Using Latent Dirichlet Allocation for Child Narrative Analysis

Khairun-nisa Hassanali and **Yang Liu**
The University of Texas at Dallas
Richardson, TX, USA
nisa,yangl@hlt.utdallas.edu

Thamar Solorio
University of Alabama at Birmingham
Birmingham, AL, USA
solorio@uab.edu

Abstract

Child language narratives are used for language analysis, measurement of language development, and the detection of language impairment. In this paper, we explore the use of Latent Dirichlet Allocation (LDA) for detecting topics from narratives, and use the topics derived from LDA in two classification tasks: automatic prediction of coherence and language impairment. Our experiments show LDA is useful for detecting the topics that correspond to the narrative structure. We also observed improved performance for the automatic prediction of coherence and language impairment when we use features derived from the topic words provided by LDA.

1 Introduction

Language sample analysis is a common technique used by speech language researchers to measure various aspects of language development. These include speech fluency, syntax, semantics, and coherence. For such analysis, spontaneous narratives have been widely used. Narrating a story or a personal experience requires the narrator to build a mental model of the story and use the knowledge of semantics and syntax to produce a coherent narrative. Children learn from a very early age to narrate stories. The different processes involved in generating a narrative have been shown to provide insights into the language status of children.

There has been some prior work on child language sample analysis using NLP techniques. Sahakian and Snyder (2012) used a set of linguistic features computed on child speech samples to create language metrics that included age prediction. Gabani et al. (2011) combined commonly used measurements in communication disorders with several NLP based features for the prediction of Language Impairment (LI) vs. Typically Developing (TD) children. The features they used included measures of language productivity, morphosyntactic skills, vocabulary knowledge, sentence complexity, probabilities from language models, standard scores, and error patterns. In their work, they explored the use of language models and machine learning methods for the prediction of LI on two types of child language data: spontaneous and narrative data.

Hassanali et al. (2012a) analyzed the use of coherence in child language and performed automatic detection of coherence from child language transcripts using features derived from narrative structure such as the presence of critical narrative components and the use of narrative elements such as cognitive inferences and social engagement devices. In another study, Hassanali et al. (2012b) used several coherence related features to automatically detect language impairment.

LDA has been used in the field of narrative analysis. Wallace et al. (2012) adapted LDA to the task of multiple narrative disentanglement, in which the aim was to tease apart narratives by assigning passages from a text to the subnarratives that they belong to. They achieved strong empirical results.

In this paper, we explore the use of LDA for child narrative analysis. We aim to answer two questions: Can we apply LDA to children narratives to identify meaningful topics? Can we represent these topics automatically and use them for other tasks, such as coherence detection and language impairment prediction? Our results are promising. We found that using LDA topic modeling can infer useful topics, and incorporating features derived from such automatic topics improves the performance of coherence classification and language impairment detection over the previously reported results.

Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), pages 111–115,
Sofia, Bulgaria, August 4-9 2013.

Coherence Scale	TD	LI	Total
Coherent	81	6	87
Incoherent	18	13	31
Total	**99**	**19**	**118**

Table 1: Number of TD and LI children on a 2-scale coherence level

2 Data

For the purpose of the experiments, we used the Conti-Ramsden dataset (Wetherell et al., 2007a; Wetherell et al., 2007b) from the CHILDES database (MacWhinney, 2000). This dataset consists of transcripts belonging to 118 adolescents aged 14 years. The adolescents were given the wordless picture story book "Frog, where are you?" and asked to narrate the story based on the pictures. The storybook is about the adventures of a boy who goes searching for his missing pet frog. Even though our goal is to perform child narrative analysis, we used this dataset from adoloscents since it was publicly available, and was annotated for language impairment and coherence. Of the 118 adolescents, 99 adolescents belonged to the TD group and 19 adolescents belonged to the language impaired group. Hassanali et al. (2012a) annotated this dataset for coherence. A transcript was annotated as coherent, as long as there was no difficulty in understanding the narrative, and incoherent otherwise. Table 1 gives the TD and LI distribution on a 2-scale coherence level. Figure 1 shows an example of a transcript produced by a TD child.

```
um the boy had the frog.
and he's playing with it.
and then he went to sleep and forgot to put a lid or anything on it.
and then in the middle of the night, the frog went out.
and in the morning he got really worried.
he was looking for his frog.
and he was like ignoring his dog like if it did  if it
he's just ignoring to him  ignoring him.
and he was looking around everywhere for the frog.
and then he looked up on a big rock.
and saw  and he fell on a  like a deer.
and then the deer went running and chucked him like in a pond.
and he heard the frog.
and then he saw the frog.
and then it was with like s and diff  another
and then the frogs gave him like a baby frog to have.
~
```

Figure 1: Sample transcript from a TD child

3 Narrative Topic Analysis Using LDA

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) has been used in NLP to model topics within a collection of documents. In this study, we use LDA to detect topics in narratives. Upon examining the transcripts, we observed that each topic was described in about 3 to 4 utterances. We therefore segmented the narratives into chunks of 3 utterances, with the assumption that each segment corresponds roughly to one topic.

