Learning to Predict Code-Switching Points

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

Predicting possible code-switching points can help develop more accurate methods for automatically processing mixed-language text, such as multilingual language models for speech recognition systems and syntactic analyzers. We present in this paper exploratory results on learning to predict potential code-switching points in Spanish-English. We trained different learning algorithms using a transcription of code-switched discourse. To evaluate the performance of the classifiers, we used two different criteria: 1) measuring precision, recall, and F-measure of the predictions against the reference in the transcription, and 2) rating the naturalness of artificially generated code-switched sentences. Average scores for the code-switched sentences generated by our machine learning approach were close to the scores of those generated by humans.

1 Introduction

Multilingual speakers often switch back and forth between languages when speaking or writing, mostly in informal settings. The mixing of languages involves very elaborated patterns and forms and we usually use the term Code-Switching (CS) to encompass all of them (Lipski, 1978). Before the Internet era, CS was mainly used in its spoken form. But with so many different informal interaction settings, such as chats, forums, blogs, and web sites like Myspace and Facebook, CS is being used more and more in written form. For English and Spanish, CS has taken a step further. It has become a hallmark of the chicano culture as it is evident by the growing number of chicano writers publishing work in Spanish-English CS.

We have not completely discovered the process of human language acquisition, especially dual language acquisition. Findings in linguistics, sociolinguistics, and psycholinguistics show that the production of code-switched discourse requires a very sophisticated knowledge of the languages being mixed. Some theories suggest bilingual speakers might have a third grammar for processing this type of discourse. The general agreement regarding CS is that switches do not take place at random and instead it is possible to identify rules that bilingual speakers adhere to.

Understanding the CS process can lead to accurate methods for the automatic processing of bilingual discourse, and corpus-driven studies about CS can also inform linguistic theories. In this paper we present exploratory work on learning to predict CS points using a machine learning approach. Such an approach can be used to reduce perplexity of language models for bilingual discourse. We believe that CS behavior can be learned by a classifier and the results presented in this paper support our belief.

One of the difficult aspects of trying to predict CS points is how to evaluate the performance of the learner since switching is intrinsically motivated and there are no forced switches (Sankoff, 1998b). Therefore, standard classification measures for this task such as precision, recall, F-measure, or accuracy, are not the best approach for measuring the effectiveness of a CS predictor. To comple-

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 973–981,
Honolulu, October 2008.

ment the evaluation of our approach, we designed a task involving human judgements on the naturalness of automatically generated code-switched sentences. Both evaluations yielded encouraging results.

The next section discusses theories explaining the CS production process. Then in Section 3 we present our framework for learning to predict CS points. Section 4 discusses the empirical evaluation of the classifiers compared to the human reference. In Section 5 we present results of human evaluations on automatically generated code-switched sentences. Section 6 describes previous work related to the processing of code-switched text. Finally, we conclude in Section 7 with a summary of our findings and directions for future work.

2 Bilingual Discourse

The combination of languages can be considered to be a continuous spectrum where on each end of the spectrum we have one of the standard languages and no blending. As one moves closer to the middle of the spectrum the amount and complexity of the blending pattern increases. The blending pattern most widely known, and studied, is code-switching, which refers to the mixing of words from two languages, but the words themselves do not suffer any syntactic or phonological alterations. The CS points can lie at sentence boundaries, but very often we will also observe CS inside sentences. According to (Sankoff, 1998b; Poplack, 1980; Lipski, 1978) when CS is used inside a sentence, it can only happen at syntactic boundaries shared by both languages, and the resulting monolingual fragments will conform to the grammar of the corresponding language. In this CS theory the relationship between both languages is symmetric –lexical items from one language can be replaced by the corresponding items in the second language and vice versa. Another prevalent linguistic theory argues the contrary: there is an asymmetric relation where the changes can occur only in one direction, which reflects the existence of a Matrix Language (ML), the dominant language, and an Embedded Language (EL), or subordinate language (Joshi, 1982). The Matrix Language Frame model, proposed and extended by Scotton-Myers, supports this asymmetric relation theory. This formalism prescribes that content morphemes can come from the ML or the EL, whereas late system morphemes, the elements that indicate grammatical relations, can only be provided by the ML (Myers-Scotton, 1997).

