Automatically evaluating L2 proficiency level to study low dimensional features — the case of “true cognates“ transfer

September 10, 2020

1 Motivation and background

Non-native speakers transfer from their mother tongue in many ways. One form of lexical transfer relates to cognates between the native and the learned language. While much attention has been paid to false cognates [REFERENCES], Rabinovich et al. (2018) have recently shown that this holds for so-called “true cognates“ as well. When faced with synonymous options that include both cognates and non-cognates, non-native speakers are more likely to use cognate words. We hypothesize that this preference is correlated with proficiency: the more proficient a speaker is, the more likely she is to go beyond this preference and select the synonymous alternative that best servers the usage pattern in context, cognate or not.

2 methodology

2.1 Some methodological hurdles

The way for testing this hypothesis is paved with at least three methodological hurdles. Although some learner corpora are annotated for level, the production settings are contrived, the size of the corpora is limited, and most importantly, they rarely reflect the level of highly proficient speakers. On the other hand, large corpora crawled from the web, representing multiple levels of speakers from a wide range of native languages (as the one reported in (Goldin et al., 2018), are not annotated for proficiency level.

The second hurdle pertains to text classification. The classical learning paradigm provides a robust tool to engineer features and to check whether they can assist us in discriminating native from non-native speakers [REFERENCES] or to use these features to tease apart classes of non-native speakers [REFERENCES]. Classifiers, however, provide a dichotomous answer: either an instance *a* is a member of class *A*, namely a true positive, or an instance *a* belongs in class *B* (or *C* or *D*, etc), namely a false positive. The third hurdle is that rare phenomena, such as cognates, yield low dimensional feature vectors when representing texts, and hence the ability to use them for classification is scant.

2.2 Breaking up dichotomous outputs to a continuum to gauge proficiency

When using linear kernel models, one can use the separating hyperplane h drawn between the classes to measure the Euclidean Distance between h and the individual instances. Each instance is classified according to its position relative to the separating hyperplane. We use logistic expression, since the sigmoid function has a nice property such that feeding it with the values of the Euclidean Distance, it outputs the confidence of the model that an instance in class.

More formally, a feature vector representing an instance $y^{(i)}$ is expressed by the distance of this vector from the the separating hyperplane using $(d(h, y^{(i)}))$. To gauge the proficiency, we use a binary classifier of native vs. non-native speakers, and each instance is just the feature vector of an individual author. An author that resides near the hyperplane, is hard for the machine to classify; in other words, it is closer to the class from which it has to be separated. Consequently, we expect this author to have a higher proficiency level. We no longer use the dichotomous output of the classifier, and render its either-or output to a continuum that will hopefully reflect the proficiency level of the writers.

2.3 Proof of concept

Recall that we are interested in using corpora that are not annotated for level. In order to check our methodology, however, we have to first prove its validity on a corpus that *is* annotated for proficiency. Additionally, we want to be sure that the variance across the non-native speakers is fine enough that it will capture subtle differences with low dimensionality.

We go then to study an assortment of commonly accepted measures of proficiency, both lexical and syntactic, and how they correlate with the automatically detected proficiency. Although some of these measures are much debated (Anat??), we expect at least some of them to correlate with the confidence probability of the various authors.

3 Datasets and experimental setups

3.1 Datasets

We first use The TOEFL Corpus (Tetreault et al., 2013) — a dataset of essays written by non-native English speakers wishing to enroll in English speaking universities. The essays were evaluated for proficiency by highly skilled humans, each given a grade in the ranges *low*, *medium* or *high*.

The corpus consists of $12,100$ essays written by native speakers of 11 native languages: 1330 graded low, 6568 graded $medium$ and the remaining 4202 are graded with $high$. Other metadata fields include the author's L1, and the prompt question for the essay.

The comparative native-speaker corpus employed is Louvain Corpus of Native English Essays (LOCNESS). It was compiled at the University of Louvain la Neuve in Belgium. We use a subset of 412 essays written by A-level native English speakers from British and American universities.

