Extracting the Native Language Signal for Second Language Acquisition

Ben Swanson
Brown University
Providence, RI
chonger@cs.brown.edu

Eugene Charniak
Brown University
Providence, RI
ec@cs.brown.edu

Abstract

We develop a method for effective extraction of linguistic patterns that are differentially expressed based on the native language of the author. This method uses multiple corpora to allow for the removal of data set specific patterns, and addresses both feature relevancy and redundancy. We evaluate different relevancy ranking metrics and show that common measures of relevancy can be inappropriate for data with many rare features. Our feature set is a broad class of syntactic patterns, and to better capture the signal we extend the Bayesian Tree Substitution Grammar induction algorithm to a supervised mixture of latent grammars. We show that this extension can be used to extract a larger set of relevant features.

1 Introduction

Native Language Identification (NLI) is a classification task in which a statistical signal is exploited to determine an author's native language (L1) from their writing in a second language (L2). This academic exercise is often motivated not only by fraud detection or authorship attribution for which L1 can be an informative feature, but also by its potential to assist in Second Language Acquisition (SLA).

Our work focuses on the latter application and on the observation that the actual ability to automatically determine L1 from text is of limited utility in the SLA domain, where the native language of a student is either known or easily solicited. Instead, the likely role of NLP in the context of SLA is to provide a set of linguistic patterns that students with certain L1 backgrounds use with a markedly unusual frequency. Experiments have shown that such L1 specific information can be incorporated into lesson plans that improve student performance (Laufer and Girsai, 2008; Horst et al, 2008).

This is essentially a feature selection task with the additional caveat that features should be individually discriminative between native languages in order to facilitate the construction of focused educational excersizes. With this goal, we consider metrics for data set dependence, relevancy, and redundancy. We show that measures of relevancy based on mutual information can be inappropriate in problems such as ours where rare features are important.

While the majority of the methods that we consider generalize to any of the various feature sets employed in NLI, we focus on the use of Tree Substitution Grammar rules as features. Obtaining a compact feature set is possible with the well known Bayesian grammar induction algorithm (Cohn and Blunsom, 2010), but its rich get richer dynamics can make it difficult to find rare features. We extend the induction model to a supervised mixture of latent grammars and show how it can be used to incorporate linguistic knowledge and extract discriminative features more effectively.

The end result of this technique is a filtered list of patterns along with their usage statistics. This provides an enhanced resource for SLA research such as Jarvis and Crossley (2012) which tackles the manual connection of highly discriminative features with plausible linguistic transfer explanations. We output a compact list of language patterns that are empirically associated with native language labels, avoid-

Proceedings of NAACL-HLT 2013, pages 85–94,
Atlanta, Georgia, 9–14 June 2013.

ing redundancy and artifacts from the corpus creation process. We release this list for use by the linguistics and SLA research communities, and plan to expand it with upcoming releases of L1 labeled corpora[1].

2 Related Work

Our work is closely related to the recent surge of research in NLI. Beginning with Koppel et al (2005), several papers have proposed different feature sets to be used as predictors of L1 (Tsur and Rappaport, 2007; Wong and Dras, 2011a; Swanson and Charniak, 2012). However, due to the ubiquitous use of random subsamples, different data preparation methods, and severe topic and annotation biases of the data set employed, there is little consensus on which feature sets are ideal or sufficient, or if any reported accuracies reflect some generalizable truth of the problem's difficulty. To combat the bias of a single data set, a new strain of work has emerged in which train and test documents come from different corpora (Brooke and Hirst, 2012; Tetreault et al, 2012; Bykh and Meurers, 2012). We follow this cross corpus approach, as it is crucial to any claims of feature relevance.

Feature selection itself is a well studied problem, and the most thorough systems address both relevancy and redundancy. While some work tackles these problems by optimizing a metric over both simultaneously (Peng et al, 2005), we decouple the notions of relevancy and redundancy to allow ad-hoc metrics for either, similar to the method of Yu and Liu (2004). The measurement of feature relevancy in NLI has to this point been handled primarily with Information Gain, and elimination of feature redundancy has not been considered.

