Part of Speech Tagging Bilingual Speech Transcripts with Intrasentential Model Switching

Paul Rodrigues
University of Maryland Center for Advanced Study of Language
7005 52nd Avenue
College Park, MD 20742

Sandra Kübler
Indiana University
1021 E 3rd St.
Bloomington, IN 47405

Abstract

This paper investigates incremental part of speech tagging for speech transcripts that contain multilingual intrasentential code-mixing, and compares the accuracy of a monolithic tagging model trained on a heterogeneous-language dataset to a model that switches between two homogeneous-language tagging models dynamically using word-by-word language identification. We find that the dynamic model, even though presented a smaller context consisting of sentence fragments, meets the accuracy of the monolithic code-mixing model which is aware of increased context. Our system is modular, and is designed to be expanded to many-language code-mixing.

Introduction

This paper discusses a part of speech (POS) tagger that is designed to be modular to multilingual code-mix settings using dynamic model switching. POS tagging is the labeling of each token in a text with a category label, such as verb, noun, or preposition. The basic language unit used in POS tagging is the sentence. However, code-mixing in bilingual speakers often happens within a sentence rather than at sentence boundaries. This forces a POS tagger to either use a multilingual language model, or to split sentences into chunks of a single language, and then label the chunks. Such chunks can be considered microtext since they are often rather short. In our corpus, the language chunks average 6.14 words in length. Language chunks also provide novel challenges for POS tagging since they are generally grammatically incomplete.

Table 1 shows an example of code switching between English and Spanish.

Our novel approach to POS tagging multilingual text uses dynamic model switching, which relies on an indicator function that emits word-by-word language identification tags, and an event controller which monitors this output and selects the appropriate tagging model to use for a given word (Rodrigues 2012). The tagger is reactive, in that the system receives a streaming queue of tokenized words from the beginning of each sentence or language boundary, and makes a tag guess for each word utilizing only current- and leftward-contextual data as input. Such a system is useful in fields such as human-computer interaction where a multilingual speech signal might require a response to the content initiated without hesitation.

Language chunk	Fragment label
i am afraid that something will	ENG
se va a caer o ninguna cosa	SPA
.	PUNC
lo ponemos ah	SPA

Table 1: Three fragments, and the language chunk label determining which POS model to utilize

Typically, a POS tagger maximizes word-tag association probability across an entire sentence. The individual probabilities of transitioning from one word to the next use a context of 2 words to condition on. In our system, the event controller segments the multilingual streaming queue into chunks consisting of a homogeneous language, as can be seen in Table 1. This chunking has an effect of reducing the word's contextual data, blinding the tagger to leftward context across chunk boundaries. For example, though the first word of the second chunk is a continuation of the sentence that started in English, it would have to be considered the first word of a new sentence to the Spanish language module. This model also requires the POS tagger to be incremental, thus limiting its accessible information to the left context of a word to be tagged since the right context is not available yet.

We perform our experiments on a POS-annotated Spanish/English speech transcription introduced by Solorio and Liu (2008). As far as the authors know, this is the only multilingual POS-annotated dataset, of any language combination, large enough to train a statistical tagging model.

To evaluate the accuracy of the tagging models alone, this paper assume an oracle – the indicator function for this paper performs word-by-word language identification without error.

As baselines, we compare our system to a tagging model which outputs only English tags, as well as a heterogeneous tagging model which outputs both English and Spanish tags. Neither of these baseline systems require an indicator func-

tion to identify the language of each word; the former model assumes that every word is an English word while the latter uses a single, hybrid model for both languages.

We start by surveying incremental tagging systems and bilingual POS tagging research. We then describe the Spanish/English speech transcription dataset we use and how we prepare it for our experiments. With this, we evaluate several POS tagging models trained and tested on the data. We establish baselines using a single-language POS model trained on the predominant language of the corpus, as well two single-language POS models trained across the language chunk barrier. We compare these to tagging with separate models for each language, activated by the language identification event controller.

Incremental Tagging

Incremental POS tagging provides a label for each word after the word is presented to the system. The process is blind to the far right context of the token, leaving the label of the current token more ambiguous, and therefore limiting the accuracy of the tagger. Right context is known to be important (Ivanova and Kübler 2008), but Beuck, Köhn, and Menzel (2011) highlight several possible advantages of incremental tagging, such as a timely response, a decisiveness hypothesis stack, and a monotonic or fixed stream of output tags. These qualities allow for word-level online processing, emitting a stream of tags, with input of a stream of words.