Until an empirical evaluation is carried out on large representative samples of discourse involving a large number of different speakers, and different language-pairs, the production of CS discourse will not be explained satisfactorily. The goal of this work is to move closer to a better understanding of CS by learning from corpora to predict possible CS points.

3 Learning When To Code-Switch

3.1 The English-Spanish Code-Switched Data Set

We recorded a conversation among three English-Spanish bilingual speakers that code-switch regularly when speaking to each other. The conversation lasts for about 40 minutes ($\sim$8k words, 922 sentences). It was manually transcribed and annotated with Part-of-Speech (POS) tags. A total of 239 switches were identified manually. English is the predominant language used, with a total of 576 monolingual sentences. We refer to this transcription as the Spanglish data set. We are currently in the process of collecting new transcriptions of this conversation in order to measure inter annotator agreement.

3.2 Approach

Machine learning algorithms have proven to be surprisingly good at language processing tasks, including optical character recognition, text classification, named entity extraction, and many more. The premise of our paper is that machine learning algorithms can also be successful at learning how to code-switch as well as humans. At the very least we want to provide encouraging evidence that this is possible. To the best of our knowledge, there is no previous work related to the problem of automatically predicting CS points. Our machine learning framework then is inspired by existing theories of CS and existing work on part-of-speech tagging code-switched text (Solorio and Liu, 2008).

In our approach, each word boundary is a potential point for switching – an instance of the learning task. It should be noted that we can only rely on the history of words preceding potential CS points in or-

Feature id	Description
1	Word
2	Language id
3	Gold-standard POS tag
4	BIO chunk
5	English Tree Tagger POS
6	English Tree Tagger prob
7	English Tree Tagger lemma
8	Spanish Tree Tagger POS
9	Spanish Tree Tagger prob
10	Spanish Tree Tagger lemma

Table 1: Features explored in learning to predict CS points.

der to extract meaningful features. Otherwise, if we look also into the future, we could just do language identification to extract the CS points. However, our goal is to provide methods that can be used in real time applications, where we do not have access to observations beyond the point of interest. Another restriction we imposed on the method is related to the size of the context used. A sentence can be code-switched in different ways, with all different versions adhering to the CS "grammar". The number of permissible CS sentences grows almost exponentially with the length of the sentence[1]. By limiting the length of the context to at most two words we are trying to avoid some sort of over fitting by having the model making assumptions over the interaction of the two languages that will be too weak, or speaker-dependent.

Previous studies have identified several socio-pragmatic functions of code-switching. The most common include direct quotation, emphasis, clarification, parenthetical comments, tags, and trigger switches. Other characteristics relevant to CS behavior are the topic being discussed, the speakers involved, the setting where the conversation is taking place, and the level of familiarity between the speakers. Having encoded information regarding the CS function and the aforementioned relevant factors might help in predicting upcoming CS points. However, annotating this information in the transcription can be time consuming and very often this information is not readily available. Therefore, at the expense of making this task even more difficult, we decided against trying to include this type of information and include only lexical and syntactic features, to evaluate a practical and cost effective method for this task. Table 1 shows the list of features. All of these features are associated with word w_n, the word immediately preceding boundary n. Feature 1 is the word form[2]. Feature 2 is language identification. If the production of CS discourse adheres to the matrix language frame model, then knowledge of the language can potentially be a good source of information. Feature 3 is the gold-standard POS tag. We also include as a feature the position of the word relative to the phrase constituent using a Beginning-Inside-Outside (BIO) scheme. For instance, the word at the beginning of the verb phrase will be labeled as B, the following words inside this verb phrase will be tagged as I, and words that were not identified as part of a phrase constituent were labeled as O. This chunking information was extracted using the English and Spanish versions of FreeLing[3]. We did not measure accuracy on the chunking information. Features 5 to 9 were generated by tagging the Spanglish conversation using the Spanish and the English versions of the Tree Tagger (Schmid, 1994). Attributes 5 to 7 are extracted from the English version, which include the POS tag, the confidence, and the lemma for that word. Similarly, features 8 to 10 were taken from the Spanish monolingual tree tagger. Features from the monolingual taggers will have some noisy labels when tagging fragments of the other language. However, considering that our feature set is small we want to explore if adding these features, which include the lemmas and probability estimates, can contribute to the learning task.

