An Empirical Analysis of Formality in Online Communication: Supplementary Material

Ellie Pavlick
University of Pennsylvania[*]
epavlick@seas.upenn.edu

Joel Tetreault
Yahoo Labs
tetreaul@yahoo-inc.com

1 Annotation

Our exact annotation instructions are given in Figure 1. Workers were asked to rank sentences on the following scale:

- 1- Very Informal
- 2- Somewhat Informal
- 3- Slightly Informal
- 4- Neither formal nor informal
- 5- Slightly Formal
- 6- Somewhat Formal
- 7- Very Formal

Additionally, workers were given an an "I cannot tell" option, which they are instructed to use if the sentence contains gibberish or is not in English. We also provide a check box for "machine generated" in which workers are asked to indicate if it appears that the sentence comes from some automatically generated or batch content, something that is more relevant for email data than for Answers. Checking the "machine generated" box does not take the place of a formality rating. Note, for cleaner computations, we eventually shift the 1 to 7 scale to a -3 to 3 scale.

Quality control. We perform a very rudimentary quality control in order to remove spammers from our annotation. We manually label a small number of sentences (43 formal and 39 informal examples) from the emails and use these items as gold standard. We chose examples which we feel fall clearly at one end of the spectrum or the other, noting that the task is inherently subjective. We grade the controls on a binary scale (e.g. if the gold standard label is "informal", any rating the worker supplies below 0 will be correct). Specifically, we approve each worker's first 10 HITs automatically. Afterward, workers receive regular notifications of their current accuracy against our controls, and a warning that we do not approve HITs which fall below a random baseline on our controls. Note we don't use the quality control as a reliable measure of a single worker's qualification, or even of the workers' collective ability to perform the task. Rather, it serves to "scare away" spammer workers, and in practice, only a few workers (6%) have any work rejected. Our workers achieve a mean accuracy 82% and a median accuracy of 90%.

Aggregation. Before taking the mean score for a sentence, we remove workers from the list who chose the "I cannot tell" option for that sentence (used in 10% of cases). We also remove workers who at any point had a HIT rejected (e.g. after more than 10 HITs, their accuracy was at-or-below random guessing). If after removing these judgements, the total number of reliable judgements for the sentence is less than 3, we throw away the item. These criteria omit 15% of our sentences, leaving a total of 1,701 remaining email sentences and 4,977 remaining answers sentences.

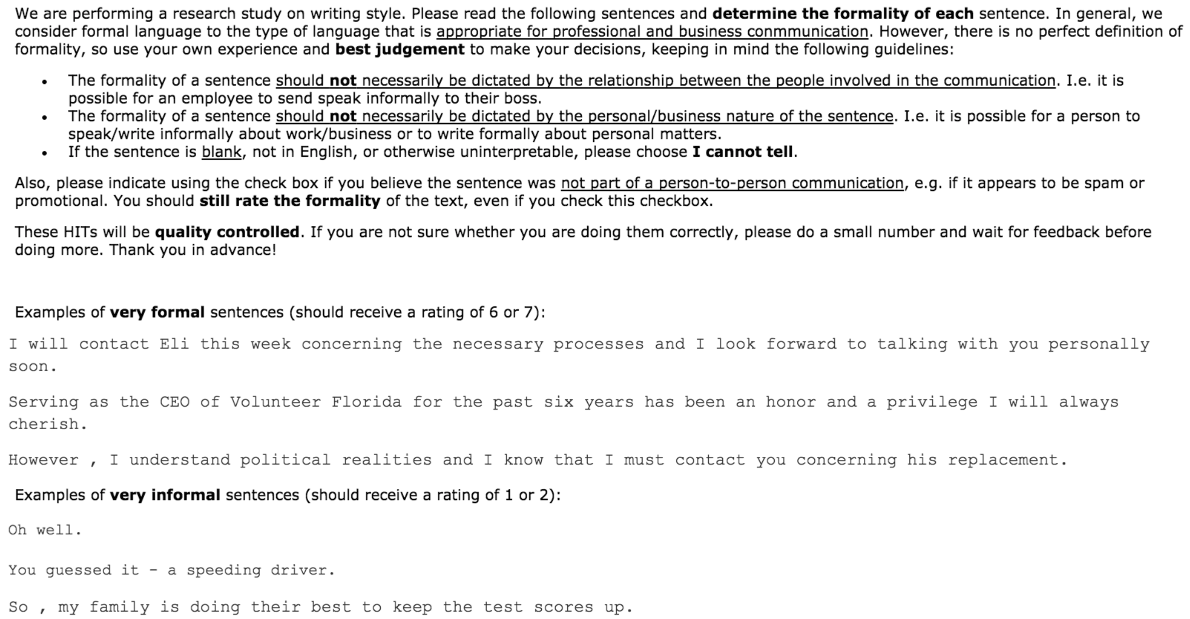
2 Features

We use Stanford CoreNLP for all of our linguistic preprocessing. Before extracting features, we replace URLs and email addresses with special tokens.

