Automatic detection of pragmatic deficits in children with autism

Emily Prud’hommeaux and Masoud Rouhizadeh
Center for Spoken Language Understanding, Oregon Health & Science University, Portland, Oregon, USA
emilypx@gmail.com, rouhizad@ohsu.edu

Abstract

Autism spectrum disorder (ASD) is characterized by atypical and idiosyncratic language, which often has its roots in pragmatic deficits. Identifying and measuring pragmatic language ability is challenging and requires substantial clinical expertise. In this paper, we present a method for automatically identifying pragmatically inappropriate language in narratives using two features related to relevance and topicality. These features, which are derived using techniques from machine translation and information retrieval, are able to distinguish the narratives from children with ASD from those of their language-matched peers and may prove useful in the development of automated screening tools for autism and neurodevelopmental disorders.

Index Terms: spoken language evaluation, child language, diagnostic tools, discourse analysis

1. Introduction

Atypical or idiosyncratic language has been associated with autism since Kanner first gave a name to the disorder [1]. This atypicality is used as a diagnostic criterion in many of the widely used screening instruments for autism [2, 3, 4]. The current view is that phonological, morphological, and syntactic skills are typically spared in autism spectrum disorders (ASD), while pragmatic expression is impaired [5, 6], which contributes to the perceived atypicality of language in individuals with ASD.

Narrative retelling tasks, in which a subject listens to a brief narrative and must retell the narrative to the examiner, are widely used in neuropsychological assessment instruments in order to test memory and language ability [7, 8, 2]. Since relating a narrative requires pragmatic competence, the study of narrative production in ASD is particularly useful for revealing pragmatically atypical language. A narrator must stay on topic and present the story in a meaningful way in order to guarantee that the listener understands the narrative. Deviations from these expected communication behaviors can be indicative of pragmatic language problems.

In this paper, we present work on identifying features associated with pragmatic language deficits in the spoken narrative retellings of children with ASD. Previous work in the neuropsychology literature on this topic has relied on extensive expert manual annotation. In contrast, the features discussed here, while grounded in observations from the literature, can be derived in an automated fashion using techniques adapted from machine translation and information retrieval.

The evaluation of pragmatic language ability on a very general scale, as is required by many of the standard diagnostic instruments for autism, is inherently subjective and difficult to perform reliably. The automated methods described here provide an entirely objective and efficient way of characterizing some of the specific pragmatic deficits observed in ASD narratives. Although there are many other factors contributing to atypical language in ASD, automated approaches like these could lead to future work in developing diagnostic screening tools for autism and other developmental disorders.

2. Background

Clinically elicited narratives are often analyzed in terms of syntactic or lexical properties, such as sentence length, vocabulary diversity, or grammatical complexity, but such features rarely distinguish children with ASD from their language-matched peers [9, 10]. The more reliable distinctive features tend to be related to pragmatics. Expert manual annotation and analysis of narratives in young people with ASD has revealed at least one consistent trend. The narratives of children and young adults with autism include significantly more bizarre, irrelevant, and off-topic content [11, 12, 13] than the retellings of their language-matched peers. In addition, ASD narratives contain more frequent instances of inappropriate story-telling [10].

Existing work on automated analysis of narratives in children has focused primarily on syntactic and structural features rather than pragmatic or semantic features. Using the CHILDES corpus [14], which includes a narrative task, Sagae et al. [15] extracted grammatical relations in order to calculate the Index of Productive Syntax, a measure of syntactic development in children [16]. Gabani et al. [17] used part-of-speech language models, as well as other features such as utterance length and subject-verb agreement, to characterize agrammaticity in narratives of children with language impairments.
Automated analysis of the content and pragmatics of narratives in the elderly has received more attention. In particular, several researchers have developed techniques for scoring narrative retellings relative to the source narrative. Latent semantic analysis [18] and unigram overlap [19] have both been used to approximate the standard scoring procedures normally applied to narrative retellings. In previous work, we automatically scored narrative retellings produced by elderly subjects, according to the published scoring guidelines, by aligning the retellings to the source narrative [20]. We have also used alignments to explore other narrative features relating to the relevance of content and ordering of events in the narrative retellings of elderly subjects [21]. In the analysis of children’s narrative described in Section 4, we will use alignments to derive a measure of topical content. In addition, we will explore measures of word atypicality that can be estimated using word centrality measures used in information retrieval.

3. Data

3.1. Study participants

In this paper, we analyze the narrative retellings of 97 children, ranging in age from 4 years to 8 years. All subjects were required to have a full-scale IQ of greater than 70, indicating no intellectual disability, and a mean length of utterance (MLU) of at least 3, indicating an adequate level of verbal fluency. IQ was measured using the WPPSI-III [22] for children under 7, and the WISC-IV [23] for children 7 and older. Exclusion criteria included any brain lesion or neurological condition unrelated to the diagnoses of interest; orofacial abnormalities; bilinguality; extreme unintelligibility; gross sensory or motor impairments; or identified mental retardation. In addition, the exclusion criteria for the typical development (TD) group included a history of psychiatric disturbance and a family history ASD or language impairment.

