Duzen oder Siezen:
Predicting the Formality of
Second-Person Address in German

Kristin Asmus
[kasmus@mit.edu]

Final Project Report
MIT 6.864 Advanced Natural Language Processing
Professors Regina Barzilay and Tommi Jaakkola

0 Abstract

Many non-English languages possess two different forms for formal and informal second-person pronouns, yet there is not much existing research in NLP characterizing this dichotomy. By using contextual features to predict the form of second-person address, many related problems such as machine translation, information extraction, and conversational AI could be improved. In this project, the dialogues between characters in German novels from Project Gutenberg are used as training data to inform models making use of bag-of-words and n-gram features to predict the formality of second-person pronouns in a test set of dialogues where these words have been erased. Such models were found to have significantly higher accuracy than random or mode baseline approaches.

1 Problem Statement and Motivation

German, like many other non-English languages, possesses two different forms for its second-person pronouns: *du* and *Sie*, corresponding to informal and formal address, respectively. Both words translate to the English *you*, but are used distinctly according to the relationship and social distance between the speaker and the addressee. This T/V dichotomy (so labeled due to the Latin equivalents, *tu* and *vos*) is a prevalent distinction in many languages, yet there does not exist much research presently in the natural language processing space attempting to characterize it.

The ability to classify such forms of address would empower advances across a number of well-defined problems in NLP. Machine translation suffers from a current lack of knowledge

when converting from languages where the T/V dichotomy is not explicit to those where it is; in these cases, extra information extracted from features of the rest of the translation could be instrumental in determining which form to use in the target language. Information extraction regarding social relationships similarly depends on these features. Another interesting application involves conversational artificial intelligence agents, which could interact more naturally with humans if they understand more about, and accordingly adhere to, our existing social conventions.

To explore this problem, the German language is a good starting point, as its formal and informal second-person forms can be clearly observed in text with minimal ambiguity, so the need for manually annotated training data can be circumvented. Many other T/V languages are pro-drop languages, which often exclude pronouns from sentences in favor of implicit address deduced from context and the form of the conjugated verb; however, most second-person formal constructs also share a verb conjugated with third person forms, making these very difficult to identify. German is a non-pro-drop language, meaning that these pronouns are necessarily explicit. Additionally, in many languages, the words for formal second-person personal pronouns have alternate meanings, which would similarly require much more rigorous disambiguation. In German, this case is largely avoided by the fact that second-person formal pronouns are always capitalized. The only ambiguous case that remains is determining the meaning when such a word is also the first in the sentence. The standard in related work is to ignore this case; the resulting loss of information is found to be nonproblematic.

To formalize the problem of characterizing the T/V dichotomy in German, the project involves observing German sentences where a second-person pronoun, formal or informal, has been replaced with a generic tag. The goal of the classifier is to predict whether a formal or informal pronoun should go where the tags are, based on the information within the sentence.

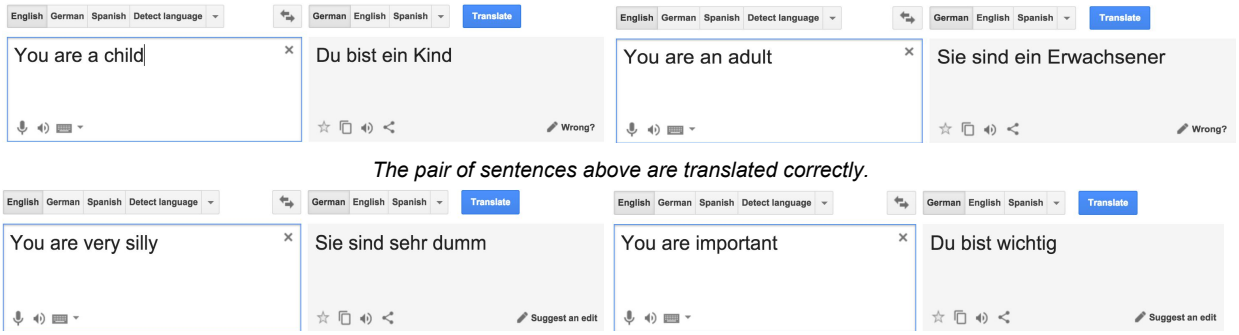
2 Related Work

There does not appear to be any existing research into the specific question of classifying the T/V dichotomy of second-person pronouns. However, advances have been made in a number of similar contexts.

