Y'all should read this!
Identifying Plurality in Second-Person Personal Pronouns in English Texts

Gabriel Stanovsky
University of Washington
Allen Institute for Artificial Intelligence
gabis@allenai.org

Ronen Tamari*
Hebrew University of Jerusalem
ronent@cs.huji.ac.il

Abstract

Distinguishing between singular and plural "you" in English is a challenging task which has potential for downstream applications, such as machine translation or coreference resolution. While formal written English does not distinguish between these cases, other languages (such as Spanish), as well as other dialects of English (via phrases such as "y'all"), do make this distinction. We make use of this to obtain distantly-supervised labels for the task on a large-scale in two domains. Following, we train a model to distinguish between the single/plural 'you', finding that although in-domain training achieves reasonable accuracy ($\geq$ 77%), there is still a lot of room for improvement, especially in the domain-transfer scenario, which proves extremely challenging. Our code and data are publicly available.[1]

1 Introduction

The second-person personal pronoun (e.g., "you" in English) is used by a speaker when referring to active participants in a dialog or an event. Various languages, such as Spanish, Hebrew, or French, have different words to distinguish between singular "you" (referring to a single participant) and plural "you" (for multiple participants). Traditionally, English has made this distinction as well. The now archaic "thou" indicated singular second-person, while "you" was reserved for plural uses. The last several hundred years, however, have seen modern formal written English largely abandoning this distinction, conflating both meanings into an ambiguous "all-purpose you" (Maynor, 2000).

In this work, we are interested in the following research question: *How can we automatically disambiguate between singular and plural "you"?*

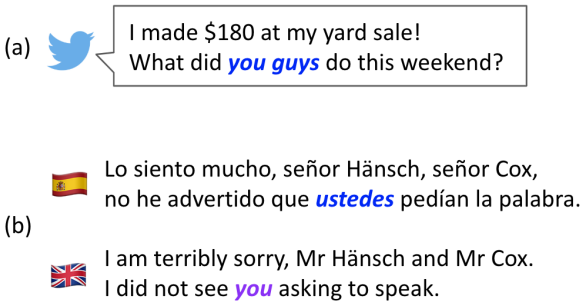

Figure 1: We use two sources for distant-supervision for singular (marked in purple) versus plural (marked in blue) second person pronouns: (a) we find colloquial uses on Twitter, and (b) through alignment with Spanish, which formally distinguishes between the cases.

While this topic has received much attention in linguistic literature (Jochnowitz, 1983; Tillery and Bailey, 1998; Maynor, 2000; Haspelmath, 2013; Molina, 2016), it has been largely unexplored in the context of computational linguistics, despite its potential benefits for downstream natural language processing (NLP) tasks. For example, distinguishing between singular and plural "you" can serve as additional signal when translating between English and a language which formally makes this distinction. See Figure 2 where an error in interpreting a plural "you" in the source English text results in a non-grammatical Hebrew translation. This example can be amended by replacing "you" with the informal "you guys".

To tackle this task, we create two large-scale corpora annotated with distantly-supervised binary labels distinguishing between singular and plural "you" in two different domains (see Figure 1). First, we regard Twitter as a noisy corpus for informal English and automatically identify speakers who make use of an informal form of the English plural "you", such as "y'all" or "you

* Work done during an internship at the Allen Institute for Artificial Intelligence.

[1] https://github.com/gabrielStanovsky/yall

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 375–380
Hong Kong, Nov 4, 2019.

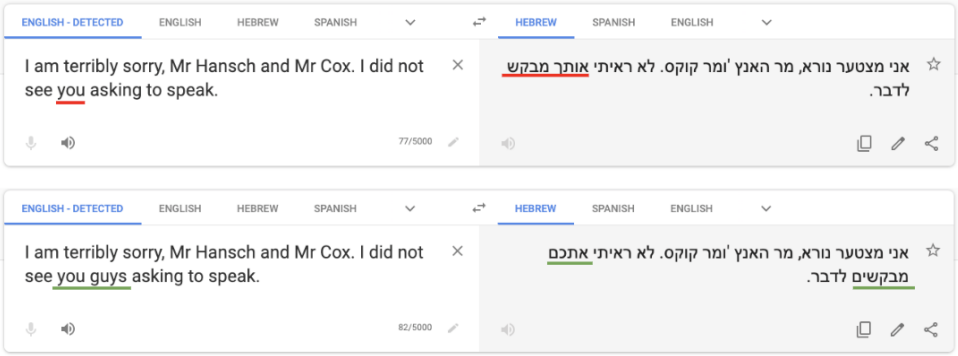

Figure 2: An example translation from English to Hebrew (Google Translate, Aug. 21, 2019). The first sentence depicts wrong interpretation of "you" resulting in a non-grammatical Hebrew translation, due to wrong inflections of pronoun and verb (marked in red). Both issues are fixed when changing "you" to "you guys" in English in the second example (marked in green).

