Combining Shallow and Linguistically Motivated Features in Native Language Identification

Serhiy Bykh, Sowmya Vajjala, Julia Krivanek and Detmar Meurers
Department of Linguistics, University of Tübingen

In a Nutshell

- We explored a wide range of features:
  - from *surface* (e.g., n-grams)
  - to deep *linguistic* features (e.g., dependency)
- We created ensemble classifiers by combining multiple single-feature classifiers, significantly increasing performance.
- Our best accuracy of *83.5%* is the *second best* score in the overall ranking of the *NLI Shared Task* (Tetreault et al., 2013).
  - *Closed* task: 82.2% (rank 5, difference to best result 83.6% not statistically significant)
  - *Open-2* task: 83.5% (rank 1)
  - *Open-1* task: 38.5% (rank 2)

Background

- Early work on NLI has explored different kinds of features ranging from word n-grams to spelling and grammar errors. (e.g., Jarvis et al., 2004; Koppel et al., 2005)
- Wong & Dras (2009) used features based on Contrastive Analysis.
- More recently, complex syntactic constructs were used as features. (e.g., Wong & Dras, 2011; Swanson & Charniak, 2012)
- Brooke & Hirst (2011) studied the effect of training data size on classifier performance.
- Tetreault et al. (2012) used ensemble models that combine multiple feature groups by building a meta-classifier of base classifiers.
- Bykh & Meurers (2012) explored a data driven approach using recurring n-grams with words and POS tag combinations.
  - We started with these features and extended our feature set to include more linguistically motivated features for this task.

Corpora used

TOEFL11 (Blanchard et al., 2013)

- Main corpus of the shared task
- 1100 essays of English learners with 11 L1 backgrounds.

NON-TOEFL11

- 5843 essays for 11 L1s for the *open-1* and *open-2* tasks
- unevenly distributed across 11 L1s, created from 5 corpora:
  - ICLE corpus (Granger et al., 2009)
  - FCE corpus (Yannakoudakis et al., 2011)
  - BALC Arabic Learner Corpus (Randall & Groom, 2009)
  - ICNALE corpus (Ishikawa, 2011)
  - TÜTEL-NLI: Tübingen Telugu NLI Corpus

Features

Recurring n-gram features

1. rc. word ng.	recurring word-based n-grams
2. rc. OCPOS ng.	recurring n-grams, where open class words are replaced by POS tags
3. rc. word dep.	rec. word-based dependencies (MATE): a head and all its immediate dependents, ordered as in the sentence **Ex:** *My own experience confirms this fact.* ⇒ (my, own, experience); (experience, confirms, fact); (this, fact)
4. rc. func. dep.	rec. function-based dependencies: each dependent is replaced by its grammatical function ⇒ (NMOD,NMOD,experience); (SBJ, confirms, OBJ); (NMOD, fact)

Complexity Features

5. complexity	• text complexity features of Vajjala & Meurers (2012): *lexical richness, syntactic complexity, . . .* • morphological and POS features from CELEX

Sublexical Morphological Features

6. stemsuffix, bin.	presence/absence of *stem+suffix*.
7. stemsuffix, cnt.	number of *stem+suffix* occurrences.
8. suffix, bin.	presence/absence of valid English *suffixes*.
9. suffix, cnt.	number of *suffix* occurrences.

Constituency Parser-based Features

10. type dep. lm.	lemma-typed Stanford dependencies ⇒poss(experience, my); amod(experience, own) etc.,
11. type dep. POS	POS-typed Stanford dependencies ⇒poss(NN, PRP$); amod(NN, JJ) etc.,
12. local trees	all syntactic trees of depth one parse: (ROOT (S (NP (PRP$ *My*) (JJ *own*) (NN *experience*)) (VP (VBZ *confirms*) (NP (DT *this*) (NN *fact*))) (. .))) ⇒ local trees: (S NP VP .), (NP PRP), (NP PRP$ JJ NN), . . .

