Data Driven Language Transfer Hypotheses

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

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

Abstract

Language transfer, the preferential second language behavior caused by similarities to the speaker's native language, requires considerable expertise to be detected by humans alone. Our goal in this work is to replace expert intervention by data-driven methods wherever possible. We define a computational methodology that produces a concise list of lexicalized syntactic patterns that are controlled for redundancy and ranked by relevancy to language transfer. We demonstrate the ability of our methodology to detect hundreds of such candidate patterns from currently available data sources, and validate the quality of the proposed patterns through classification experiments.

1 Introduction

The fact that students with different native language backgrounds express themselves differently in second language writing samples has been established experimentally many times over (Tetreault et al., 2013), and is intuitive to most people with experience learning a new language. The exposure and understanding of this process could potentially enable the creation of second language (L2) instruction that is tailored to the native language (L1) of students.

The detectable connection between L1 and L2 text comes from a range of sources. On one end of the spectrum are factors such as geographic or cultural preference in word choice, which are a powerful L1 indicator. On the other end lie linguistic phenomena such as language transfer, in which the preferential over-use or under-use of structures in the L1 is reflected in the use of corresponding patterns in the L2. We focus on language transfer in this work, based on our opinion that such effects are more deeply connected to and effectively utilized in language education.

The inherent challenge is that viable language transfer hypotheses are naturally difficult to construct. By the requirement of contrasting different L1 groups, hypothesis formulation requires deep knowledge of multiple languages, an ability reserved primarily for highly trained academic linguists. Furthermore, the sparsity of any particular language pattern in a large corpus makes it difficult even for a capable multilingual scholar to detect the few patterns that evidence language transfer. This motivates data driven methods for hypothesis formulation.

We approach this as a representational problem, requiring the careful definition of a class of linguistic features whose usage frequency can be determined for each L1 background in both L1 and L2 text (e.g. both German and English written by Germans). We claim that a feature exhibiting a sufficiently non-uniform usage histogram in L1 that is mirrored in L2 data is a strong language transfer candidate, and provide a quantified measure of this property.

We represent both L1 and L2 sentences in a universal constituent-style syntactic format and model language transfer hypotheses with contiguous syntax sub-structures commonly known as Tree Substitution Grammar (TSG) fragments (Post and Gildea, 2009)(Cohn and Blunsom, 2010). With these features we produce a concise ranked list of candidate language transfer hypotheses and their usage statistics that can be automatically augmented as increasing amounts of data become available.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 169–173,
Gothenburg, Sweden, April 26-30 2014.

2 Related Work

This work leverages several recently released data sets and analysis techniques, with the primary contribution being the transformations necessary to combine these disparate efforts. Our analysis methods are closely tied to those described in Swanson and Charniak (2013), which contrasts techniques for the discovery of discriminative TSG fragments in L2 text. We modify and extend these methods to apply to the universal dependency treebanks of McDonald et al. (2013), which we will refer to below to as the UTB. Bilingual lexicon construction (Haghighi et al., 2008) is also a key component, although previous work has focused primarily on nouns while we focus on stopwords. We also transform the UTB into constituent format, in a manner inspired by Carroll and Charniak (1992).

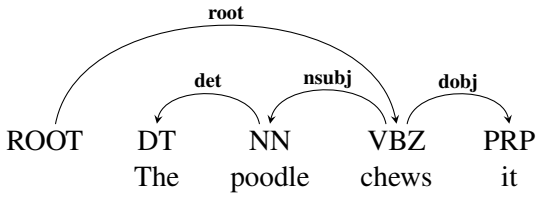
There is a large amount of related research in Native Language Identification (NLI), the task of predicting L1 given L2 text. This work has culminated in a well attended shared task (Tetreault et al., 2013), whose cited report contains an excellent survey of the history of this task. In NLI, however, L1 data is not traditionally used, and patterns are learned directly from L2 text that has been annotated with L1 labels. One notable outlier is Brooke and Hirst (2012), which attempts NLI using only L1 data for training using large online dictionaries to tie L2 English bigrams and collocations to possible direct translations from native languages. Jarvis and Crossley (2012) presents another set of studies that use NLI as a method to form language transfer hypotheses.

