Simulating Errors Typical of Disordered Language

Eric Morley
Center for Spoken Language Understanding

1 Introduction

Atypical or impaired language is a core feature of both Autism spectrum disorder (ASD) and specific language impairment (SLI). At present, clinicians tend to use highly-structured tests to elicit and analyze language from children suspected of having ASD or SLI. These tests, however, may be unable to elicit certain types of errors that could inform diagnosis. Analyzing natural language samples may yield more information than the results of such highly-structured tests. This is rarely done, however, because transcribing speech and then coding linguistic abnormalities is so expensive.

Severe data sparsity typically hinders the application of statistical techniques from NLP to automatically coding linguistic abnormalities in transcripts from children suspected of having ASD or SLI. Producing such transcripts in large quantities is extremely expensive as it requires accurate diagnoses of ASD or SLI in addition to producing the coded transcripts. Here we compare two models that take attested strings, and corrupt them with simulated morphological and syntactic errors that are typical of children with ASD or SLI.

We use these models to create training data for a discriminative model that identifies morphological and syntactic errors. Foster and Andersen provide a good overview of previous work in which artificial grammatical errors are used to help train a discriminative model for classifying grammatical and ungrammatical sentences [2].

Our work builds on Hassanali and Liu, who report classifying six types of grammatical errors common in language from children with SLI with a high degree of accuracy, in that we aim to classify a wider variety of errors, and also to classify errors typical of children with ASD [3]. Ultimately, our goal to automatically code linguistic abnormalities in natural language elicited from children suspected of having ASD or SLI.

2 Language Impairments in Individuals with ASD or SLI

There are a wide variety of language impairments associated with ASD and SLI, and since the two disorders are extremely heterogeneous, individuals with either condition tend to exhibit only a subset of potential impairments. In general, the impairments associated with ASD are at the pragmatic or semantic level [5], although ASD is also associated with errors at the morphological and syntactic level [1]. SLI can result in phonological, morphological, syntactic, semantic, or pragmatic impairments. Typical morphological and syntactic errors in individuals with ASD or SLI include errors marking the past tense, as well as marking the third person singular in present tense verbs [8]. Here, we are concerned with identifying and coding errors only at the morphological and syntactic levels.

3 The ADOS and Data

This investigation uses transcripts of the Autism Diagnosis Observations Schedule (ADOS), which is a semi-structured test that can help diagnose autism [6]. Errors in the transcripts have been hand-annotated in the SALT format [7]. This format allows one to see both what the child said, and

the grammatically correct version of what the child said. In some cases, the corrections are quite straightforward, for example *a* replacing *an*. In other cases, however, the corrections may involve seemingly arbitrary word changes for example "getting" being corrected to "turning to".

We disambiguate the annotations by hand to produce a parallel corpus containing RAW utterances (as the child said them), and CORRECTED utterances. We only include utterances for which the RAW and CORRECTED versions differ in the parallel corpus. We manually segment the following morphemes in each utterance: past tense; third person singular; nominal plural; and possessive marker ('s and s'). We only segment words in which the morpheme is separable from the root (ex. *worked*→*work +ed*, but *went*→*went*).

Our ASD and SLI corpora contain 512 and 617 utterances with errors, respectively. Collecting and annotating more transcripts is extremely expensive, and transcripts cannot simply be shared between institutions due to privacy issues.

4 Methods

We propose two methods to corrupt attested utterances with plausible morphological and syntactic corruptions: 1) hand-built rules informed by the literature on errors observed in individuals with ASD and SLI; and 2) a machine-translation model built with Moses [4]. We evaluate both how effectively these methods corrupt grammatical sentences, and the performance of the discriminative models that have been trained on the two different types of simulated training data.

References

[1] I.M. Eigsti, L. Bennetto, and M.B. Dadlani. Beyond pragmatics: Morphosyntactic development in autism. *Journal of autism and developmental disorders*, 37(6):1007–1023, 2007.

[2] J. Foster and Ø.E. Andersen. Generrate: generating errors for use in grammatical error detection. In *Proceedings of the fourth workshop on innovative use of nlp for building educational applications*, pages 82–90. Association for Computational Linguistics, 2009.

[3] K. Hassanali and Y. Liu. Measuring language development in early childhood education: a case study of grammar checking in child language transcripts. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 87–95. Association for Computational Linguistics, 2011.

[4] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[5] C. Lord and R. Paul. Language and communication in autism. *Handbook of autism and pervasive developmental disorders*, 2:195–225, 1997.

[6] C. Lord, M. Rutter, S. Goode, J. Heemsbergen, H. Jordan, L. Mawhood, and E. Schopler. Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of autism and developmental disorders*, 19(2):185–212, 1989.

[7] J. Miller and R. Chapman. Systematic analysis of language transcripts. *Madison, WI: Language Analysis Laboratory*, 1985.

[8] D. Williams, N. Botting, and J. Boucher. Language in autism and specific language impairment: Where are the links? *Psychological bulletin*, 134(6):944, 2008.