Feature Extraction for Native Language Identification

Vincent Kríž, Martin Holub, Pavel Pecina

{kriz, holub, pecina}@ufal.mff.cuni.cz

Charles University in Prague,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Headlines

- the task of **Native Language Identification (NLI)**
- **TOEFL11 corpus** (Blanchard et al., 2013)
- supervised learning: **combining the SVM learner with a language modeling approach** to feature extraction
- **cross-entropy scores as features** for supervised learning
- **results achieved with reduced feature space** and comparing with results of the First Shared Task in NLI

Introduction

Native Language Identification

- automatic identification of the writers' native language (L1)
- based on a sample of their writing in a second language (L2)

Contrastive Analysis Hypothesis (Lado, 1957)

- speakers and writers of the same L1 can sometimes be identified by similar L2 errors
- linguistic interference

NLI as a text classification task

- raw texts $\rightarrow$ feature vectors $\rightarrow$ classified texts

Motivation

Educational settings

- more targeted feedback to language learners about their errors (Smith and Swan, 2001)

Authorship analysis (Stamatatos, 2009)

- criminal law (identifying writers of harassing messages)
- civil law (copyright disputes)
- literary research (attributing anonymous or disputed literary works to known authors)

Related Work

Approaches to the task

- Support Vector Machines (SVM)
- n-grams, function words, POS, spelling errors, writing quality (grammatical errors, style markers)

- Tree Substitution (TSG) structures (Swanson and Charniak, 2012)
- recurring n-grams (Bykh et al., 2013)
- string kernels & multiple kernel learning (Ionescu et al, 2014)

Tetreault et al. (2012)

- extensive study
- includes language modeling and entropy-based features

Related Work

The First NLI Shared Task (Tetreault et al., 2013)

- new corpus TOEFL11 (Blanchard et al., 2013)
- common set of L1s as well as evaluation standards
- a direct comparison of approaches

We experiment with exactly the same data, using the same cross-validation splits as the participants of the Shared Task, so we can provide the exact comparison with the published results.

Data

TOEFL11 (Blanchard et al., 2013)

- a corpus of non-native English writings – contains 1,100 essays per L1 language with an average of 348 word tokens per essay
- consists of essays on 8 different topics (*prompts*)
- written by non-native speakers of three *proficiency levels* (low, medium, high)
- the essays' authors have 11 different native languages:

L1	ID
Arabic	ARA
Chinese	CHI
French	RFE
German	GER

L1	ID
Hindi	HIN
Italian	ITA
Japanese	JAP
Korean	KOR

L1	ID
Telugu	TEL
Turkish	TUR
Spanish	SPA

Use of Language Modeling

Language modeling fundametals

- **n-gram** is a contiguous sequence of n items from a given sequence of text
- **language model** (LM) estimates the probabilities of possible n-grams
- estimated probability distributions should be **smoothed** (assigning non-zero probability to unseen n-gram)

Our approach

- a small set of **cross-entropy based features** computed over different language models
- significant reduction of the usual feature space based on n-grams
- features are then used by a SVM classifier

Cross-entropy scoring

Basic idea

- 11 special LMs of English, based on the same L1 language in the training data ($M_1, \ldots, M_{11}$)
- compare M_i to a general LM of English (M_G)
- the cross-entropy of text t given a language model M is

$$\mathrm{H}(t, M) = -\sum_x p(x) \log q(x).$$

Normalized cross-entropy score

$$D_G(t, M_i) = \mathrm{H}(t, M_i) - \mathrm{H}(t, M_G) = -\sum_x p(x) \log \frac{q_i(x)}{q_G(x)}$$

M_i with distributions q_i, M_G with the distribution q_G

Features

Cross-entropy Based Features

Family	ID	Description
Tokens	T	token based LM
Characters	C	character based LM
Suffixes	S_n	LMs on suffixes of the length $\{2, ..., 6\}$
POS tags	P	POS tags based LM

Statistical features (ST)

- **Text length characteristics**: # of sentences, tokens, characters
- **Lexical variety family**: number of unique tokens, proportion between # of unique tokens and # of all tokens in texts