3 Methodology

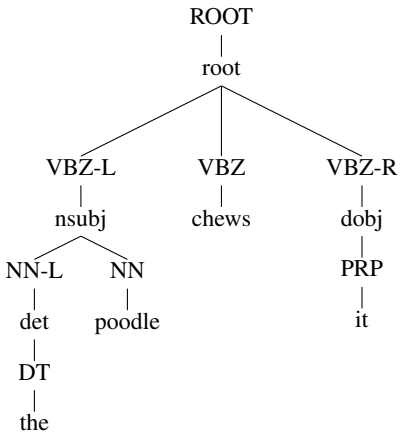
The first of the four basic requirements of our proposed method is the definition of a class of features $\mathcal{F}$ such that a single feature $F \in \mathcal{F}$ is capable of capturing language transfer phenomenon. The second is a universal representation of both L1 and L2 data that allows us to count the occurrences of any F in an arbitrary sentence. Third, as any sufficiently expressive $\mathcal{F}$ is likely to be very large, a method is required to propose an initial candidate list $C \subset \mathcal{F}$. Finally, we refine C into a ranked list H of language transfer hypotheses, where H has also been filtered to remove redundancy.

In this work we define $\mathcal{F}$ to be the set of Tree Substitution Grammar (TSG) fragments in our data, which allows any connected syntactic structure to be used as a feature. As such, our universal representation of L1/L2 data must be a constituent tree structure of the general form used in syntactic parsing experiments on the Penn Treebank. The UTB gets us most of the way to our goal, defining a dependency grammar with a universal set of part of speech (POS) tags and dependency arc labels.

Two barriers remain to the use of standard TSG induction algorithms. The first is to define a mapping from the dependency tree format to constituency format. We use the following dependency tree to illustrate our transformation.



Under our transformation, the above dependency parse becomes



We also require a multilingual lexicon in the form of a function $M_L(w)$ for each language L that maps words to clusters representing their meaning. In order to avoid cultural cues and reduce noise in our mapping, we restrict ourselves to clusters that correspond to a list of L2 stopwords. Any L2 words that do not appear on this list are mapped to the unknown "UNK" symbol, as are all foreign words that are not good translations of any L2 stopword. Multiple words from a single language can map to the same cluster, and it is worth noting that this is true for L2 stopwords as well.

To determine the mapping functions M_L we train IBM translation models in both directions between the L2 and each L1. We create a graph in which nodes are words, either the L2 stopwords or any L1 word with some translation probability to

or from one of the L2 stopwords. The edges in this graph exist only between L2 and L1 words, and are directed with weight equal to the IBM model's translation probability of the edge's target given its source. We construct M_L by removing edges with weight below some threshold and calculating the connected components of the resulting graph. We then discard any cluster that does not contain at least one word from each L1 and at least one L2 stopword.

To propose a candidate list C, we use the TSG induction technique described in Swanson and Charniak (2013), which simultaneously induces multiple TSGs from data that has been partitioned into labeled types. This method permits linguistically motivated constraints as to which grammars produce each type of data. For an experimental setup that considers n different L1s, we use $2n+1$ data types; Figure 1 shows the exact layout used in our experiments. Besides the necessary n data types for each L1 in its actual native language form and n in L2 form, we also include L2 data from L2 native speakers. We also define $2n + 1$ grammars. We begin with n grammars that can each be used exclusively by one native language data type, representing behavior that is unique to each native language (grammars A-C in Figure 1) . This is done for the L2 as well (grammar G). Finally, we create an interlanguage grammar for each of our L1 types that can be used in derivation of both L1 and L2 data produced by speakers of that L1 (grammars D-F).