We categorize the subjects into four diagnostic groups: typical development (TD, n = 40); specific language impairment not meeting the criteria for autism (SLI, n = 17); ASD not meeting criteria for a language impairment (ALN, n = 21); and ASD meeting the criteria for a language impairment (ALI, n = 21). A subject received a diagnosis of a language impairment if his CELF Core Language Index [24] was more than one standard deviation below the mean. A diagnosis of ASD was indicated if the child reached the ASD cut-off score on two instruments: the Autism Diagnostic Observation Schedule (ADOS) [2], a semi-structured series of activities designed to allow an examiner to observe behaviors associated with autism; and the Social Communication Questionnaire (SCQ) [4], a parental questionnaire. A team of experienced clinicians generated for each child a consensus diagnosis according the instrumental findings and the diagnostic criteria listed in the DSM-IV-TR [25].

3.2. NEPSY Narrative Memory

The narrative retelling task analyzed here is the NEPSY Narrative Memory task, which was administered to the subjects as part of an earlier study on prosody in autism and language disorders. The NEPSY Narrative Memory (NNM) task is a narrative retelling test in which the subject listens to a brief narrative and then must retell the narrative to the examiner. The examination scoring materials include a list of 17 story elements, each corresponding to a word or phrase expressing a detail or event in the narrative. Figure 1 shows the NNM narrative, with the portions of the text corresponding to the 17 story elements placed in brackets and numbered according to the published scoring guidelines. The score for a retelling re-
ported under standard administration of the task is simply the total number of story elements recalled in the retelling. The elements do not need to be recalled verbatim or in the correct order, and all elements contribute equally to the final score.

The retelling of the NNM narrative for each experimental subject was recorded and later transcribed at the word level, tokenized, and downcased, and all punctuation and pause-fillers were removed. In Figures 2 and 3, we provide two short sample retellings produced by a child with SLI and a child with ALI. The child with SLI struggles to tell the story, but stays on-topic, first setting the stage by describing the characters and then by outlining the resolution of the story, when Pepper brings the shoe to Anna so that she can rescue Jim. The child with ALI, on the other hand, very quickly strays from the source narrative. The contrast between these two retellings illustrates some of the features of ASD retellings described in Section 2.

4. Measuring topicality and relevance

4.1. Word alignment

The first topicality and relevance feature we will discuss is a measure of how well a retelling can be aligned to the source narrative. In previous work on scoring narrative retellings in the elderly, we generated word-level alignments between a retelling and the source narrative, from which we then extracted several narrative features, including a measure of relevance [21]. These word alignments are similar to those used in machine translation (MT), where each word can be considered a translation of the word to which it is aligned.

Standard word alignment packages for machine translation, such as the Berkeley aligner [26] and Giza++ [27], start with a parallel corpus consisting of sentences in one language, the source language, and translations of those sentences in another language, the target language. These packages then use a pipeline of expectation maximization techniques to determine in an unsupervised fashion which words in each source sentence are aligned to (i.e., are translations of) the words in each corresponding target sentence. The techniques used to generate word alignments between two languages in MT can also be used to align a collection of narrative retellings to a source narrative. Figure 4 shows a snippet of the word-level alignment between the SLI retelling in Figure 3 in the source narrative shown in Figure 1.

In order to automatically generate word alignments, we need to collect an in-domain parallel corpus and then train a word alignment model on that corpus. In order to ensure a perfect match for the domain, we build a parallel corpus consisting only of data from the NNM narrative retellings. From this corpus, a word alignment model that is tailored to this specific alignment task can be learned.

The corpus includes: (1) a source-to-retelling corpus containing the bracketed portions of the source narrative, representing the story elements, and the retellings for all 97 study participants, as well as roughly 160 retellings from other individuals not participating in the study; (2) a full pairwise retelling-to-retelling corpus; (3) a manually derived corpus of the phrases in the non-experimental retellings aligned with their corresponding story elements in the source narrative; (4) a word identity corpus, consisting of each word that appeared in a retelling aligned to itself. Example lines from each of these corpora are provided in Figure 5.

We use the Berkeley aligner [26] to build an alignment model on this four-part parallel corpus. The Berkeley aligner estimates a set of initial word-to-word alignment probabilities using a simple mixture model relying on co-occurrence statistics. These initial probability estimates serve as the input probabilities for an HMM, in which expectation maximization is used not only to refine the word-to-word alignment probabilities but also to estimate probabilities modeling the reordering of the words in a retelling relative to the source narrative. We use the alignment model generated by the Berkeley aligner to align every retelling to the bracketed portions of the source narrative corresponding from the story elements.

Recall that previous work has found that the retellings of children with ASD contain more off-topic and bizarre content and more instances of inappropriate story telling. The feature we extract from the alignments captures how well the child stays on topic during his retelling by measuring the amount of a retelling’s content that is related to the content of the source narrative. First we determine which words in a retelling transcript are content words rather than function words. We then count the number of these content words that are aligned to some word in the source narrative. The percentage of content words in a child’s retelling that cannot be aligned to a word in the source narrative is an estimate of how much of the retelling is unrelated to the source narrative, which in turn is a measure of how far off-topic the child has veered.

