Part-of-Speech Tagging for English-Spanish Code-Switched Text

Thamar Solorio and **Yang Liu**
Human Language Technology Research Institute
The University of Texas at Dallas
Richardson, TX 75080, USA
tsolorio,yangl@hlt.utdallas.edu

Abstract

Code-switching is an interesting linguistic phenomenon commonly observed in highly bilingual communities. It consists of mixing languages in the same conversational event. This paper presents results on Part-of-Speech tagging Spanish-English code-switched discourse. We explore different approaches to exploit existing resources for both languages that range from simple heuristics, to language identification, to machine learning. The best results are achieved by training a machine learning algorithm with features that combine the output of an English and a Spanish Part-of-Speech tagger.

1 Introduction

Worldwide the percentage of bilingual speakers is fairly large, and it keeps increasing at a high rate. In the U.S., 18% of the total population speaks a language other than English at home, the majority of which speaks Spanish (U.S. Census Bureau, 2003). A significant percentage of this Spanish-English bilingual population code-switch between the two languages in what is often referred as Spanglish, the mix of Spanish and English. Spanish and English are not the only occurrence of language mixtures. Examples of other popular combinations include Arabic dialects, French and German, Spanish and Catalan, Maltese and English, and English and French. Typically when there are linguistic borders, or when the country has more than one official language, we can find instances of code-switching.

Despite the wide use of code-switched discourse among bilinguals, this linguistic phenomenon has received little attention in the fields of Natural Language Processing and Computational Linguistics. Part-of-Speech (POS) tagging is a well studied problem in these fields. For languages such as English, German, Spanish, and Chinese there are several different POS taggers that reach high accuracies, especially in news text corpora. However, to our knowledge, there is no previous work on developing a POS tagger for text with mixes of languages.

In this paper we present results on the problem of POS tagging English-Spanish code-switched discourse by taking advantage of existing taggers for both languages. This rationale follows the evidence from studies of code-switching on different language pairs, which have shown code-switching to be grammatical according to both languages being switched. We use different heuristics to combine POS tag information from existing monolingual taggers. We also explore the use of different language identification methods to select POS tags from the appropriate monolingual tagger. However, the best results are achieved by a machine learning approach using features generated by the monolingual POS taggers.

The next section presents the facts about code-switching, including some previous work done mainly in linguistics. Then in Section 3 we discuss previous work related to the automated processing of code-switched discourse. In Section 4 we describe the English-Spanish code-switched data set gathered for the experimental evaluation. Section 5 presents the heuristics-based approaches for POS tagging that we explored. In Section 6 we describe our machine learning approach and show

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 1051–1060,
Honolulu, October 2008.

results on POS tagging code-switched text. An in depth analysis of results is presented in Section 7, and we conclude this paper with a summary of the findings and directions for future work in Section 8.

2 Rules of Code-switching

In the linguistic, sociolinguistic, psychology, and psycholinguistic literature, bilingualism and the inherent phenomena it exhibits have been studied for nearly a century (Espinosa, 1917; Ervin and Osgood, 1954; Gumperz, 1964; Gumperz and Hernandez-Chavez, 1971; Gumperz, 1971; Sankoff, 1968; Lipski, 1978). Despite the numerous previous studies of linguistic characteristics of bilingualism, there is no clear consensus on the terminology related to language alternation patterns in bilingual speakers. The alternation of languages within a sentence is known as code-mixing, but it has also been referred as intrasentential code-switching, and intrasentential alternation (Poplack, 1980; Grosjean, 1982; Ardila, 2005). Alternation across sentence boundaries is known as intersentential code-switching, or just code-switching. In the rest of this paper we will refer to the mixing of languages as code-switching. When necessary, we will differentiate the type of code-switching by referring to alternations within sentences as intrasentential code-switching and alternations across sentence boundaries as intersentential code-switching.

Linguistic phenomena in bilingual speakers have been analyzed on different language pairs, including English-French, English-Dutch, Finish-English, Arabic-French, and Spanish-English, to name a few. There is a general agreement that code-switched patterns are not generated randomly; according to these studies, they follow specific grammatical rules. Furthermore, some studies suggest that, if these rules are violated, the resulting discourse will sound unnatural (Toribio, 2001b; Toribio, 2001a). The following shows the rules governing code-switching discourse described in several studies (Poplack, 1980; Poplack, 1981; Sankoff, 1981; Sankoff, 1998a).