Both corpora include confounding variables which are of no interest for our study. The LOCNESS corpus consists of two English varieties, British and American. Additionally, the essays in this corpus are much longer and essay length is a good predictor of level [REFERENCE, Shuly?]. The TOEFL corpus consists of writers from various linguistic and cultural backgrounds, while the only variable that matters for this study proficiency level.

To overcome these confounders, we shuflled all the sentences in LOCNESS and all the sentences in each level in TOEFL, and generated artificial chunks of shuffled sentences of 1000 words each, respecting sentence boundaries.

The resulting non-native dataset consists of 4145 chunks: 1661 — *high*, 2194 — *medium*, 299 — low. The native speaker's corpus was assumed to be homogeneous in terms of proficiency, at least when compared to non-native speakers. We generated 1000-word chunks as well. The resulting data comprises 355 chunks.

We balance our corpus by randomly down-sampling the non-native group to match the native group size. We performed this process 10 times to eliminate bias that might be a result originating in a specific sample.

3.2 Feature selection

Our goal is to use classification in order to provide each instance with a proficiency level index, and use this index to study sparser features and how they correlate with level. Ever since (Mosteller and Wallace, 1963) who used a set of 70 function words to solve some open authorship question in The Federlaist Papers, function words were adopted to solve many classification problems. They have several advantages: they are highly frequent, they do not carry much content and hence do not bias the model by topic, and they indirectly reflect grammar and therefore add another level of abstraction to the classification. The main caveat with using function words is that it is hard to interest the results.

Recall, however, that we are not interested in the role individual function words play in separation, we are interested in a robust method that will provide good classification results, and using the sgimoid function inherent to logistic regression provide reliable metric for proficiency level. It is only then that we move on to look at linguistically meaningful features. We use the list developed in Pennebaker et al. (2001), which has proven useuful for tasks such as ours (cf. Koppel and Ordan (2011).

The list consists of 468 words. Every chunk is represented as a 468-dimensional vector with the relative frequency of each function words for each element in the vector. We further compute the TF-IDF values to for each term to weight our factors.

4 Results

4.1 Binary classification

We use 10-fold to evaluate the accuracy of the binary classification, with mean accuracy above 99%, regardless of proficiency level. It seems, however, that the results still reflect proficiency: the more advanced a speaker is, the less confident the classifier is about its prediction. As yet we are not sure about the statistical significance of these differences.

The average absolute value of the distances for the non-native graded as low, medium and high are 4.578, 4.016, 3.106, respectively.

We went further and checked the correlation of the machine confidence values represented by distances with other, well-established proficiency measures of lower dimensions, such as type-token ratio (TTR).

	low	medium	high
distance	4.578	4.016	3.106
probability	0.985	0.974	0.939
TTR	2.421	2.357	2.238

4.2 Non-native inner group classification

It seems that the differences between natives and non-natives are so marked in our dataset, that the differences between the non-natives are almost marginal, at least with respect to the model's confidence. We therefore set forth to learn more more about the inner-group variance.

For this we run three pairwise comparisons over three levels, and we run predictions over the left-out subgroup. We applied same settings we used for the native vs. non-native classification.

We expected now more nuanced results in terms of (a) the performance of the model, and (b) the confidence of the model in its predictions. The results are as follows:

For classifying the chunks graded **low vs. high** the classifier achieved accuracy of 97%. around 47% of the medium graded chunks were classified as low, and the remaining 53 as high. For classifying the chunks graded **medium vs. high** the classifier achieved accuracy of 84%. around 98% of the low graded chunks were classified as medium, and the remaining 2 as high. For classifying the chunks graded **low vs. medium** the classifier achieved accuracy of 80%. around 95.5% of the high graded chunks were classified as medium, and the remaining 4.5 as low.

Need to organize this in a table or some other way...

References

Gili Goldin, Ella Rabinovich, and Shuly Wintner. Native language identification with user generated content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3591–3601, 2018.

Moshe Koppel and Noam Ordan. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, 2011.

Frederick Mosteller and David L Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342, 2018.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 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*, pages 48–57, 2013.