Measuring idiosyncratic interests in children with autism

Masoud Rouhizadeh[†], **Emily Prud'hommeaux**[°], **Jan van Santen**[†], **Richard Sproat**[§]
[†]Oregon Health & Science University
[°]Rochester Institute of Technology
[§] Google, Inc.
{rouhizad,vansantj}@ohsu.edu, emilypx@rit.edu, rws@xoba.com

Abstract

A defining symptom of autism spectrum disorder (ASD) is the presence of restricted and repetitive activities and interests, which can surface in language as a perseverative focus on idiosyncratic topics. In this paper, we use semantic similarity measures to identify such idiosyncratic topics in narratives produced by children with and without ASD. We find that neurotypical children tend to use the same words and semantic concepts when retelling the same narrative, while children with ASD, even when producing accurate retellings, use different words and concepts relative not only to neurotypical children but also to other children with ASD. Our results indicate that children with ASD not only stray from the target topic but do so in idiosyncratic ways according to their own restricted interests.

1 Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by impaired communication and social behavior. One of the core symptoms is a preoccupation with specific restricted interests (American Psychiatric Association, 2013), and several commonly used diagnostic instruments for ASD instruct examiners to evaluate the degree to which subjects display this characteristic (Lord et al., 2002; Rutter et al., 2003). In verbal individuals with ASD, such a preoccupation can be expressed as a tendency to fixate on a particular idiosyncratic topic.

Previous research relying on expert annotation of spoken language in children with ASD has found that their spoken narratives and conversations include significantly more instances of irrelevant content and more topic digressions (Loveland et al., 1990; Losh and Capps, 2003; Lam et al., 2012). Similar results at the lexical level have been reported using automated annotations (Prud'hommeaux and Rouhizadeh, 2012; Rouhizadeh et al., 2013). There has been little work, however, in characterizing the precise direction of the departure from a target topic, leaving open the question of whether children with ASD are instigating similar, potentially reasonable topic changes or whether they are introducing idiosyncratic topics consistent with their own restricted interests.

In this paper, we attempt to automatically identify topic digressions in the narrative retellings of children with ASD and to determine whether these digressions are influenced by their idiosyncratic or restricted interests. From a corpus of spoken retellings of the same brief narrative, we extract several measures designed to capture different facets of semantic similarity between a pair of retellings. We find that the retellings of children with typical development (TD) semantically resemble one another much more than they resemble retellings by children with ASD. This indicates that TD children are adhering to a common target topic, while children with ASD are introducing topic changes. More strikingly, the similarity between pairs of ASD retellings is even lower, suggesting that children with ASD are straying from the target topic in individual and idiosyncratic ways. Although we do not yet have manual annotations to confirm that these topic shifts correspond to the particular restricted interests of each study participant, our methods and results show the potential of using automated analysis for revealing diagnostically relevant linguistic features.

2 Data

Thirty-nine children with typical development (TD) and 21 high-functioning children with ASD,

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 212–217,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

ranging in age from 4 to 9 years, participated in this study. ASD was diagnosed via clinical consensus according to the DSM-IV-TR criteria (American Psychiatric Association, 2000) and the established thresholds on two widely-used diagnostic instruments: the Autism Diagnostic Observation Schedule (Lord et al., 2002) and the Social Communication Questionnaire (Rutter et al., 2003). No children met the criteria for a language impairment, and there were no significant between-group differences in age or full-scale IQ.

To elicit retellings, we used the Narrative Memory subtest of the NEPSY (Korkman et al., 1998), a large battery of tasks testing neurocognitive functioning in children. In the NEPSY Narrative Memory (NNM) subtest, the subject listens to a brief narrative about a boy and his dog and then must retell the narrative to the examiner. Figure 1 shows two sample retellings from our corpus. The NNM was administered by a trained clinician to each study participant, and each participant's retelling was recorded, transcribed, and evaluated according to the published scoring guidelines.

Under standard administration of the NNM, a retelling is scored according to how many story elements from a predetermined list it contains. The guidelines for scoring do not require verbatim recall for most elements and generally allow the use of synonyms and paraphrases. As is typically reported when comparing matched groups (Diehl et al., 2006), we observed no significant difference in the standard NNM free recall score between the TD group (mean = 6.25, sd = 3.43) and the ASD group (mean = 4.90, sd = 3.72). It might seem that a low similarity score between two retellings simply indicates that one retelling includes fewer story elements. However, given the equivalent number of story elements recalled by the two groups, we can assume that a low similarity score indicates a difference in the quality rather than the quantity of information in the retellings.

