Michael Fry December 1st, 2013

Modeling cross-linguistic child language acquisition: Lessons for machine learning

While computers facilely best humans in things like chess, sorting mail, and crunching numbers, *language* remains something machines simply do not *do* as well as humans do. Most children seem to learn language effortlessly, but no artificial intelligence is remotely close to replicating human language use convincingly (i.e. no Loebner grandprizes have been won). Despite advances in areas such as machine translation, natural language processing, and targeted advertising, human language performance is unmatched by machines. Lately, with the hopes of gaining insight into computational modeling of language, a growing body of research (see Yang (2011) for an overview) has been focusing on models of child language acquisition (an independently studied area with much work already done, e.g. Bruner (1973)). This area is intuitively informative in that mistakes and generalizations made by a child learning language provide a window into the cognitive processes occuring in that child's mind. If those cognitive processes can be isolated, they are more easily adapted to computational algorithms and may lead to the development of more effective machine learning. Unfortunately, there is a severe lack of such potential processes because the majority of research in computational modeling of child language learning has been done with English (Yang, 2011). To this avail, I propose to develop a computational model which replicates the development (common errors and generalizations) of child learners of Mandarin Chinese. This work will be the first step towards the development of a *single* model capable of capturing diverse aspects of child language learning *across* languages.

Models of child language acquisition have been in the literature for quite some time. One classic model is that of Rumelhart and McClelland (1986). They utilized neural networks to replicate the U-shaped curve seen in children learning the English past tense. In brief, the U-shaped curve of past tense learning consists of three contiguous stages:

1. Children produce irregular and regular past tense forms with *high accuracy* (e.g. eat → ate, walk → walked) presumably because they are learning exemplars.
2. Children 'become aware' of the past tense rule ($verb+/-ed/$) and overgeneralize it to irregular past tense forms, resulting in *lower accuracy* of past tense forms (e.g. eat → eated/ated, go → goed/wented)
3. Children learn to isolated irregular verbs from regular past tense verbs and regain *high accuracy* of past tense forms.

This development is termed U-shaped because, when past tense form accuracy is graphed, it is shaped like a parabola (a 'U'). The programming paradigm (neural networks) is modeled after how the human brain functions. It consists of interconnected units (analagous to neurons) where the connections are weighted and transmit activation values between the units. Over time, these connections are 'trained,' resulting in weights that (in this case) map a *present tense verb* input to its *past tense form* output. Naturally, computational modeling has come a long way since 1986, but the work of Rumelheart and McClelland has been some of the most insightful.

In their book *Computational Approaches to Morphology and Syntax* (Roark & Sproat, 2007), the authors outline several current approaches to machine learning of Syntax and Morphology. In Linguistics, Syntax is the study of how words are combined to form phrases and sentences, and Morphology is the study of how parts of words (roots, prefixes, suffixes, etc.) are combined to make words. Both of these processes are essential for any child to learn a language, and children make errors and overgeneralize in both. Computationally, most current models that 'learn' significant sequences of words (Syntax) or morphemes[1] (Morphology) utilize probabilities (Schone & Jurasky, 2001; Yarowsky & Wicentowski, 2001). The technical details of such models are well beyond the scope of this proposal, but, essentially, reoccuring patterns in a large corpus of words have a higher probability of being significant (whether in sentence structure, word structure, or otherwise).

To begin my research, I will utilize such probabilistic learning models to replicate two known errors that Mandarin children make: the incorrect substitution of rising and falling tones within words (Li, 1977) and the omission of subjects of verbs (Wang et al., 1992). To achieve this, a corpus of Mandarin infant-directed speech will 'train' a network, and the output of the network will be tested at different levels of its development. I plan to develop separate models that replicate specific errors/overgeneralizations, and later unify them into a coherent whole. My previous experience working with computational models (Tupper & Fry, 2012) and teaching Mandarin children has provided me with a strong base to build the current work on. Upon completion of these first steps, I will investigate whether the model can capture errors/overgeneralizations in English child language learners as well.

I acknowledge that it would be ignorant to assume computers *must* 'learn' language in the same way children do (computers certainly do not play chess the same), but it would be foolish to assume that there is nothing applicable to computational models that could be drawn from child language learning. The fact is that children learn language insurmountably better than machines, in spite of the fact that machines have access to unfathomably more examples of language use (e.g. Google books n-gram viewer). Ultimately, this project will provide valuable information useful for developing more natural machine translation, more accurate natural language processing and will make testable predictions for future research in child language acquisition[2].

[1] morphemes are the meaningful chunks of a word, e.g. resusability has 4 morphemes, $re - use - able - ity$)

[2] Works cited available upon request