- Switches can take place only between full word boundaries. This is also known as the free morpheme constraint.
- Monolingual constructs within the sentence will follow the grammatical rules of the monolingual fragment.
- Permissible switch points are those that do not violate the order of adjacent constituents on both sides of the switch point of either of the languages. This is called the equivalence constraint.

Although these rules are somewhat controversial, and most of the studies on this area have been conducted on small samples, we cannot ignore the fact that patterns bearing the above rules have emerged in different bilingual communities with different backgrounds.

3 Automated Processing of Code-Switched Discourse

A previous work related to the processing of code-switched text deals with language identification on English-Maltese code-switched SMS messages (Rosner and Farrugia, 2007). In addition to dealing with intrasentential code-switching, they have to deal with text where misspellings and ad hoc word contractions abound. What Rosner and Farrigua have found to work best for language identification in this noisy domain is a combination of a bigram Hidden Markov Model, trained on language transitions, and a trigram character Markov Model for handling unknown words. In another related work, Franco and Solorio present preliminary results on training a language model for Spanish-English code-switched text (Franco and Solorio, 2007). To evaluate their language model, they asked a human subject to judge sentences generated by a PCFG induced from training data and the language model. However, they only used one human judge.

Regarding the automated POS tagging and parsing of code-mixed utterances there is little prior work. To the best of our knowledge, there is no parser, nor POS tagger, currently available for the syntactic analysis of this type of discourse. There are theoretical approaches that propose formalisms to represent the structure of code-switched utterances and describe a framework for parsing and generating mixed sentences, for example for Marathi and English (Joshi, 1982), or Hindi and English (Goyal *et al.*, 2003). Sankoff proposed a production model of bilingual discourse that accounts for the

equivalence constraint and the unpredictability of code-switching (Sankoff, 1998a; Sankoff, 1998b). His real-time production model draws on the alternation of fragments from two virtual monolingual sentences. It also accounts for other types of code-switching such as repetition-translation and insertional code-switching. But no statistical assessment has been conducted on real corpora.

Our goal is to develop a POS tagger for code-switched utterances, which is the first step of the syntactic analysis of any language. Among the challenges we face is the lack of a representative sample of code-mixed discourse. Most POS taggers are built using large collections, usually at least a million words, such as the Brown corpus (Kucera and Francis, 1967), the Wall Street Journal corpus (Paul and Baker, 1992), or the Switchboard corpus (Godfrey *et al.*, 1992). Currently, there is no annotation of code-switched text of comparable size. But in contrast to the lack of linguistic resources available for Spanish-English code-mixed discourse, English and Spanish have sufficient resources, especially English. Thus, rather than starting from scratch, we will draw on existing taggers for both languages, which will reduce the amount of code-switched data needed. Some examples of POS taggers that perform reasonably well on monolingual text of each language can be found in (Brants, 2000; Brill, 1992; Carreras and Padró, 2002; Charniak, 1993; Ratnaparkhi, 1996; Schmid, 1994). However, these tools are designed to work on monolingual text, therefore if applied as they are to code-switched text, their accuracy will decrease by a large margin. In the following sections we will explore different methods for combining monolingual taggers.

4 Data Set

Data collections that have instances of Spanish-English code-switching, Spanglish for short, are not easily found since code-switching is primarily used in spoken form. To gather data we recorded a conversation among three staff members of a southwest university in the U.S. The three speakers come from a highly bilingual background and code-switch regularly when speaking among themselves, or other bilingual speakers.

This recording session has around 39 minutes of continuous speech (922 sentences, about 8k words) and was transcribed and annotated with POS tags by a human annotator. The annotations were later revised by a different annotator but no inter-annotator agreement was measured. The POS tag set used in the annotation is the combination of the tag sets from the English and the Spanish Tree Taggers (see Section 5). The vocabulary of the transcription has a total of 1,516 different word forms[1]. In the conversation a total of 239 switches were identified manually, out of which 129 are intrasentential code-switches, and the rest are intersentential. English is the predominant language used, with a total of 6,020 tokens and 576 monolingual sentences. In contrast, the transcription has close to 2k tokens in Spanish. Table 1 shows examples of code-switching taken from the recorded conversation; (a) and (b) are instances of intrasentential code-switching, and (c) shows intersentential code-switching.

