Towards Machine Understanding of Interlanguages: A Language Model for Spanglish

Olac Fuentes, Thamar Solorio and Nigel Ward
Computer Science Department
University of Texas at El Paso
El Paso, TX 79968

1 Introduction

Spanglish is the blend of Spanish and English. This interlanguage phenomenon is very common, especially among Hispanic communities in the United States (Ardila 2005). Although it has been described as the most interesting contemporary linguistic phenomenon in this country, it has received surprisingly little attention in the linguistic research community, and it has been almost completely ignored by the NLP community. This is perplexing, as it is the dominant dialect of Spanish spoken in the United States, with an estimated 15 million speakers. In fact, it may be the most commonly used language for every day communication in several major cities in the U.S., including Miami, El Paso, and several others.

The main motivation for this research project is to advance the state-of-the-art of speech recognition systems by making possible to handle interlanguages. An immediate practical application of such systems are spoken dialog systems. Spoken dialog systems have gained a large popularity in the last decade. Most medium to large companies have now an automated voice system to handle all incoming calls. Unfortunately, this type of systems are unable to parse utterances from mixes of languages; when a Spanglish speaker is trying to interact with an automated voice system, it is very likely the user will end up frustrated, given that from his/her point of view his/her utterances are clear and intelligible, but regardless of how clear they are, current speech systems will not be able to parse them. Then these speakers will need to wait until their call is transferred to a live agent. The problem does not stop there, since there is no guarantee that a live agent will be able to communicate effectively with the person at the other end of the line, and there is also the problem of determining if the transfer should be made to an agent that speaks Spanish or English.

Besides the application described above, a systems that can generate Spanglish transcriptions can also be useful for second language learning. Being able to characterize the different phenomena present in bilinguals opens the possibility to develop tools to assist in the writing of traditional English and Spanish, perhaps by using a statistical machine translation approach such as the one proposed by Brockett, Dolan & Gammon (2006);
2 Practical Considerations

Building a language model for Spanglish involves different challenges. Some of these challenges are inherent to language models; it is well known that language models are very sensitive to the training corpus with regard to the size of the corpus, as well as to its contents. It should be noted tough that the difficulty in building a good language model of Spanglish is far more complex than just gathering a large enough sample. This is due to the primary difference between Spanglish and traditional English or Spanish: Spanglish is a spectrum of interlanguages, there are different degrees to which people mix English and Spanish; and it is also possible that geography plays a key role in the way Spanglish is used. However, there are also invariants across the different Spanglish speaking populations; for instance, we know that native Spanish speakers will rarely borrow English words that are phonetically different or too far apart from Spanish. They tend to borrow words that have a very familiar pronunciation. Thus, regardless of the speakers’ country of origin, Mexico, Puerto Rico, or Cuba, there will be a large enough intersection on the Spanglish patterns used by the speakers.

Our research goal is to develop a flexible language model for Spanglish. By flexible we mean that we want the language model to be easily adaptable to a specific instance of Spanglish, or to a specific degree of mixture across the two traditional languages.

References
