Published in final edited form as:
Proc Conf. 2013 June ; 2013: 709–714.

Distributional semantic models for the evaluation of disordered language

Masoud Rouhizadeh[†], **Emily Prud'hommeaux**[○], **Brian Roark**[†], and **Jan van Santen**[†]

Masoud Rouhizadeh: rouhizad@ohsu.edu; Emily Prud'hommeaux: emilypx@gmail.com; Brian Roark: roarkbr@gmail.com; Jan van Santen: vansantj@ohsu.edu

[†]Center for Spoken Language Understanding, Oregon Health & Science University

[○]Center for Language Sciences, University of Rochester

Abstract

Atypical semantic and pragmatic expression is frequently reported in the language of children with autism. Although this atypicality often manifests itself in the use of unusual or unexpected words and phrases, the rate of use of such unexpected words is rarely directly measured or quantified. In this paper, we use distributional semantic models to automatically identify unexpected words in narrative retellings by children with autism. The classification of unexpected words is sufficiently accurate to distinguish the retellings of children with autism from those with typical development. These techniques demonstrate the potential of applying automated language analysis techniques to clinically elicited language data for diagnostic purposes.

1 Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by impaired communication and social behavior. Although the symptoms of ASD are numerous and varied, atypical and idiosyncratic language has been one of the core symptoms observed in verbal individuals with autism since Kanner first assigned a name to the disorder (Kanner, 1943). Atypical language currently serves as a diagnostic criterion in many of the most widely used diagnostic instruments for ASD (Lord et al., 2002; Rutter et al., 2003), and the phenomenon is especially marked in the areas of semantics and pragmatics (Tager-Flusberg, 2001; Volden and Lord, 1991).

Because structured language assessment tools are not always sensitive to the particular atypical semantic and pragmatic expression associated with ASD, measures of atypical language are often drawn from spontaneous language samples. Expert manual annotation and analysis of spontaneous language in young people with ASD has revealed that children and young adults with autism include significantly more bizarre and irrelevant content (Loveland et al., 1990; Losh and Capps, 2003) in their narratives and more abrupt topic changes (Lam et al., 2012) in their conversations than their language-matched typically developing peers. Most normed clinical instruments for analyzing children's spontaneous language, however, focus on syntactic measures and developmental milestones related to the acquisition of vocabulary and syntactic structures. Measures of semantic and pragmatic atypicality in spontaneous language are rarely directly measured. Instead, the degree of language atypicality is often determined via subjective parental reports (e.g., asking a parent

whether their child has ever used odd phrases (Rutter et al., 2003)) or general impressions during clinical examination (e.g., rating the child's degree of “stereotyped or idiosyncratic use of words or phrases” on a four-point scale (Lord et al., 2002)). This has led to a lack of reliable and objective information about the frequency of atypical language use and its precise nature in ASD.

In this study, we attempt to automatically detect instances of contextually atypical language in spontaneous speech at the lexical level in order to quantify its prevalence in the ASD population. We first determine manually the off-topic, surprising, or inappropriate words in a set of narrative retellings elicited in a clinical setting from children with ASD and typical development. We then apply word ranking methods and distributional semantic modeling to these narrative retellings in order to automatically identify these unexpected words. The results indicate not only that children with ASD do in fact produce more semantically unexpected and inappropriate words in their narratives than typically developing children but also that our automated methods for identifying these words are accurate enough to serve as an adequate substitute for manual annotation. Although unexpected off-topic word use is just one example of the atypical language observed in ASD, the work presented here highlights the potential of computational language evaluation and analysis methods for improving our understanding of the linguistic deficits associated with ASD.

2 Data

Participants in this study included 37 children with typical development (TD) and 21 children with autism spectrum disorder (ASD). ASD was diagnosed via clinical consensus according to the DSM-IV-TR criteria (American Psychiatric Association, 2000) and the established threshold scores on two diagnostic instruments: the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2002), a semi-structured series of activities designed to allow an examiner to observe behaviors associated with autism; and the Social Communication Questionnaire (SCQ) (Rutter et al., 2003), a parental questionnaire. None of the children in this study met the criteria for a language impairment, and there were no significant between-group differences in age (mean=6.4) or full-scale IQ (mean=114).