We highlight MeLT, SVMTool, and HunPOS, three publicly available incremental tagging systems.

The Maximum-Entropy Lexicon-Enriched Tagger (MeLT) (Denis and Sagot 2009) is a maximum entropy tagger that uses the MEGA Model Optimization Package (MegaM) (Daumé III 2004) for parameter optimization. The tagger was originally developed for French, and it has been designed to integrate a morphosyntactic lexicon. MeLT+ is an extension of the MeLT tool that allows for the addition of supplementary independent features that are passed to MegaM (Kübler et al. 2010).

SVMTool (Giménez and Márquez 2004) is a discriminative POS tagger that uses support vector machine (SVM) classifiers to determine the POS of a word. The system can extract some lexical, POS, and orthographic features from examples, and external features can be supplied. SVM^{light} (Joachims 1997) is used for the SVM architecture. An adjustable sliding window around a target word is used to extract features from the series. A 5-token window is used by default, with the third token being the target word position. This provides feature extraction on the two tokens prior to the word, as well as the two tokens after the word. This window can be set to configure to ignore right-context of a word. SVMTool has high monolingual performance. Charniak et al. (2000) reached 97.16% accuracy on the English-language WSJ corpus and European Language Resources Association (2000) reached 96.89% accuracy on the Spanish-language LEXESP corpus.

Köhn (2009) describes an incremental version of HunPOS (Halácsy, Kornai, and Oravecz 2007), an open-source implementation of the TnT POS tagger (Brants 2000). This tagger uses hidden Markov models on known words and falls back to character n-grams for unknown words.

SVMTool was found to correlate in accuracy with HunPOS when SVMTool was configured to run with a 0 word lookahead, but SVMTool had slightly higher accuracy. [1] We chose SVMTool 1.3.1 over MeLT+ due to its highly competitive POS tagging accuracy, amount of documentation, and the ease of use. Information regarding our configuration of SVMTool can be found in the section on the experimental setup.

Bilingual Part of Speech Tagging

Solorio and Liu (2008) built several bilingual speech taggers by utilizing the output of a decision tree POS tagger, TreeTagger (Schmid 2004). They used single- and hybrid-language tagging models, lemma lexicons, and probability scores, in order to determine the parameter combinations that would deliver the highest accuracy on their Spanish/English code-mixing conversation data. Some experimental trials used a pre-built Spanish model, trained on the Spanish CRATER corpus, while other trials used a pre-built English model, which was trained on the Penn Treebank. These datasets each contain a tag to mark foreign words, which was used as a special condition in one rule-based approach in the paper. Their Spanish/English code-mixing dataset is used for the experiments in this paper, and will be detailed later.

The authors established two baselines by running their dataset through the pre-built monolingual TreeTagger models. The Spanish model received 25.99% accuracy, and the English model received 54.59% accuracy. The difference between the two results was attributed to the dataset having a larger number of English words than Spanish.

The authors utilized TreeTagger's probability score output for the first experiment. For each word, the tag was selected from the model that yielded the highest probability score. Where there was a tie, a default language was used. With the English model as the default, the accuracy was 51.51%, but when Spanish was the default, the accuracy was 49.16%.

In the second experiment, the authors built on the first experiment but used exceptions to short-circuit the algorithm and make a choice before relying on the probability scores. If either model tagged the word as a foreign word, the other language's tag was used. If either TreeTagger model failed to lemmatize the word, the other language's tag was used. If neither of these short-circuits applied, the tag was assigned the highest probability tag produced from the two language models. This reached 64.27% tagging accuracy.

A third experiment used language identification to select the appropriate tag of the word from the two tagging models. Using a custom lexicon received 86.03% accuracy. Using a 5-character n-gram language model, and selecting the language with the lowest perplexity, received 81.46%.[2] Select-

[1] A summary of Kohn's findings is available in English at http://arne-koehn.de/bachelorarbeit/.

[2] Though smoothing and back-off parameters are highly influential in short-string language identification, neither were presented in the paper.