Tree Substitution Grammars have recently been successfully applied in several domains using the induction algorithm presented by Cohn and Blunsom (2010). Our hierarchical treatment builds on this work by incorporating supervised mixtures over latent grammars into this induction process. Latent mixture techniques for NLI have been explored with other feature types (Wong and Dras, 2011b; Wong and Dras, 2012), but have not previously led to measurable empirical gains.

[1] bllip.cs.brown.edu/download/nli_corpus.pdf

3 Corpus Description

We first make explicit our experimental setup in order to provide context for the discussion to follow. We perform analysis of English text from Chinese, German, Spanish, and Japanese L1 backgrounds drawn from four corpora. The first three consist of responses to essay prompts in educational settings, while the fourth is submitted by users in an internet forum.

The first corpus is the International Corpus of Learner English (ICLE) (Granger et al, 2002), a mainstay in NLI that has been shown to exhibit a large topic bias due to correlations between L1 and the essay prompts used (Brooke and Hirst, 2011). The second is the International Corpus of Crosslinguistic Interlanguage (ICCI) (Tono et al, 2012), which is annotated with sentence boundaries and has yet to be used in NLI. The third is the public sample of the Cambridge International Corpus (FCE), and consists of short prompted responses. One quirk of the FCE data is that several responses are written in the form of letters, leading to skewed distributions of the specialized syntax involved with use of the second person. The fourth is the Lang8 data set introduced by Brooke and Hirst (2011). This data set is free of format, with no prompts or constraints on writing aids. The samples are often very short and are qualitatively the most noisy of the four data sets.

One distinctive experimental decision is to treat each sentence as an individual datum. As document length can vary dramatically, especially across corpora, this gives increased regularity to the number of features per data item. More importantly, this creates a rough correspondence between feature co-occurrence and the expression of the same underlying linguistic phenomenon, which is desirable for automatic redundancy metrics.

We automatically detect sentence boundaries when they are not provided, and parse all corpora with the 6-split Berkeley Parser. As in previous NLI work, we then replace all word tokens that do not occur in a list of 614 common words with an unknown word symbol, UNK.

While these are standard data preprocessing steps, from our experience with this problem we propose additional practical considerations. First, we filter the parsed corpora, retaining only sentences that are

parsed to a Clause Level[2] tag. This is primarily due to the fact that automatic sentence boundary detectors must be used on the ICLE, Lang8, and FCE data sets, and false positives lead to sentence fragments that are parsed as NP, VP, FRAG, etc. The wild internet text found in the Lang8 data set also yields many non-Clause Level parses from non-English text or emotive punctuation. Sentence detection false negatives, on the other hand, lead to run-on sentences, and so we additionally remove sentences with more than 40 words.

We also impose a simple preprocessing step for better treatment of proper nouns. Due to the geographic distribution of languages, the proper nouns used in a writer's text naturally present a strong L1 signal. The obvious remedy is to replace all proper nouns with UNK, but this is unfortunately insufficient as the structure of the proper noun itself can be a covert signal of these geographical trends. To fix this, we also remove all proper noun left sisters of proper nouns. We choose to retain the rightmost sister node in order to preserve the plurality of the noun phrase, as the rightmost noun is most likely the lexical head.

From these parsed, UNKed, and filtered corpora we draw 2500 sentences from each L1 background at random, for a total of 10000 sentences per corpus. The exception is the FCE corpus, from which we draw 1500 sentences per L1 due to its small size.

4 Tree Substitution Grammars

A Tree Substitution Grammar (TSG) is a model of parse tree derivations that begins with a single ROOT nonterminal node and iteratively rewrites nonterminal leaves until none remain. A TSG rewrite rule is a tree of any depth, as illustrated in Figure 1, and can be used as a binary feature of a parsed sentence that is triggered if the rule appears in any derivation of that sentence.

