Contextualized Embeddings for Enriching Linguistic Analyses on Politeness

Ahmad Aljanaideh and **Eric Fosler-Lussier**
Department of Computer Science and Engineering
The Ohio State University

Marie-Catherine de Marneffe
Department of Linguistics
The Ohio State University

Abstract

Linguistic analyses in natural language processing (NLP) have often been performed around the static notion of words where the context (surrounding words) is not considered. For example, previous analyses on politeness have focused on comparing the use of static words such as personal pronouns across (im)polite requests without taking the context of those words into account. Current word embeddings in NLP do capture context and thus can be leveraged to enrich linguistic analyses. In this work, we introduce a model which leverages the pre-trained BERT model to cluster contextualized representations of a word based on (1) the context in which the word appears and (2) the labels of items the word occurs in. Using politeness as case study, this model is able to automatically discover interpretable, fine-grained context patterns of words, some of which align with existing theories on politeness. Our model further discovers novel finer-grained patterns associated with (im)polite language. For example, the word *please* can occur in impolite contexts that are predictable from BERT clustering. The approach proposed here is validated by showing that features based on fine-grained patterns inferred from the clustering improve over politeness-word baselines.

1 Introduction

Linguistic analyses have often been computationally performed around the static notion of words or word categorization methods (e.g. LIWC) where the context of words is not taken into account. Previous work on politeness (Danescu-Niculescu-Mizil et al., 2013) and gender (Bamman et al., 2014) have focused on comparing the use of non-contextual broad word categories such as personal pronouns across different categories/groups. For example, Bamman et al. (2014) found that women use more pronoun words than men. Danescu-Niculescu-Mizil et al. (2013) found that requests which contain a hedge word (e.g. *think*) are more likely to be perceived as polite than impolite. However, words often occur in many different contexts and thus analyzing them statically hides cues which can potentially enrich our understanding of the phenomenon being studied. As opposed to static word embeddings which provide the same representation for a word regardless of its context (i.a. Mikolov et al., 2013a; Mikolov et al., 2013b; Pennington et al., 2014), the BERT (Devlin et al., 2018) and ELMo (Peters et al., 2018) models provide methods for extracting pre-trained contextualized word representations. By leveraging contextualized representations, linguistic theories can be validated and enriched.

Given a dataset annotated for a downstream task, we build a model which automatically discovers fine-grained context patterns of words. Discovering such patterns provides insight into the phenomenon the task attempts to model. We use pre-trained BERT embeddings and exploit the fact that words which occur in similar contexts tend to have similar representations. Our model takes as input contextualized representations for a given word and splits them into different clusters based on the context in which the word appeared and the labels of the items the word belongs to. By doing so, we are able to identify the different contexts in which the word appears and how they correlate with the categories of the task being studied.

Proceedings of the 28th International Conference on Computational Linguistics, pages 2181–2190
Barcelona, Spain (Online), December 8-13, 2020

We use politeness as case study of a linguistic phenomenon. Existing computational work on politeness developed feature-based (Danescu-Niculescu-Mizil et al., 2013) and neural (Aubakirova & Bansal, 2016, Niu & Bansal, 2018) models which detect if a natural language request (e.g. *Can you please tell me how to do that?*) is polite or impolite. Danescu-Niculescu-Mizil et al. (2013) developed a computational tool driven by existing theories in the literature on politeness (Brown & Levinson, 1987). These theories highlighted linguistic constructions that speakers use to reduce the burden on the addressee by sounding indirect (e.g. *Could you please [...]*). Danescu-Niculescu-Mizil et al. (2013) showed further that, for some words, their position in the request plays a role in whether the request will be perceived as polite or impolite. For example, even though *please* is considered polite, they show that requests starting with *please* (e.g. *Please explain this to me.*) are more likely to be perceived as impolite while requests with *please* in the middle (e.g. *Could you please help me out?*) are more likely to be perceived as polite. We take an exploratory approach that is driven to discover fine-grained word patterns that encode the surrounding context as opposed to simply the position of the word. The proposed model automatically discovers patterns that have already been discussed in the literature in addition to multiple novel ones. For example, the model discovers fine-grained context patterns of the word *please* one of which include occurrences of *please* in the middle of impolite requests (e.g. *Would you please stop?*) signaling that our model does not simply encode the position of the word but also the surrounding context. We validate the proposed model by showing that features based on the fine-grained patterns that it discovers outperform current feature-based politeness models.