Token	POS	Language
I	PP	ENG
am	VBP	ENG
afraid	JJ	ENG
that	CC	ENG
something	NN	ENG
will	MD	ENG
,	,	ENG
se	SE	SPA
va	VLfin	SPA
a	PREP	SPA
caer	VLinf	SPA
o	CC	SPA
ninguna	QU	SPA
cosa	NC	SPA
.	.	SPA
Lo	ART	SPA
ponemos	VLfin	SPA
ahí	ADV	SPA
UNK	UK	SPA
.	.	SPA

Table 2: Two sample sentences from the Spanish/English Conversation Corpus (Solorio and Liu 2008)

ing the POS tagging result using human-labeled language tags reached 89.72% accuracy.

In addition to rule-based classifiers, the authors experimented with machine learning. Using the gold-standard POS tag in their Spanish/English code-mixing dataset as their label, they used the Logit Boost algorithm to examine 21 feature set variations. The word, as well as the POS tags, POS tag probability, and lemma output from both the English and Spanish TreeTagger models served as available features. The classifiers were evaluated using 10-fold cross-validation on the constructed datasets. The top result combined the word, the two tags, the two probability scores, and the two lemmas. This reached 93.19% accuracy. Without the probability scores, they reached 92.95%, showing these scores were not particularly informative.

The previous Solorio and Liu (2008) experiments highlighted used POS tagging models that were pre-trained on large monolingual corpora and distributed with TreeTagger, but a final pilot experiment used a POS model that was trained on the Spanish/English code-mixing conversation corpus. The authors utilized a Conditional Random Field (CRF) tagger, and used as features the capitalization of the word, as well as the previous token.[3] This yielded 81% accuracy.

Spanish/English Conversation Corpus

We use the same corpus of Spanish/English conversation data as described in Solorio and Liu (2008). An example of this corpus is shown in Table 2.

[3]The authors did not mention if they used the current token as a feature.

Solorio and Liu (2008) provides some information about the corpus. The corpus is 39 minutes of transcribed bilingual Spanish/English conversation between 3 colleagues at a southwestern United States university. The transcribed corpus contains 922 sentences and 239 language switches, with 129 of the language switches being performed intrasententially. As far as we are aware, this is the largest annotated corpus of multilingual code-mixing.

Table 3 lists the tags used in the Spanish/English Conversation Corpus. Several anomalies are present in the corpus, and they appear to be annotation errors. For example tags in the set [DT, IN, JJ, NNS] each have a single occurrence in the Spanish utterances, but do not appear in the Spanish CRATER tagset.[4] Unfortunately, the standards documentation for this corpus could not be found. Without documentation, we made no effort to correct these POS anomalies in our corpus.

Experiment Setup

Preprocessing the Spanish/English Conversation Corpus

Several changes were made to the dataset to correct understood errors and better approximate input from an automatic speech recognition (ASR) system. Before correction, non-lexical utterances (e.g. *Mmm-hmm*), interjections (e.g. *Nn-hmm*, *eh*), unknown tokens (UNK), and punctuation (/[.,?!]—\.\.\./) were labeled with the language of the previous tagged language. We created a PUNC language for the punctuation, and removed these other tokens completely from our corpus. Additionally, we lowercased all tokens in the corpus. The removal of punctuation and the lowercasing of letters was performed in order to closer approximate ASR input to our system. Our tagging system used end of sentence punctuation for token series cut-off. We added a period (/[.]/) to separate one language chunk from another, and labeled this token with the POS PUNC.

In two instances, the corpus had a misspelled language. One string was not tokenized correctly, containing the end of one sentence and the beginning of another. One token was listed with two possibilities for the heard word, instead of using the UNK token. We corrected the typographical errors, tokenized the string and removed the unknown token.

Table 4 presents corpus information after preprocessing. Including punctuation, 5171 tokens remained in the English subset of the corpus, with 4488 of those consisting of words, and 916 of these words being unique word/POS pairs. Including punctuation, 1887 tokens remained in the Spanish subset, with 1537 of these being words, of which 477 were unique word/POS pairs. The corpus averaged 6.14 words per chunk, where the chunk is demarcated as either an end of sentence marker or language switch point. English chunks averaged 7.57 words, and Spanish chunks averaged 5.39 words. The corpus had 921 sentences, of which 252 had intrasentential code-mixing.