Related NLI work compares a plethora of suggested feature sets, ranging from character n-grams to latent topic activations to labeled dependency arcs, but TSG rules are best able to represent complex lexical and syntactic behavior in a homogeneous feature type. This property is summed up nicely by the desire for features that capture rather

[2]S, SINV, SQ, SBAR, or SBARQ

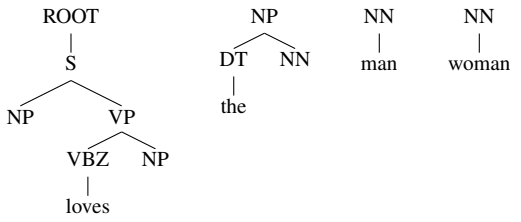

Figure 1: A Tree Substitution Grammar capable of describing the feelings of people of all sexual orientations.

than cover linguistic phenomena (Johnson, 2012); while features such as character n-grams, POS tag sequences, and CFG rules may provide a usable L1 signal, each feature is likely covering some component of a pattern instead of capturing it in full. TSG rules, on the other hand, offer remarkable flexibility in the patterns that they can represent, potentially capturing any contiguous parse tree structure.

As it is intractable to rank and filter the entire set of possible TSG rules given a corpus, we start with the large subset produced by Bayesian grammar induction. The most widely used algorithm for TSG induction uses a Dirichlet Process to choose a subset of frequently reoccurring rules by repeatedly sampling derivations for a corpus of parse trees (Cohn and Blunsom, 2010). The rich get richer dynamic of the DP leads to the use of a compact set of rules that is an effective feature set for NLI (Swanson and Charniak, 2012). However, this same property makes rare rules harder to find.

To address this weakness, we define a general model for TSG induction in labeled documents that combines a Hierarchical Dirichlet Process (Teh et al, 2005), with supervised labels in a manner similar to upstream supervised LDA (Mimno and McCallum, 2008). In the context of our work the document label η indicates both its authors native language L and data set D. Each η is associated with an observed Dirichlet prior ν_η, and a hidden multinomial θ_η over grammars is drawn from this prior. The traditional grammatical model of nonterminal expansion is augmented such that to rewrite a symbol we first choose a grammar from the document's θ_η and then choose a rule from that grammar.

For those unfamiliar with these models, the basic idea is to jointly estimate a mixture distribution over grammars for each η, as well as the parameters of these grammars. The HDP is necessary as the size

of each of these grammars is essentially infinite. We can express the generative model formally by defining the probability of a rule r expanding a symbol s in a sentence labeled η as

$$\begin{aligned}\theta_\eta &\sim Dir(\nu_\eta)\\ z_{i\eta} &\sim Mult(\theta_\eta)\\ H_s &\sim DP(\gamma, P_0(\bullet|s))\\ G_{ks} &\sim DP(\alpha_s, H_s)\\ r_{i\eta s} &\sim G_{z_{i\eta}s}\end{aligned}$$

This is closely related to the application of the Hierarchical Pitman Yor Process used in (Blunsom and Cohn, 2010) and (Shindo et al, 2012), which interpolates between multiple coarse and fine mappings of the data items being clustered to deal with sparse data. While the underlying Chinese Restaurant Process sampling algorithm is quite similar, our approach differs in that it models several different distributions with the same support that share a common prior.

By careful choice of the number of grammars K, the Dirichlet priors ν, and the backoff concentration parameter γ, a variety of interesting models can easily be defined, as demonstrated in our experiments.

5 Feature Selection

5.1 Dataset Independence

The first step in our L1 signal extraction pipeline controls for patterns that occur too frequently in certain combinations of native language and data set. Such patterns arise primarily from the reuse of essay prompts in the creation of certain corpora, and we construct a hard filter to exclude features of this type.

A simple first choice would be to rank the rules in order of dependence on the corpus, as we expect an irregularly represented topic to be confined to a single data set. However, this misses the subtle but important point that corpora have different qualities such as register and author proficiency. Instead we treat the set of sentences containing an arbitrary feature X as a set of observations of a pair of categorical random variables L and D, representing native language and data set respectively.

To see why this treatment is superior, consider the outcomes for the two hypothetical features shown in Figure 2. The left table has a high data set dependence but exhibits a clean twofold preference for L_1 in both data sets, making it a desirable feature to retain. Conversely, the right table shows a feature where the distribution is uniform over data sets, but has language preference in only one. This is a sign of either a large variance in usage or some data set specific tendency, and in either case we can not make confident claims as to this feature's association with any native language.

	L_1	L_2
D_1	1000	500
D_2	100	50

	L_1	L_2
D_1	1000	500
D_2	750	750

Figure 2: Two hypothetical feature profiles that illustrate the problems with filtering only on data set independence, which prefers the right profile over the left. Our method has the opposite preference.