Table 1: Excerpts taken from the Spanglish data set.

Spanglish	English Translation
(a)*Entonces le dió el virus y no se lo atendió and the virus spread through his body.*	(a)Then he got the virus and he didn't receive treatment and the virus spread through his body.
(b)*Cuando yo lo vi he looked pretty bad.*	(b)When I saw him he looked pretty bad.
(c)*I think she was taller than he was.*	(c)I think she was taller than he was.
Y un carácter muy bonito también ella.	And a very nice character she as well.
Very easy going.	Very easy going.

5 Rule-based Methods for Exploiting Existing Resources

In this section we present several heuristics-based methods for POS tagging code-switched text. First, we describe the monolingual taggers used in this work. Then we present the different approaches explored and contrast their performance.

[1]This transcription and the audio file are freely available for research purposes by contacting the first author.

5.1 Monolingual Taggers

We used the Tree Tagger (Schmid, 1994) for this work because of the following considerations:
1. It has both English and Spanish versions. The English tagger uses a slightly modified version of the Penn Treebank tag set and was trained and evaluated on different portions of the Penn Treebank, reaching a POS tagging accuracy of 96.36%. The Spanish one uses a different tagset with 75 different POS tags[2] and was trained on the Spanish CRATER corpus.
2. The transition probabilities are estimated using a modified version of the ID3 decision tree algorithm (Quinlan, 1986), which provides more freedom to learn contextual cues than n-grams.
3. Both taggers include a special tag for foreign words, PE for Spanish and FW for English. We do not expect this tag to identify correctly all foreign words, but when available this information will be exploited.
4. The Tree tagger generates probability estimates on the tags that can be used as features.
5. Finally, when the tagger fails to lemmatize a word it outputs the special token ⟨*unknown*⟩. This information can be used as a hint of words that do not belong to that particular language.

5.2 Heuristic-based Systems

For all heuristics the complete Spanglish data set was given to both taggers as a single text, then the final tag for each word was selected from the output of the taggers according to the different heuristics. Table 2 shows the tagging accuracies of the different heuristics we explored, which are explained below.
1. Using the monolingual tagger. Here we simply give the Spanglish text to the Spanish and the English tree tagger. We expect from both taggers a performance degradation due to the inclusion of foreign words in code-switching, as compared against their accuracy on monolingual texts. Another complicating factor to keep in mind is that we are dealing with spoken language. Hesitations, fillers, disfluencies, and interruption points, such as *Umm*, *Mmmhmm*, and *Uh-huh*, are frequently observed in speech and it is well known that they complicate the POS tagging task. The tagging accuracy from using the individual taggers is rather low, 26% for the Spanish tagger and 54% for the English one. The large difference between the two taggers can be attributed to the fact that the majority of the words in the corpus are in English.

[2]The authors were unable to identify the source of the Spanish tagset.

Table 2: Accuracy on POS tagging Spanglish text using simple heuristics for combining the output of the English and Spanish tagger.

	Heuristic	**Accuracy (%)**
1	Spanish Tree Tagger	25.99
	English Tree Tagger	54.59
2	Highest prob tag or English	51.51
	Highest prob tag or Spanish	49.16
3	Prob + special tags + lemmas	64.27
4	Dictionary-based Language Id	**86.03**
	Character 5-grams Language Id	81.46
	Human Language Id	89.72

2. Using confidence thresholds. The Tree Tagger can output probabilities for each tag, showing the confidence of the tagger on each particular tag. To use this information we choose for each word the tag from the tagger with the highest confidence. When there is a tie we use either the English or the Spanish tag. Table 2 shows the results for the two cases. The "Highest prob tag or English" heuristic gives an accuracy of 51%, which is almost as accurate as using only the English tagger. The "Highest prob tag or Spanish" achieves an accuracy of 49%, which is an improvement over using only the monolingual Spanish tagger, but it is still below the accuracy of the English monolingual tagger. This is also possibly due to the task being easier for the English tagger.
3. Combining confidence thresholds with knowledge from special tags and lemmas. This heuristic uses confidence thresholds combined with decisions based on the special tags, described in Section 5.1, and the unknown lemmas found. Let $POS_E(w_i)$ and $POS_S(w_i)$ be the POS tags assigned to word w_i by the English and Spanish tagger respectively; and let $Prob_E(w_i)$ and $Prob_S(w_i)$ be the confidence scores of POS tags for word w_i computed by the English and Spanish tree taggers, respectively. For each word w_i in the text, the final POS tag,