We used the software by Blei et al.[1] to perform LDA. Prior to performing LDA, we removed the stop words from the transcripts. We chose α to be 0.8 and K to be 20, where α is the parameter of the Dirichlet prior on the per-document topic distributions and K denotes the number of topics considered in the model.

We chose to use the transcripts of TD children for generating the topics, because the transcripts of TD children have fewer disfluencies, incomplete utterances, and false starts. As we can observe from Table 1, a higher percentage of TD children produced coherent narratives when compared to children with LI.

Table 2 gives the topic words for the top 10 topics extracted using LDA. The topics in Table 2 were manually labeled after examination of the topic words extracted using LDA. We found that some of the topics extracted by LDA corresponded to subtopics. For example, searching for the frog in the house has subtopics of the boy searching for the frog in room and the dog falling out of the window, which were part of the topics covered by LDA. The subtopics are marked in italics in Table 2.

The following narrative components were identified as important features for the automatic prediction of coherence by Hassanali et al. (2012a).

1. Instantiation: introduce the main characters of the story: the boy, the frog, and the dog, and the frog goes missing
2. 1st episode: search for the frog in the house
3. 2nd episode: search for the frog in the tree
4. 3rd episode: search for the frog in the hole in the ground
5. 4th episode: search for the frog near the rock
6. 5th episode: search for the frog behind the log
7. Resolution: boy finds the frog in the river and takes a frog home

Upon examining the topics extracted by LDA, we observed that all the components mentioned above

[1] http://www.cs.princeton.edu/ blei/lda-c/index.html

Topic No	Topic Words Used by TD Population	Topic Described
1	went,frog,sleep,glass,put,caught,jar,yesterday,out,house	Introduction
2	frog,up,woke,morning,called,gone,escaped,next,kept,realized	Frog goes missing
3	window,out,fell,dog,falls,broke,quickly,opened,told,breaking	*Dog falls out of window*
4	tree,bees,knocked,running,popped,chase,dog,inside,now,flying	Dog chases the bees
5	deer,rock,top,onto,sort,big,up,behind,rocks,picked	Deer behind the rock
6	searched,boots,room,bedroom,under,billy,even, floor,tilly,tried	*Search for frog in room*
7	dog,chased,owl,tree,bees,boy,came,hole,up,more	Boy is chased by owl from a tree with beehives
8	jar,gone,woke,escaped,night,sleep,asleep,dressed,morning,frog	Frog goes missing
9	deer,top,onto,running,ways,up,rocks,popped,suddenly,know	Boy runs into the deer
10	looking,still,dog,quite,cross,obviously,smashes,have,annoyed	*Displeasure of boy with dog*

Table 2: Top 10 topic words extracted by LDA on the story telling task. Subtopics are shown in italics.

were present in these topics. Many of the LDA topics corresponded to a picture or two in the storybook.

4 Using LDA Topics for Coherence and Language Impairment Classification

We extended the use of LDA for two tasks, namely: the automatic evaluation of coherence and the automatic evaluation of language impairment. For the experiments below, we used the WEKA toolkit (Hall et al., 2009) and built several models using the naive Bayes, Bayesian net classifier, Logistic Regression, and Support Vector Machine (SVM) classifier. Of all these classifiers, the naive Bayes classifier performed the best, and we report the results using the naive Bayes classifier in Tables 3 and 4. We performed all the experiments using leave-one-out cross-validation, wherein we excluded the test transcript that belonged to a TD child from the training set when generating topics using LDA.

4.1 Automatic Evaluation of Coherence

We treat the automatic evaluation of coherence as a classification task. A transcript could either be classified as coherent or incoherent. We use the results of Hassanali et al. (2012a) as a baseline. They used the presence of narrative episodes, and the counts of narrative quality elements such as cognitive inferences and social engagement devices as features in the automatic prediction of coherence. We add the features that we automatically extracted using LDA.

We checked for the presence of at least six of the ten topic words or their synonyms per topic in a window of 3 utterances. If the topic words were present, we took this as a presence of a topic; otherwise we denoted it as an absence of a topic. In total, there were 20 topics that we extracted using LDA, which is higher compared to the 8 narrative structure topics that were annotated for by Hassanali et al. (2012a).

Table 3 gives the results for the automatic classification of coherence. As we observe in Table 3, there is an improvement in performance over the baseline. We attribute this to the inclusion of subtopics that were extracted using LDA.

4.2 Automatic Evaluation of Language Impairment

We extended the use of LDA to create a summary of the narratives. For the purpose of generating the summary, we considered only the narratives generated by TD children in the training set. We generated a summary, by choosing 5 utterances corresponding to each topic that was generated using LDA, thereby yielding a summary that consisted of 100 utterances.

