Evaluating NLP Features for Automatic Prediction of Language Impairment Using Child Speech Transcripts

Khairun-nisa Hassanali[1], *Yang Liu*[1], *Thamar Solorio*[2]

[1]Computer Science Department, The University of Texas at Dallas, Richardson, TX, USA
[2]Department of Computer and Information Sciences, University of Alabama at Birmingham, Birmingham, AL, USA

nisa@hlt.utdallas.edu, yangl@hlt.utdallas.edu, solorio@uab.edu

Abstract

Language impairment (LI) in children is pervasive in all walks of life. Automatic prediction of LI is useful as a first pass for speech language pathologists in identifying prospective children with LI. Previous work in the automatic prediction of LI has explored various features, mostly shallow and surface level features. In this paper, we evaluate deeper Natural Language Processing (NLP) features such as syntactic, semantic and entity grid model features, along with narrative structure and quality features in the prediction of LI using child language transcripts. Our experiments show that narrative structure and quality features along with a combination of other features are helpful in the prediction of LI in storytelling narratives.

Index Terms: language impairment, machine learning, natural language processing

1. Introduction

Language impairment has been extensively studied in the communication disorder field. Children that perform according to the expected norm are called Typically Developing (TD) children whereas children who are lagging behind in some aspect of language, but are otherwise developing typically, are said to have language impairment. It is important to identify the children with LI at an early age so that they can get help.

Some traditional methods of detecting LI include cutoff methods on standardized, norm-referenced language assessment tests such as Mean Length of Utterance (MLUm) and Number of Different Words (NDW). Children scoring at the lower end of the distribution, typically more than 1.25 or 1.5 Standard Deviations (SD) below the mean, are identified as having LI [1]. This cutoff-based approach has several well-documented weaknesses that may result in both over- and under-identification of children as language impaired [2].

Gabani et al. [3] explored the use of automated methods for analyzing transcripts of monolingual English speaking children to predict the presence or absence of LI. They exploit corpus-based approaches inspired by the fields of natural language processing and machine learning. They use mostly shallow and surface level features that focus on different aspects of language such as language productivity, morphosyntactic skills, vocabulary knowledge, probabilities from language models and sentence complexity. They compare results against a cut off baseline and find their methods are superior, reaching F-measures of above 85%. Prud'hommeaux et al. [4] automatically extracted lexical and syntactic features from children's transcripts to classify these transcripts as being producted by children with LI, autism or TD children. They report F-1 measure of 0.79 for prediction of LI.

NLP techniques are now mature enough to allow for a deeper analysis into various aspects of language use such as syntax and coherence. We explore the use of deeper NLP feature sets in the prediction of LI on both spontaneous and storytelling narrative data. The feature sets we explore include general word and text features, syntax based features, referential and semantic features and entity grid model features. We also annotate the storytelling narratives for narrative structure and quality features and use these features in predicting LI on story retells. Our study shows that deeper NLP features in addition to surface level features in children's speech are useful for LI prediction, especially in storytelling narratives.

2. Data

The dataset we use for the experiments contains transcripts of adoloscents, aged 14 years, for two tasks: a storytelling task and a spontaneous personal narrative task.

The first task is based on Mayers 24-page wordless picture storybook "Frog, Where Are You?" [5]. The story is about the search for a missing frog. The second task requested participants to identify the most annoying person they know and describe some of the annoying behaviors. The transcripts were annotated for LI status. The TD group consisted of 99 speakers while the LI group consisted of 19 speakers [6].

3. Features

We approach the problem of predicting LI as a binary classification task. A transcript is classified to be produced by either a TD child or a child with LI. We describe the feature sets below:

3.1. Baseline (Gb) Features

We use the features used by Gabani et al. [3] as a baseline. These are language productivity features, morphosyntactic skills, vocabulary knowledge features, speech fluency features, probabilities from language models, standard scores, sentence complexity and error pattern features.