Uncovering contextual information on how words are used can enhance our understanding of the task/phenomena being studied. It can also help in enriching the linguistic theories associated with the task. Striving for informativeness and interpretability can make models more useful in downstream applications.

2 Discovering Fine-grained Context Patterns

Given a word w appearing in multiple labeled text items, we build a model to automatically discover clusters such that each cluster contains use cases of the word which (1) are contextually similar in the way the word w is used, and (2) belong to items which are for the most part of the same label. For example, the word *think* can appear in different contexts across items which express positive sentiment such as *I think this movie is great!* and *Great movie! Do you think it will win the Oscars?*. The goal is to automate the process of splitting multiple use cases of a word into different clusters. This helps uncover information about how a word can be used and how its usage relates to the labels of the task at hand, thus enriching our understanding of the phenomenon being studied. The resulting clusters can also be used to extract fine-grained, contextualized unigram features for the target task. We now describe the process by which the clustering is performed.

We not only want to encourage contextual similarity of the word's usage across elements of each cluster, but also encourage clusters to have high purity determined based on the labels of items the word appeared in. Striving for such criteria simplifies the process of interpreting the resulting clusters. We use a Decision Tree (Quinlan, 1986)-like approach to obtain clusters with such criteria.

We apply the following process on each word in a given vocabulary *V*. First, the pre-trained BERT-base representations (Devlin et al., 2018) of the word in the training set of items are extracted. For every word, BERT-base produces 12 representations that correspond to 12 layers. We average the representations of the last 6 layers for every word. Given that BERT embeddings are contextualized, the same word will have a different representation depending on the context in which it appears. We label each contextualized representation with the same label as the item it was extracted from. We denote the representations of the word in the training set as the *parent cluster*. Elements (contextualized representations) of the parent cluster are recursively split into two *child clusters*. In order to determine the cluster membership of each element, we sample two *centroid* representations from the parent cluster and map each of its representations to the closer of the two centroids. The proximity measure that we use is the euclidean distance. We try all possible pairs of centroids and select the pair which results in the binary split which

maximizes the information gain (IG). The information gain is calculated as in (1):

$$IG_w = E(pc_w) - \sum_{cc_w}(\frac{\#cc_w}{\#pc_w}E(cc_w)) \quad (1)$$

where pc_w represents the labels of the elements of the parent cluster, cc_w represents the labels of the elements in one of the child clusters, $E(pc_w)$ and $E(cc_w)$ represent the entropy values corresponding to the parent cluster and any of the child clusters respectively, and $\#pc_w$ and $\#cc_w$ represent the number of elements in the parent cluster and any of the child clusters.

We recursively repeat the splitting for each child cluster until a cluster of 100% purity is obtained (i.e. a cluster constituted only of elements that have the same label), or a certain depth *d* is reached. The depth is set based on the common logarithm of the word's frequency in the training set. For words with a frequency of less than 10, we set *d* to 1. For words with a frequency between 10 and 100, we set *d* to 2. For words with a frequency between 100 and 1000, we set *d* to 3. For words with a frequency higher than 1000, we set *d* to 4. The intuition is that travelling down the tree for frequently occurring words will allow the discovery of fine-grained patterns that correspond to their usage. Therefore, it is wise to strive for extracting a higher number of clusters for frequently occurring words than for words with a low frequency. Moreover, it is possible that a split results in some clusters with 100% purity and others with less purity. For very frequently occurring words, we randomly sample 20k pairs of centroids as trying all possible ones is computationally expensive. This process is illustrated in Algorithm 1.

Algorithm 1

Input: BERT embeddings for word *w* instances in the training set
Output: *n* clusters

```
Label each word embedding with the label of the item the word belongs to
parent-cluster = BERT(w)
SplitCandidates = []
for all pairs of representations (c1,c2) in parent-cluster do
    cluster1 = {}, cluster2 = {}
    for each representation x in parent-cluster do
        if euclid_dist(c1, x) < euclid_dist(c2, x) then
            map x to cluster1
        else
            map x to cluster2
        end if
    end for
    Add cluster1 and cluster2 to SplitCandidates
end for
Select the best split(pair of clusters) in SplitCandidates which provides the maximum information gain (IG) calculated using
equation 1.
Re-do for each child cluster recursively.
The output of the algorithm are n leave clusters.
```

3 Application: Politeness Detection

The task that we choose as case study for our model is the task of detecting politeness in natural language requests. In Section 3.1, we describe the dataset used to train and evaluate our model. In Section 3.2, we describe existing politeness features. In Section 3.3, we describe the model instantiation and evaluate the model qualitatively.