Ratio Features

13. dep. num.	number of dependents (MATE) realized by a verb lemma normalized by this lemma's count **Ex:** *take*=10 ⇒ f1: *take:2-dependents=3/10* ⇒ f2: *take:3-dependents=7/10*
14. dep. var.	number of possible dependent-POS combinations for a verb lemma, normalized by this lemma's count **Ex:** *take:2-deps=3/10* ⇒ f1: *take:JJ-NN=3/10* *take:3-deps=7/10* ⇒ f2: *take:JJ-NN-VB=2/10* ⇒ f3: *take:NN-NN-VB=5/10*
15. dep. POS	POS-based dependent frequency for a verb lemma **Ex:** *f1, f2, f3* from *14.* ⇒ *take:JJ=(1/2+1/3)/10* ⇒ *take:NN=(1/2+1/3+2/3)/10* ⇒ *take:VB=(1/3+1/3)/10*
16. lm. realiz.	• lemma counts of a specific POS normalized by the total count of this POS **Ex:** *A document with 30 verbs and 50 nouns includes the lemma* can *2 times as a verb and 5 times as a noun.* ⇒ f1: *can:VB=2/30*, f2: *can:NN=5/50* • Type-Lemma ratio: lemmas of same category normalized by total lemma count • Type-Token ratio: tokens of same category normalized by total token count • Lemma-Token Ratio: lemmas of same category normalized by tokens of same category

Experimental setup & Results

- We submitted five system results for each of the three tasks.
- The ensembles are meta-classifiers created based on the probability distributions of the base classifiers.
- All systems consisted of classifier ensembles, except system 2.

Feature type	systems 1	2	3	4	5	Single feature results on T11 *dev* set: closed	open1	open2
1. rc. word ng.	x	x	-	x	-	***81.3***	***42.0***	***80.3***
2. rc. OCPOS ng.	x	-	x	x	-	67.6	26.6	64.8
3. rc. word dep.	x	-	x	x	-	67.7	30.9	69.4
4. rc. func. dep.	x	-	x	x	-	62.4	28.2	61.3
5. complexity	x	-	x	x	x	37.6	19.7	36.5
6. stemsuffix, bin.	x	-	x	x	x	50.3	21.4	48.8
7. stemsuffix, cnt.	x	-	x	-	x	48.2	19.3	47.1
8. suffix, bin.	x	-	x	x	x	20.4	9.1	17.5
9. suffix, cnt.	x	-	x	-	x	19.0	13.0	17.7
10. type dep. lm.	x	-	x	-	x	67.3	25.7	67.5
11. type dep. POS	x	-	x	-	x	46.6	27.8	27.6
12. local trees	x	-	x	-	x	49.1	26.2	25.7
13. dep. num.	x	-	x	x	-	39.7	19.6	41.8
14. dep. var.	x	-	x	x	-	41.5	18.6	40.1
15. dep. POS	x	-	x	x	-	47.8	21.5	47.4
16. lm. realiz.	x	-	x	x	-	70.3	30.3	66.9

Task	Overall system results				
$Closed_{test}$	**82.2**	79.6	81.0	81.5	74.7
$Closed_{dev}$	**85.4**	81.3	83.5	84.9	76.3
$Closed^{10foldCV}_{train \cup dev}$	**82.4**	78.9	80.7	81.7	74.1
$Open1_{test}$	36.4	**38.5**	33.2	37.8	21.2
$Open1_{test}$*	37.0	**38.5**	35.4	37.8	29.9
$Open2_{test}$	83.5	81.0	79.3	82.5	64.8
$Open2_{test}$*	**84.5**	81.0	83.3	82.9	79.8

The starred Open task results finished computing after submission.

Discussion

- Best single feature group: surface-based recurring n-grams
- Ensemble models combining a range of linguistically motivated features clearly outperform individual feature models.
  - Even individually weak features significantly contribute.

Future Work

- Qualitatively analyze feature types in depth and study the correlations between them
- Explore more linguistic features like syntactic alternations as proposed in Krivanek (2012)