3 Semantic similarity measures

We expect that two different retellings of the same narrative will lie in the same lexico-semantic space and will thus have high similarity scores. In this work we use well-known similarity measures with two modifications. Children with autism tend to use more off-topic and unexpected words. Such words always have high inverse document frequency (IDF) scores since they are very specific to a particular retelling. By including IDF weights, a similarity measure would be biased toward off-topic words rather than actual content words in the story elements. Conventional IDF weights are therefore not useful for our particular purpose. Instead, we remove closed-class function words to avoid their bias in our similarity measures. In addition, we lemmatize our narrative corpus to reduce the sparsity due to inflectional variation.

3.1 Word overlap measures

3.1.1 Jaccard similarity coefficient

The Jaccard similarity coefficient (Sim_{Jac}) (Jaccard, 1912) is a simple word overlap measure between a pair of narratives n and m defined as the size of intersection of the words in narratives n and m, relative to the size of word union of n and m:

$$Sim_{Jacc}(n,m) = \frac{|n \cap m|}{|n \cup m|} \quad (1)$$

3.1.2 Cosine similarity score

Cosine similarity score Sim_{Cos} is the similarity between two narratives by cosine of the angle between their vector. We use a non-weighted cosine similarity based on the following formula, where $tf_{w,n}$ is the term frequency of word w in narrative n:

$$Sim_{Cos}(n,m) = \frac{\sum\limits_{w \in n \cap m} tf_{w,n} \times tf_{w,m}}{\sqrt{\sum\limits_{w_i \in n} (tf_{w_i,n})^2} \sqrt{\sum\limits_{w_j \in m} (tf_{w_j,m})^2}} \quad (2)$$

3.1.3 Relative frequency measure

Relative frequency measure (Sim_{RF}) (Hoad and Zobel, 2003) is an author identity measure for identifying plagiarism at the document level. This measure normalizes the frequency of the words appearing in both narratives n and m by the overall length of the two narratives, as well as the relative frequency of the words common to the two narratives. We used a simplified variation of this measure, described by Metzler et al. (2005) and formulated as follows:

$$Sim_{RF}(n,m) = \frac{1}{1 + \frac{max(|n|,|m|)}{min(|m|,|m|)}} \times \sum_{w \in n \cap m} \frac{1}{1 + |tf_{w,n} - tf_{w,m}|} \quad (3)$$

Jim went up a tree with a ladder. He lost his shoe he got stuck he hung from a branch. Pepper took his shoe. He showed it to his sister and she helped him down. Let me look at this picture with my trusty vision gadget.

The boy got stuck and someone rescued him and pepper was a really smart dog. Dogs have a great sense of smell too, like T-rex. T-rex could smell things that were really far away. T-rex could be over there and the meat could be way back there under the couch Well, that guy got stuck on the tree and then he, and then Pepper, his shoe fell out of the tree. Anna rescued it. Pepper brought his shoe back and Anna rescued them.

Figure 1: Two topically different NNM retellings with similar free recall scores (6 and 5, respectively).

3.1.4 BLEU

BLEU (Papineni et al., 2002) is commonly used measure of n-gram overlap for automatically evaluating machine translation output. Because it is a precision metric, the BLEU score for any pair of narratives n and m will depend on which narrative is considered the "reference". To create a single BLEU-based overlap score for each pair of narratives, we calculate $Sim_{BLEU(n,m)}$ as the mean of $BLEU(m,n)$ and $BLEU(n,m)$.

3.2 Knowledge-based measures

It is reasonable to expect people to use synonyms or semantically similar words in their narratives retellings. It is therefore possible that children with autism are discussing the appropriate topic but choosing unusual words within that topic space in their retellings. We therefore use a set of measures that consider the semantic overlap of two narratives using WordNet (Fellbaum, 1998) similarities (Achananuparp et al., 2008), in order to distinguish instances of atypical but semantically appropriate language from true examples of poor topic maintenance. Because WordNet-based similarity measures only consider word pairs with the same part-of-speech, we POS-tagged the data using a perceptron tagger (Yarmohammadi, 2014).

3.2.1 WordNet-based vector similarity

In a modified version ofWordNet-based vector similarity, Sim_{WV}), (Li et al., 2006), we first create vectors v_n and v_m for each narrative n and m, where each element corresponds to a word in the type union of n and m. We assign values to each element e in v_n using the following formulation:

$$S(e,n) = \begin{cases} 1 & \text{if } e \in n \\ \max\limits_{w_i \in n} LS(e,w_i) & \text{otherwise} \end{cases} \quad (4)$$

where LS is Lin's universal similarity (Lin, 1998). In other words, if the element e is present in n, $S(e,n)$ will be 1. If not, the most similar word to e will be chosen from words in n using Lin's universal similarity and $S(e,n)$ will be that maximum score. The same procedure is applied to v_m, and finally the similarity score between n and m is derived from the cosine score between v_n and v_m.