$POS_F(w_i)$, will be assigned as follows:

1. **If** $POS_E(w_i) = FW$, **then** $POS_F(w_i) \leftarrow POS_S(w_i)$
2. **Else if** $POS_S(w_i) = PE$, **then** $POS_F(w_i) \leftarrow POS_E(w_i)$
3. **Else if** $POS_E(w_i) = \langle unknown \rangle$, **then** $POS_F(w_i) \leftarrow POS_S(w_i)$
4. **Else if** $POS_S(w_i) = \langle unknown \rangle$, **then** $POS_F(w_i) \leftarrow POS_E(w_i)$
5. **Else if** $Prob_E(w_i) > Prob_S(w_i)$, **then** $POS_F(w_i) \leftarrow POS_E(w_i)$
6. **Else** $POS_F(w_i) \leftarrow POS_S(w_i)$

This heuristic performs better than the other methods explored so far, yielding an accuracy of 64.27%. It seems that knowledge of the taggers can be used to improve results. However, POS tagging accuracy is still poor.

4. Selecting POS tags based on automated language identification. We used two different strategies for automatically identifying the language at the word level. One is based on dictionary look-up and the other is character-based language models. For the first approach, every word in the text is searched in the English and Spanish dictionaries. If a word is found in the English dictionary, then we identify that word as belonging to English and the POS tags from the English tagger are used for that word and the following ones, until a word is found in the Spanish dictionary. Similarly, for a word not found in the English dictionary, but found in the Spanish dictionary, we use the Spanish tags until an English word is found. Note that this simple heuristic will always label words that belong to both languages as English, which is also the case for words not found in either dictionary. This dictionary-based method has a language identification accuracy of 94% on the Spanglish corpus.

The character language models were trained on the Agence France Presse (AFP) portions of the Gigaword for English and Spanish, respectively. For each of the words in the Spanglish corpus, we first decide its language by choosing the one with the lowest perplexity, calculated using character n-gram language models, then we use the corresponding POS tag. We experimented with different language model orders, with n ranging from 2 to 6, and found that we achieve the highest accuracy, 81.46%, on POS tagging using a 5-gram language model. This 5-gram method reached a language identification accuracy of 85% for the Spanglish corpus. However, the language identification method using dictionary look-up achieved the best POS tagging result so far: 86.03%. The Spanglish conversation is dominated by every-day language that is easily found in dictionaries, while the text used to train the character based n-gram language models includes vocabulary that is not commonly used in conversations. This can explain why the simple dictionary look-up approach yielded better results for our corpus. Performing manual identification of the language and sending to the appropriate tagger just the corresponding fragments yields a very high POS tagging accuracy, 89.72%. This shows that it is important to deal with the language switches for boosting accuracy. However relying on human annotated language tags would be expensive and for some tasks unfeasible.

6 Machine Learning for POS Tagging Code-Switched Discourse

From Table 2 we can see that, with the exception of the language identification heuristic, accuracies are low for the previous experiments. However, we believe that we can improve results further by using Machine Learning (ML) algorithms trained specifically for this task. In this section we describe the ML setting and present a comparison of the different algorithms we tested.

6.1 Approach

The key point is that the features selected for describing the learning instances are the output from the English and the Spanish taggers. This scheme is similar to a stacked classifier approach (Wolpert, 1992), where the final classifier takes as input the predictions made by the different learners on the first pass and is trained to select the right tag from them, or a different one if the right answer is not available.

The gold-standard POS tags are used as the class label, and instances in this learning task are described by the following attributes:

1. The word (word)
2. English POS tag (E_t)
3. English POS tagger lemma (E_l)
4. English POS tagger confidence (E_p)
5. Spanish POS tag (S_t)
6. Spanish POS tagger lemma (S_l)
7. Spanish POS tagger confidence (S_p)

Feature 1 is just the lexical word form as it appears in the transcript. Features 2 to 4 are generated by the English Tree tagger, while features 5 to 7 are generated by the Spanish Tree tagger. Thus all features are automatically extracted.

