Using Language Models to Identify Language Impairment in Spanish-English Bilingual Children

Thamar Solorio
Department of Computer Science
The University of Texas at Dallas
tsolorio@hlt.utdallas.edu

Yang Liu
Department of Computer Science
The University of Texas at Dallas
yangl@hlt.utdallas.edu

1 Introduction

Children diagnosed with Specific Language Impairment (SLI) experience a delay in acquisition of certain language skills, with no evidence of hearing impediments, or other cognitive, behavioral, or overt neurological problems (Leonard, 1991; Paradis et al., 2005/6). Standardized tests, such as the Test for Early Grammatical Impairment, have shown to have great predictive value for assessing English speaking monolingual children. Diagnosing bilingual children with SLI is far more complicated due to the following factors: lack of standardized tests, lack of bilingual clinicians, and more importantly, the lack of a deep understanding of bilingualism and its implications on language disorders. In addition, bilingual children often exhibit code-switching patterns that will make the assessment task even more challenging. In this paper, we present preliminary results from using language models to help discriminating bilingual children with SLI from Typically-Developing (TD) bilingual children.

2 Our Approach

We believe that statistical inference can assist in the problem of accurately discriminating language patterns indicative of SLI. In this work, we use Language Models (LMs) for this task since they are a powerful statistical measure of language usage and have been successfully used to solve a variety of NLP problems, such as text classification, speech recognition, hand-writing recognition, augmentative communication for the disabled, and spelling error detection (Manning and Schütze, 1999). LMs estimate the probability of a word sequence $W = \langle w_1, ...w_k \rangle$ as follows (using the chain rule):

$$p(W) = \prod_{i=1}^{k} p(w_i | w_1, \ldots, w_{i-1})$$

which can be approximated using an N-gram as:

$$p(W) \approx \prod_{i=1}^{k} p(w_i | w_{i-N+1}, w_{i-N+2}, ..., w_{i-1})$$

Since in our problem we are interested in differentiating syntactic patterns, we will train the LMs on Part-of-Speech (POS) patterns instead of words. Using a 3-gram we have:

$$p(T) = \prod_{i=1}^{k} p(t_i | t_{i-2}, t_{i-1})$$

where $T = \langle t_1, t_2, ..., t_k \rangle$ is the sequence of POS tags assigned to the sequence of words W.

The intuition is that the language patterning of an SLI child will differ from those of TD children at two different levels: one is at the syntactic level, and the second one is at the interaction between both languages in patterns such as code-switching. Given that the tagset for each language is different, by using the POS tags we will incorporate into the model the syntactic structure together with the switch points across languages.

We train two LMs with the POS sequences: M_T, with data from the TD children and M_I, with data from the SLI bilingual children. Once both LMs are trained, then we can use them to make predictions over new speech samples of bilingual children. To determine whether an unobserved speech sample is likely to belong to a child suffering from SLI, we will measure the perplexity of the two LMs over the POS patterns of this new speech sample. We make the final decision using a threshold:

$$d(s) = \begin{cases} SLI & if\ (PP_T(s) - PP_I(s)) > 0 \\ TD & otherwise \end{cases}$$

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 116–117,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

where $PP_T(s)$ is the perplexity of the model M_T over the sample s, and $PP_I(s)$ is the perplexity of the model M_I over the same sample s. In other words, if the perplexity of the LM trained on syntactic patterns of children with SLI is smaller than that of the LM trained on POS patterns of TD children, then we will predict that the sample belongs to a child with SLI.

In a related work, (Roark et al., 2007) explored the use of cross entropy of LMs trained on POS tags as a measure of syntactic complexity. Their results were inconsistent across language tasks, which may be due to the meaning attached to cross entropy in this setting. Unlikely patterns are a deviation from what is expected; they are not necessarily complex or syntactically rich.

3 Preliminary Results

We empirically evaluated our approach using transcripts that were made available by a speech pathologist in our team. The TD samples were comprised of 5 males and 4 females between 48 and 72 months old. The children were identified as being bilingual by their parents, and according to parental report, these children live in homes where Spanish is spoken an average of 46.3% of the time. Language samples of SLI bilinguals were collected from children being served in the Speech and Hearing Clinic at UTEP. The samples are from two females aged 53 and 111 months. The clients were diagnosed with language impairment after diagnostic evaluations which were conducted in Spanish. The transcriptions were POS tagged with the bilingual tagger developed by (Solorio et al., 2008).

Table 1 shows the preliminary results using cross validation. With the decision threshold outlined above, out of the 9 TD children, the models were able to discriminate 7 as TD; from the 2 SLI children both were correctly identified as SLI. Although the results presented above are not conclusive due to the very small size corpora at hand, they look very promising. Stronger conclusions can be drawn once we collect more data.

Table 1: Perplexity and final output of the LMs for the discrimination of SLI and TD.

Sample	$PP_T(s)$	$PP_I(s)$	$d(s)$
TD_1	14.73	23.12	TD
TD_2	11.37	16.17	TD
TD_3	18.35	36.58	TD
TD_4	30.23	22.27	SLI
TD_5	9.42	15.50	TD
TD_6	17.37	36.75	TD
TD_7	20.32	33.19	TD
TD_8	16.40	24.47	TD
TD_9	24.35	23.71	SLI
SLI_1	20.21	19.10	SLI
SLI_2	19.70	12.43	SLI
average TD	18.06	25.75	TD
average SLI	19.95	15.76	SLI

4 Final Remarks

This paper presents very promising preliminary results on the use of LMs for discriminating patterns indicative of SLI in Spanish-English bilingual children. As more data becomes available, we expect to gather stronger evidence supporting our method. Our current efforts involve collecting more samples, as well as evaluating the accuracy of LMs on monolingual children with and without SLI.

Acknowledgements

Thanks to Bess Sirmon Fjordbak for her contribution to the project and the three anonymous reviewers for their useful comments.

References

L. B. Leonard. 1991. Specific language impairment as a clinical category. *Language, Speech, and Hearing Services in Schools*, 22:66–68.

C. D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

J. Paradis, M. Crago, and F. Genesee. 2005/6. Domain-general versus domain-specific accounts of specific language impairment: Evidence from bilingual childrens acquisition of object pronouns. *Language Acquisition*, 13:33–62.

B. Roark, M. Mitchell, and K. Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *BioNLP 2007: Biological, translational, and clinical language processing*, pages 1–8, Prague, June. ACL.

T. Solorio, Y. Liu, and B. Medina. 2008. Part-of-speech tagging English-Spanish code-switched text. *Submitted to Natural Language Engineering*.