Word Formation Variation as Features for Native Language Identification

Julia Krivanek Detmar Meurers
Universität Tübingen

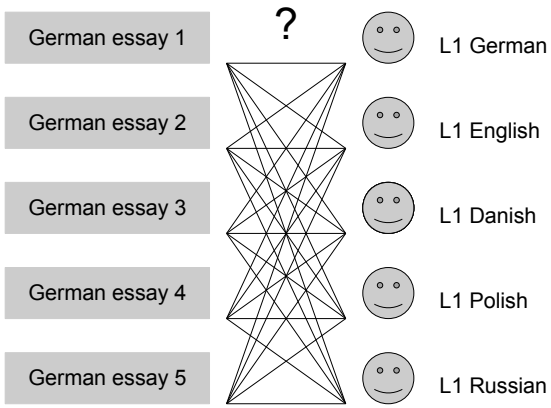
Learner Corpus Research Conference (LCR 2013)
Bergen, 29. September 2013

Introduction

- A key question in current corpus-based research concerns the role of linguistic abstraction:
  - When are linguistic categories relevant and when are surface-based characterizations just as successful?
- We need an experimental sandbox to try out different types of linguistic modeling and study their impact.
- How about Automatic Native Language Identification?
  - Transfer is known to involve many linguistic dimensions (lexicon, syntax, pragmatics, . . .).
  - Let's run classification experiments to quantify the effect of linguistic abstractions.

Native Language Identification (NLI)

- Automatically determine the native language of a writer based on a text they wrote in a second language.



Current NLI approaches

- Shallow approaches using surface n-grams
  (e.g., Brooke & Hirst 2012; Bykh & Meurers 2012; Jarvis et al. 2012)
  - high classification accuracy
  - large feature sets impossible to interpret qualitatively
  - some dependence on domain (genre, topic, . . .)
- Error pattern approaches (Wong & Dras 2009; Bestgen et al. 2012)
  - focus on one aspect of learner language
  - often requires manual error annotation
- NLI Shared Task 2013 (Tetreault et al. 2013):
  - English essays by writers with 11 native languages
  - approaches often use a combination of features, directly or using meta-classifiers

Linguistic Variation as Features for NLI

- Word-based features encode form and meaning together.
  - requires very high number of features to be applicable to unseen data, across domains
- Can we abstract away from the meaning to be expressed to choices in the linguistic system?
  - Idea: Study where the linguistic system provides multiple ways to express the same meaning or function.
    - method related to variationist sociolinguistics
- How about using valence alternations for NLI? (Krivanek 2012; Meurers et al. 2013)

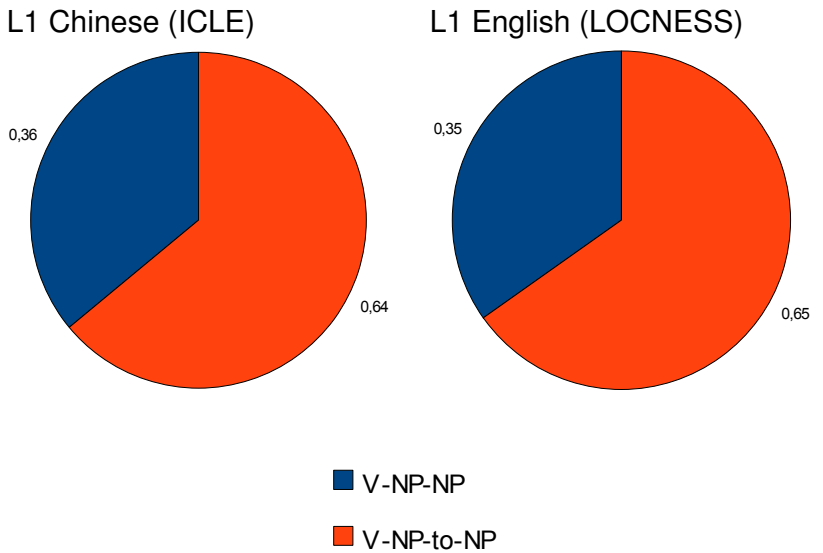
(1) a. *He gave the book to John.*
b. *He gave John the book.*

Dative Alternation

Popular topic in linguistics (Levin 1993), but so far little corpus-based SLA work (but cf. Callies & Zaytseva 2011).