The L-D dependence can be measured with Pearson's χ^2 test, although the specifics of its use as a filter deserve some discussion. As we eliminate the features for which the null hypothesis of independence is rejected, our noisy data will cause us to overzealously reject. In order to prevent the unneccesary removal of interesting patterns, we use a very small p value as a cutoff point for rejection. In all of our experiments the χ^2 value corresponding to $p < .001$ is in the twenties; we use $\chi^2 > 100$ as our criteria for rejection.

Another possible source of error is the sparsity of some features in our data. To avoid making predictions of rules for which we have not observed a sufficient number of examples, we automatically exclude any rule with a count less than five for any L-D combination η. This also satisfies the common requirements for validity of the χ^2 test that require a minimum number of 5 expected counts for every outcome.

5.2 Relevancy

We next rank the features in terms of their ability to discriminate between L1 labels. We consider three relevancy ranking metrics: Information Gain (IG), Symmetric Uncertainty (SU), and χ^2 statistic.

	IG	SU	χ^2
r	.84	.72	.15

Figure 3: Sample Pearson correlation coefficients between different ranking functions and feature frequency over a large set of TSG features.

$$IG(L, X_i) = H(L) - H(L|X_i)$$

$$SU(L, X_i) = 2\frac{IG(L, X_i)}{H(L) + H(X_i)}$$

$$\chi^2(X_i) = \sum_m \frac{(n_{im} - \frac{N_i}{M})^2}{\frac{N_i}{M}}$$

We define L as the Multinomial distributed L1 label taking values in $\{1, ..., M\}$ and X_i as a Bernoulli distributed indicator of the presence or absence of the ith feature, which we represent with the events X_i^+ and X_i^- respectively. We use the Maximum Likelihood estimates of these distributions from the training data to compute the necessary entropies for IG and SU. For the χ^2 metric we use n_{im}, the count of sentences with L1 label m that contain feature X_i, and their sum over classes N_i.

While SU is often preferred over IG in feature selection for several reasons, their main difference in the context of selection of binary features is the addition of $H(X_i)$ in the denominator, leading to higher values for rare features under SU. This helps to counteract a subtle preference for common features that these metrics can exhibit in data such as ours, as shown in Figure 3. The source of this preference is the overwhelming contribution of $p(X_i^-)H(L|X_i^-)$ in $IG(L, X_i)$ for rare features, which will be essentially the maximum value of $\log(M)$. In most classification problems a frequent feature bias is a desirable trait, as a rare feature is naturally less likely to appear and contribute to decision making.

We note that binary features in sentences are sparsely observed, as the opportunity for use of the majority of patterns will not exist in any given sentence. This leads to a large number of rare features that are nevertheless indicative of their author's L1. The χ^2 statistic we employ is better suited to retain such features as it only deals with counts of sentences containing X_i.

The ranking behavior of these metrics is highlighted in Figure 4. We expect that features with profiles like X_a and X_b will be more useful than those like X_d, and only χ^2 ranks these features accordingly. Another view of the difference between the metrics is taken in Figure 5. As shown in the left plot, IG and SU are nearly identical for the most highly ranked features and significantly different from χ^2.

	L_1	L_2	L_3	L_4	IG	SU	χ^2
X_a	20	5	5	5	**.0008**	**.0012**	**19.29**
X_b	40	20	20	20	.0005	.0008	**12.0**
X_c	2000	500	500	500	**.0178**	**.0217**	**385.7**
X_d	1700	1800	1700	1800	**.0010**	**.0010**	5.71

Figure 4: Four hypothetical features in a 4 label classification problem, with the number of training items from each class using the feature listed in the first four columns. The top three features under each ranking are shown in bold.

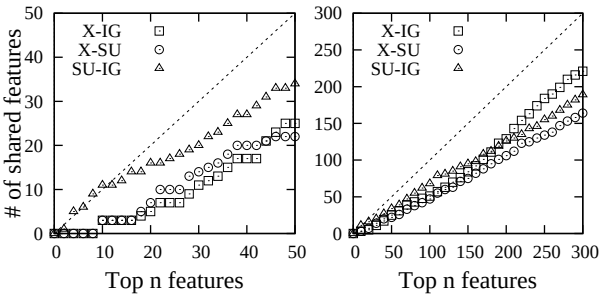

Figure 5: For all pairs of relevancy metrics, we show the number of features that appear in the top n of both. The result for low n is highlighted in the left plot, showing a high similarity between SU and IG.