3.1 Politeness Data

We use the two datasets provided by Danescu-Niculescu-Mizil et al. (2013). The first dataset includes requests taken from a collaborative Wikipedia forum, a place in which edits to Wikipedia articles are discussed. The second dataset includes requests taken from the Stack-Exchange forum, a place for users to ask questions on any topic. Each request is composed of a pair of sentences, one of which is a question where the speaker asks the addressee for a favor, and the other one can simply be any sentence (e.g. *I think the edits [...]*). Each request was labeled by five annotators indicating the request's level of politeness. The

annotations were then averaged and normalized to obtain a politeness score. Requests in the top quartile (in terms of scores) were labeled as *polite* while those in the bottom quartile were labeled as *impolite*. The final Wikipedia and Stack-Exchange datasets contain 2,176 and 3,302 requests respectively and both were balanced across the two classes. We use the 70-10-20 training/development/testing split from Niu & Bansal (2018).[1]

3.2 Politeness Features

Danescu-Niculescu-Mizil et al. (2013) developed a classifier which uses features inspired by theories of politeness (Brown & Levinson, 1987). Those theories include linguistic constructions speakers use to reduce the burden on the addressee (e.g. *Would you please ..*) and constructions speakers use to express positiveness towards the addressee (e.g. Thank you!, *Nice work!*). Table 1 shows the features that correspond to these politeness strategies. For each feature, we show the % of polite requests in the training set. Few of those features do encode some context (position of the word). For some words, Danescu-Niculescu-Mizil et al. (2013) found a relationship between the position of their occurrence and whether they were more likely to appear in polite or impolite requests. For example, requests starting the sentence with a first person pronoun (Row#1) are more likely to be perceived as polite than requests which contain the first person pronoun in any position in the sentence (Row#2). Both markers along with the plural first pronoun (Row#3) were more associated with polite requests (62%, 57% and 56%, respectively). Similarly, requests starting with a second person pronoun (Row#4) were much more likely to be perceived as impolite (22%) while requests which contain the second person pronoun in any position in the sentence (Row#5) were more associated with polite requests (55%). Requests starting with *please* (Row#6) were more likely to be perceived as impolite while requests containing *please* in the middle (Row#7) were more likely to be perceived as polite. Requests containing counterfactual modals (Row#8, e.g. *Could*) and indicative modals (Row#9, e.g. *can*) are more associated with polite requests (85% and 71%, respectively). Requests which contain hedges (Rows#10-11) such as *think* or *maybe* are more likely to be perceived as polite. Those two features were detected by matching words in each request against a list of hedge words from Hyland (2018). Requests containing direct questions (Row#12) or direct starts (Row#13) are more likely to be perceived as impolite. Direct questions were detected based on whether the request contains a question which uses a *wh*-interrogative (*what*, *who*, *why*, *how*), while direct starts were detected based on whether the request starts with *so*, *then*, etc. Other feature examples include gratitude, apologizing, greeting, and deference (Rows#14-17), all are associated with polite requests. Factuality (Row#18) is more associated with impolite requests, while the use of *by the way* (Row#19) is more associated with polite requests. Unsurprisingly, positive sentiment (Row#20) is associated with polite requests while negative sentiment (Row#21) is associated with impolite requests. Both features were detected by matching words in the request against sentiment word lists provided by Liu et al. (2005).

3.3 Model Instantiation and Qualitative Analysis of the Clusters

We train our model using the Wikipedia training set of requests. We run Algorithm 1 for every word in the training set. The algorithm outputs a set of clusters for every word. Each cluster contains contextualized representations of the word across multiple items.

3.3.1 Cluster Analysis

We perform qualitative analyses on clusters corresponding to words discussed in the literature (Danescu-Niculescu-Mizil et al., 2013). Table 2 shows examples of clusters. We show a subset of clusters for each word. For each cluster, we show its number of items, the % of polite items, and a set of examples. Each cluster is identified by the word, a subscript which indicates its id and a superscript which indicates the majority label of its members (- for impolite and + for polite). Below, we show what our clustering technique highlights for some of the feature words discussed by Danescu-Niculescu-Mizil et al. (2013).