3.2. New Features

This section describes the new features we explored for the predicting of LI. We use the Coh-Metrix tool[1] to compute the first five feature sets described below. Coh-Metrix is a tool that provides an implementation of 54 features in the psycholinguistic literature that is known to correlate with coherence of human written texts.

3.2.1. Readability (RM) Features

This is the Flesh reading ease score.

3.2.2. Situational Model (SM) Features

The features in this category are based on the micro-world that a text is about. The features are:

- Causal dimension features: Number of causal verbs and causal particles in the text and ratio of causal particles to causal verbs.
- Intentional dimension features: Number of incidental actions, events and particles.
- Temporal dimension features: Repetition score for tense and aspect.

3.2.3. General Word and Text (GWT) Features

The features in this category are based on surface text properties. The features are:

- Basic count features: Number of words, number of utterances, syllables per word and mean number of words per utterance.
- Frequency features: Mean frequency of content words, log mean frequency of content words, minimum frequency of content words per utterance, log minimum frequency of content words per utterance. Content words are nouns, adverbs, adjectives, main verbs, and other categories with rich conceptual content.
- Concreteness features: Mean concreteness of content words and mean of lowest concreteness content words in each sentence. Concreteness measures how concrete or abstract a content word is.
- Hypernymy features: Mean hypernym value of nouns in text and mean hypernym value of main verbs in text.

3.2.4. Syntactic (Syn) Features

The syntax based features assess syntactic complexity in a text. The features are:

- Constituent features: Number of Noun Phrases (NP), number of modifiers per NP, mean number of words before main verb, higher level constituents and number of negations.
- Pronouns, types and tokens features: Number of personal pronouns per 1000 words, ratio of pronouns to NPs and type-to-token ratio.
- Connectives features: Number of connectives, number of positive temporal, causal, additive and logical connectives, number of negative temporal, causal, additive and logical connectives.
- Logical operators features: Number of logical operators.
- Sentence syntax similarity features: syntactic similarities between adjacent utterances.

3.2.5. Referential and Semantic (RS) Features

The referential and semantic features measure argument overlap and use LSA (Latent Semantic Analysis) to calculate conceptual similarity. These features are:

- Number of anaphora references between utterances
- Proportion of adjacent utterances that share arguments
- Proportion of adjacent utterances that share word stems
- Proportion of content words in adjacent utterances that share a content word

3.2.6. Entity Grid (EG) Model Features

We use the Entity Grid model that was developed by Barzilay and Lapata [7] to measure local coherence. Here the text is represented by a matrix with the rows corresponding to each utterance in the transcript and the columns corresponding to each entity mentioned anywhere in the text. The value of a cell is the entity's grammatical role in that utterance (Subject, Object, Neither or Absent).

Since TD children have better language capabilities, we expected the TD and children with LI to have different distributions of entity transitions. We used the Brown

[1] http://cohmetrix.memphis.edu/CohMetrixWeb2/HelpFile2.htm

Coherence Toolkit[2] to construct the entity grids. The features we used were the fractions of each type of transitions in the entire entity grid for the transcript.

3.2.7. Narrative Structure and Quality Features

We considered the following narrative structure and narrative quality features for the prediction of LI in story retells. These features are based on human annotation. We asked 6 native English speakers to annotate the story retell transcripts for the following features:

1. Coherence: This feature indicates the overall coherence of narrative and takes on two values coherent (0) or incoherent (1).
2. Narrative Structure: This feature set consists of 7 binary features that denote the presence or absence of the instantiation of the story, the 5 search episodes in the story, and the resolution of the story.
3. Maintenance of search theme: This feature is a score between 0 to 4 based on the number of times the search theme of the story is mentioned.
4. Cognitive Inferences: Number of references to the mental state of a character such as character motivation and causality.
5. Social Engagement Devices: Number of social engagement devices such as character speech and sound effects.
6. Affective Devices: Number of references to the affective state of a character.
7. Hedges: Number of references to certainity or uncertainity
8. Intensifiers: Number of references to intensifiers such as repetitions and adjectives.