Krishnan and Eisenstein induce formality in social networks by modeling both social network structure and formality associated with address terms such as names or placeholders. Though it is much more challenging to model social network structure implicitly from interactions in a novel than from a signed-in social network site, such factors could likely inform the T/V distinction very well. Faruqui and Pado attempt to classify the degree of formality in English using a standard supervised learning binary classification task with a logistic regression model. They examine three primary feature types: word features (as some words are closely associated with formal or informal contexts), semantic class features (clustering morphologically similar words to combat data sparsity), and features derived from politeness theory (other

semantic structures or multi-word expressions associated with formality). Their model represents a good basis for characterizing local features within a smaller context than a whole novel. Peterson, Hohensee, and Xia developed a Maximum Entropy classifier to determine the formality of email communications, and use similar linguistic features to the set described by Faruqui and Pado, with the addition of several informal style indicators likely specific to digital communications. Interestingly, with no explicit knowledge of the underlying social network, Peterson et al. modeled social distance based on the nature of communications (emails were labeled for business or personal purposes) and number of interactions, and found meaningful correlations between these simpler representations of the network and the formality measure.

Google Translate also appears to apply some measure of informed prediction to translations from implicit to explicit T/V languages, though their algorithm is not publicly documented and the accuracy of observed results was found to be quite inconsistent. Examples are shown in Figure 1.



The pair of sentences above are translated correctly.

The pair of sentences above use the incorrect pronoun formality.

Figure 1. In these Google Translate screenshots, formality is translated correctly for the upper pair: “Du” is used to informally address a child, while “Sie” is used to formally address an adult. The lower pair, however, loses this grasp, with “very silly” failing to convey a relationship where the speaker is most likely addressing someone informally, and “important” not taken into account to indicate that the addressee should probably be addressed in a formal manner.

3 Corpus

Because second person address is not a form commonly found in many published media, such as news articles, many of the standard sources for natural language processing corpora would not be suitable for this project. The construct is most frequently used in the context of spoken conversation, but a close translation to print is present in the dialogue between characters within novels.

Project Gutenberg is a large, free-to-access, digital archive of cultural works, where the majority are public domain books available in plain text. This corpus has been used previously in many natural language processing studies. Though the majority of these 50,000+ texts are in English, there are still hundreds of German novels to draw from. The site does not condone automated access, but instead enables direct downloads or mirroring of the project’s resources.

In total, all 997 German full-text novels were downloaded for this project. Of these, a short Python script discarded the 223 that did not contain any instances of second-person pronouns. The 774 remaining novels include a collective total of 98,838 uses of formal second-person pronouns ("Sie", "Ihnen", "Ihrer", "Ihr") and 148,803 uses of informal second-person pronouns ("du", "dich", "dir", "deiner", "dein").

4 Baselines and Architecture

In order to compute baseline measures of accuracy in predicting the formality of second-person pronouns, first, the corpus is divided into a training set and a test set, where each novel has a 50% probability of being assigned to either set. Next, the model is trained (if necessary) to optimize its parameters on the sentences from the training set of novels that involve the second-person pronouns in question. In the test set, a preliminary step scans through each novel and replaces any instances of formal or informal second-person pronouns with a specially formatted tag. This placeholder indicates the presence of a second-person address, but carries no information regarding whether the replaced pronoun was formal or informal. The trained model is then used to predict, for each sentence containing the tag, which level of formality the pronoun belonging there should have expressed. These predictions are finally evaluated against the gold standard represented by the original text, with accuracy represented as the ratio of predicted pronouns matching the original pronoun's formality to the total number of predicted pronouns. To obtain results, each classifier is run 10 times with independent training and test sets, and with the mean of the accuracies representing the classifier's performance.

Two baseline measures to compare against the results of the model developed for this project are random selection and majority selection, which are used in most similar research contexts.

The first baseline is determined by a random classifier. Since this model does not require training any parameters, the novels in the training set are skipped. The model makes a prediction for each of the placeholder tags completely randomly, with a 50-50 chance of classifying any given tag as formal or informal. Results will be discussed in depth in later sections, but predictably, the accuracy of this classifier hovered around 50%.

The majority selection model, also known as a mode classifier, trains by iterating over every novel in the training set to count up instances of formal versus informal second-person pronouns. It then passes the most common form along to the prediction step, where every the formality second-person pronoun tag is predicted to be the same as that mode. Because the training and test sets are selected randomly for each round, either formality could potentially be more common; however, as informal pronouns are more frequent than formal in this corpus, this model usually assigns all tags an informal prediction.