The example from Table 2 after processing can be found in Table 5.

[4]This tagset can be found across individual corpus links on *http://www.comp.lancs.ac.uk/linguistics/crater/spanishfiles.html*

POS Tag	English Count	Spanish Count
ADJ	1	48
ADV	1	157
ART		118
CARD		5
CC	199	88
CCAD		21
CCNEG		1
CD	76	
CQUE		42
CSUBF		1
CSUBI		6
CSUBX		24
DM		9
DT	369	1
EX	4	
IN	285	1
INT		11
ITJN	168	1
JJ	168	1
JJR	2	
MD	54	
NC	6	
NEG	16	48
NN	467	5
NNS	101	1
NP	121	27
NPS	1	
ORD		1
PAL		14
PDEL		4
PDT	11	
POS		13
PP	748	3
PPC	1	91
PPO		6
PPX		62
PP_s	49	
PREP	1	130
QU		15
QUE		1
RB	369	11
...	...	...

Table 3: POS tag frequency counts for English and Spanish.

Language	Tokens	Word Tokens	Unique Word Token-POS Pairs
English	5171	4488	916
Spanish	1887	1537	477

Table 4: Spanish/English Conversation Corpus after Preprocessing

Token	POS	Language
i	PP	ENG
am	VBP	ENG
afraid	JJ	ENG
that	CC	ENG
something	NN	ENG
will	MD	ENG
se	SE	SPA
va	VLfin	SPA
a	PREP	SPA
caer	VLinf	SPA
o	CC	SPA
ninguna	QU	SPA
cosa	NC	SPA
.	.	PUNC
lo	ART	SPA
ponemos	VLfin	SPA
ah	ADV	SPA
.	.	PUNC

Table 5: Two sample sentences from theSpanish/English Conversation Corpus (Solorio and Liu 2008), after preprocessing

SVMTool Configuration

The default window length in SVMTool is a 5-token n-gram, with the core token for calculations declared as the middle token. Typically, it is beneficial to disambiguate a token based on contextual information available on both sides of the token. Given that we are investigating a processing model with a streaming flow of data and immediate response, we did not desire the right context of the token. We adjusted this window to be a 3-token n-gram, with its core token for classification as the final word. We used the default model in SVMTool, referred to as Model 0, which is a one-pass model, with a left-to-right tagging direction.

SVMTool is configured with default features in Model 0 that would be uninformative to our classifier, given the nature of our dataset.

Due to our focus on an immediate incremental tagger, we removed any feature that referred to rightward tokens.

As a speech transcript, numbers were written out as words, so we removed the Boolean features *Does the word contain a number?* and *Does the word start with a number?*. Additionally, since we lowercased all text when we preprocessed the corpus, we could remove *Does the word start with an uppercase character?*, *Are all the words uppercase?*, *Are there any uppercase characters?*, and *Are there several uppercase characters?*. Since we removed all commas, we removed the feature *Does the word contain a comma?*.

The POS tagger uses two separate classifiers. One for known words, or words seen in the training data, and one for unknown words, or words not seen in the training data. The features we use for all experiments in this paper are listed in Table 6 (for known words) and Table 7 (for unknown words).

For known words, we extract a series of POS tag tuples that include all consecutive subsets of the POS tag and the

Features (known words)
Part of Speech n-grams:
$\{POS_{-2}\}$
$\{POS_{-1}\}$
$\{POS_0\}$
$\{POS_{-2}, POS_{-1}\}$
$\{POS_{-1}, POS_0\}$
$\{POS_{-2}, POS_{-1}, POS_0\}$
Ambiguity Class: k(0)
Maybe Tag Set: m(0)

Table 6: Features used for known words

Features (unknown words)
All features from Table 6
Prefixes, Character 4-gram: $\{s_1\}$
$\{s_1, s_2\}$
$\{s_1, s_2, s_3\}$
$\{s_1, s_2, s_3, s_4\}$
Suffixes, Character 4-gram: $\{s_1\}$
$\{s_1, s_2\}$
$\{s_1, s_2, s_3\}$
$\{s_1, s_2, s_3, s_4\}$
Integer: Word Length
Boolean: Does the word contain a period?
Boolean: Does the word contain a hyphen?