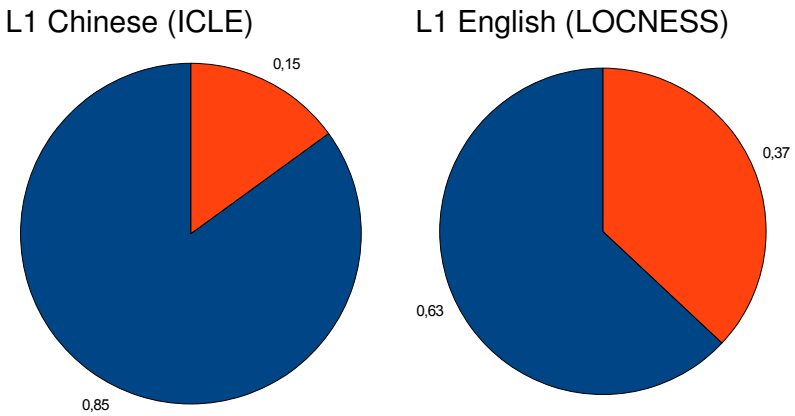
Syntactic alternations as features

A non-distinctive alternation: Dative Alternation Drop



Syntactic alternations as features

A distinctive alternation: Locative Preposition Drop



V-PPloc (Martha climbed up the mountain.)
V-NP (Martha climbed the mountain.)

Syntactic alternations as features

(Krivanek 2012; Meurers, Krivanek & Bykh 2013)

- Theory-driven approach:
  - 21 alternations from Levin (1993) as features

  ⇒ effective, but features not common enough in short texts
- Adding a data-driven twist:
  - define classes consisting of all verbs with the same set of syntactic realization alternatives occurring in a corpus
  - features encode variants chosen in text for a given class

  ⇒ improves accuracy, features qualitatively interpretable
- How about taking such a variationist perspective further?

Taking the next step

- We target German as L2.
  - language is morphologically richer than English
- Focus on morphological variation:
  - word formation
- Lexical features are attractive since there are many opportunities to observe words even in a short text.

Inflection

	singular		plural	
1. pers.	ich	schwimm**e**	wir	schwimm**en**
2. pers.	du	schwimm**st**	ihr	schwimm**t**
3. pers.	er/sie/es	schwimm**t**	sie	schwimm**en**

- Inflection directly reflects morpho-syntactic requirements

⇒ not likely to be informative for our purposes

Word Formation

- Language offers several options for forming new words:
  - with/without derivational morphemes
  - with different part-of-speech or gender as source and as target
- Example:

schreien$_{to\ shout}$ → das$_{the-neut}$ Schreien$_{shouting}$
der$_{the-masc}$ Schrei$_{shout}$
die$_{the-fem}$ Schrei**erei**$_{yelling}$
das$_{the-neut}$ **Ge**schrei$_{yelling}$

⇒ Define *word formation variables* and use *variants* as features

Variables and their variants as features for NLI

Morpheme alternation

Variants	Examples		
no affix	*Frau*$_{woman}$	+ *Welt*$_{world}$	→ *Frauenwelt*$_{woman's\ world}$
suffix	*Feminist*$_{feminist}$	+ ***in***$_{female}$	→ *Feministin*$_{feminist}$
prefix	***un***$_{in}$	+ *gerecht*$_{just}$	→ *ungerecht*$_{injust}$
verb particle	***auf***$_{up}$	+ *geben*$_{give}$	→ *aufgeben*$_{give\ up}$