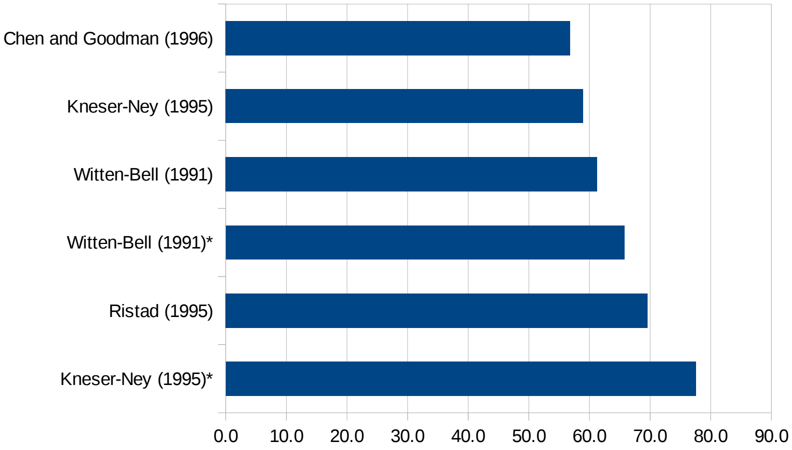
Prompt and proficiency (PR)

Results & Discussion

Experiments and results

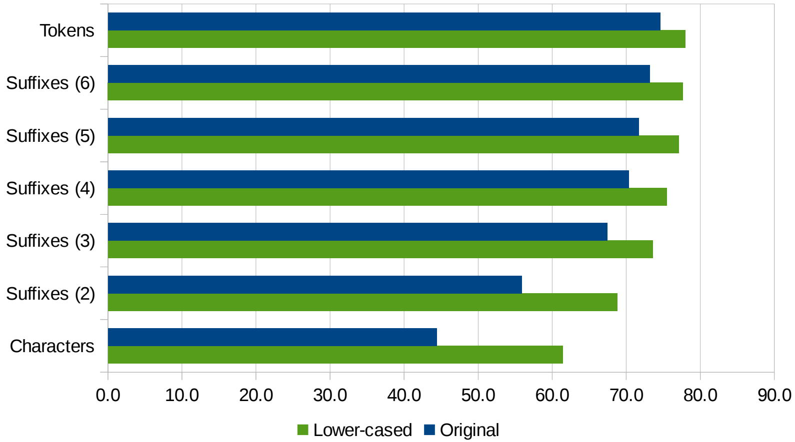
1. different smoothing methods
2. effect of lower-cased letters
3. performance of different feature families
4. different n-gram range used by LMs
5. different combinations of feature families

Smoothing methods – comparison



* indicates models with interpolation

Effect of lower-cased letters



Language models built on different n-gram families

ID	Feature family	Maximum n-gram order					
		3	4	5	6	7	8
C	Characters	61.4	70.5	73.0	74.1	74.6	**74.9**
S_2	Suffixes (2)	68.8	68.4	68.3	68.3	68.3	68.2
S_3	Suffixes (3)	73.6	73.2	73.2	73.2	73.1	73.0
S_4	Suffixes (4)	**75.5**	75.3	75.4	75.5	75.4	75.4
S_5	Suffixes (5)	77.1	76.9	77.2	77.1	77.1	77.1
S_6	Suffixes (6)	77.7	77.8	77.8	77.8	77.7	77.8
T	Tokens	**78.0**	78.0	77.9	78.0	77.9	78.0
P	POS tags	**53.1**	53.2	52.0	50.4	49.1	48.2

Classification accuracy using background language models built on different n-gram families. Each system uses 11 cross-entropy based features over the specified language model.

Feature families – combinations

C	T	S_4	P	PR	ST	Accuracy
x	x	x	x	x	x	82.43 ± 0.5
x	x	x	x	x		82.18 ± 0.8
x	x	x		x		82.16 ± 0.6
	x	x	x	x		81.97 ± 0.5
x	x	x		x	x	81.91 ± 0.6
x	x	x				81.31 ± 0.4
	x	x				81.07 ± 0.5
x	x					80.94 ± 0.7
x		x	x	x	x	78.29 ± 0.7
	x					77.99 ± 0.7

C – characters, T – tokens, S_4 – suffixes of length 4, P – POS tags, PR – proficiency and prompt, ST – statistical features.

Comparison with the best Shared Task systems

System	# of feat.	Acc.	Approach
Gebre et al.	-	84.6	T, C, POS, spelling errors
Jarvis et al.	400,000	84.5	T, L, POS
Lynum	867,479	83.9	T, C, S
Malmasi et al.	-	82.5	T, function words, POS, syntax
Our system	**55***	**82.4**	**LMs** using T, C, POS, S
Bykh et al.	-	82.4	T, POS, syntax, S

* traditional n-grams are hidden in the language models

T - tokens, C - characters, POS - part of speech tags, L - lemmas

Conclusion

- new NLI system for identifying the native language (L1) of a non-native English writer
- significantly **reduced feature space** ($10^5 \rightarrow 55$)
- using language modeling – improved performance
  - **different smoothing methods**
  - combination of language models based on **different types of n-grams**
  - using **normalized cross-entropy score**
- **resulting accuracy 82.4 %** comparable to the state-of-the-art