We also explored using a larger context. In this case, we extract the same features shown in Table 1 for the two words preceding the word boundary, resulting in 20 attributes representing each instance.

Evaluation for this task is not straightforward. Within a sentence, there are several CS points that will result in a natural sounding code-switched sentence, but none of these CS points are mandatory.

[1] Almost exponentially because not all sentences will be considered grammatical.

[2] Strictly speaking these should be called tokens, not words since punctuation marks are considered as well.

[3] http://garraf.epsevg.upc.es/freeling/

CS has a lot to do with the speaker's preferences, the topic being discussed, and the background of the participants involved. Using the standard approach for measuring performance of classifiers can be misleading, especially if the reference data set is small and/or has only a small number of speakers. It is unrealistic to just consider F-measure, or accuracy, as truthfully reflecting how well the learners generalize to the task. Therefore, we evaluated the classifier's performance using two different criteria, which are discussed in the next sections.

4 Evaluation 1: Using the Reference Data Set

This is the standard evaluation of machine learning classifiers. We randomly divided the data into sentences and grouped them into 10 subsets to perform a cross-validation. Tables 2 and 3 show results for Naive Bayes (NB) and Value Feature Interval (VFI) (Demiroz and Guvenir, 1997). Using WEKA (Witten and Frank, 1999), we experimented with different subsets of the attributes and two context windows: using only the preceding word and using the previous two words. The results presented here are overall averages of 10-fold cross validation. We also report standard deviations. It should be noted that the Spanglish data set is highly imbalanced, around 96% of the instances belong to the negative class. Therefore, our comparisons are based on Precision, Recall, and F-measure, leaving accuracy aside, since a weak classifier predicting that all instances belong to the negative class will reach an accuracy of 96%.

The performance measures shown on Tables 2 and 3 show that NB outperforms VFI in most of the configurations tested. In particular, NB yields the best results when using a 1 word context with no lexical forms nor lemmas as attributes (see Table 2 row 3). This is a fortunate finding –for most practical problems there will always be words in the test set that have not been observed in the training set. For our small Spanglish data set that will certainly be the case. In contrast, VFI achieves higher F-measures when using a context of two words and all the features are used.

Analyzing the predictions of the learners we noted that the NB classifier is heavily biased by the language attribute, close to 80% of the positive predictions made by NB are after seeing a word in Spanish. This preference seems to support the assumption of the asymmetry between the two languages and the existence of an ML[4]. This however is not the case for VFI, only a little over 50% of the positive predictions belong to this scenario. Another interesting finding is the learner's tendency to predict a code-switch after observing words like "Yeah", "anyway", "no", and "shower". The first two seem to fit the pattern of idiomatic expressions. According to Montes-Alcalá this type of CS includes linguistic routines and fillers that are difficult to translate accurately (Montes-Alcalá, 2007), which might be the case of "anyway", and unconscious changes, which can explain the case of "Yeah". The case of "shower" and "no" are more difficult to explain, they might be overfitting patterns from the learners. We also found out that VFI learned to predict that a CS will take place right after seeing the sequence of words *le dije* (I said). This sequence of words is frequently used when the speaker is about to quote his/herself, and this quotation is one of the well-documented CS functions (Montes-Alcalá, 2007).