Table 7: Features used for unknown words.

two previous POS tags. Additionally, we store an ambiguity class (i.e., the list of all POS tags) for the current token, as well as the set of possible tags for the token.

For unknown words, we include the same POS n-gram features used to classify the known words, but include several additional features based on the characters in the word. Table 7 shows these features. A word's initial characters and final characters are represented with a 1, 2, 3, and 4-character tuple, to store possible prefixes and suffixes. A word's character length, and the Boolean functions *Does the word contain a period?* and *Does the word contain a hyphen?* are also included.

Evaluation Settings

We evaluate our system using 10-fold cross-validation on the Spanish/English code-mixing corpus. A model was trained on each of these folds, which created 10 tagging models for each evaluation. The results reported here are the averages of the 10 executions. When the tagging model is applied to the same language, folds were created by dividing the data for the language into 10 non-overlapping partitions. For each execution, 9 folds were used for training, and the remaining fold was used for testing. We also examine the accuracy of out-of-domain tagging models, where the model was trained on either English or Spanish and evaluated on the other. For these, 10 partitions of a random 90% of the language chunks were selected for training, and 10 partitions of the other language were cycled through for testing. When a Mixed language model was trained, and evaluated against a single language, a random 90% of the sentences in the Mixed corpus served as training data, while the remaining 10% were filtered, leaving only chunks from the language to be evaluated.

We tuned the 10 models used in each evaluation, by configuring SVMTool to search a parameter space for the optimal SVM soft margin parameter, *C*, for the known word and unknown word classifiers. SVMTool internally tested 3 ranges of *C* values, and within each of these ranges, up to 10 values for *C* were tested. The value for *C* incremented on a log scale with initial values for each search falling in the range [.01...1.0]. SVMTool performed an internal 10-fold cross validation on the training corpus and chose the optimal parameters.

Experiments

For each experiment, we report four numbers for accuracy. *Known* are tokens that have been seen in the training set. *Ambiguous Known* are those that have several possible POS tags occurring in the training set (this is a subset of *Known*). *Unknown* are those tokens that have not been seen in the training set. *Overall* is weighted score based on the number of words in *Known* and *Unknown*.

Baseline 1: English-Trained, Tested on English, Spanish, and Mixed Corpus

In order to establish a baseline for our experiments, we wanted to determine how well a model trained on the predominant language in the dataset could perform on the dataset.

We trained a model on the English subset, and tested it on the Spanish subset, as well as on the full Mixed corpus. On Spanish, we received an *Overall* accuracy of 24.11% , and on Mixed we received 77.27%. The results can be seen in Table 8.

Baseline 2: Bilingual POS Model Tagging Across Two Languages

A slightly more complex method for dealing with multilingual text is to train a multilingual POS tagging model, in which both languages are dealt with in one model.

For some cross-lingual tasks, we might want to have the same POS tagset for both languages. A tag represents a class of tokens that operate the same way syntactically. Labeling words across language barriers with the same tags would permit tools to understand which tokens behave syntactically similar. Petrov, Das, and McDonald (2011) introduced a coarse-grained POS tagset that is designed to be applied to all languages. This tagset consists of 12 POS tags, and would work well to equate parts of speech in a multilingual tagging situation. Without a need to switch between different tagsets, model selection is less necessary. Additionally, with such a small tagset, accuracy would be easier to achieve. However, many of the syntactic characteristics specific to individual languages are lost, which results in increased difficulty for subsequent analyses, such as parsing.

Model	Known	Ambig. Known	Unknown	Overall
English on Mixed	96.18%	87.88%	3.81%	77.27%
English on Spanish	80.02%	1.58%	56.76%	24.11%

Table 8: Performance using an English-trained model on the Spanish subset, as well as the unfiltered corpus.

In our setting, we have two possibilities to create a multilingual model: One possibility is to use a set of tags that contains both tagsets, i.e., it contains every tag in the Mixed corpus. As some of the POS tags are shared between the two languages (cf. Table 3), we do not retain the information to which language the word belongs. The other possibility is to explicitly mark the language on the tag with a prefix such as ENG or SPA. This would create two different tags for the shared tags.