Domain	Example	Plurality
Twitter	# goodnight #twittersphere <3 I love **y'all**! Including @anonimized. Even if she hates me. <3	*Plural*
(masked)	# goodnight #twittersphere <3 I love **you**! Including @anonimized. Even if she hates me. <3	
Twitter	! @anonimized, Happy anniversary of entering the world! Look how much **you** have done!	*Singular*
Europarl	I am terribly sorry, Mr Hansch and Mr Cox. I did not see **you** asking to speak.	*Plural*
Europarl	I should be very grateful, Mrs Schroedter, if **you** would actually include this proposed amendment in the part relating to subsidiarity in your positive deliberations.	*Singular*

Table 1: Examples from our two domains. Twitter is informal, includes hashtags, mentions (anonymized here), and plural "you" (e.g., "y'all" in the first example), which we mask as a generic "you" as shown in the second row. In contrast, Europarl is formal and "you" is used for plural (third example), as well as singular uses (last example).

guys", which are common in American English speaking communities (Katz, 2016). We record a *plurality* binary label, and mask the tweet by replacing these with the generic "you". Second, we use the Europarl English-Spanish parallel corpus (Koehn, 2005), and identify cases where the formal plural Spanish second-person pronoun aligns with "you" in the English text.

Finally, we fine-tune BERT (Devlin et al., 2018) on each of these corpora. We find that contextual cues indeed allow our model to recover the correct intended use in more than 77% of the instances, when tested in-domain. Out-of-domain performance drops significantly, doing only slightly better than a random coin toss. This could indicate that models are learning surface cues which are highly domain-dependent and do not generalize well.

Future work can make use of our corpus and techniques to collect more data for this task, as well as incorporating similar predictors in downstream tasks.

2 Task Definition

Given the word "you" marked in an input text, the task of *plurality identification* is to make a binary decision whether this utterance refers to:

- A single entity, such as the examples in rows 2 or 4 in Table 1.
- Multiple participants, such as those referred to in the third line in Table 1.

In the following sections we collect data for this task and develop models for its automatic prediction.

3 Distant Supervision: The *y'all* Corpus

Manually collecting data for this task on a large-scale is an expensive and involved process. Instead, we employ different techniques to obtain distantly supervised labels in two domains, as

	Twitter	Europarl
Train	58963	11249
Dev	7370	1405
Test	7370	1405
Total	**73703**	**14059**

Table 2: Number of instances in our two corpora. Each of the partitions (train, dev, test) is equally split between plural and singular second-person personal pronouns.

elaborated below. These are then randomly split between train (80% of the data), development, and test (10% each). See Table 2 for details about each of these datasets, which we make publicly available.

3.1 The Twitter Domain

As mentioned in the Introduction, English speaking communities tend to maintain the singular versus plural "you" distinction by introducing idiosyncratic phrases which specifically indicate a plural pronoun, while reserving "you" for the singular use-case. We operationalize this observation on a large Twitter corpus (Cheng et al., 2010) in the following manner:

- First, we identify speakers who use an informal plural "you" at least once in any of their tweets.[2]
- Following, we assume that these users speak an English dialect which distinguishes between singular and plural second-person pronouns, interpreting their "you" as a singular pronoun. See the first two tweets in Table 1, for an example of these two uses by the same user.
- Finally, we mask out the plural pronoun in each of their tweets by replacing it with a generic "you" (see the second row in Table 1). This allows us to test whether models can subsequently rely on contextual cues to recover the original intention.

This process yields about 36K plural instances, which we augment with 36K singular instances from the same users, to obtain a corpus which is balanced between the two classes.

[2]We use a fixed list of informal plural "you", including *you guys*, *y'all* and *youse*. See https://en.wikipedia.org/wiki/You#Informal_plural_forms for the complete list.