Casing

- Number of capitalized words, not including 'I'

[*] Research performed while at Yahoo Labs.



We are performing a research study on writing style. Please read the following sentences and **determine the formality of each** sentence. In general, we consider formal language to the type of language that is appropriate for professional and business conmmunication. However, there is no perfect definition of formality, so use your own experience and **best judgement** to make your decisions, keeping in mind the following guidelines:

- The formality of a sentence should **not** necessarily be dictated by the relationship between the people involved in the communication. I.e. it is possible for an employee to send speak informally to their boss.
- The formality of a sentence should **not** necessarily be dictated by the personal/business nature of the sentence. I.e. it is possible for a person to speak/write informally about work/business or to write formally about personal matters.
- If the sentence is blank, not in English, or otherwise uninterpretable, please choose **I cannot tell**.

Also, please indicate using the check box if you believe the sentence was not part of a person-to-person communication, e.g. if it appears to be spam or promotional. You should **still rate the formality** of the text, even if you check this checkbox.

These HITs will be **quality controlled**. If you are not sure whether you are doing them correctly, please do a small number and wait for feedback before doing more. Thank you in advance!

Examples of **very formal** sentences (should receive a rating of 6 or 7):

I will contact Eli this week concerning the necessary processes and I look forward to talking with you personally soon.

Serving as the CEO of Volunteer Florida for the past six years has been an honor and a privilege I will always cherish.

However , I understand political realities and I know that I must contact you concerning his replacement.

Examples of **very informal** sentences (should receive a rating of 1 or 2):

Oh well.

You guessed it - a speeding driver.

So , my family is doing their best to keep the test scores up.

Figure 1: Annotation guidelines shown to Turkers.

- Binary indicator for whether the sentence is all lower case
- Binary indicator for whether the first word is capitalized

Punctuation

- Number of '!' in the sentence
- Number of '...' in the sentence
- Number of '?' in the sentence

Constituency

- The depth of the constituency parse tree, normalized by the length of the sentence
- The number of times each parse tree production rule occurs in the sentence, normalized by the length of the sentence. We do not include terminal symbols (i.e. lexical items) in the productions.

Dependency Given a dependency parse that consists of tuples connecting a governor word (gov) to a dependent word (dep) via a dependency type (typ), we include the following as 1-hot features:

- (gov, typ, dep) tuples with gov and dep backed off to their POS tags
- (gov, typ) tuples with gov and dep backed off to their POS tags
- (typ, dep) tuples with gov and dep backed off to their POS tags
- (gov dep) tuples with gov and dep backed off to their POS tags

Entity

- Entity types (e.g. 'PERSON', 'LOCATION') occuring in the sentence, as 1-hot features
- For PERSON entities, the average length (in characters) of the mentions

Lexical

- Number of words in the sentence
- Average word length, in characters
- Average word log frequency according the Google Ngram corpus, not including stop words
- Average formality score, as computed in Pavlick and Nenkova (2015)
- Number of contractions, normalized by the length of the sentence

Ngrams

- Unigrams, as 1-hot features
- Bigrams, as 1-hot features
- Trigrams, as 1-hot features

Part-of-speech

- The number of occurrences of each POS tag in the sentence, normalized by the length of the sentence

Readability

- Length of the sentence, in words
- Length of the sentence, in characters
- Flesch-Kincaid Grade Level, defined as: $(0.39 \times \frac{\#words}{\#sentences}) + (11.8 \times \frac{\#syllables}{\#words}) - 15.59$

Subjectivity

- Number of passive constructions (i.e. "be" verb followed by VBN), normalized by the length of the sentence
- Number of hedge words, normalized by the length of the sentence. Based on a list of hedge words taken from a combination of online sources.
- Number of 1st person pronouns, normalized by the length of the sentence
- Number of 3rd person pronouns, normalized by the length of the sentence
- Subjectivity score of the sentence, according the the TextBlob[1] sentiment module
- Binary indicator for whether the sentiment is positive or negative, according the the TextBlob sentiment module

Word2Vec

- We compute the sentence vector to be the average of the precomputed w2v word vectors[2] in the sentence. Words for which there is not pre-computed vector are skipped.

References

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224, Denver, Colorado, May–June. Association for Computational Linguistics.

[1] `https://textblob.readthedocs.org/en/dev/`

[2] `https://code.google.com/p/word2vec/`