6.2 Results

We evaluated experimentally the idea of using ML with different learning algorithms in WEKA (Witten and Frank, 1999). We selected some of the most widely known algorithms, including Support Vector Machines (SVM) with a polynomial kernel of exponent one (Schölkopf and Smola, 2002), Weka's modified version of Quinlan's C4.5 (J48) (Quinlan, 1986), Additive Logistic Regression with Decision Stumps (Logit Boost) (Friedman *et al.*, 1998) and Naive Bayes. The only parameter we modified was for J48 –we enabled the option for reducing error pruning.

Table 3: POS tagging accuracy of Spanglish text with different Machine Learning algorithms. Oracle shows the accuracy achieved when always selecting the right POS tag from the output of both Tree Taggers. Language Id shows accuracy of identifying the language and then choosing the output of the corresponding tagger.

ML Algorithms	Mean Accuracy (%)	Variance
Naive Bayes	88.50	1.9280
SVM	**93.48**	1.2784
Logit Boost	**93.19**	1.4437
J48	91.11	2.1527
Oracle	90.31	-
Language Id	85.80	-

Table 3 shows the average accuracy of 10-fold cross-validation for each classifier together with the variance. SVM and Logit Boost performed the best and the difference between the two algorithms is not significant according to the paired t-test (P-value = 0.1). For comparison, we show the accuracy of the language identification approach together with the oracle accuracy. The oracle is the accuracy achieved when always selecting the right POS tag, when it is available, from the output of both Tree Taggers. We did not expect the oracle's accuracy to be an upper bound on the accuracy for the ML learning algorithm. Our intuition is that the ML algorithm can be trained to identify when the taggers have made a mistake and what the right answer should be. As the results show, the ML approach can indeed outperform the oracle, and the language identification method.

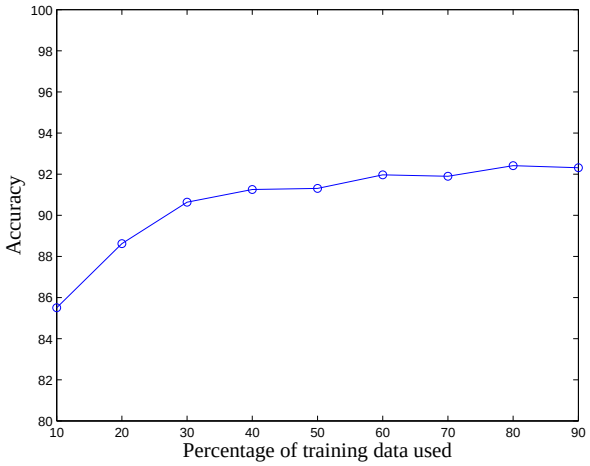

Figure 1: Effect of different amounts of training data on accuracy

In Figure 1 we show the effect of the amount of training data on the accuracy using Logit Boost. We selected Logit Boost for this and the following experiments since its accuracy is comparable to SVMs but it is computationally less expensive. We randomly partitioned the transcription into 10 subgroups. Then we used one subgroup as the test set and the rest for training. Starting with one subgroup in the training set, we incrementally added one subgroup to the training set and evaluate the tagging performance of the test set. We repeated this process several times, choosing randomly a new test set each time. The percentages shown are the average over all the experiments. With only 10% of the sentences for training we are reaching very good accuracy already, as high as that from the strategy based on language identification. The curve flattens after

Table 4: Accuracy of Logit Boost with different subsets of attributes. 'X' marks attributes included. E_t, E_l, E_p, and S_t, S_l, and S_p are the POS tag, lemma and confidence output by the English and the Spanish POS tagger, respectively.

