical database for English. *Commun. ACM*, 38(11):39–41, November.

Melzack Ronald and Jones Carol. 1967. *The Day Tuk Became a Hunter and Other Eskimo Stories*. Dodd, Mead New York.

Karin Kipper Schuler. 2005. *Verbnet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Philadelphia, PA, USA. AAI3179808.