The final step is to filter and rank the TSG fragments produced in C, where filtering removes redundant features and ranking provides some quantification of our confidence in a feature as a language transfer hypothesis. Swanson and Charniak (2013) provides a similar method for pure L2 data, which we modify for our purposes. For redundancy filtering no change is necessary, and we use their recommended Symmetric Uncertainty method. For a ranking metric of how well a fragment fits the profile of language transfer we adopt the expected per feature loss (or risk) also described in their work. For an arbitrary feature F, this is defined as

$$\mathcal{R}(F) = \frac{1}{|T_F|} \sum_{t \in T_F} P_F(L \neq L_t^*)$$

where T_F is the subset of the test data that contains the feature F, and L_t^* is the gold label of test datum t. While in their work the predictive distribution $P_F(L)$ is determined by the observed counts of F in L2 training data, we take our estimates directly from the L1 data of the languages under study. This metric captures the extent to which the knowledge of a feature F's L1 usage can be used to predict its usage in L2.

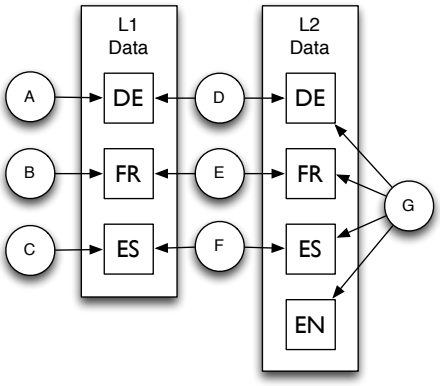


Figure 1: The multi-grammar induction setup used in our experiments. Squares indicate data types, and circles indicate grammars. Data type labels indicate the native language of the speaker, and all L2 data is in English.

The final result is a ranked and filtered list of hypotheses H. The elements of H can be subjected to further investigation by experts and the accompanying histogram of counts contains the relevant empirical evidence. As more data is added, the uncertainty in the relative proportions of these histograms and their corresponding $\mathcal{R}$ is decreased. One additional benefit of our method is that TSG induction is a random process, and repeated runs of the sampling algorithm can produce different features. Since redundancy is filtered automatically, these different feature lists can be combined and processed to potentially find additional features given more computing time.

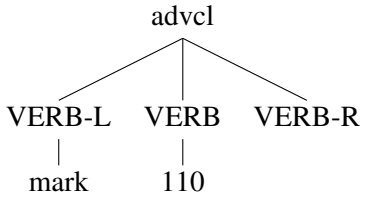
4 Results

Limited by the intersection of languages across data sets, we take French, Spanish, and German as our set of L1s with English as the L2. We use the UTB for our native language data, which provides around 4000 sentences of human annotated text for each L1. For our L2 data we use the ETS Corpus of Non-Native English (Blanchard et al., 2013), which consists of over 10K sentences per L1 label drawn from TOEFL® exam essays. Fi-

nally, we use the Penn Treebank as our source of native English data, for a total of seven data types; four in English, and one in each L1.

When calculating metrics such as redundancy and $\mathcal{R}(F)$ we use all available data. For TSG sampling, we balance our data sets to 4000 sentences from each data type and sample using the Enbuske sampler that was released with Swanson and Charniak (2013). To construct word clusters, we use Giza++ (Och and Ney, 2003) and train on the Europarl data set (Koehn, 2005), using .25 as a threshold for construction on connected components.

We encourage the reader to peruse the full list of results[1], in which each item contains the information in the following example.



	ES	DE	FR
L1	4.2	0.0	0.0
L2	2.3	0.3	0.3

This fragment corresponds to an adverbial clause whose head is a verb in the cluster 110, which contains the English word "is" and its various translations. This verb has a single left dependent, a clause marker such as "because", and at least one right dependent. Its prevalence in Spanish can explained by examining the translations of the English sentence "I like it because it is red".