A greedy search approach for attribute selection using WEKA showed that out of the 20 attributes (when using a two word context), the subset with the highest predictive value included the language identification for word w_{n-1} and w_{n-2}, the confidence threshold from the English tagger for word w_{n-2}, the lemma from the Spanish Tree tagger for w_{n-1}, and the lexical form of the word w_{n-1}. We expected the chunk information to be useful and this does not seem to be the case. Another unexpected outcome is that higher F-measures are reached by adding features generated by the monolingual Tree taggers. Even though these features are noisy, they still carry useful information.

We only show results from NB and VFI. Initial experiments with a subset of the data showed that these algorithms were the most promising for this task. They both yielded higher F-measures, even when compared against Support Vector Machines (SVMs), C4.5, and neural networks. On this experiment all the discriminative classifiers reached a classification accuracy close to 96%, but an F-

[4]We remind the reader that in this paper ML stands for Matrix Language.

Features Used											Naive Bayes		
					English Tree tagger			Spanish Tree tagger					
C	Word Form	Lang id	POS tag	BIO chunk	POS	Prob	Lem	POS	Prob	Lem	**P**	**R**	**F_1**
1		X	X	X							0.09(0.01)	0.01(0.00)	0.02(0.00)
1	X	X	X	X							0.23(0.01)	0.32(0.02)	0.27(0.02)
1*		X	X	X	X	X		X	X		0.19(0.00)	0.53(0.00)	0.28(0.00)
1	X	X	X	X	X	X	X	X	X	X	0.18(0.00)	0.59(0.00)	0.27(0.00)
2		X	X	X							0.13(0.00)	0.35(0.00)	0.19(0.00)
2	X	X	X	X							0.16(0.00)	0.46(0.00)	0.23(0.00)
2		X	X	X	X	X		X	X		0.14(0.00)	0.55(0.01)	0.23(0.00)
2	X	X	X	X	X	X	X	X	X	X	0.16(0.00)	0.59(0.01)	0.25(0.00)

Table 2: Prediction results of CS points with NB using different features. Column C indicates the size of the context used, 1 indicates a 1 word context, and 2 indicates two words preceding the word boundary. Columns P, R, and F_1, show precision, recall, and F-measure, respectively. Numbers in parenthesis show standard deviations. The row marked with a '*' shows the configuration used for the generation of CS sentences presented in Section 5.

measure on the positive class of around 0%. NB and VFI estimate predictions for each class separately, which makes them robust to imbalanced data sets. In addition, generative models are known to be better for smaller data sets since they reach their higher asymptotic error much faster than discriminative models (Ng and Jordan, 2002). This might explain why Naive Bayes outperformed strong classifiers such as SVMs by a large margin.

The overall prediction performance is not very high. However, we should remark that for this particular task expecting a high F-measure is unrealistic. Consider for example, a case where the learners predict a CS point where the speaker decided not to switch, this does not imply that particular point is not a good CS point. And similarly, if the classifier missed an existing CS point in the reference data set the resulting sentence might still be grammatical and natural sounding. This motivated the use of an alternative evaluation, which we discuss below.

5 Evaluation 2: Using Human Evaluators

The goal of this evaluation is to explore how humans perceive our automatically generated CS sentences, and in particular, how do they compare to the original sentences and to the randomly generated ones. We selected 30 spontaneous and naturally occurring CS sentences from different sources. Some of them were selected from the Spanglish Times Magazine[5], some others from blogs found in (Montes-Alcalá, 2007). Other sentences were taken from a paper discussing CS on e-mails (Montes-Alcalá, 2005). All of the sentences are true occurrences of written CS, from speakers different from the ones in the Spanglish data set. The sentences were translated to standard English and Spanish and were manually aligned. We will use this parallel set of sentences to predict CS points with our models. Based on the model predictions we will generate code-switched sentences by combining monolingual fragments.

It should be noted that the Spanglish data set is a transcription of spoken CS. In contrast, this new evaluation set contains only written CS. Recent studies suggest written CS will adhere to the rules of spoken CS (Montes-Alcalá, 2005), but there is still some controversy on this issue. From our perspective, both samples come from informal conversational interactions. It is expected that both will have similar patterns and therefore will provide a good source for our evaluation.