Variables and their variants as features for NLI

Derived category alternation

Variants	Examples		
noun	*anerkennen*$_{recognize}$	+ ung	→ ***Anerkennung***$_{recognition}$
verb	*auf*$_{up}$	+ geben$_{give}$	→ ***aufgeben***$_{give\ up}$
adjective	*entsprechen*$_{correspond}$		→ ***entsprechend***$_{corresponding}$
adverb	*möglich*$_{possible}$	+ weise	→ ***möglicherweise***$_{possibly}$

Variables and their variants as features for NLI

Source category alternation

Variants	Examples		
noun	***Feminist***$_{feminist}$	+ in$_{female}$	→ *Feministin*$_{feminist}$
verb	***anerkennen***$_{recognize}$	+ ung	→ *Anerkennung*$_{recognition}$
adjective	***möglich***$_{possible}$	+ *weise*	→ *möglicherweise*$_{possibly}$
adverb	***so***$_{as}$	+ *bald*$_{soon}$	→ *sobald*$_{as\ soon\ as}$

- Combining the variants of the three variables, one obtains 29 distinct features.
- For each feature, count number of occurrences per text, normalized by derived category.

Word Formation Variation as Features for NLI
Julia Krivanek and Detmar Meurers
Introduction
Native Language Identification
Current NLI approaches
Linguistic variation as features for NLI
Morphological Variation
Inflection
Word Formation
Experiment
Features used
Setup and Results
Qualitative Results
Summary
References

Setup and Results

- Corpus used:
  - 185 German essays from Falko corpus (Reznicek et al. 2012)
  - 5 native languages (English, Polish, Russian, Danish, German)
  - advanced learners of German and native control group
  - average text length: 470 words
  - POS and morphology: RFTagger (Schmid & Laws 2008)
- Classification setup:
  - WEKA SMO Classifier (Witten & Frank 2005)
  - Leave-one-out evaluation
- Accuracy: **55.1%** (20% random baseline)
  - morphological information can clearly contribute to NLI
  - can be integrated into ensemble classifier for high accuracy NLI (Bykh, Vajjala, Krivanek & Meurers 2013)

Word Formation Variation as Features for NLI
Julia Krivanek and Detmar Meurers
Introduction
Native Language Identification
Current NLI approaches
Linguistic variation as features for NLI
Morphological Variation
Inflection
Word Formation
Experiment
Features used
Setup and Results
Qualitative Results
Summary
References

Qualitative Results: Confusion Matrix

	ger	eng	pol	rus	dan
ger	**34**	0	1	1	1
eng	3	**20**	2	3	9
pol	1	1	**26**	6	3
rus	2	6	8	**16**	5
dan	4	11	2	2	**18**

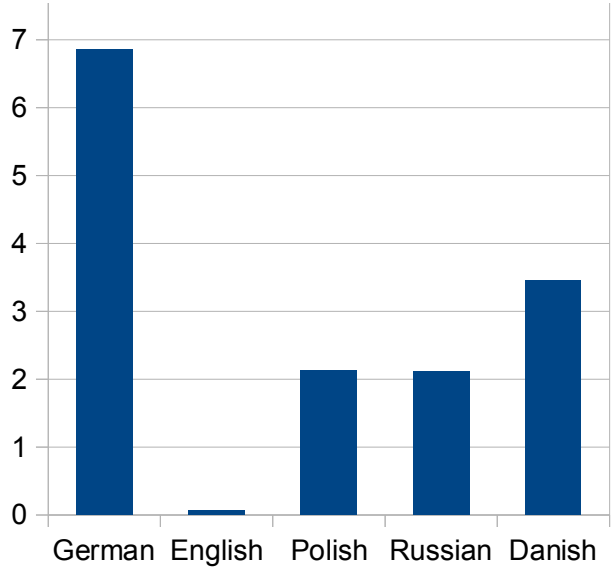
- German control group is recognized the best.
- Most confusions arise
  - within Slavic language group
  - within Germanic language group

⇒ potential usefulness of cascading classification (cf. Vajjala & Loo 2013)