ES Me gusta porque es rojo.
DE Ich mag es, weil es rot ist.
FR Je l'aime parce qu'il est rouge.

Only in the Spanish sentence is the last pronoun dropped, as in "I like it because is red". This observation, along with the L1/L2 profile which shows the count per thousand sentences in each language provides a strong argument that this pattern is indeed a form of language transfer.

Given our setup of three native languages, a feature with $\mathcal{R}(F) < .66$ is a candidate for language transfer. However, several members of our filtered list have $\mathcal{R}(F) > .66$, which is to say that their

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

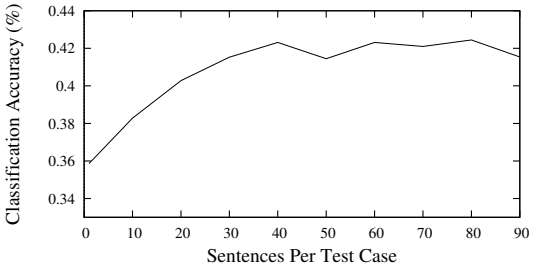


Figure 2: Creating test cases that consist of several sentences mediates feature sparsity, providing clear evidence for the discriminative power of the chosen feature set.

L2 usage does not mirror L1 usage. This is to be expected in some cases due to noise, but it raises the concern that our features with $\mathcal{R}(F) < .66$ are also the result of noise in the data. To address this, we apply our features to the task of cross language NLI using only L1 data for training. If the variation of $\mathcal{R}(F)$ around chance is simply due to noise then we would expect near chance (33%) classification accuracy. The leftmost point in Figure 2 shows the initial result, using boolean features in a log-linear classification model, where a test case involves guessing an L1 label for each individual sentence in the L2 corpus. While the accuracy does exceed chance, the margin is not very large.

One possible explanation for this small margin is that the language transfer signal is sparse, as it is likely that language transfer can only be used to correctly label a subset of L2 data. We test this by combining randomly sampled L2 sentences with the same L1 label, as shown along the horizontal axis of Figure 2. As the number of sentences used to create each test case is increased, we see an increase in accuracy that supports the argument for sparsity; if the features were simply weak predictors, this curve would be flat. The resulting margin is much larger, providing evidence that a significant portion of our features with $\mathcal{R}(F) < .66$ are not selected due to random noise in $\mathcal{R}$ and are indeed connected to language transfer.

The number and strength of these hypotheses is easily augmented with more data, as is the number of languages under consideration. Our results also motivate future work towards automatic generation of L1 targeted language education exercises, and the fact that TSG fragments are a component of a well studied generative language model makes them well suited to such generation tasks.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. Technical report, Educational Testing Service.

Julian Brooke and Graeme Hirst. 2012. Measuring Interlanguage: Native Language Identification with L1-influence Metrics. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 779–784, Istanbul, Turkey, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1016.

Glenn Carroll and Eugene Charniak. 1992. Two experiments on learning probabilistic dependency grammars from corpora. Technical Report CS-92-16, Brown University, Providence, RI, USA.

Trevor Cohn and Phil Blunsom. 2010. Blocked inference in bayesian tree substitution grammars. pages 225–230. Association for Computational Linguistics.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*, pages 771–779.

Scott Jarvis and Scott Crossley, editors. 2012. *Approaching Language Transfer Through Text Classification: Explorations in the Detection-based Approach*, volume 64. Multilingual Matters Limited, Bristol, UK.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit*.

Ryan T. McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *ACL (2)*, pages 92–97.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Matt Post and Daniel Gildea. 2009. Bayesian learning of a tree substitution grammar. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 45–48. Association for Computational Linguistics.

Ben Swanson and Eugene Charniak. 2013. Extracting the native language signal for second language acquisition. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–94, Atlanta, Georgia, June. Association for Computational Linguistics.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, GA, USA, June. Association for Computational Linguistics.