4. Experiments and Results

We constructed the naive Bayes, logitboost, Bayesian network and Support Vector Machine (SVM) classifiers for both spontaneous and storytelling narratives. For this purpose, we used the WEKA toolkit [8] using leave-one out cross validation. Here we consider LI as the class of interest.

Table 1 and Table 2 give the evaluation of the general word and text features, syntactic features, referential and semantic features and entity grid model features using the spontaneous and storytelling narratives respectively. Table 3 gives the results of using narrative quality and structure features in the prediction of LI in storytelling narratives since we have narrative structure and quality features only for storytelling narratives. We also

Feature	**P**	**R**	**F-1**
GWT	**0.571**	0.421	**0.485**
Syn	0.533	0.421	0.471
RS	0.321	0.474	0.383
Entity Grid	0.162	**0.579**	0.253
RS + Syn	**0.533**	0.421	0.471
RS + GWT	0.423	**0.579**	**0.489**
Syn + GWT	0.471	0.421	0.444
RS + Syn + GWT	0.471	0.421	0.444
RS + Syn + GWT + EG	**0.533**	0.421	0.471
Gb (baseline)	0.65	**0.684**	**0.667**
Gb + RS + GWT	0.65	**0.684**	**0.667**
Gb + Syn + GWT	0.667	0.632	0.649
Gb + RS + Syn + GWT	0.667	0.632	0.649
Gb + EG	0.65	**0.684**	**0.667**
Gb + RS + Syn + GWT + EG	0.667	0.632	0.649
Gb + All	**0.706**	0.632	**0.667**
Gb + All - Syn	0.632	0.632	0.632
Gb + All - RS	**0.706**	0.632	**0.667**
Gb + All - GWT	**0.706**	0.632	**0.667**

Table 1: Evaluation of NLP features on the spontaneous narratives.

Feature	**P**	**R**	**F-1**
GWT	0.25	0.316	0.279
Syn	0.267	**0.632**	**0.375**
RS	**0.346**	0.474	0.4
Entity Grid	0.304	0.368	0.333
RS + Syn	0.224	0.579	0.324
RS + GWT	**0.353**	**0.632**	**0.453**
Syn + GWT	0.262	0.579	0.361
RS + Syn + GWT	0.244	0.579	0.344
RS + Syn + GWT + EG	0.27	0.526	0.357
Gb (baseline)	0.824	0.737	0.778
Gb + RS + GWT	0.824	0.737	0.778
Gb + Syn + GWT	0.824	0.737	0.778
Gb + RS + Syn + GWT	0.824	0.737	0.778
Gb + EG	0.778	0.737	0.757
Gb + RS + Syn + GWT + EG	0.778	0.737	0.757
Gb + All	**1**	0.79	0.882
Gb + All - Syn	0.762	**0.842**	0.8
Gb + All - RS	**1**	**0.842**	**0.914**
Gb + All - GWT	**1**	0.79	0.882

Table 2: Evaluation of NLP features on the storytelling narratives.

[2] http://www.cs.brown.edu/~melsner/manual.html

combine Gabani et al.'s features along with our new features and report the results in these tables. We report the results of the classifiers that performed the best. The results in Table 1, Table 2, and Table 3 were obtained using the Bayesian network classifier.

In all these tables, P stands for precision, R stands for recall, F-1 stands for F-1 measure, Spon stands for spontaneous, ST stands for storytelling and Gb stands for Gabani's features, All stands for general word and text features, syntactical features, referential and semantic features, situational model features and readability metric features (i.e. the features described in Section 3.2.1 to 3.2.6). Since the usage of the readability and situational model features by themselves did not yield good results, we do not report the results for these two feature sets by themselves. The readability and situational model features are only used in "All" and we report these results in Table 1 and Table 2.