word	E_t	E_l	E_p	S_t	S_l	S_p	**Accuracy**
X	X	X	X	–	–	–	88.80
–	X	X	X	–	–	–	86.22
X	–	–	–	X	X	X	78.59
–	–	–	–	X	X	X	65.28
X	X	X	–	X	X	–	**92.95**
X	X	–	X	X	–	X	92.53
X	X	–	–	X	–	–	91.22
X	X	X	–	–	X	–	89.76
X	–	X	X	–	X	X	77.08
–	–	X	–	–	X	–	74.18
X	X	–	–	–	–	–	85.76
X	–	–	–	–	–	–	71.17
–	–	–	X	–	–	X	24.96
X	X	X	X	X	–	–	92.55
X	X	X	X	–	–	X	88.89
X	–	X	–	X	–	–	78.74
X	X	X	X	–	X	–	89.62
–	X	–	–	X	X	–	90.76
–	–	X	X	–	X	X	75.94
X	–	X	–	X	X	X	80.24
X	–	–	X	X	X	X	79.13

60% of the training data is used. We do not gain much by adding more training data after this.

Results shown in Table 3 demonstrate that POS tagging can be learned effectively based on the attributes described in Subsection 6.1, even if we are not explicitly adding contextual information. To determine the extent to which each attribute is contributing to the learning task, we performed another set of experiments where we selected different subsets of the attributes. Table 4 shows the results with Logit Boost. Overall, the attributes taken from the English POS tagger are more valuable for this learning task. If we only take the word form and the features from the English Tree tagger (first row in Table 4) we are reaching an accuracy that outperforms all heuristics. Still, there is some valuable information provided by the Spanish POS tagger output since the highest accuracy is achieved by including the Spanish-based attributes in combination with the English-based ones. Surprisingly, we can manage to outperform the oracle by using only three attributes: the lexical word form and the POS tags from the English and Spanish tagger (see row 7 in table), or the POS tags from the monolingual taggers together with the lemma from the Spanish tagger (see row 4 from bottom to top). We also experimented adding as an attribute the output of the language identification method, but found no significant changes in the accuracy.

7 Discussion

We analyzed the different results gathered through the experiments and we present here the most relevant insights.

The first discovery, is that a lot of the errors made by the oracle, and the other methods as well, are due to the difficulties inherent in dealing with spontaneous speech where fillers, interruption points, hesitations, and the like abound. About as much as 20% of the errors made by the oracle are due to these features. Another roughly 20% is due to unknown tokens in the transcription, such as mumbling, slang words such as "gonna" and "wanna", or other sounds unintelligible for the human transcriber. For the rest of the analysis we decided to ignore these types of mistakes for all methods and focus only on the remaining mistakes. In the case of the oracle we are left with 445 erroneously POS tagged words. From those, about 50%, or 233 to be exact, are errors in sentences with code-switches. We consider this to be a strong indication of the complexity that intrasentential switches add to the task of POS tagging. For the taggers, these sentences are incomplete, or ill-formed, since they have fragments with foreign words and thus, they fail to identify them. The rest of the oracle mistakes can not be attributable to a single cause. Some are fragmented sentences, and some are due to errors inherent of the tagger, but nothing is particulary salient about them.

The language identification methods share, of course, the same mistakes made by the oracle, plus 342 more, for a total of 787 (in the case of the dictionary-based language identification). The challenge of POS tagging code-switched text is more evident for this method. Out of the mistakes made by the language identification method, 540 lie in sentences with code-switching, that is, nearly 70% of the mistakes. For 307 of these mistakes the right

POS tag was available from one of the taggers. Some typical examples of these errors are words that belong to both languages, such as “a”, “no”, “me” and “con”.

The ML approach outperformed both the language identification method and the oracle. Analyzing the predictions made by SVM we verified that out of the 445 errors made by the oracle, SVM correctly tagged 223, the majority of which are words in sentences with code-switching (142 words). When compared against the errors from the method based on language identification, SVM correctly tagged 481 words out of the 787, 374 of which are words in sentences with code-switches. In summary, the ML approach is more robust to code-switched sentences. Note that we did find some errors made by the ML approach that are not shared by the oracle or the language identification method, a total of 105. Some of these mistakes are due to inconsistencies on the human-annotated tags. For instance, in most cases slang words such as “gonna” and “wanna” are labeled as unknown words, but we found that these words were labeled as verbs in a few cases. Not surprisingly this caused the ML algorithm to fail, since these class labels were misleading. The majority of the mistakes, however, seem to be due to systematic mistakes by the POS taggers.