The narrative retelling task analyzed here is the Narrative Memory subtest of the NEPSY (Korkman et al., 1998), a large and comprehensive battery of tasks that test neurocognitive functioning in children. The NEPSY Narrative Memory (NNM) sub-test is a narrative retelling test in which the subject listens to a brief narrative, excerpts of which are shown in Figure 1, and then must retell the narrative to the examiner. The NNM was administered to each participant in the study, and each participant's retelling was recorded and transcribed. Using Amazon's Mechanical Turk, we also collected a large corpus of retellings from neurotypical adults, who listened to a recording of the story and provided written retellings. We describe how this corpus was used in Section 3, below.

Two annotators, blind to the diagnosis of the experimental subjects, identified every word in each retelling transcript that was unexpected or inappropriate given the larger context of the story. For instance, in the sentence *T-rex could smell things*, both *T-rex* and *smell* were marked as unexpected, since there is no mention of either concept in the story. In a

seemingly more appropriate sentence, *the boy sat up off the bridge*, the word *bridge* is considered unexpected since the boy is trapped up in a tree rather than on a bridge.

3 Methods

We start with the expectation that different retellings of the same source narrative will share a common vocabulary and semantic space. The presence of words outside of this vocabulary or semantic space in a retelling may indicate that the speaker has strayed from the topic of the story. Our approach for automatically identifying these unexpected words relies on the ranking of words according to the strength of their association with the target topic of the corpus. The word association scores used in the ranking are informed by the frequency of a word in the child's retelling relative to the frequency of that word in other retellings in the larger corpus of retellings. These association measures are similar to those developed for the information retrieval task of topic modeling, in which the goal is to identify topic-specific words – i.e., words that appear frequently in only a subset of documents – in order to cluster together documents about a similar topic. Details about how these scores are calculated and interpreted are provided in the following sections.

The pipeline for determining the set of unusual words in each retelling begins by calculating word association scores, described below, for each word in each retelling and ranking the words according to these scores. A threshold over these scores is determined for each child using leave-one-out cross validation in order to select a set of potentially unexpected words. This set of potential unexpected words is then filtered using two external resources that allow us to eliminate words that were not used in other retellings but are likely to be semantically related to topic of the narrative. This final set of words is evaluated against the set of manually identified words in order determine the accuracy of our unexpected word classification.

3.1 Word association measures

Before calculating the word association measures, we tokenize, downcase, and stem (Porter, 1980) the transcripts and remove all punctuation. We then use two association measures to score each word in each child's retelling: *tf-idf*, the term frequency-inverse document frequency measure (Salton and Buckley, 1988), and the log odds ratio (van Rijsbergen et al., 1981). We use the following formulation to calculate *tf-idf* for each child's retelling *i* and each word in that retelling *j*, where c_{ij} is the count of word *j* in retelling *i*; f_j is the number of retellings from the full corpus of child and adult retellings containing that word *j*; and *D* is the total number of retellings in the full corpus (Manning et al., 2008):

$$tf-idf_{ij} = \begin{cases} (1+\log c_{ij})\log\frac{D}{f_j} & \text{if } c_{ij} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

The log odds ratio, another association measure used in information retrieval and extraction tasks, is the ratio between the odds of a particular word, *j*, appearing in a child's retelling, *i*, as estimated using its relative frequency in that retelling, and the odds of that word

appearing in all other retellings, again estimated using its relative frequency in all other retellings. Letting the probability of a word appearing in a retelling be p_1 and the probability of that word appearing in all other retellings be p_2, we can express the odds ratio as follows:

$$\text{odds ratio} = \frac{\text{odds}(p_1)}{\text{odds}(p_2)} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

A large *tf-idf* or log odds score indicates that the word *j* is very specific to the retelling *i*, which in turn suggests that the word might be unexpected or inappropriate in the larger context of the NNM narrative. Thus we expect that the words with higher association measure scores are likely to be the words that were manually identified as unexpected in the context of the NNM narrative.

3.2 Application of word association measures

As previously mentioned, both of these word association measures are used in information retrieval (IR) to cluster together documents about a similar target topic. In IR, words that appear only in a subset of documents from a large and varied corpus of documents will have high word association scores, and the documents containing those words will likely be focused on the same topic. In our task, however, we have a single cluster of documents focused on a single topic: the NNM narrative. Topic-specific words ought to occur much more frequently than other words. As a result, words with high *tf-idf* and log odds scores are likely to be those unrelated to the topic of the NNM story. If a child veers away from the topic of the NNM story and uses words that do not occur frequently in the retellings produced by neurotypical speakers, his retellings will contain more words with high word association scores. We predict that this set of high-scoring words is likely to overlap significantly with the set of words identified by the manual annotators as unexpected or off-topic.