Figure 4: Snippet of the word-level alignment between the SLI sample retelling and the source narrative.
Jim Pepper, a big black dog, liked to walk in the woods and climb the oak tree with branches too high for him to reach. One day, he took a ladder from home and carried it to the oak tree. Climbing up, he looked out over his neighborhood. His foot slipped, and his shoe fell off. Pepper took the shoe in his mouth and ran away. Jim felt sad, thinking his friend didn't want to stay with him when he was in trouble. Pepper took the shoe to Anna, Jim's sister. She barked and barked until Anna put the ladder up and rescued Jim.

He had a friend named Pepper. Pepper was a black dog but James forgot Pepper got his shoe. He didn't know if it was a big dog or not. Pepper was his friend, and Pepper rescued his shoe when he brought it to Anna. That's all James knew, and then Anna rescued him.

### 4.2. Log odds score

Another way to evaluate the relevance and topicality of a retelling is to measure the degree to which the retelling contains unexpected or unlikely words given the context of the source narrative. We expect that different retellings of the same source narrative will share a common vocabulary. The presence of words outside of this vocabulary in a retelling may indicate that the speaker has strayed from the story or is failing to select the appropriate words, both of which are signs of pragmatic language difficulties.

Identifying these outliers is similar to the information retrieval (IR) task of clustering documents together based on a common topic. Documents with a similar vocabulary are likely to be about the same topic; documents containing numerous words that are not found in the other documents are likely to be about a different topic. There are several document lexical similarity measures used in IR to cluster documents. Here we will use one of those similarity measures, the log odds score, to characterize the relevance and topicality of a retelling relative to the other retellings.

In order to establish the topic of the NNM narrative relative to other topics, we include transcribed conversational data from four interactive activities of the Autism Diagnostic Observation Schedule (ADOS) [2], an instrument used in the diagnosis of ASD. Each child’s retelling is compared to a collection of documents that includes every other retelling as well as each of the 4 ADOS activities for every other child. The log odds score for each word in a child’s retelling will be 0 or close to 0 if the word is as likely to appear in the retelling as in any other document. As the log odds score increases, so does the probability that the word is specific to the retelling and unlikely to appear in another retelling or any of the ADOS activities.

In order to generate a single log odds score for a complete retelling, we calculate the log odds score for each word in the retelling and take the mean of that set of scores. A retelling with a high mean log odds score likely contains many words that are unusual or unexpected given the context. We predict that this will indicate that the child has strayed off-topic or is failing to communicate the story in a typical or appropriate way.

### 5. Diagnostic classification

We now proceed to use these two features of topicality and relevance in order to perform diagnostic classification. Recall that many of the features of narratives that purport to distinguish typical development from ASD fail to distinguish the groups when they are matched for language ability. For this reason, we will perform three pairwise classification tasks: 1) typical development (TD) vs. ASD with no language impairment (ALN); 2) specific language impairment (SLI) vs. ASD with a language impairment (ALI); and 3) Non-ASD (TD and SLI) vs. ASD (ALN and ALI).

We perform classification using the support vector machine (SVM) implemented in LibSVM [28] within the Weka API [29]. Classification accuracy is evaluated using a leave-pair-out cross-validation scheme, in which ev-
children with ASD. Finally, we plan to experiment with identifying unusual and unexpected words in the retellings of transcripts using graph-based methods, to more accurately explore other narrative coherence features, such as element ordering, and other document similarity measures that rely on extensive manual annotation and clinical expertise. The features we developed are grounded in clinical observations of language in ASD but were extracted from transcripts in an unsupervised and automatic fashion. The results indicate that searching for additional features of topicality and relevance that can be derived using NLP techniques may further improve diagnostic classification accuracy.

Although we were able to use these automatically derived features to distinguish children with ASD from their language-matched peers, we have not yet manually determined whether these features correlate with human judgements of relevance, topicality, and coherence, or whether such judgements are able to distinguish the subjects in our data set. We plan to manually code the transcripts according to the guidelines described in previous work in the neuropsychology literature on narrative coherence in ASD. In addition, we would like to explore other narrative coherence features, such as element ordering, and other document similarity measures that are widely used in information retrieval, such as those derived using graph-based methods, to more accurately identify unusual and unexpected words in the retellings of children with ASD. Finally, we plan to experiment with using ASR to generate the transcripts in order to produce a fully automated narrative-based screening tool.

### 7. Acknowledgments
This research was supported in part by NIH NIDCD grant R01DC012033 and NSF grant 0826654. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not reflect the views of the NIH or NSF.

### 8. References

<table>
<thead>
<tr>
<th>Features</th>
<th>TD v. ALN</th>
<th>SLI v. ALI</th>
<th>Non-ASD v. ASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log odds</td>
<td>0.66</td>
<td>0.61</td>
<td>0.68</td>
</tr>
<tr>
<td>Unaligned content</td>
<td>0.71</td>
<td>0.63</td>
<td>0.69</td>
</tr>
<tr>
<td>Log odds + unaligned content</td>
<td>0.71</td>
<td>0.63</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Table 1: Classification accuracy for relevance and topicality features: AUC.