We observe from Table 1, when using the spontaneous narratives, the best result using only the Coh-Metrix features is obtained when combining the GWT category and the RS category with an F1-score of 0.489. When using the storytelling narratives, the best result of 0.453 is obtained by combining the GWT and RS features. In Table 1, we observe that the addition of these deep NLP features to Gabani's features does not lead to an improvement in results for spontaneous narratives.

When we consider storytelling narratives, we see from Table 2 that the addition of situational model and readability features improves the performance of the classifiers with the best result of 0.914 using Gabani's features, general word and text features, syntactic features, situational model and readability features. Further, the inclusion of situational model and readability features gives an improved performance over the baseline for all combinations that have the situational model features for the storytelling narratives. This can be attributed to the difference in structure of a spontaneous narrative and a storytelling narrative. Since situational model features look at the micro-world of what a text is about, which is more clear in a storytelling narrative as opposed to a spontaneous narrative, the performance is better for the storytelling narratives.

We also observe, from Table 1 and Table 2, that while the general word and text feature and syntactic category yield better results for spontaneous narratives over storytelling narratives, the RS and entity grid feature category perform better for storytelling narratives.

In Table 3, we observe that adding narrative structure and narrative quality features to Gabani et al's feature set results in an increase of 8.7% over the baseline for story retells. We performed feature selection on the top scoring narrative structure and narrative quality features. Instantiation of the narrative, use of cognitive inferences and the use of social engagement devices were identified as the top scoring narrative structure and quality features. We also evaluated combining the narrative features with the best feature configuration in Table 2, but did not find any improvement.

Feature	**P**	**R**	**F-1**
Gabani's	0.824	0.737	0.778
Narrative	0.285	0.263	0.313
Narrative + Gabani's	**0.889**	**0.842**	**0.865**

Table 3: Evaluation of narrative structure and quality features on story telling sessions

5. Conclusions and Future Work

In this paper, we looked at the use of several deep NLP features including syntactic, referential and semantic, entity grid model and situational model features. We also looked at the usage of narrative structure and narrative quality features in combination with the baseline features for story retells. We observed a difference in results on spontaneous and storytelling narratives with better results achieved on storytelling narratives. In the future, we plan to explore more features that exploit the characteristics of both spontaneous and storytelling narratives.

6. Acknowledgements

This research is supported by an NSF award IIS-1017190 and 1018124.

7. References

[1] Tomblin, J.B., Records, N.L., Buckwalter, P., Zhang, X., Smith, E. and O'Brien, M., "Prevalence of specific language impairment in kindergarten children", Journal of Speech, Language, and Hearing Research, 40(6):1245–1260, 1997.

[2] Plante, E. and Vance, R., "Selection of preschool language tests: A data-based approach", Language, Speech, and Hearing Services in Schools, 25(1):15–24, 1994.

[3] Gabani, K., Solorio, T., Liu, Y., Hassanali, K. and Dollaghan, C., "Exploring a corpus-based approach for detecting language impairment in monolingual English-speaking children", Artifical Intelligence in Medicine, 53(3):161–170, 2011.

[4] Prudhommeaux, E.T., Roark, B., Black, L.M., Santen, J.V., "Classification of atypical language in autism", in Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, Association for Computational Linguistics, 2011, pp. 88–96.

[5] Mayer, M., "Frog, where are You?", Dial Press New York, 1969.

[6] Conti-Ramsden, G., Botting, B. and Faragher, B., "Psycholinguistic markers for specific language impairment (SLI)", The Journal of Child Psychology and Psychiatry and Allied Disciplines, 42(6):741–748, Cambridge Univ Press, 2001.

[7] Barzilay, R. and Lapata, M., "Modeling local coherence: An entity-based approach", Computational Linguistics, 34(1):1–34, MIT Press, 2008.

[8] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I., "The WEKA data mining software: An update", ACM SIGKDD Explorations Newsletter, 11(1):10–18, ACM, 2009.