I: While Danescu-Niculescu-Mizil et al. (2013) showed that requests which contain the first person pronoun *I* are slightly more likely to be perceived as polite, our model was able to discover an impolite

[1] Available at https://github.com/WolfNiu/polite-dialogue-generation.

Row#	Politeness strategy	%polite	Example
1	1st person pronoun start	62	**I**'ve replied to most of the concerns. Could you please look over them?
2	1st person pronoun	57	By the way, are you honestly ok with **me** bothering you like this?
3	1st person pronoun plural	56	When **we** go to the American Indian museum.
4	2st person pronoun start	22	**You** sure kill educating information. I wounder why?
5	2nd person pronoun	55	Could **you** post me a link to where it occurred?
6	Please start	31	**Please** don't mention other languages when we're discussing ...
7	Please	81	Would you **please** explain it to me?
8	Counterfactual modal	85	**Could** you give me more precise examples ?
9	Indicative modal	71	**Can** you show me how to do that?
10	Hedges	55	What do you **think** about creating a template for convoys ...
11	Hedges (nsubj)	60	**I think** you can ...
12	Direct Question	27	**What** hobbies do you have?
13	Direct Start	14	**So** you exchanged my suggestion with an...
14	Gratitude	95	**Thank** you so much!
15	Apologizing	85	**Sorry** but I am out of the loop ...
16	Greeting	86	**Hey** Leah, ...
17	Deference	96	**Nice** job!
18	Factuality words	25	**In fact**, ...
19	By the way	67	**By the way** , i noticed you said you ...
20	Positive sentiment	60	**Great** work!
21	Negative sentiment	41	I feel **sad** all the time ...

Table 1: Examples of politeness strategies from Danescu-Niculescu-Mizil et al. (2013).

context of using *I*. Cluster I_0^- contains contextualized representations of *I* that belong to mostly impolite requests using the first person pronoun *I*. The pattern exhibited in those requests is that they start with a sentence which contains the first person pronoun (e.g. *I didn't see your internal link*) and ends with a sentence asking a question in a direct tone (e.g. *What are you talking about*). On the other hand, Cluster I_1^+ contains contextualized representations of *I* that belong to mostly polite requests which contain this pronoun. Items belonging to this cluster usually started out politely (e.g. with a greeting), and ended with a question asked in a polite tone (e.g. *Would you like me to grant your account[...]?*).

You: While Danescu-Niculescu-Mizil et al. (2013) showed that requests which contain the second person pronoun *you* were more likely to be perceived as polite, our model was able to discover both highly polite and highly impolite contexts in which the second person pronoun *you* is used. Cluster You_0^- contains contextualized representations of *you* that belong to impolite requests which contain the second person pronoun *you*. An item belonging to this cluster mostly contained a question being asked in a tough/direct tone (e.g. *Are you familiar with such concepts?*). Cluster You_1^+ contains contextualized representations of *you* that belong to polite requests using the second person pronoun *you*. Items which belong to this cluster start out with a polite tone (e.g. *Thank you*) and then follow using a sentence containing *you* (e.g. *How do you think is the best way to resolve these issues?*).

Please: Cluster $Please_0^-$ contains contextualized representations of *please* that belong to impolite requests with *please* occurring at the start of the sentence. Cluster $Please_1^+$ contains contextualized representations of *please* that belong to polite requests which start out with a marker that is associated with polite requests (e.g. with a greeting or use of first person pronoun) and end with a sentence which contain *please* in the middle. Danescu-Niculescu-Mizil et al. (2013) did show that requests with *please* in the middle were more likely to be perceived polite while requests starting with *please* were more likely to be perceived as impolite. However, the proposed model discovered that when *please* occurs in the middle, the request it is appearing in can still be perceived as impolite (Cluster $Please_2^-$). In that case, the speaker would be using *please* with an a direct tone (e.g. *Would you please stop?*), or would be starting the request with a direct tone (*You have listed it as no source but the author asserts own work*) and then follows with a sentence with *please* in the middle (e.g. *Could you please clarify this?*). (Cluster $Please_2^-$).