5 Context-Based Classifier

To improve upon those baselines, a context-based classifier was implemented. This classifier attempts to predict the correct pronoun formality based on the context of the rest of the sentence containing said pronoun. Because of this sentence-by-sentence perspective, any pronouns within the same sentence are all assigned the same formality prediction. This is an acceptable assumption because only in rare cases would one address two individuals with different forms of address all in the same sentence; an analysis of the corpus proves this, counting that only 0.2% of the sentences containing second-person pronouns include mixed formality.

The classifier presently makes use of bag-of-words n-gram features, with a Random Forest Classifier to combine them. Scikit-learn's ensemble and feature_extraction modules were utilized to implement this classifier. Specifically, a CountVectorizer object was used to produce bag-of-words feature vectors for, in the training set, every sentence involving one of the pronouns to classify. These features included both unigram and bigram features with a vocabulary size of up to 20,000 words. Importantly, prior to counting, any "stopwords" - the most common words in a language that are found uniformly throughout texts are not deemed informative - according to NLTK's list for German are filtered out so as not to be counted for the feature vector. This enables the classifier to focus on words that are more unique and indicative of a particular sentence. Once the feature vectors for the training set are generated, the Random Forest Classifier is used to fit these vectors to the correct classifications originally found in the text. Finally, this trained classifier is used to predict the formality classification for every sentence in the test set containing the replacement tag discussed in the previous section.

6 Evaluation and Results

To evaluate the different classifiers, each was run 10 times with independently drawn training and test sets. The accuracy of each run was determined by comparing the predicted formality for each pronoun (for baseline classifiers) or sentence (for the context-based classifier) to that indicated explicitly by the form of the second-person pronoun used in the original text. This method was repeated over four corpora, where the largest represents all 774 useful novels downloaded from Project Gutenberg, and the other three corpora are subsets: Medium contains 150 novels drawn at random, Small contains 60, and Tiny contains 30. The results for these executions are displayed in Figure 2.

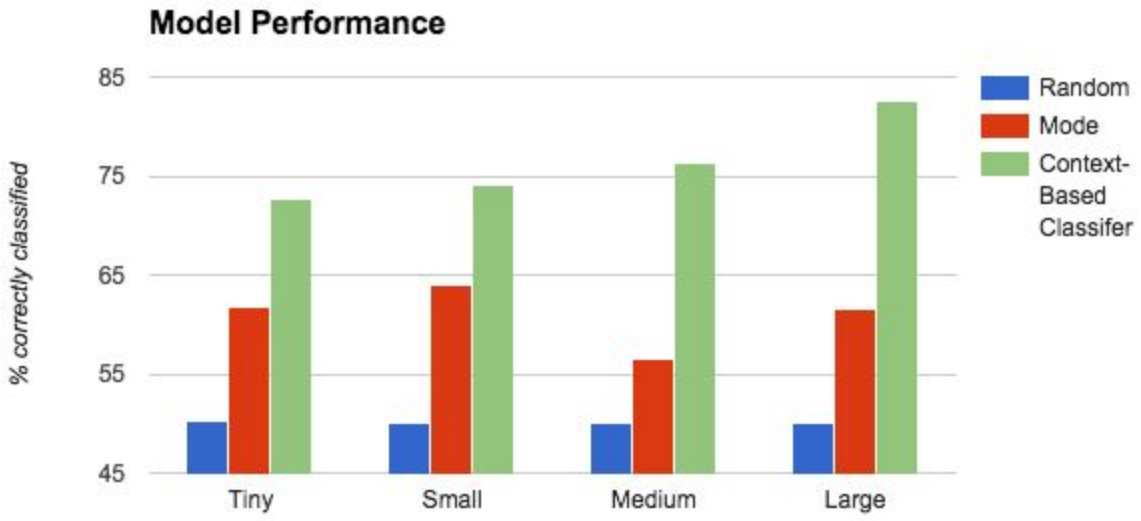


Figure 2. Mean performance over 10 independent trials with each model on corpora of various sizes.

The random classifier's accuracy is almost exactly 50% for all corpora. The mode classifier hovers between 55% and 65%, reflecting the actual distribution of informal versus formal pronouns in the given corpus. Finally, the context-based classifier clearly outperforms the baselines across the board, with accuracy significantly improving with the size of the corpus. Due to runtime constraints, the full corpus was only run with the context-based classifier once instead of averaged over ten runs like the others; however, the model achieved 82% accuracy during this single run. Since the final classifier predicts on a sentence-by-sentence basis instead of for each individual pronoun, the maximum possible performance for this model would be 99.8%. A breakdown of the model's false positive, false negative, and correct classifications over 10 runs on the Medium corpus is charted in Figure 3.