One last remark is regarding our decision to find a method for successfully exploiting the existing taggers for POS tagging Spanglish text. Our original motivation came from the lack of linguistic resources to process Spanglish text. However, we did train from scratch a sequential model for POS tagging Spanglish, namely Conditional Random Fields (CRFs) (Lafferty *et al.*, 2001). We used MALLET (McCallum, 2002) for this experiment and the same training/testing partitions used in the experiment reported in Table 3. The CRF POS tagger was trained using capitalization information and the previous token as context. The average accuracy of this CRF was 81%, which is lower than the language identification heuristic. We believe that this low accuracy is due to the lack of a representative sample of annotated Spanglish. It will be interesting to see if when more data becomes available the ML algorithms still yield the best results.

8 Conclusions

Code-switching is a fresh and exciting research area that has received little attention in the language processing community. Research on this topic has many interesting applications, including automatic speech recognition, machine translation, and computer assisted language learning. In this paper we present preliminary work towards developing a POS tagger for English-Spanish code-switched text that, to the best of our knowledge, is the first effort towards this end.

We explored different heuristics for taking advantage of existing linguistic resources for English and Spanish with unimpressive results. A simple word-level language identification strategy outperformed all heuristics tested. But the best results, even better than the oracle, were achieved by using machine learning using the output of monolingual POS taggers as input features.

In the error analysis we showed that most of the mistakes made by the language identification method, and the oracle itself, occur in sentences with intrasentential code-switching, showing the difficulty of the task. In contrast, our machine learning approach was less sensitive to the complexity of this alternation pattern.

There is still a lot of work to do in this area. Our ongoing efforts include gathering a larger corpus, with different speakers and conversational styles, as well as written forms of code-switching from blogs and Internet forums. In addition, we are exploring the use of context information. The features we are currently using to represent each word do not take into account the context surrounding the word. We want to test if by using contextual features we can further improve our results.

In this study we focused on code-switching, but borrowing is another complex language alternation pattern that we want the POS tagger to handle. We are working on developing a special method for identification and morphological analysis of borrowings. This method will help increase the accuracy of the POS tagger.

Spanish-English is not the only popular combination of languages. An interesting line of future work would be to explore if the method presented here can be adapted to different language combi-

nations. Moreover, multilingual communities will code-switch among more than two codes and this poses fascinating research challenges as well.

Acknowledgements

We are grateful to Ray Mooney, Melissa Sherman and the four anonymous reviewers for insightful comments and suggestions. Special thanks to Brenda Medina and Nicolle Whitman for helping with some experiments. This research is supported by the National Science Foundation under grant 0812134.

References

Ardila, A. (2005) Spanglish: an anglicized dialect. *Hispanic Journal of Behavioral Sciences*, **27(1)**: 60–81.

Brants, T. (2000) TnT - a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference ANLP-2000* Seatle, WA.

Brill, E. (1990) A simple rule-based part-of-speech tagger. In *3rd Conference on Applied Natural Language Processing*, pp. 152–155. Trento, Italy.

Carreras, X. and Padró, L. (2002) A flexible distributed architecture for natural language analyzers. In *Third International Conference on Language Resources and Evaluation, LREC-02*, pp. 1813–1817. Las Palmas de Gran Canaria, Spain.

Charniak, E. (1993) Equations for part-of-speech tagging. In *11th National Conference on AI*, pp. 784–789.

Ervin, S. and Osgood, C. (1954) Second language learning and bilingualism. *Journal of abnormal and social phsychology*, Supplement 49, pp. 139–146.

Espinosa, A. (1917) Speech mixture in New Mexico: the influence of English language on New Mexican Spanish. In H. Stevens and H. Bolton, (eds.), *The Pacific Ocean in History*, pp. 408–428.

Franco, J.C. and Solorio T. (2007) Baby-steps towards building a Spanglish language model. In *8th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing-2007*, pp. 75–84. Mexico City, Mexico.

Friedman, J., Hastie, T., and Tibshirani, R. (1998) Additive logistic regression: a statistical view of boosting. Technical Report, Stanford University.

Godfrey, J., Holliman, E. and McDaniel, J. (1992) Switchboard: Telephone speech corpus for research development. In *ICASSP*, pp. 517–520, San Francisco, CA, USA.

