UPPSALA
UNIVERSITET

Learning with Learner Corpora: using the TLE for Native Language Identification

Allison Adams and Sara Stymne

2017-05-22

UPPSALA UNIVERSITET

Contents

UPPSALA UNIVERSITET

What is NLI?

Native Language Identification(NLI):

- Determine L1 by analyzing L2 texts
- Most often treated as a text classification problem
- Language learner data can be used to train classifiers

NLI: a guessing game

Spanish German Arabic Chinese

"In my conclusion, older people enjoy their life more than young person do. But when young person got enough experience of the life, they will begin to love it. "

"In opposite of the younger people older people enjoy their live, too. But not like the younger it do. They cannot do much more things, because they have families and a job. "

"The young have their own enjoy, so the older have their own enjoy. I couldn't to say this enjoy life more than this other, because their likes are very differents."

"the mind or the think is deferent when you yong you just think about fun ,but when you get older you will think about the real thing like your sheldrens or jop or how to be arich man like this."

NLI: a guessing game

Chinese

"In my conclusion, older people enjoy their life more than young person do. But when young person got enough experience of the life, they will begin to love it. "

German

"In opposite of the younger people older people enjoy their live, too. But not like the younger it do. They cannot do much more things, because they have families and a job. "

Spanish

"The young have their own enjoy, so the older have their own enjoy. I couldn't to say this enjoy life more than this other, because their likes are very differents."

Arabic

"the mind or the think is deferent when you yong you just think about fun ,but when you get older you will think about the real thing like your sheldrens or jop or how to be arich man like this."

Prior Research

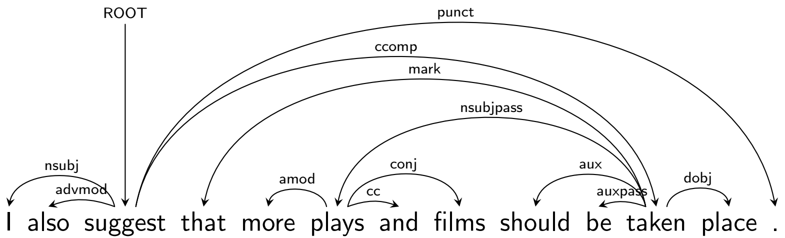
- Koppel et al, 2005:
  - Used function words, POS bi-grams, spelling/grammatical errors as features
  - Combined all features using SVM, achieved an accuracy of 80,2%
- Wong & Dras, 2011:
  - Included syntactic features
  - 2 parsers, Charniak and CFG
- Swanson & Charniak, 2012:
  - Used tree substitution grammars
- Brooke & Hirst, 2012:
  - Tested range of features, incl dependencies
  - Found feature ensemble most effective
- Tetreault et al, 2012:
  - Tested range of features, incl dependencies
  - Overall accuracy of 90.1%

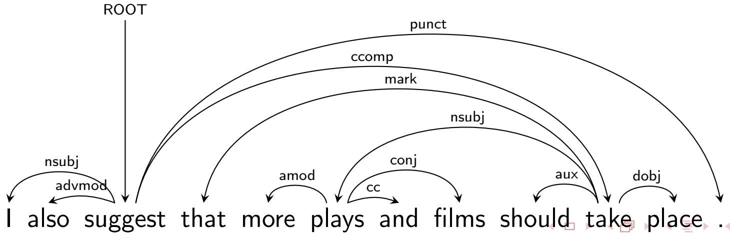
The Treebank of Learner English

- Manually annotated syntactic treebank for ESL
- PoS tags and UD trees for 5,124 sentences
- Cambridge First Certificate in English (FCE) corpus
- Parallel corpus:
  - Uses FCE error annotation scheme, adapted to UD
  - Full analyses provided for corrected & original versions

UPPSALA UNIVERSITET

TLE as a parallel corpus





The TOEFL11 Corpus

- 12,100 learner essays, 11 language backgrounds
- 1,100 essays per language item
- Proficiency score distributions correspond to real-life distribution in test results

Language	Low	Medium	High
Arabic	296	605	199
Chinese	98	727	275
French	63	577	460
German	15	412	673
Hindi	29	429	642
Italian	164	623	313
Japanese	233	679	188
Korean	169	678	253
Spanish	79	563	458
Telugu	94	659	347
Turkish	90	616	394
Total	1330	6568	4202

Table: Score level distributions in TOEFl11

Parsing with the TLE

- 3 parsers trained on TLE
  - Original: error-filled version of TLE
  - Corrected: error-corrected version of TLE
  - Merged: every other sentence from original/corrected treebank
- 2 parsers trained on EWT
  - Full English Web Treebank: UD treebank of English containing documents from five genres: weblogs, newsgroups, emails, reviews, and Yahoo! Answers
  - 50% of EWT
- Hunpos trained on EWT used to acquire POS tags
- Dependencies extracted from parsed corpus
- Classified using Scikit Learn's linear SVC

Parser accuracies for all three test sets

Train Set	Test Set	LAS	UAS
TLE_{corr}	*corrected*	**94.5**	**95.6**
	original	**90.1**	**92.2**
	merged	**92.5**	**94.1**
TLE_{orig}	*corrected*	85.7	88.5
	original	85.1	88.0
	merged	85.2	88.0
TLE_{merged}	*corrected*	85.0	88.1
	original	85.0	88.0
	merged	85.4	88.0
EWT	*corrected*	80.7	86.0
	original	80.6	86.1
	merged	80.8	86.0
EWT 50%	*corrected*	79.8	85.4
	original	79.3	85.0
	merged	80.0	85.5

Types of dependency relations used in feature set

dep(lemma, lemma)	(lemma, lemma)
dep(PoS, lemma)	(PoS, lemma)
dep(lemma, PoS)	(lemma, PoS)
dep(PoS, PoS)	(PoS, PoS)

nsubj(think, you)	(think, you)
nsubj(VERB, you)	(VERB, you)
nsubj(think, PRON)	(think, PRON)
nsubj(VERB, PRON)	(VERB, PRON)

Evaluation

	Acc	P	R
Original	70.5	70.7	70.6
Corrected	70.2	70.3	70.3
Merged	70.5	70.6	70.5
EWT	70.5	70.7	70.6
EWT 50%	70.0	70.1	70.0

Table: Accuracy, precision, recall for native language identification with the three TLE models and the two EWT models.

Evaluation

Language	Original	Corrected	Merged
Arabic	**68.0**	66.1	67.0
Chinese	74.3	**74.7**	73.5
French	70.1	71.0	**71.2**
German	81.5	**82.5**	81.7
Hindi	**64.7**	64.2	64.3
Italian	75.9	**76.2**	75.8
Japanese	**71.3**	70.5	**71.3**
Korean	63.7	62.7	**64.5**
Spanish	62.2	62.7	**62.9**
Telugu	**71.5**	71.1	71.3
Turkish	**72.1**	70.7	71.6

Table: Accuracy scores by language for all three models

Suggestions for future research

- Use TLE (or TLE+EWT) for POS tagger training
- Test method using feature ensembles (e.g. with character n-grams or language models)
- Investigate why the corrected TLE results in increased parser performance on uncorrected learner essays with a detailed error analysis

Conclusions

- We tested a potential use of the TLE for NLI
- We used 3 versions of the TLE to train dependency parsing models using MaltParser
- Each model was used to extract dependency relations from learner texts, which were used as features in a classification task
- Parsing results are much better when trained on grammatical relations, but native language classification is slightly better using a parser trained on the original treebank containing ungrammatical relations
- The contrastive systems trained on the EWT demonstrate the importance of both corpus size and domain

Thank you