5.1 Automatically Generated Code-Switching Sentences

In this subsection we describe how to generate code-switched sentences randomly and with the learned models described in the previous sections. For the

[5]http://www.spanglishtimes.com/

Features Used											Voting Feature Intervals		
					English Tree tagger			Spanish Tree tagger					
C	Word Form	Lang id	POS tag	BIO chunk	POS	Prob	Lem	POS	Prob	Lem	**P**	**R**	**F_1**
1		X	X	X							0.12(0.00)	0.68(0.00)	0.21(0.00)
1	X	X	X	X							0.12(0.00)	0.65(0.01)	0.20(0.00)
1*		X	X	X	X	X		X	X		0.12(0.00)	0.72(0.01)	0.21(0.00)
1	X	X	X	X	X	X	X	X	X	X	0.13(0.00)	0.65(0.00)	0.22(0.00)
2		X	X	X							0.13(0.00)	0.60(0.00)	0.21(0.00)
2	X	X	X	X							0.15(0.00)	0.52(0.01)	0.23(0.00)
2		X	X	X	X	X		X	X		0.13(0.00)	0.68(0.00)	0.22(0.00)
2	X	X	X	X	X	X	X	X	X	X	0.15(0.00)	0.51(0.00)	0.24(0.00)

Table 3: Prediction results of CS points with VFI using different features. The notation on this table is the same as in Table 2

classifier-based approach, we POS tagged each parallel set of sentences, with the monolingual English and Spanish Tree Taggers, and we extracted the same set of features described shown in Table 1. We decided to train the models with a context size of one word, even though both learners reached higher F-measures when using a two-word context. This decision was based on the observation that having a two-word context will pose restrictions on possible CS points, since we would not be able to switch unless we have inserted into the sentence at least two tokens from the same language.

We trained the NB and VFI models with the Spanglish data set (using features 2–6, 8, and 9, see Table 1) and generated CS predictions for each parallel file. A code-switched sentence is generated by adding the first token of the sentence in language 1 (L1), and continue adding more tokens from L1 until a CS point is found. When a CS prediction is found, the following tokens are selected from the second language (L2), and we continue adding tokens from L2 until the classifier has predicted a change. Different versions of the sentences are generated by changing the definition of L1 and L2.

For the randomly generated CS sentences, switching decisions are made randomly with a probability proportional to the positive predictions made by the classifiers (in this case NB). That is, for the Spanish sentences switch points are predicted randomly with a 30% chance of switching while for English switch points are predicted with a 10% chance.

Generator	Average Score
Human	3.64
NB	3.33
Random	2.68
VFI	2.50

Table 4: Average score of 18 judges over the set of 28 code-switched sentences rated.

In total we generated 180 CS sentences: 30 sentences per generator scheme (we have three generators: NB, VFI, and random), and two versions from each generator corresponding to the two possible configurations of L1-L2 (Spanish-English, English-Spanish). We noticed that in some cases same sentences are generated by different methods and sometimes there are no switches. We narrowed down the sentences by randomly choosing the combination of L1-L2 for each generator. This reduced the number of sentences from having 6 versions, to having only 3 versions of each sentence. From the resulting 30 sets, we removed 2 sets because one or more of the generator schemes produced a monolingual sentence. Therefore, we used 28 sets for human evaluations.

5.2 Human Evaluation Results

We had a total of 18 subjects participating in the experiment. All of them identified themselves as being able to read and write Spanish and English, and the majority of them said to have used CS at least

some times. We showed to the human subjects the 28 sets of sentences. This time we included the original version of the sentence. Therefore, each judge was given 4 versions of each of the 28 code-switched sentences: the one generated from NB predictions, the one from VFI, the randomly generated, and the original one. Then we asked them to rate each sentence with a number from 1 to 5 indicating how natural and human-like the sentence sounds. A rating of 5 means that they strongly agree, 4 means they agree, 3 not sure, 2 disagree, 1 strongly disagree.