3.2.2 WordNet-based mutual similarity

In a modified version of WordNet-based mutual similarity (Sim_{WM}) (Mihalcea et al., 2006), we find the maximum similarity score $S(w_i,m)$ for each word w_i in narrative n with words in narrative m as described in Equation 4. The same procedure is applied to narrative m, and Sim_{WM} is calculated as follows:

$$Sim_{WM}(n,m) = \frac{1}{2}\left(\frac{\sum\limits_{w_i \in n} S(w_i,m)}{|n|} + \frac{\sum\limits_{w_j \in m} S(w_j,n)}{|m|}\right) \quad (5)$$

4 Results

For each of the semantic similarity measures, we build a similarity matrix comparing every possible pair of children. Because this pairwise similarity matrix is diagonally symmetrical, we need only consider the top right section of the matrix above the diagonal in our analyses. Table 1 shows the mean semantic overlap scores between the narratives for each of the three sub-matrices described above. We see that for both the word-overlap and the knowledge-based semantic similarity measures described in Section 3, TD children are most similar to other TD children. ASD children are less similar to TD children than TD children are to one another; and children with ASD are even less similar to other ASD children than to TD children.

Our goal is to explore the degree of similarity, as measured by the semantic overlap measures, within and across diagnostic groups. With this in mind, we consider the following three sub-matrices for each similarity matrix: one in which each TD child is compared with every other

	TD.TD	TD.ASD	ASD.ASD
Sim_{Jac}	0.19	0.14	0.11
Sim_{Cos}	0.42	0.34	0.28
Sim_{RF}	2.07	1.52	1.08
Sim_{BLEU}	0.36	0.29	0.24
Sim_{WV}	0.54	0.47	0.42
Sim_{WM}	0.80	0.69	0.59

Table 1: Average semantic overlap scores for each group.

measure	*statistic*	***p-values***		
		TD.TD vs ASD.ASD	TD.TD vs TD.ASD	TD.ASD vs ASD.ASD
Sim_{Jac}	t	.014	.022	.022
	w	.012	.002	.002
Sim_{Cos}	t	.025	.043	.027
	w	.025	.001	.001
Sim_{RF}	t	.056	.072	.046
	w	.012	.002	.002
Sim_{BLEU}	t	.032	.039	.034
	w	.036	.002	.002
Sim_{WV}	t	.014	.008	.028
	w	.01	.01	.01
Sim_{WM}	t	.018	.007	.042
	w	.018	.002	.002

Table 2: Monte Carlo significance test p-values for each similarity measure.

TD child (the TD.TD sub-matrix); one in which each ASD child is compared with every other ASD child (the ASD.ASD sub-matrix); and one in which each child is compared with the children in the diagnostic group to which he does not belong (the TD.ASD sub-matrix).

Note that we have no a priori reason to assume that the similarity scores are from any particular distribution. In order to calculate the statistical significance of these between-group differences, we therefore apply a Monte Carlo permutation method, a non-parametric procedure commonly used in non-standard significance testing situations. For each pair of sub-matrices (e.g., TD.TD vs ASD.ASD) we calculate two statistics that compare the cells in one sub-matrix with the cells in other sub-matrices: the t-statistic, using the Welch Two Sample t-test; and the w-statistic, using the Wilcoxon rank sum test. We next take a large random sample with replacement from all possible permutations of the data by shuffling the diagnosis labels of the children 1000 times. We then calculate two above statistics for each shuffle and count the number of times the observed values exceed the values produced by the 1000 shuffles.

Applying the Monte Carlo permutation method, we calculate the statistical significance of the following comparisons: TD.TD vs ASD.ASD; TD.TD vs TD.ASD; and TD.ASD vs ASD.ASD. Table 2 summarizes the results of these significance tests. In all cases, the differences are significant at $p < 0.05$ except for the first two comparisons in the t-test permutation of Sim_{RF}, which narrowly eluded significance.

5 Conclusions and future work

High-functioning children with ASD have long been described as "little professors", using pedantic or overly-adult language (Asperger, 1944). Low lexical overlap similarity measures by themselves might indicate that children with ASD are using semantically appropriate but infrequent or sophisticated words that were not used by other children. We note, however, that the knowledge-based overlap measures follow the same pattern as the purely lexical overlap measures. This suggests that it not the case that children with ASD are simply using rare synonyms of the more common words used by TD children. Instead, it seems that the children with ASD are moving away from the target topic and following their own individual and idiosyncratic semantic paths. These findings

provide additional quantitative evidence not only for the common qualitative observation that young children with ASD have difficulty with topic maintenance but also for the more general behavioral symptom of idiosyncratic and restricted interests.