Goyal, P., Mital, Manav R., Mukerjee, A., Raina, Achla M., Sharma, D., Shukla, P. and Vikram, K. (2003) A bilingual parser for Hindi, English and code-switching structures. In *Computational Linguistics for South Asian Languages –Expanding Synergies with Europe, EACL-2003 Workshop*, Budapest, Hungary.

Grosjean, F. (1982) *Life with Two Languages: An Introduction to Bilingualism*. Harvard University Press.

Gumperz, J. J. and Hernandez-Chavez, E. (1971) *Cognitive aspects of bilingual communication.* Oxford University Press, London.

Gumperz, J. J. (1964) Linguistic and social interaction in two communities. In John J. Gumperz (ed.), *Language in social groups*, pp. 151–176, Stanford. Stanford University Press.

Gumperz, J. J. (1971) Bilingualism, bidialectism and classroom interaction. In *Language in social groups*, pp. 311–339, Stanford. Stanford University Press.

Joshi, A. K. (1982) Processing of sentences with intrasentential code-switching. In Ján Horecký (ed.), *COLING-82*, pp. 145–150, Prague.

Kucera, H. and Francis, W. N. (1967) *Computational analysis of present-day American English*. Brown University Press.

Lafferty, J. and McCallum, A. and Pereira F. (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *18th ICML*, pp. 282–289. MA, USA.

Lipski, J. M. (1978) Code-switching and the problem of bilingual competence. In M. Paradis (ed.), *Aspects of bilingualism*, pp. 250–264, Columbia, SC. Hornbeam.

McCallum, A. (2002) MALLET: A Machine Learning for Language Toolkit. Retrieved January 7, 2008 from `http://mallet.cs.umass.edu`

Paul, D. B. and Baker, J. M. (1992) The design of the Wall Street Journal-based CSR corpus. In *HLT'91: workshop on speech and Natural Language* pp. 357–362, Morristown, NJ, USA.

Poplack, S., Sankoff, D. and Miller, C. (1988) The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, **26(1)**: 47–104.

Poplack, S. (1980) Sometimes I'll start a sentence in Spanish y termino en español: toward a typology of code-switching. *Linguistics*, **18(7/8)**: 581–618.

Poplack, S. (1981) Syntactic structure and social function of code-switching. In R. Duran (ed.), *Latino discourse and communicative behavior*, pp. 169–184, Norwood, NJ. Ablex.

Quinlan, J. R. (1986) Induction of decision trees. *Machine Learning*, **1**: 81–106.

Ratnaparkhi, A. (1996) A maximum entropy model for part-of-speech tagging. In *EMNLP*, pp. 133–142, Philadelphia, PA, May.

Rosner, M. and Farrugia, P. (2007) A tagging algorithm for mixed language identification in a noisy domain.

In *INTERSPEECH 2007*, pp. 190–193, Antwerp, Belgium.

Sankoff, D. (1968) Social aspects of multilingualism in New Guinea. Ph.D. thesis, McGill University.

Sankoff, D. (1981) A formal grammar for codeswitching. *Papers in Linguistics: International Journal of Human communications*, **14(1)**: 3–46.

Sankoff, D. (1998a) A formal production-based explanation of the facts of code-switching. *Bilingualism, Language and Cognition*, **1**: 39–50. Cambridge University Press.

Sankoff, D. (1998b) The production of code-mixed discourse. In *36th ACL*, volume I, pp. 8–21, Montreal, Quebec, Canada.

Schmid, H. (1994) Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing, Manchester, UK*.

Schölkopf, B. and Smola, A. J. (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press.

Toribio, A. J. (2001a) Accessing Spanish-English code-switching competence. *International Journal of Bilingualism*, **5(4)**:403–436.

Toribio, A. J. (2001b) On the emergence of bilingual code-switching competence. *Bilingual Language and Cognition*, **4(3)**:203–231.

U.S. Census Bureau. (2003) Language use and English speaking ability: 2000. Retrieved October 30, 2006 from `http://www.census.gov/prod/2003pubs/c2kbr-29.pdf`.

Witten, I. H. and Frank, E. (1999) *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

Wolpert, D. H. (1992) Stacked Generalization. *Neural Networks*, **5(2)**:241–259.