The average results are presented in Table 4. The sentences generated by NB were scored considerably higher than those from VFI and random, and closer to the human sentences. According to the paired t-test the difference between the NB score and the random one is significant (p=0.01). However the average score for VFI is lower than random. More experiments are needed to see if by choosing the setting where VFI had the highest F-measure would make a difference in this respect. Overall the subjects rated the human-generated CS sentences lower than what we were expecting, although it is clear that they consider these sentences more natural sounding than the rest. This low rating might be related to the attitude several evaluators expressed toward CS. In the evaluation form we asked the judges to express their opinion on CS and several of them indicated feelings along the lines of "we shouldn't code-switch".

There are several ways in which two parallel sentences can be combined in CS, and possibly several will sound natural, but from our results, it is clear that the NB algorithm was indeed able to generate a human-like CS behavior that was successfully differentiated from randomly-generated sentences.

By looking at the set of automatically generated code-switched sentences, we realized that the majority of the sentences are grammatical and natural sounding. We believe that for a large number of the sentences it would be hard for a human to distinguish the sentences that were automatically generated from the human-generated ones. One of the give away clues is when a multi-word expression is CS, or a tag line. Table 5 shows three examples from the sentences evaluated. In the table there is an example in sentence 1c where the noun phrase is code-switched, the sentence is grammatical according to Spanish rules, but it sounds very odd to have the noun *carta* followed by the adjective in English, "astrological". Other interesting features are present in example 3 where for the same noun phrase "produce section" we have both, the female marking determiner *la* and the masculine *el*. The same thing happens for the noun phrase "check-out line". We would need to have a larger occurrence of these instances in our test set to determine if on average one form is preferred over the other.

In another experiment, we measured the prediction performance of NB and VFI on the 30 code-switched sentences used in this part of the evaluation. The best results, an F-measure of 0.418, were achieved by NB when a context of 1 word was used, and no words, nor lemmas were included as features. This is the same setting used for the generation process. In contrast, VFI reached an F-measure of 0.351 on this same setting. 30 sentences represent a very small dataset but the results are very promising since the speakers are different in the training and testing dataset. Moreover, these results support the claim that written and spoken CS obey similar rules.

6 Related Work

There is little prior work on computational linguistic approaches to code-switched discourse. Most of the previous work includes formalisms to 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). His real-time production model draws on the alternation of fragments from two virtual monolingual sentences. But no statistical assessment has been conducted on real corpora.

Another related work deals with language identification on English-Maltese code-switched SMS messages (Rosner and Farrugia, 2007). What the authors 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.

1a. Naive Bayes:
By unlocking the information in your astrological chart, *puedo ver la respuesta!* Ask me!
1b. VFI:
Puedo ver la answer by unlocking the information in your *carta astrológica!* Ask me !
1c. Random:
By unlocking the information *de tu carta* astrological, I can see the answer! Ask me !
1d. Human:
By unlocking the information in your astrological chart, *puedo ver* the answer! *Pregúntame!*
1e. English version:
By unlocking the information in your astrological chart, I can see the answer! Ask me!
2a. Naive Bayes:
Pero siendo this a new year, *es tiempo de empezar de nuevo que no?*
2b. VFI:
But this being a new year, *es tiempo de empezar* over isn't it ?
2c. Random:
But this being a new *año*, it's *tiempo* to start over isn't it?
2d. Human:
Pero this being a new year, it's a time to start over *que no?*
2e. English version:
But this being a new year, it's time to start over isn't it?
3a. Naive Bayes:
Juan confirmed me that it was very obvious, *y no solamente en el* produce section, *en la* check-out line as well.
3b. VFI:
Me confirmó Juan que it was very obvious, *y no solamente en el* produce section, *también en la* check-out line.
3c. Random:
Juan confirmed *que fue* very obvious, *y* not *solamente en el área de* produce, in the check-out line as well.
3d. Human:
Me confirmó Juan que fue muy obvio, y no solamente en la produce section, *también en el* check-out line.
3e. English version:
Juan confirmed me that it was very obvious, and not only on the produce section, in the check-out line as well.