Corpus-defined Tags To test an unmodified tagset, a monolithic heterogeneous-language tagging model was trained on the Mixed corpus without making a distinction in the model for the language of the text. Since some tags are used for both languages, the assigned POS may be ambiguous with regard to the language (e.g. CC, coordinating conjunction).

In the heterogeneous-language model, tokens and POS tags that overlap between the languages, cause the evaluation to be more permissive–the tagger would not need to predict the language in order to predict the tag, and the tag can occur in the context used by either language.

Language-distinct Part of Speech Tags Tag labels are chosen for mnemonic purposes, and are arbitrary to the computer. Though the overlapping tags between the two tagsets highlighted in Table 3 can be exploited for some tasks, for some other tasks the overlap could be harmful. The performance of the monolithic model can be hurt or aided by the precise strings chosen for each tag by the dataset creators.

The following experiment removes this overlap by replacing each language-tag pair in the dataset with a unique label, in our case consisting of the original POS tag in combination with a language label. An example can be found in Table 9.

Results The results for the tagset with the original and the language-distinct, unique tags can be found in Table 10. For the model using the original tagset, *Known* words were correctly tagged in 94.96% of the instances, ambiguous words were correctly tagged in 84.65% of the instances, and *Unknown* were correctly tagged 49.75% of the time. This led to a weighted *Overall* accuracy of 89.96%. For the experiment with the unique tags, we find that while *Known* and *Unknown* words maintained roughly the same accuracy, the accuracy on words that had multiple tag possibilities increased considerably, rising from 84.65% to 89.25%. As expected, the weighted average showed a slight decrease from 89.96% to 89.50% accuracy. This is due to the increased difficulty of the task since there are now more tags for the tagger to choose from.

Token	POS	Language
i	ENG-PP	ENG
am	ENG-VBP	ENG
afraid	ENG-JJ	ENG
that	ENG-CC	ENG
something	ENG-NN	ENG
will	ENG-MD	ENG
se	SPA-SE	SPA
va	SPA-VLfin	SPA
a	SPA-PREP	SPA
caer	SPA-VLinf	SPA
o	SPA-CC	SPA
ninguna	SPA-QU	SPA
cosa	SPA-NC	SPA
.	PUNC-.	PUNC
lo	SPA-ART	SPA
ponemos	SPA-VLfin	SPA
ah	SPA-ADV	SPA
.	PUNC-.	PUNC

Table 9: Two sample sentences from the Spanish/English Conversation Corpus (Solorio and Liu 2008) with unique POS tags

Separate homogeneous models: Do distinct tagging models perform better?

In the last experiment, we examined POS tagging with two languages stored in one model. This section utilizes two tagging models, and a language identification step that determines which model should process which string. For this paper, models were trained and evaluated using gold standard language identification tags found in the corrected corpus.

SVMTool requires a sentence terminating punctuation mark to reset the token sequence. We terminate mid-sentence language switch points by the addition of a period (e.g. Table 11), and analyze each chunk independently. This his may impair the selection of initial tokens in a language chunk. Results can be found in Table 12.

The results for the separate models show an increase for *Known* words, but also for *Unknown* Spanish words over the bilingual model. The *Overall* accuracy across the whole data set increased from 89.50% for the bilingual model with unique labels to 90.45%.

While it is difficult to compare these results to Solorio and Liu (2008) because of the differences in the POS tagger as well as the corrected data, our results compare favorably to the 89.72% accuracy they obtain using a non-incremental POS tagger and their gold standard language identification tags.

Model	Known	Ambiguous Known	Unknown	Overall
original labels	94.96%	84.65%	49.75%	89.96%
unique labels	94.42%	89.25%	50.06%	89.50%

Table 10: Performance using a single, multilingual model

	Token	POS	Language
English Chunk	i	PP	ENG
	am	VBP	ENG
	afraid	JJ	ENG
	that	CC	ENG
	something	NN	ENG
	will	MD	ENG
	.	.	PUNC
Spanish Chunk	se	SE	SPA
	va	VLfin	SPA
	a	PREP	SPA
	caer	VLinf	SPA
	o	CC	SPA
	ninguna	QU	SPA
	cosa	NC	SPA
	.	.	PUNC

Table 11: Sample sentences from the Spanish/English Conversation Corpus (Solorio and Liu 2008), after preprocessing and chunking

Model	Known	Ambig. Known	Unknown	Overall
English	95.08%	83.94%	49.57%	91.16%
Spanish	95.49%	84.86%	52.75%	88.50%
Weighted Average	-	-	-	90.45%

Table 12: Performance using two homogeneous models and model switching.