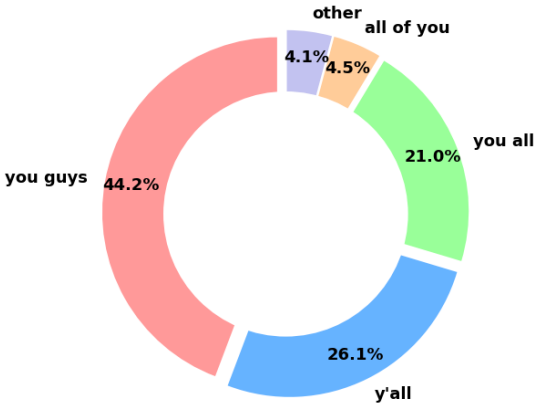

Figure 3: Histogram distribution of informal plural "you" forms in the development partition of our Twitter corpus.

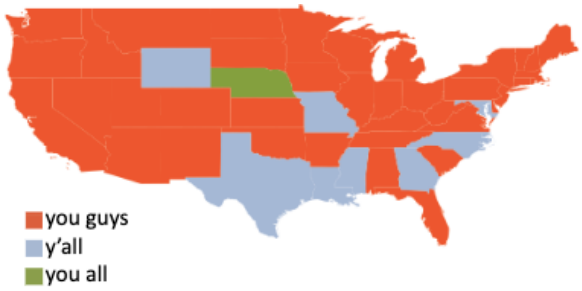

Figure 4: Variation in the most common phrase used for plural "you" in our Twitter corpus across states in the continental United States.

Data analysis Our Twitter corpus was composed of U.S. based users, and included geolocations for 36.8K of the plural tweets. This allows for several interesting observations. First, Figure 3 shows the distribution of informal plural "you" phrases in our corpus (before masking). Second, using the tweets geo-location, we can plot the geographical variation in usage. Figure 4 shows the most common term for plural "you" in each state in the continental United States. While most of the U.S. prefers "you guys", southern states (such as Texas or Louisiana) prefer "y'all". Katz (2016) reached similar conclusions through a large-scale online survey. Interestingly, our survey of Twitter usage differs from theirs for several Midwestern states, such as Wyoming or Nebraska.

Quality estimation. We evaluate the assumption we made above (i.e., that users reserve "you" for the singular case) by asking an English native-speaker to annotate a sample of 100 singular "you" instances from our Twitter corpus. In 70% of the instances, the annotator agreed that these indeed

represent a singular "you", leaving the other 30% to either ambiguous or inconsistent usage (i.e., sometimes "you" *is* used for a plural use-case). Overall, while there is a non-negligible amount of noise in the Twitter domain, in Section 5 we show that the signal is strong enough for models to pick up on and achieve good accuracy.

3.2 The Europarl Domain

Another method to obtain distant supervision for this task is through an alignment with a language which distinguishes between the two usages of the pronoun. To that end, we use the Spanish and English parallel texts available as part of Europarl (Koehn, 2005), a large corpus containing aligned sentences from meeting transcripts of the European parliament.

As these texts originate from a formal setting, we expect to find much less colloquial phrases. Indeed, the term "y'all" does not appear at all, while "you guys" appears only twice in about 2 million sentences. Instead, we rely on the gold alignment with Spanish sentences, which have a formal plural "you" - *ustedes* or *vosotros*. We find Spanish sentences which have exactly one plural "you" and which aligns with an English sentence containing exactly one "you". This process yields about 7K sentences which we mark with a "plural" label. Similarly to the Twitter domain, we augment these with the same amount of singular "you", found in the same manner; by tracing a Spanish singular "you" to a single English "you". This process yields a balanced binary corpus.

Quality estimation We sampled 100 instances from the Europarl domain to estimate the quality of our binary labels. Unlike the Twitter domain, in Europarl we rely on gold alignments and cleaner text. As a result, we found that about 90% of the labels agree with a human annotator, while the remaining 10% were ambiguous.

3.3 Discussion

The distinction between plural and singular "you" in English is an instance of a more general phenomenon. Namely, semantic features are expressed in varying degrees of lexical or grammatical explicitness across different languages.

For instance, languages vary in grammatical tense-marking (Wolfram, 1985), from languages with no morphological tense, such as Mandarin (Wang and Sun, 2015), to languages with 9 different tense-marking morphological inflections (Comrie, 1985). Similarly, languages vary in gender-marking in pronouns, from genderless Turkish, to languages with six genders or more (Awde and Galaev, 1997).

The two data collection methods we presented here, finding colloquial explicit utterances on social media, and alignment with another language, may also be applicable to some of these phenomena and present an interesting avenue for future work.

test → **train ↓**	**Europarl**	**Twitter**
Europarl	77.1	56.8
Twitter	56.3	**83.1**
Joint	**77.5**	82.8

Table 3: Accuracy (percent of correct predictions) of our fine-tuned BERT model, tested both in- and out-of-domain. Rows indicate train corpus, columns indicate test corpus. Bold numbers indicate best performance on test corpus.