The overlap measures presented in this paper could be used as features for machine learning classification of ASD in combination with other linguistic features we have explored, including the use of off-topic lexical items (Rouhizadeh et al., 2013), features associated with poor pragmatic competence (Prud'hommeaux et al., 2014), and repetitive language measures (van Santen et al., 2013). Recall, however, that a clinician must consider a wide range of social, communication, and behavioral criteria when making a diagnosis of ASD, making it unlikely that language features alone could perfectly predict a diagnosis of ASD. The more significant potential in our approaches is more likely to lie in the area of language deficit detection and remediation.

A focus of our future work will be to manually annotate the data to determine the frequency and nature of the topic excursions. It is our expectation that children with ASD do not only veer from the target topic more frequently than typically developing children but also pursue topics of their own individual specific interests. We also plan to apply our methods to ASR output rather than manual transcripts. Despite the high word error rates typically observed with this sort of audio data, we anticipate that our methods, which rely primarily on content words, will be relatively robust.

The work presented here demonstrates the utility of applying automated analysis methods to spoken language collected in a clinical settings for diagnostic and remedial purposes. Carefully designed tools using such methods could provide helpful information not only to clinicians and therapists working with children with ASD but also to researchers exploring the specific linguistic and behavioral deficits associated with ASD.

Acknowledgments

This work was supported in part by NSF grant #BCS-0826654, and NIH NIDCD grants #R01-DC007129 and #1R01DC012033-01. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF or the NIH.

References

Palakorn Achananuparp, Xiaohua Hu, and Xiajiong Shen. 2008. The evaluation of sentence similarity measures. In *Data Warehousing and Knowledge Discovery*, pages 305–316. Springer.

American Psychiatric Association. 2000. *DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, Washington, DC.

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders (5th ed.)*. American Psychiatric Publishing, Washington, DC.

Hans Asperger. 1944. Die "autistischen psychopathe" im kindesalter. *Archiv fur Psychiatrie und Nervenkrakheiten*, 117:76–136.

Joshua J. Diehl, Loisa Bennetto, and Edna Carter Young. 2006. Story recall and narrative coherence of high-functioning children with autism spectrum disorders. *Journal of Abnormal Child Psychology*, 34(1):87–102.

Christian Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Timothy C Hoad and Justin Zobel. 2003. Methods for identifying versioned and plagiarized documents. *Journal of the American society for information science and technology*, 54(3):203–215.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Marit Korkman, Ursula Kirk, and Sally Kemp. 1998. *NEPSY: A developmental neuropsychological assessment*. The Psychological Corporation, San Antonio.

Yan Grace Lam, Siu Sze, and Susanna Yeung. 2012. Towards a convergent account of pragmatic language deficits in children with high-functioning autism: Depicting the phenotype using the pragmatic rating scale. *Research in Autism Spectrum Disorders*, 6(2):792–797.

Yuhua Li, David McLean, Zuhair A Bandar, James D O'shea, and Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on*, 18(8):1138–1150.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304.

Catherine Lord, Michael Rutter, Pamela DiLavore, and Susan Risi. 2002. *Autism Diagnostic Observation Schedule (ADOS)*. Western Psychological Services, Los Angeles.

Molly Losh and Lisa Capps. 2003. Narrative ability in high-functioning children with autism or asperger's syndrome. *Journal of Autism and Developmental Disorders*, 33(3):239–251.

Katherine Loveland, Robin McEvoy, and Belgin Tunali. 1990. Narrative story telling in autism and down's syndrome. *British Journal of Developmental Psychology*, 8(1):9–23.

Donald Metzler, Yaniv Bernstein, W Bruce Croft, Alistair Moffat, and Justin Zobel. 2005. Similarity measures for tracking information flow. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 517–524.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational :inguistics*, pages 311–318.

Emily Prud'hommeaux and Masoud Rouhizadeh. 2012. Automatic detection of pragmatic deficits in children with autism. In *Proceedings of the 3rd Workshop on Child, Computer and Interaction (WOCCI)*, pages 1–6.

Emily Prud'hommeaux, Eric Morley, Masoud Rouhizadeh, Laura Silverman, Jan van Santen, Brian Roark, Richard Sproat, Sarah Kauper, and Rachel DeLaHunta. 2014. Computational analysis of trajectories of linguistic development in autism. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 266–271.

Masoud Rouhizadeh, Emily Prud'hommeaux, Brian Roark, and Jan van Santen. 2013. Distributional semantic models for the evaluation of disordered language. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Michael Rutter, Anthony Bailey, and Catherine Lord. 2003. *Social Communication Questionnaire (SCQ)*. Western Psychological Services, Los Angeles.

Jan van Santen, Richard Sproat, and Alison Presmanes Hill. 2013. Quantifying repetitive speech in autism spectrum disorders and language impairment. *Autism Research*, 6(5):372–383.

Mahsa Yarmohammadi. 2014. Discriminative training with perceptron algorithm for pos tagging task. Technical Report CSLU-2014-001, Center for Spoken Language Understanding, Oregon Health & Science University.