Percentage of Results		
Predicted ↓ Actual →	Formal	Informal
Formal	34.0%	12.1%
Informal	10.6%	43.2%

Figure 3. Table detailing percentages of correct, false positive, and false negative results.

The green cells in Figure 3 represent results that were correctly classified, for a total of 77.2% with the Medium corpus, and the red cells represent incorrectly classified results. The percentage of formal pronouns classified correctly was 76.2%, almost equivalent to the 78.1%

of informal pronouns that were classified correctly. This indicates that the model's performance is not biased more strongly on one classification than the other.

7 Conclusions

In the final model, unstructured contextual features were used to improve significantly on the performance of the baseline classifiers. Various aspects of the model improved accuracy to differing degrees: while the addition of bigrams was a marginal improvement, the elimination of stopwords drastically increased performance. Among the most decisive features were words including "Herr" ("Mr."), "Frau ("Mrs."), "Fraulein" ("Miss"), and "Junge" ("Boy"), indicating that the model was successful in picking out words that are indicative of formality. Though the results presented involve an expected 50-50 split of the test and training sets, given that the model's accuracy correlates positively with the size of the corpus, it is possible that performance could be further improved by allocating a larger portion of the novels to the training set.

Though these models explored only a few types of features that affect the T/V dichotomy, the project has formalized a solid foundation of a problem, corpus, and baselines for future efforts to build upon.

8 Future Work

Much potential remains to improve upon the context-based classifier. In particular, it would be interesting to observe how the addition of more structured sentence features could affect results. Features based on part-of-speech could possibly increase accuracy, as there is likely a fairly strong association between formality and sentence adjectives or names/titles. Similarly, extracting features from parse trees could enable words more directly associated with the pronoun to be classified to be weighted more heavily. This would require breaking out of the sentence classifying context to assign each pronoun independently, or using a voting algorithm to determine the sentence's formality based on the classifications of the individual second-person pronouns it contains. Due to limited training data for German in these contexts and the fact that such parsing is very resource intensive, these features have not been added to the first version of the classifier; however, they could likely improve accuracy if implemented in future iterations.

Another much more challenging problem, but one which could drastically improve classification, would be to model the social graph of entire novels from a broader perspective than sentence-by-sentence. If a system could accurately identify the speaker and addressee for each dialogue containing a second-person address, it could learn to predict future addresses between any pair based on their previous interactions.

9 References

Vinodh Krishnan and Jacob Eisenstein. “You’re Mr. Lebowski, I’m the dude”: Inducing address term formality in signed social networks. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1616–1626, 2015.

Kelly Peterson, Matt Hohensee, and Fei Xia. Email formality in the workplace: A case study on the Enron corpus. In Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 86–95, 2011.

Manaal Faruqui and Sebastian Pado. 2011. ”I Thou Thee, Thou Traitor”: Predicting Formal vs. Informal Address in English Literature. In Proceedings of the Association for Computational Linguistics (ACL), pages 467–472, Portland, OR.

Manaal Faruqui and Sebastian Pado. 2012. Towards a model of formal and informal address in english. In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), pages 623–633.

Google. “Google Translate.” Accessed Dec 7, 2015. translate.google.com/.

Project Gutenberg. “Free ebooks by Project Gutenberg.” Accessed Nov 1, 2015. www.gutenberg.org/wiki/Main_Page.

Kaggle. “Bag of Words Meets Bags of Popcorn.” Accessed Nov 20, 2015. www.kaggle.com/c/word2vec-nlp-tutorial/details/part-1-for-beginners-bag-of-words.

Scikit Learn. “Count Vectorizer.” Accessed Nov 20, 2015. scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

10 Appendix

Project Proposal:
https://docs.google.com/document/d/10mHTlB0RJXF807F4Y0WJ8Pbj1aDLI1vpS-BaJvQQLEQ/edit?usp=sharing

Slides for Poster Session:
https://docs.google.com/presentation/d/1PthjCBMc3tXvmDnR46uvCnihio_YhhuF8Lo2M2o7DBc/edit?usp=sharing

Code: https://github.mit.edu/kasmus/nlp-project