5.3 Redundancy

The second component of thorough feature selection is the removal of redundant features. From an experimental point of view, it is inaccurate to compare feature selection systems under evaluation of the top n features or the number of features with ranking statistic at or beyond some threshold if redundancy has not been taken into account. Furthermore, as our stated goal is a list of discriminative patterns, multiple representations of the same pattern clearly

degrade the quality of our output. This is especially necessary when using TSG rules as features, as it is possible to define many slightly different rules that essentially represent the same linguistic act.

Redundancy detection must be able to both determine that a set of features are redundant and also select the feature to retain from such a set. We use a greedy method that allows us to investigate different relevancy metrics for selection of the representative feature for a redundant set (Yu and Liu, 2004). The algorithm begins with a list S containing the full list of features, sorted by an arbitrary metric of relevancy. While S is not empty, the most relevant feature X^* in S is selected for retention, and all features X_i are removed from S if $R(X^*, X_i) > \rho$ for some redundancy metric R and some threshold ρ.

We consider two probabilistic metrics for redundancy detection, the first being SU, as defined in the previous section. We contrast this metric with Normalized Pointwise Mutual Information (NPMI) which uses only the events $A = X_a^+$ and $B = X_b^+$ and has a range of [-1,1].

$$\text{NPMI}(X_a, X_b) = \frac{\log(P(A|B)) - \log(P(A))}{-\log(P(A,B))}$$

Another option that we explore is the structural redundancy between TSG rules themselves. We define a 0-1 redundancy metric such that $R(X_a, X_b)$ is one if there exists a fragment that contains both X_a and X_b with a total number of CFG rules less than the sum of the number of CFG rules in X_a and X_b. The latter constraint ensures that X_a and X_b overlap in the containing fragment. Note that this is not the same as a nonempty set intersection of CFG rules, as can be seen in Figure 6.

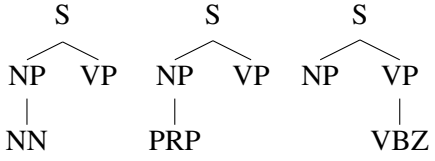

Figure 6: Three similar fragments that highlight the behavior of the structural redundancy metric; the first two fragments are not considered redundant, while the third is made redundant by either of the others.

6 Experiments

6.1 Relevancy Metrics

The traditional evaluation criterion for a feature selection system such as ours is classification accuracy or expected risk. However, as our desired output is not a set of features that capture a decision boundary as an ensemble, a per feature risk evaluation better quantifies the performance of a system for our purposes. We plot average risk against number of predicted features to view the rate of quality degradation under a relevancy measure to give a picture of a each metric's utility.

The per feature risk for a feature X is an evaluation of the ML estimate of $P_X(L) = P(L|X^+)$ from the training data on T_X, the test sentences that contain the feature X. The decision to evaluate only sentences in which the feature occurs removes an implicit bias towards more common features.

We calculate the expected risk $\mathcal{R}(X)$ using a 0-1 loss function, averaging over T_X.

$$\mathcal{R}(X) = \frac{1}{|T_X|} \sum_{t \in T_X} P_X(L \neq L_t^*)$$

where L_t^* is the gold standard L1 label of test item t. This metric has two important properties. First, given any true distribution over class labels in T_X, the best possible $P_X(L)$ is the one that matches these proportions exactly, ensuring that preferred features make generalizable predictions. Second, it assigns less risk to rules with lower entropy, as long as their predictions remain generalizable. This corresponds to features that find larger differences in usage frequency across L1 labels.

The alternative metric of per feature classification accuracy creates a one to one mapping between features and native languages. This unnecessarily penalizes features that are associated with multiple native languages, as well as features that are selectively dispreferred by certain L1 speakers. Also, we wish to correctly quantify the distribution of a feature over all native languages, which goes beyond correct prediction of the most probable.