Table 5: Examples of automatically generated CS sentences.

7 Conclusions

We presented preliminary results on learning to predict CS points with machine learning. One of the possible applications of our method involves fine-tuning the weights in a multilingual language model, for instance, as part of a speech recognizer for Spanglish. With this in mind, we restricted the possible features in the learning scenario allowing only lexical and syntactic features that could be automatically generated from the text. Empirical evaluations on a Spanglish conversation showed that Naive Bayes and VFI can predict with acceptable F-measures possible CS points, considering the difficulty of the task. Prediction of CS points can help improve multilingual language models.

Evaluation of our approach cannot be done based only on the gold-standard set since there is no single right answer in this task. Therefore, we complemented the evaluation by involving judgements from bilingual speakers. We generated CS sentences by taking the predictions from the classifiers to merge parallel sentences. On average, the sentences generated from the NB model were rated closer to the original sentences, and a lot higher than the ones from a random generator. Most of the sentences sounded human-like. But because the process is automatic we did find some awkward constructions, for example plural vs singular noun-verb agreement, or multi-word phrases that were code-switched in the middle. Perhaps a multi-word recognition feature could improve results.

One of the advantages of technological development and economic globalization is that more people from different regions of the world with different cultures, and therefore, different languages will

be in closer contact. As a result, code-switching will become more popular. It is important to start addressing this type of bilingual communication from a computational linguistics point of view. This work is one of the few attempts to fill the gap.

Some directions for future work include: exploring the extent to which our results can be improved by including a multi-word expression recognition system. We also want to investigate the integration of our approach to multilingual language models and move beyond CS to address other deeper linguistic phenomena. Lastly, we would like to explore similar approaches in other popular language combinations.

Acknowledgements

This research is supported by the National Science Foundation under grant 0812134. We are grateful to Ray Mooney, Melissa Sherman and the three anonymous reviewers for insightful comments and suggestions. Special thanks to the human judges that helped with the sentence evaluations.

References

G. Demiroz and H. A. Guvenir. 1997. Classification by voting feature intervals. In *European Conference on Machine Learning, ECML-97*, pages 85–92.

P. Goyal, Manav R. Mital, A. Mukerjee, Achla M. Raina, D. Sharma, P. Shukla, and K. Vikram. 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.

A. Joshi. 1982. Processing of sentences with intrasentential code-switching. In Ján Horecký, editor, *COLING-82*, pages 145–150, Prague, July.

J. Lipski. 1978. Code-switching and the problem of bilingual competence. In M. Paradis, editor, *Aspects of bilingualism*, pages 250–264. Hornbeam.

C. Montes-Alcalá. 2005. Mándame un e-mail: cambio de códigos español-inglés online. In Luis Ortiz and Manel Lacorte, editors, *In Contacto y contextos lingüísticos: El español en los Estados Unidos y en contacto con otras lenguas*. Iberoamericana/Vervuert.

C. Montes-Alcalá. 2007. Blogging in two languages: Code-switching in bilingual blogs. In Jonathan Holmquist, Augusto Lorenzino, and Lotfi Sayahi, editors, *In Selected Proc. of the Third Workshop on Spanish Sociolinguistics*, pages 162–170, Somerville, MA. Cascadilla Proceedings Project.

C. Myers-Scotton. 1997. *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford University Press, 2nd edition.

A. Ng and M. Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and Naive Bayes. In *Advances in Neural Information Processing Systems (NIPS) 15*. MIT Press.

S. Poplack. 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.

M. Rosner and P. J. Farrugia. 2007. A tagging algorithm for mixed language identification in a noisy domain. In *INTERSPEECH 2007*, pages 190–193, Antwerp, Belguim, August.

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

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

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

T. Solorio and Y. Liu. 2008. Part-of-speech tagging for English-Spanish code-switched text. In *EMNLP-2008*, Honolulu, Hawai, October.

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