Duzen oder Siezen

Predicting the Formality of Second-Person Address in German

Kristin Asmus

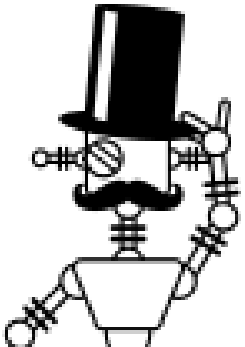
Problem Statement

- Predicting whether to use “du” (informal) or “Sie” (formal) form of second-person pronouns in German texts
- Analogous to many other languages with T/V dichotomy
  - German is non-pro-drop and has few ambiguous cases
- Based on sentence context

Motivation and Related Work

Improvements in other areas:

- Machine Translation
  - from implicit to explicit T/V
- Information Extraction
- Conversational AIs
  - understanding social conventions



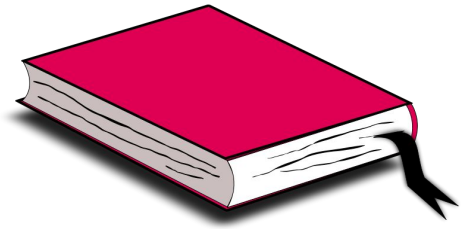
- No existing research found on this particular problem
- Google Translate solves by attempting to match existing texts, but is inconsistent
- Similar Research:
  - Address formality in English
  - Politeness
  - Social power relationships

Corpus

- Second person address is uncommon in written text
- Dialogue between characters in novels

Project Gutenberg

- 997 German novels from the public domain
- 774 include use of second person pronouns
- 98,838 instances of formal and
  148,803 instances of informal pronouns



Baseline Models

Random

- Ignores training texts
- Predicts formal/informal pronoun according to 50-50 chance
- Accuracy: 49.98%

Mode

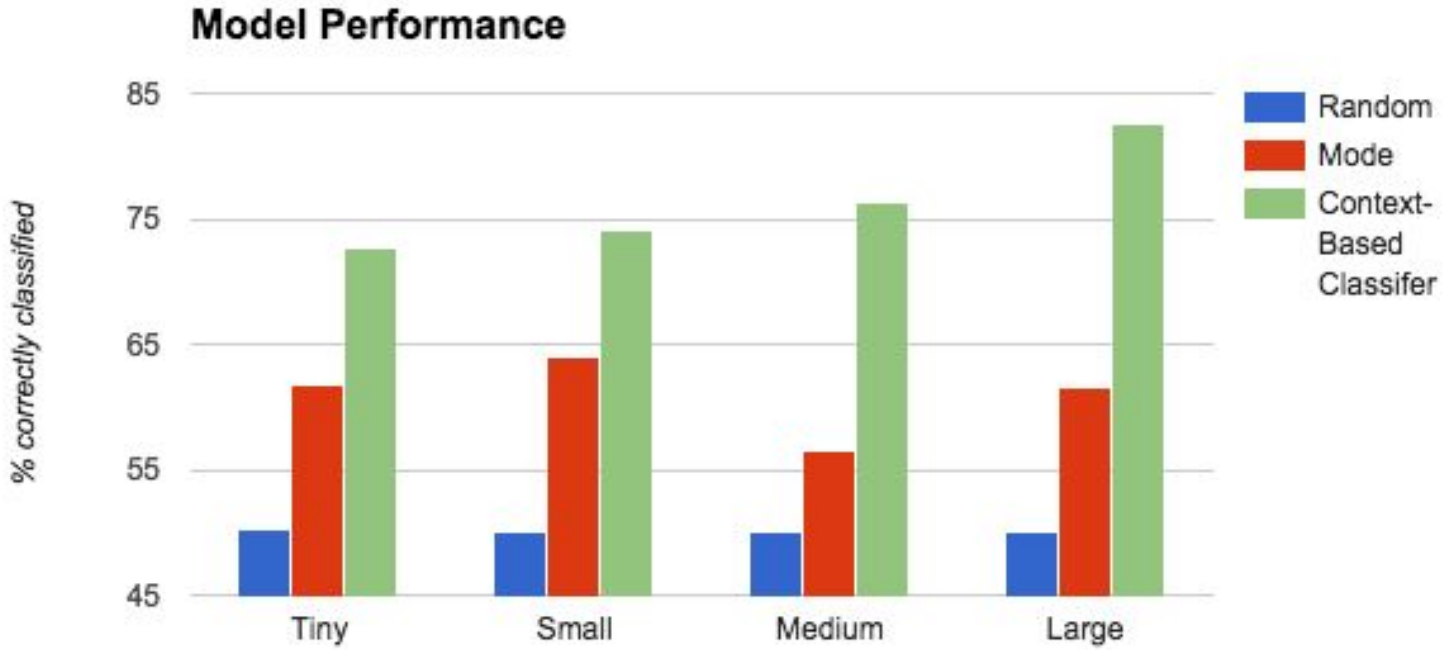
- Finds most common pronoun encountered in training texts via simple count
- Always predicts that form
  - Usually informal is more common
- Accuracy: 56.54%

* accuracy determined by average of 10 runs on sample of 150 novels

Context-Based Classifier

- Classify on a sentence-by-sentence basis using contextual features
  - only .2% of sentences include both formal and informal pronouns
- Random Forest Classifier
- Bag-of-words and N-Gram Features
  - unigram and bigram
  - stopwords (common, uninformative words) removed
  - trained with vocab size of 20,000

Results


Model Performance
% correctly classified
85
75
65
55
45
Tiny
Small
Medium
Large
Random
Mode
Context-Based Classifer

Conclusions

- Unstructured contextual features improved upon baselines significantly
- Removing stopwords increased bag of words power dramatically
- Including bigrams in bag of words made no significant impact

Future Features and Challenges:

- Structured features from parse trees
  - Limited training data for German and very resource intensive
- Modeling social graph of entire novels
  - Identifying speakers and addressees to learn/predict relationships