We observed that different words were used to represent the same concept. For example, “look” and “search” were used to represent the concept of searching for the frog. Since the narration was based on a picture storybook, many of the children used different terms to refer to the same animal. For example, “the deer” in the story has been interpreted to be “deer”, “reindeer”, “moose”, “stag”, “antelope” by different children. We created an extended topic vocabulary using Wordnet to include words that were semantically similiar to the topic keywords. In addition, for an utterance to be

Feature Set	Coherent			Incoherent			Accuracy (%)
	Precision	Recall	F-1	Precision	Recall	F-1	
Narrative (Hassanali et al., 2012a) (baseline)	0.869	0.839	0.854	0.588	0.645	0.615	78.814
Narrative + automatic topic features	**0.895**	**0.885**	**0.89**	**0.688**	**0.71**	**0.699**	**83.898**

Table 3: Automatic classification of coherence on a 2-scale coherence level

in the summary, we put in the additional constraint that neighbouring utterances within a window of 3 utterances also talk about the same topic. We used this summary for constructing unigram and bigram word features for the automatic prediction of LI.

The features we constructed for the prediction of LI were as follows:

1. Bigrams of the words in the summary
2. Presence or absence of the words in the summary regardless of the position
3. Presence or absence of the topics detected by LDA in the narratives
4. Presence or absence of the topic words that were detected using LDA

We used both the topics detected and the presence/absence of topic words as features since the same topic word could be used across several topics. For example, the words "frog", "dog", "boy", and "search" are common across several topics. We refer to the above features as "new features".

Table 4 gives the results for the automatic prediction of LI using different features. As we can observe, the performance improves to 0.872 when we add the new features to Gabani's and the narrative structure features. When we use the new features by themselves to predict language impairment, the performance is the worst. We attribute this to the fact that other feature sets are richer since these features take into account aspects such as syntax and narrative structure.

We performed feature analysis on the new features to see what features contributed the most. The top scoring features were the presence or absence of the topics detected by LDA that corresponded to the introduction of the narrative, the resolution of the narrative, the search for the frog in the room, and the search for the frog behind the log. The following bigram features generated from the summary contributed the most: "deer rock", "lost frog", and "boy hole". Using a subset of these best features did not improve the performance when we added them to the narrative features and Gabani's features.

Feature	P	R	F-1
Gabani's (Gabani et al., 2011)	0.824	0.737	0.778
Narrative (Hassanali et al., 2012a)	0.385	0.263	0.313
New features	0.308	0.211	0.25
Narrative + Gabani's	0.889	0.842	0.865
Narrative + Gabani's + new features	**0.85**	**0.895**	**0.872**

Table 4: Automatic classification of language impairment

5 Conclusions

In this paper, we explored the use of LDA in the context of child language analysis. We used LDA to extract topics from child language narratives and used these topic keywords to create a summary of the narrative and an extended vocabulary. The topics extracted using LDA not only covered the main components of the narrative but also covered subtopics too. We then used the LDA topic words and the summary to create features for the automatic prediction of coherence and language impairment. Due to higher coverage of the LDA topics as compared to manual annotation, we found an increase in performance of both automatic prediction of coherence and language impairment with the addition of the new features. We conclude that the use of LDA to model topics and extract summaries is promising for child language analysis.

Acknowledgements

This research is supported by NSF awards IIS-1017190 and 1018124.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Keyur Gabani, Thamar Solorio, Yang Liu, Khairunnisa Hassanali, and Christine A. Dollaghan. 2011. Exploring a corpus-based approach for detecting language impairment in monolingual English-speaking children. *Artificial Intelligence in Medicine*, 53(3):161–170.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Khairun-nisa Hassanali, Yang Liu, and Thamar Solorio. 2012a. Coherence in child language narratives: A case study of annotation and automatic prediction of coherence. In *Proceedings of WOCCI 2012 - 3rd Workshop on Child, Computer and Interaction*.

Khairun-nisa Hassanali, Yang Liu, and Thamar Solorio. 2012b. Evaluating NLP features for automatic prediction of language impairment using child speech transcripts. In *Proceedings of INTERSPEECH*.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Lawrence Erlbaum Associates.

Sam Sahakian and Benjamin Snyder. 2012. Automatically learning measures of child language development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 95–99. Association for Computational Linguistics.

Bryon C. Wallace. 2012. Multiple narrative disentanglement: Unraveling infinite jest. In *Proceeding of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–10.

Danielle Wetherell, Nicola Botting, and Gina Conti-Ramsden. 2007a. Narrative in adolescent specific language impairment (SLI): a comparison with peers across two different narrative genres. *International Journal of Language & Communication Disorders*, 42(5):583–605.

Danielle Wetherell, Nicola Botting, and Gina Conti-Ramsden. 2007b. Narrative skills in adolescents with a history of SLI in relation to non-verbal IQ scores. *Child Language Teaching and Therapy*, 23(1):95.