Using cross validation with each corpus as a fold, we plot the average $\mathcal{R}(X)$ for the best n features against n for each relevancy metric in Figure 7. This clearly shows that for highly ranked features χ^2 is

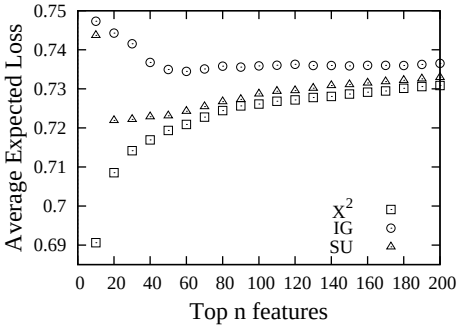

Figure 7: Per-feature Average Expected Loss plotted against top N features using χ^2, IG, and SU as a relevancy metric

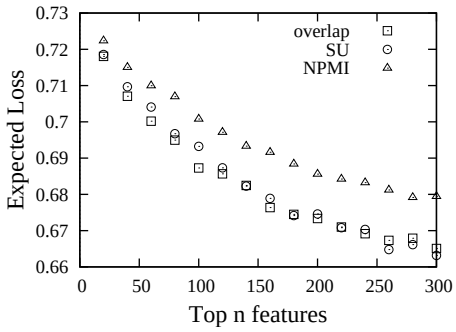

Figure 8: The effects of redundancy filtering on classification performance using different redundancy metrics. The cutoff values (ρ) used for SU and NPMI are .2 and .7 respectively.

able to best single out the type of features we desire. Another point to be taken from the plot is that it is that the top ten features under SU are remarkably inferior. Inspection of these rules reveals that they are precisely the type of overly frequent but only slightly discriminative features that we predicted would corrupt feature selection using IG based measures.

6.2 Redundancy Metrics

We evaluate the redundancy metrics by using the top n features retained by redundancy filtering for ensemble classification. Under this evaluation, if redundancy is not being effectively eliminated performance should increase more slowly with n as the set of test items that can be correctly classified remains relatively constant. Additionally, if the metric is overzealous in its elimination of redundancy, useful patterns will be eliminated leading to diminished increase in performance. Figure 8 shows the tradeoff between Expected Loss on the test set and the number of features used with SU, NPMI, and the overlap based structural redundancy metric described above. We performed a coarse grid search to find the optimal values of ρ for SU and NPMI.

Both the structural overlap hueristic and SU perform similarly, and outperform NPMI. Analysis reveals that NPMI seems to overstate the similarity of large fragments with their small subcomponents. We choose to proceed with SU, as it is not only faster in our implementation but also can generalize to feature types beyond TSG rules.

6.3 TSG Induction

We demonstrate the flexibility and effectiveness of our general model of mixtures of TSGs for labeled data by example. The tunable parameters are the number of grammars K, the Dirichlet priors ν_η over grammar distributions for each label η, and the concentration parameter γ of the smoothing DP.

For a first baseline we set the number of grammars $K = 1$, making the Dirichlet priors ν irrelevant. With a large $\gamma = 10^{20}$, we essentially recover the basic block sampling algorithm of Cohn and Blunsom (2010). We refer to this model as M1. Our second baseline model, M2, sets K to the number of native language labels, and sets the ν variables such that each η is mapped to a single grammar by its L1 label, creating a naive Bayes model. For M2 and the subsequent models we use $\gamma = 1000$ to allow moderate smoothing.

We also construct a model (M3) in which we set $K = 9$ and ν_η is such that three grammars are likely for any single η; one shared by all η with the same L1 label, one shared by all η with the same corpus label, and one shared by all η. We compare this with another $K = 9$ model (M4) where the ν are set to be uniform across all 9 grammars.