Word Formation Variation as Features for NLI
Julia Krivanek and Detmar Meurers
Introduction
Native Language Identification
Current NLI approaches
Linguistic variation as features for NLI
Morphological Variation
Inflection
Word Formation
Experiment
Features used
Setup and Results
Qualitative Results
Summary
References

Qualitative Results: Overuse/Underuse

Verb particle feature (e.g., auf$_{up}$ + geben$_{give}$)



Summary

- Native Language Identification makes it possible to probe into the linguistic properties involved in Transfer.
- We argued for the use of variation within the linguistic system as meaningfully interpretable features for NLI.
  - syntactic variation as example (Meurers et al. 2013),
- We discussed new research on morphological variation:
  - targeting **German** learner texts
  - with features encoding **word formation** variants
- Results confirm that morphological variation can provide valuable information for NLI.
  - qualitatively interpretable features
  - can be integrated into ensemble classifiers for high quantitative results

References I

Bestgen, Y., S. Granger & J. Thewissen (2012). Error Patterns and Automatic L1 Identification. In S. Jarvis & S. A. Crossley (eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, Multilingual Matters, pp. 127–153.

Brooke, J. & G. Hirst (2012). Robust, Lexicalized Native Language Identification. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbai, India, pp. 391–408.

Bykh, S. & D. Meurers (2012). Native Language Identification Using Recurring N-grams – Investigating Abstraction and Domain Dependence. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbay, India, pp. 425–440. URL http://purl.org/dm/papers/bykh-meurers-12.html.

Bykh, S., S. Vajjala, J. Krivanek & D. Meurers (2013). Combining Shallow and Linguistically Motivated Features in Native Language Identification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-8) at NAACL-HLT 2013*. Atlanta, GA. URL http://purl.org/dm/papers/Bykh.Vajjala.ea-13.html.

Callies, M. & E. Zaytseva (2011). The Corpus of Academic Learner English (CALE): A new resource for the study of lexico-grammatical variation in advanced learner varieties. In *Hedeland, Hanna and Thomas Schmidt and Kai Wörner*, Multilingual Resources and Multilingual Applications (Hamburg Working Papers in Multilingualism B 96), pp. 51–56.

Jarvis, S., G. Castañeda-Jiménez & R. Nielsen (2012). Detecting L2 Writers' L1s on the Basis of Their Lexical Styles. In S. Jarvis & S. A. Crossley (eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, Multilingual Matters, pp. 34–70.

Krivanek, J. (2012). Investigating Syntactic Alternations as Characteristic Features of Learner Language. Master's thesis, University of Tübingen.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*, vol. 348. Chicago, IL: University of Chicago Press.

Meurers, D., J. Krivanek & S. Bykh (2013). On the Automatic Analysis of Learner Corpora: Native Language Identification as Experimental Testbed of Language Modeling between Surface Features and Linguistic Abstraction. In *Proceedings of 4th International Conference on Corpus Linguistics (CILC 2012)*. To appear.

References II

Reznicek, M., A. Lüdeling, C. Krummes & F. Schwantuschke (2012). *Das Falko-Handbuch. Korpusaufbau und Annotationen Ver. 2.0*. URL http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko.

Schmid, H. & F. Laws (2008). Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*. Stroudsburg, PA, vol. 1, pp. 777–784.

Tetreault, J., D. Blanchard & A. Cahill (2013). A Report on the First Native Language Identification Shared Task. In *Proceedings of the Eighth Workshop on Building Educational Applications Using NLP*. Atlanta, GA, USA: Association for Computational Linguistics.

Vajjala, S. & K. Loo (2013). Role of Morpho-syntactic features in Estonian Proficiency Classification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA8), Association for Computational Linguistics*. URL http://aclweb.org/anthology/W13-1708.pdf.

Witten, I. H. & E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Amsterdam; Boston, MA: Morgan Kaufmann, 2nd ed.

Wong, S.-M. J. & M. Dras (2009). Contrastive analysis and native language identification. In *Australasian Language Technology Association Workshop 2009*. pp. 53–61.