Can: Cluster Can_0^+ contains contextualized representations of *can* that belong to mostly polite requests which start out politely (e.g. with a greeting) and then use the modal *can* with a second person pronoun. Previous work did show that the use of a modal with a second person pronoun is highly associated

with polite language (i.a. Danescu-Niculescu-Mizil et al., 2013,Aubakirova & Bansal, 2016). However, requests that use the modal *can* with a second person pronoun can still be perceived as impolite (Cluster Can_1^-). In that case, the speaker would start the request with a tough tone (e.g. *Are you saying that Scotty was being "hot-headed"?*) and then would ask the user directly (*can you please tell us what Scotty did that was "hot-headed"?*). The fact that occurrences of *can you* got assigned to different clusters indicates that the model looks beyond the local context of the word.

Analyses of *can* and *please* show that the proposed model is able to discover patterns which go beyond the local context of using a word. That is, context from the earlier part of the request is encoded. The fact that the proposed model uses contextualized BERT embeddings helps in identifying subtle patterns embodied by multiple markers (e.g. greeting in addition to using *please* in the middle, or starting the sentence with a direct tone and then following it with *please* in the middle).

Maybe: While Danescu-Niculescu-Mizil et al. (2013) showed that hedges are more likely to appear in polite requests than impolite ones, our proposed model was able to find fine-grained clusters with a majority of impolite items that do contain hedges. The hedge *Maybe* can be used as a part of an impolite request (e.g. *Maybe it would be better if we never talked directly to each other ever again*) (row $Maybe_0^-$) or a polite one (e.g. *Maybe we can meet*) (row $Maybe_1^+$) where the speaker signals an invite for collaboration.

These analyses qualitatively show that our approach re-discovers existing politeness markers (e.g. Clusters $Please_0^-$, $Please_1^+$, Can_0^+). The two impolite markers discovered by Aubakirova & Bansal (2016) using neural network visualization techniques, namely indefinite pronouns (e.g. *Am I missing something?*) and repeated punctuation (e.g. *Hello?????*), were also among the clusters generated for the words *something* and *?*. Our model further finds finer-grained markers that have not been discussed in the literature (e.g. Clusters $Please_2^-$, Can_1^-, $Maybe_0^-$, $Maybe_1^+$). The fact that our model is driven to discover fine-grained contextual patterns of word usages helped in identifying subtle ways of using and combining politeness words.

3.3.2 Cluster Error Analysis

We now perform error analysis to assess the quality of the clusters produced by our model. We looked at words in the development set of requests which were incorrectly mapped to a cluster: either the word appeared in a polite request but was mapped to a cluster with a majority of impolite elements, or conversely the word appeared is an impolite request but was mapped to a cluster with a majority of polite elements. The clusters that we use for this analysis are selected from Table 2. Table 3 shows examples of such incorrectly mapped items. For each cluster, we also show the % of items that got classified correctly (accuracy).

For Cluster I_0^-, 57% of items that were assigned to it were predicted correctly. The model incorrectly assigned the contextualized representation of *I* in the polite request *as you know even I use my Mobile-Opera for editing* to this cluster since the request starts with a direct tone (*Someone has Blocked the whole Opera mini Browser*). Cluster I_1^+ has an accuracy of 86%. The model incorrectly assigned the contextualized representation of *I* in the impolite request *I am just wondering why you selected the word winnings?* to this cluster since the request starts with a greeting.

Cluster You_0^- exhibited an accuracy of 81%. The model incorrectly assigned the contextualized representation of *you* in the polite request *What method (code) do you use [...]* to this cluster since the request contains a question being asked in a direct tone. Cluster You_1^+ exhibited an accuracy of 86%. The model incorrectly assigned the contextualized representation of *you* in the impolite request *Can you see the problem here?* to this cluster.

For Cluster $Please_0^-$, 67% of items that were assigned to it were predicted correctly. The model incorrectly assigned the contextualized representation of *please* in the polite request *Please take a look at [...]* in this cluster since the request contains the word *please* occurring at the start of the request. Cluster $Please_1^+$ exhibited a higher accuracy (83%). The model incorrectly assigned the contextualized representation of *please* in the impolite request *Which is it, please?* in this cluster since it contains mostly polite requests with *please* occurring in the middle or at the end.