We evaluate these systems on the percent of their resulting grammar that rejects the hypothesis of language independence using a χ^2 test. Slight adjustments were made to α for these models to bring their output grammar size into the range of approximately 12000 rules. We average our results for each model over single states drawn from five indepen-

	$p < .1$	$p < .05$	$p < .01$	$p < .001$
M1	56.5(3.1)	54.5(3.0)	49.8(2.7)	45.1(2.5)
M2	55.3(3.7)	53.7(3.6)	49.1(3.3)	44.7(3.0)
M3	**59.0**(**4.1**)	**57.2**(**4.1**)	**52.4**(**3.6**)	**48.4**(**3.3**)
M4	58.9(3.8)	57.0(3.7)	51.9(3.4)	47.2(3.1)

Figure 9: The percentage of rules from each model that reject L1 independence at varying levels of statistical significance. The first number is with respect to the number rules that pass the L1/corpus independence and redundancy tests, and the second is in proportion to the full list returned by grammar induction.

dent Markov chains.

Our results in Figure 9 show that using a mixture of grammars allows the induction algorithm to find more patterns that fit arbitrary criteria for language dependence. The intuition supporting this is that in simpler models a given grammar must represent a larger amount of data that is better represented with more vague, general purpose rules. Dividing the responsibility among several grammars lets rare patterns form clusters more easily. The incorporation of informed structure in M3 further improves the performance of this latent mixture technique.

7 Discussion

Using these methods, we produce a list of L1 associated TSG rules that we release for public use. We perform grammar induction using model M3, apply our data dependence and redundancy filters, rank for relevancy using χ^2 and filter at the level of $p < .1$ statistical significance for relevancy. Each entry consists of a TSG rule and its matrix of counts with each η. We provide the total for each L1 label, which shows the overall prediction of the proportional use of that item. We also provide the χ^2 statistics for L1 dependence and the dependence of L1 and corpus.

It is speculative to assign causes to the discriminative rules we report, and we leave quantification of such statements to future work. However, the strength of the signal, as evidenced by actual counts in data, and the high level interpretation that can be easily assigned to the TSG rules is promising. As understanding the features requires basic knowledge of Treebank symbols, we provide our interpretations for some of the more interesting rules and summarize their L1 distributions. Note that by describing a rule as being preferred by a certain set of L1 labels, our claim is relative to the other labels only; the true cause could also be a dispreference in the complement of this set.

One interesting comparison made easy by our method is the identification of similar structures that have complementary L1 usage. An example is the use of a prepositional phrase just before the first noun phrase in a sentence, which is preferred in German and Spanish, especially in the former. However, German speakers disprefer a prepositional phrase followed by a comma at the beginning of the sentence, and Chinese speakers use this pattern more frequently than the other L1s. Another contrastable pair is the use of the word "because" with upper or lower case, signifying sentence initial or medial use. The former is preferred in Chinese and Japanese text, while the latter is preferred in German and even more so in Spanish L1 data.

As these examples suggest, the data shows a strong division of preference between European and Asian languages, but many patterns exist that are uniquely preferred in single languages as well. Japanese speakers are seen to frequently use a personal pronoun as the subject of the sentence, while Spanish speakers use the phrase "the X of Y", the verb "go", and the determiner "this" with markedly higher frequency. Germans tend to begin sentences with adverbs, and various modal verb constructions are popular with Chinese speakers. We suspect these patterns to be evidence of preference in the specified language, rather than dispreference in the other three.

Our strategy in regard to the hard filters for L1-corpus dependence and redundancy has been to prefer recall to precision, as false positives can be easily ignored through subsequent inspection of the data we supply. This makes the list suitable for human qualitative analysis, but further work is required for its use in downstream automatic systems.

8 Conclusion

This work contributes to the goal of leveraging NLI data in SLA applications. We provide evidence for

our hypothesis that relevancy metrics based on mutual information are ill-suited for this task, and recommend the use of the χ^2 statistic for rejecting the hypothesis of language independence. Explicit controls for dependence between L1 and corpus are proposed, and redundancy between features are addressed as well. We argue for the use of TSG rules as features, and develop an induction algorithm that is a supervised mixture of hierarchical grammars. This generalizable formalism is used to capture linguistic assumptions about the data and increase the amount of relevant features extracted at several thresholds.

This project motivates continued incorporation of more data and induction of TSGs over these larger data sets. This will improve the quality and scope of the resulting list of discriminative syntax, allowing broader use in linguistics and SLA research. The prospect of high precision and recall in the extraction of such patterns suggests several interesting avenues for future work, such as determination of the actual language transfer phenomena evidenced by an arbitrary count profile. To achieve the goal of automatic detection of plausible transfer the native languages themselves must be considered, as well as a way to distinguish between preference and dispreference based on usage statistics. Another exciting application of such a refined list of patterns is the automatic integration of its features in L1 targeted SLA software.