Applying these word association scoring approaches to each word in each child's retelling yields a list of words from each retelling ranked in order of decreasing *tf-idf* or log odds score. We use cross-validation to determine, for each measure, the operating point that maximizes the unexpected word identification accuracy in terms of F-measure. For each child, the threshold is found using the data from all of the other children. This threshold is then applied to the ranked word list of the held-out child. All words above this threshold are potential unexpected words, while all words below this threshold are considered to be expected and appropriate in the context of the NNM narrative. Table 1 shows the recall, precision, and F-measure using the two word association measures discussed here. We see that these two techniques result in high recall at the expense of precision. The next stage in the pipeline is therefore to use external resources to eliminate any semantically appropriate words from the set of potentially unexpected or inappropriate words generated via thresholding on the *tf-idf* or log odds score.

3.3 Filtering with external resources

Recall that the corpus of retellings used to generate the word association measures described above, is very small. It is therefore quite possible that a child may have used an entirely appropriate word that by chance was never used by another child or one of the neurotypical adults. One way of increasing the lexical coverage of the corpus of retellings is through semantic expansion using an external resource. For each word in the set of potential unexpected words, we located the WordNet synset for that word (Fellbaum, 1998). If any of the WordNet synonyms of the potentially unexpected word was present in the source narrative or in one of the adult retellings, that word was removed from the set of unexpected words.

In the final step, we used the CHILDES corpus of transcripts of children's conversational speech (MacWhinney, 2000) to generate topic estimates for each remaining potentially unexpected word. For each of these words, we located every utterance in the CHILDES corpus containing that potentially unexpected word. We then measured the association of that word with every other open-class word that appeared in an utterance with that word using the log likelihood ratio (Dunning, 1993). The 20 words from the CHILDES corpus with the highest log likelihood ratio (i.e., the words most strongly associated with the potentially unexpected word), were assumed to collectively represent a particular topic. If more than two of the words in the vector of words representing this topic were also present in the NNM source narrative or the adult retellings, the word that generated that topic was eliminated from the set of unexpected words.

We note that the optimized threshold described in Section 3.2, above, is determined after filtering. There is therefore potentially a different threshold for each condition tested, and hence we do not necessarily expect precision to increase and recall to decrease after filtering. Rather, since the threshold is selected in order to optimize F-measure, we expect that if the filtering is effective, F-measure will increase with each additional filtering condition applied.

4 Results

We evaluated the performance of our two word ranking techniques, both individually and combined by taking either the maximum of the two measures or the sum, against the set of manually annotations described in Section 2. In addition, we report the results of applying these word ranking techniques in combination with the two filtering techniques. We compare these results with a simple baseline method in which every word used in a retelling that is never used in another retelling is considered to be unexpected. Table 1 shows the precision, accuracy, and F-measure of these approaches. We see that all of the more sophisticated unexpected word identification approaches outperform the baseline by a wide margin, and that *tf-idf* and log odds perform comparably under the condition without filtering and both filtering conditions. Filtering improves F-measure under both word ranking schemes, and combining the two measures results in further improvements under both filtering conditions. Although applying topic-estimate filtering yields the highest precision, the simple WordNet-based approach results in the highest F-measure and a reasonable balance between precision and recall.

Recall that the purpose of identifying these unexpected words was to determine whether children with ASD produce unexpected and inappropriate words at a higher rate than children with typical development. This appears to be true in our manually annotated data. On average, 7.5% of the words types produced by children with ASD were marked as unexpected, while only 2.5% of words produced by children with TD were marked as unexpected, a significant difference ($p < 0.01$, using a one-tailed t-test). This significant between-group difference in rate of unexpected word use holds even when using the automated methods of unexpected word identification, with the best performing unexpected word identification method estimating a mean of 6.6% in the ASD group and 2.5% in the TD group ($p < 0.01$).

5 Conclusions and future work

The automated methods presented here for ranking and filtering words according to their distributions in different corpora, which are adapted from techniques originally developed for topic modeling in the context of information retrieval and extraction tasks, demonstrate the utility of automated approaches for the analysis of semantics and pragmatics. We were able to use these methods to identify unexpected or inappropriate words with high enough accuracy to replicate the patterns of unexpected word use manually observed in our two diagnostic groups. This work underscores the potential of automated techniques for improving our understanding of the prevalence and diagnostic utility of linguistic features associated with ASD and other communication and language disorders.