Conclusion

This paper discussed several experiments on Part of Speech (POS) tagging for bilingual speech transcription. We established a baseline using a model trained on the predominant language in the corpus. We performed two experiments using heterogeneous tagging models, and then showed how using an event controller to select a homogeneous-language model performs moderately better than the heterogeneous language model.

The predominant language in our dataset was English. We trained a model on the English chunks extracted from the dataset and evaluated the Mixed language transcription, establishing our baseline at 77.27% *Overall* accuracy.

We created a heterogeneous-language tagging model on the dataset across language boundaries, and reached 89.96% accuracy. We observed that there was a tagset overlap between the English and Spanish tagsets, and noted that this method would be useful only if one wished to exploit the common tags cross-lingually and if the necessary task does not require the language distinction of each word.

Reconfiguring the dataset to remove overlapping POS tags between the two languages caused a slight decline to 89.50%.

Switching between two homogeneous-language tagging models using an oracle language identification system as the indicator function allowed us to reach 90.45% accuracy. This is an improvement in *Overall* accuracy over the baseline (77.27%), as well as the heterogeneous-language tagging models (89.50%, 89.96%).

Future Work

Unfortunately, there are limited resources for code-mixing research. These experiments were performed on the only known annotated corpus of code-mixing. Should another POS-annotated corpus become available, it would be interesting to add more languages to the system. Similarly, we lack code-mixing treebanks. Output of this system concatenates the language fragments together to reform the sentence. Should a treebank of code-mixing become available, the concatenation of the different models would require much more complexity.

Acknowledgments

Segments of this article first appeared in the first author's dissertation (Rodrigues 2012). We would like to thank Markus Dickinson for commenting on the chapter, and C. Anton Rytting for reading a draft of this paper.

References

Beuck, N.; Köhn, A.; and Menzel, W. 2011. Decision strategies for incremental pos tagging. In *Proceedings of the 18th*

Nordic Conference of Computational Linguistics (NODALIDA 2011), volume 11 of *NEALT Proceedings Series*.

Brants, T. 2000. TnT: a statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, 224–231. Association for Computational Linguistics.

Charniak, E.; Blaheta, D.; Ge, N.; Hall, K.; Hale, J.; and Johnson, M. 2000. Bllip 1987-89 WSJ corpus release 1. Linguistic Data Consortium LDC2000T43, ISBN 1-58563-165-5.

Daumé III, H. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at http://pub.hal3.name#daume04cg-bfgs, implementation available at http://hal3.name/megam/.

Denis, P., and Sagot, B. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of The Pacific Asia Conference on Language, Information and Computation (PACLIC 23)*.

European Language Resources Association. 2000. *Lexesp Corpus ELRA-U-W 0052*.

Giménez, J., and Márquez, L. 2004. SVMTool: A general POS tagger generator based on support vector machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

Halácsy, P.; Kornai, A.; and Oravecz, C. 2007. Hunpos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 209–212. Association for Computational Linguistics.

Ivanova, S., and Kübler, S. 2008. Pos tagging for german: How important is the right context? In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Joachims, T. 1997. Text categorization with support vector machines: Learning with many relevant features. Technical Report LS VIII-Report, Universität Dortmund.

Köhn, A. 2009. Inkrementelle part-of-speech-tagger. Bachelors thesis.

Kübler, S.; Scheutz, M.; Baucom, E.; and Israel, R. 2010. Adding context information to part of speech tagging for dialogues. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*.

Petrov, S.; Das, D.; and McDonald, R. 2011. A universal part-of-speech tagset. ArXiv.

Rodrigues, P. 2012. *Processing Highly Variant Language Using Incremental Model Selection*. Ph.D. Dissertation, Indiana University.

Schmid, H. 2004. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*.

Solorio, T., and Liu, Y. 2008. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 973—981. ACL.