4 Model

We fine-tune the BERT-large pretrained embeddings (Devlin et al., 2018)[3] on the training partition of each of our domains (Twitter and Europarl), as well as on a concatenation of both domains (Joint). We then classify based on the [CLS] token in each of these instances. We use a fixed learning rate of $2e-5$ and a batch size of 24. Training for 10 Epochs on a Titan X GPU took about 3 hours for the Twitter domain, 2 hours for the Europarl domain and roughly 5 hours for the Joint model.

5 Evaluation

We test models trained both in and out of domain for both parts of our dataset (Twitter and Europarl) as well as a joint model, trained on both parts of the dataset. We use accuracy (percent of correct predictions), as our dataset is binary and both classes are symmetric and evenly distributed. Our main findings are shown in Table 3. Following, several observations can be made.

[3]Using Hugging Face's implementation: https://github.com/huggingface/pytorch-transformers

In-domain performance does significantly better than chance. For both domains, BERT achieves more than 77% accuracy. Indicating that the contextual cues in both domains are meaningful enough to capture correlations with plural and singular uses.

Out-of-domain performance is significantly degraded. We see significant drop in performance when testing either model on the other part of the dataset. Both models do only slightly better than chance. This may indicate that the cues for plurality are vastly different between the two domains, probably due to differences in vocabulary, tone, or formality.

Training jointly on the two domains maintains good performance, but does not improve upon it. A model trained on both the Twitter and Europarl domains achieves the in-domain performance of each of the individual in-domain models, but does not improve over them. This may indicate that while BERT is expressive enough to model both domains, it only picks up on surface cues in each and does not generalize across domains. As a result, robustness is questionable for out-of-domain instances.

6 Related Work

Several previous works have touched on related topics. A few works developed models for understanding the second-person pronoun within coreference resolution frameworks (Purver et al., 2009; Zhou and Choi, 2018). Perhaps most related to our work is Gupta et al. (2007), who have tackled the orthogonal problem of disambiguation between generic (or editorial) "you" and referential "you".

To the best of our knowledge, we are the first to deal with plurality identification in second-person personal pronouns in English.

7 Conclusion and Future Work

We presented the first corpus for the identification of plurality in second-person personal pronouns in English texts. Labels were collected on a large scale from two domains (Twitter and Europarl) using different distant-supervision techniques.

Following, a BERT model was fine-tuned on the labeled data, showing that while models achieve reasonable in-domain performance, they significantly suffer from domain transfer, degrading performance close to random chance. Interesting avenues for future work may be to extend this data to new domains, develop more complex models for the task (which may achieve better cross-domain performance), and integrating plurality models in downstream tasks, such as machine translation or coreference resolution.

Acknowledgements

We thank the anonymous reviewers for their many helpful comments and suggestions.

References

Nicholas Awde and Muhammad Galaev. 1997. *Chechen-English English-Chechen: Dictionary and Phrasebook*. Hippocrene Books.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *CIKM*.

Bernard Comrie. 1985. *Tense*, volume 17. Cambridge university press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Surabhi Gupta, Matthew Purver, and Dan Jurafsky. 2007. Disambiguating between generic and referential you in dialog. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 105–108.

Martin Haspelmath. 2013. Politeness distinctions in second-person pronouns. In *The Atlas of Pidgin and Creole Language Structures*, pages 68–71. Oxford Univ. Pr.

George Jochnowitz. 1983. Another view of you guys. *American Speech*, 58(1):68–70.

Josh Katz. 2016. *Speaking American: How Y'all, Youse, and You Guys Talk: A Visual Guide*. Houghton Mifflin Harcourt.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation.

Natalie Maynor. 2000. Battle of the pronouns: Y'all versus you-guys. *American speech*, 75(4):416–418.

Ralph Molina. 2016. *Y'all, This Paper Is Crazy Interesting: A Study of Variation in US English*. Ph.D. thesis.

Matthew Purver, Raquel Fernández, Matthew Frampton, and Stanley Peters. 2009. Cascaded lexicalised classifiers for second-person reference resolution.

In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 306–309. Association for Computational Linguistics.

Jan Tillery and Guy Bailey. 1998. Yall in oklahoma.

William SY Wang and Chaofen Sun. 2015. *The Oxford handbook of Chinese linguistics*. Oxford University Press.

Walt Wolfram. 1985. Variability in tense marking: A case for the obvious. *Language Learning*, 35(2):229–253.

Ethan Zhou and Jinho D Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34.