In future work, we plan to use a development set to determine the optimal number of topical words to select during the topic estimate filtering stage of the pipeline in order to maintain improvements in precision without a loss in recall. We would also like to investigate using part-of-speech, word sense, and parse information to improve our approaches for both semantic expansion and topic estimation. Although the rate of unexpected word use alone is unlikely to provide sufficient power to classify the two diagnostic groups investigated here, we expect that it can serve as one feature in an array of features that capture the broad range of semantic and pragmatic atypicalities observed in the spoken language of children with autism. Finally, we plan to apply these same methods to identify the confabulations and topic shifts often observed in the narrative retellings of the elderly with neurodegenerative conditions.

Acknowledgments

This work was supported in part by NSF Grant #BCS-0826654, and NIH NIDCD grant #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

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

Dunning, Ted. Accurate methods for the statistics of surprise and coincidence. Computational linguistics. 1993; 19(1):61–74.

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

Kanner, Leo. Autistic disturbances of affective content. Nervous Child. 1943; 2:217–250.

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

Grace Lam, Yan; Sze, Siu; Yeung, Susanna. 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. 2012; 6:792797.

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

Losh, Molly; Capps, Lisa. Narrative ability in high-functioning children with autism or asperger's syndrome. Journal of Autism and Developmental Disorders. 2003; 33(3):239–251. [PubMed: 12908827]

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

MacWhinney, Brian. The CHILDES project: Tools for analyzing talk. Lawrence Erlbaum Associates; Mahwah, NJ: 2000.

Manning, ChristopherD; Raghavan, Prabhakar; Schütze, Hinrich. Introduction to information retrieval. Cambridge University Press; 2008.

Porter MF. An algorithm for suffix stripping. Program. 1980; 14(3):130–137.

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

Salton, Gerard; Buckley, Christopher. Term-weighting approaches in automatic text retrieval. Information Processing and Management. 1988; 24(5):513–523.

Tager-Flusberg, Helen. Understanding the language and communicative impairments in autism. International Review of Research in Mental Retardation. 2001; 23:185–205.

van Rijsbergen CJ, Harper DJ, Porter MF. The selection of good search terms. Information Processing and Management. 1981; 17(2):77–91.

Volden, Joanne; Lord, Catherine. Neologisms and idiosyncratic language in autistic speakers. Journal of Autism and Developmental Disorders. 1991; 21:109–130. [PubMed: 1864825]

Jim was a boy whose best friend was Pepper. Pepper was a big black dog. [...] Near Jim's house was a very tall oak tree with branches so high that he couldn't reach them. Jim always wanted to climb that tree, so one day he took a ladder from home and carried it to the oak tree. He climbed up [...] When he started to get down, his foot slipped, his shoe fell off, and the ladder fell to the ground. [...] Pepper sat below the tree and barked. Suddenly Pepper took Jim's shoe in his mouth and ran away. [...] Pepper took the shoe to Anna, Jim's sister. He barked and barked. Finally, Anna understood that Jim was in trouble. She followed Pepper to the tree where Jim was stuck. Anna put the ladder up and rescued Jim.

Figure 1.
Excerpts from the NNM narrative.

Table 1

Accuracy of unexpected word identification.

Unexpected word identification method	P	R	F1
Baseline	46.3	74.0	57.0
TF-IDF	72.1	79.5	75.6
Log-odds	70.5	79.5	74.7
Sum(TF-IDF, Log-odds)	72.2	83.3	77.4
Max(TF-IDF, Log-odds)	69.9	83.3	76.0
TF-IDF+WordNet	83.8	80.5	82.1
Log-odds+WordNet	82.1	83.1	82.6
Sum(TF-IDF, Log-odds)+WordNet	84.2	83.1	83.7
Max(TF-IDF, Log-odds)+WordNet	83.3	84.4	83.9
TF-IDF+WordNet+topic	85.7	77.9	81.7
Log-odds+WordNet+topic	83.8	80.5	82.1
Sum(TF-IDF, Log-odds)+WordNet+topic	86.1	80.5	83.2
Max(TF-IDF, Log-odds)+WordNet+topic	85.1	81.8	83.4