References

Phil Blunsom and Trevor Cohn. 2010. Unsupervised Induction of Tree Substitution Grammars for Dependency Parsing. *Empirical Methods in Natural Language Processing*.

Julian Brooke and Graeme Hirst. 2011. Native language detection with 'cheap' learner corpora. *Conference of Learner Corpus Research*.

Julian Brooke and Graeme Hirst. 2012. Measuring Interlanguage: Native Language Identification with L1-influence Metrics. *LREC*

Julian Brooke and Graeme Hirst. 2012. Robust, Lexicalized Native Language Identification. *COLING*.

Serhiy Bykh and Detmar Meurers. 2012. Native Language Identification Using Recurring N-grams - Investigating Abstraction and Domain Dependence. *COLING*.

Trevor Cohn, Sharon Goldwater, and Phil Blunsom. 2009. Inducing Compact but Accurate Tree-Substitution Grammars. In *Proceedings NAACL*.

Trevor Cohn, and Phil Blunsom. 2010. Blocked inference in Bayesian tree substitution grammars. *Association for Computational Linguistics*.

Gilquin, Gaëtanelle and Granger, Sylviane. 2011. From EFL to ESL: Evidence from the International Corpus of Learner English. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap (Book Chapter)*.

Joshua Goodman. 2003. Efficient parsing of DOP with PCFG-reductions. *In Bod et al. chapter 8.*.

S. Granger, E. Dagneaux and F. Meunier. 2002. *International Corpus of Learner English*, (ICLE).

Horst M., White J., Bell P. 2010. First and second language knowledge in the language classroom. *International Journal of Bilingualism*.

Scott Jarvis and Scott Crossley 2012. Approaching Language Transfer through Text Classification.

Mark Johnson 2011. How relevant is linguistics to computational linguistics?. *Linguistic Issues in Language Technology*.

Ekaterina Kochmar. 2011. Identification of a writer's native language by error analysis. *Master's Thesis*.

Koppel, Moshe and Schler, Jonathan and Zigdon, Kfir. 2005. Determining an author's native language by mining a text for errors. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*.

Laufer, B and Girsai, N. 2008. Form-focused Instruction in Second Language Vocabulary Learning: A Case for Contrastive Analysis and Translation. *Applied Linguistics*.

David Mimno and Andrew McCallum. 2008. Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression. *UAI*.

Hanchuan Peng and Fuhui Long and Chris Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. *Association for Computational Linguistics*.

Matt Post and Daniel Gildea. 2009. Bayesian Learning of a Tree Substitution Grammar. *Association for Computational Linguistics*.

Tono, Y., Kawaguchi, Y. & Minegishi, M. (eds.) . 2012. *Developmental and Cross-linguistic Perspectives in Learner Corpus Research.*.

Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. *CACLA*.

Shindo, Hiroyuki and Miyao, Yusuke and Fujino, Akinori and Nagata, Masaaki 2012. Bayesian Symbol-Refined Tree Substitution Grammars for Syntactic Parsing. *Association for Computational Linguistics*.

Ben Swanson and Eugene Charniak. 2012. Native Language Detection with Tree Substitution Grammars. *Association for Computational Linguistics*.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2005. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*.

Joel Tetreault, Daniel Blanchard, Aoife Cahill, Beata Beigman-Klebanov and Martin Chodorow. 2012. Native Tongues, Lost and Found: Resources and Empirical Evaluations in Native Language Identification. *COLING*.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. *Proceedings of the Australasian Language Technology Association Workshop*.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting Parse Structures for Native Language Identification. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Sze-Meng Jojo Wong and Mark Dras. 2011. Topic Modeling for Native Language Identification. *Proceedings of the Australasian Language Technology Association Workshop*.

Sze-Meng Jojo Wong, Mark Dras, Mark Johnson. 2012. Exploring Adaptor Grammars for Native Language Identification. *EMNLP-CoNLL*.

Lei Yu and Huan Liu. 2004. Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research*.