Cluster id	n	% polite	Example	Label
I_0^-	236	33	**I** didn't see your internal link, I put it back. url isn't orphaned, what are you talking about?	-
			As **I** have said before, it is an issue of breadth vs. depth. Are you familiar with such concepts?	-
I_1^+	108	99	On a related note, I went through your recent article - work, and saw nothing wrong with your contributions. **I** was wondering, would you like me to grant your account	+
			Hi leah, is it working again? **I** had some difficulties with [...]	+
You_0^-	166	11	As I have said before, it is an issue of breadth vs. depth. Are **you** familiar with such concepts?	-
			I don't know why claim that **you** reverted vandalism when to me they appeared to be good faith edits. What exactly was the vandalism?	-
You_1^+	205	100	Thank you very much for your patience and for listening. How do **you** think is the best way to resolve these issues?	+
			Nice photo! Could **you** help answer a question at the reference desk related to the picture?	+
$Please_0^-$	11	0	It's not my job to add the "bronze age" material in the article on url, so **please** refrain from removing existing material just because you haven't made any effort to add the "bronze age" content. ok?	-
			Hi, although I agree with you, **please** check the following url in process. Why can't the darn thing work, in the first place?	-
$Please_1^+$	52	100	I have posted a question at url which you may be able to answer. Can you **please** return to that discussion to answer it?	+
			Hello again . could you move this to url, **please**?	+
$Please_2^-$	20	35	So far there are four editors on url who have called you on your edit-warring. Would you **please** stop?	-
			Are you saying that Scotty was being "hot-headed"? If so, can you **please** tell us what Scotty did that was "hot-headed"?	-
			I have absolutely no idea what you are talking about. **Please** be more precise?	-
			You have listed it as no source but the author asserts own work. Could you **please** clarify this?	+
Can_0^+	76	96	Hi, url to URL broke the templates. **Can** you fix that, please?	+
			I wasn't sure what you mean by making the title a clickable link. **Can** you show me how to do that?	+
			Hi. I uploaded three very nice images in the url, **can** you please tell me what breed could that dog be?	+
Can_1^-	12	25	Are you saying that Scotty was being "hot-headed"? If so, **can** you please tell us what Scotty did that was "hot-headed"?	-
			I don't think so, giano. **Can** we just once attempt to let this pass by without calling for the heads of people?	-
$Maybe_0^-$	11	0	**Maybe** it would be better if we never talked directly to each other ever again. Agreed?	-
			Maybe others haven't noticed - or are too busy. If you are thinking what I'm thinking, why am I reverting so many of your edits?	-
$Maybe_1^+$	8	100	I'm in Outer Mongolia from the 27th till the 10th, UB, Hovsgol, UB. **Maybe** we can meet?	+
			PS I am planning on starting an article on url. **Maybe** you'd like to help?	+

Table 2: Examples of clusters generated by our model. For each example cluster, we show its number of elements, the % of polite elements, and a set of example requests. We also show the label (+ for polite and - for impolite) for each example request.

Cluster id	% Polite	Accuracy	Example	Label
I_0^-	33	57	Someone has Blocked the whole Opera mini Browser, as you know even **I** use my Mobile-Opera for editing. Can you please help?	+
I_1^+	99	86	Hello [username], I am contacting you because in golf they put Money not Winnings as the standard on most leaderboards used at the end of tournaments. **I** am just wondering why you selected the word winnings?	-
You_0^-	11	81	My question about the truncation operation is to understand the generic algorithm for applying a ring (or un-ringing) a Dynkin node. What method (code) do **you** use to get the higher dimensions truncated based on the Coxeter Dynkin ringed nodes?,	+
You_1^+	100	86	"... then you have to at least meet me half-way. For instance, <url> described the episode as displaying "an exhilarating flair for rapid change of comic gear" and made commented positively on the scene." Can **you** see the problem here?	-
$Please_0^-$	0	67	**Please** take a look at Space Shuttle Discovery, for FPC. Will you reconsider?	+
$Please_1^+$	100	83	On a more practical note: One either does or doesn't agree to take part in mediation. Which is it, **please**?	-
Can_0^+	96	77	url is a validly notable topic which has been ruthlessly suppressed at wp, probably because of egoism rather than religionist pov warriors but possibly both. **Can** you advise me on how to go about appealing deletions?	-
Can_1^-	25	100		
$Maybe_0^+$	100	100		

Table 3: Examples of incorrectly mapped tokens from requests in the development set. "% Polite" is the percentage of polite items in the original cluster obtained at training. "Accuracy" gives the performance of assigning items from the development set to each cluster.

Cluster Can_0^+ exhibited an accuracy of 77%. The model incorrectly assigned the contextualized representation of *can* in the impolite request *url is validly [...] Can you advice me[...]?* to cluster Can_0^+ which mostly contain polite items.

The overall accuracy for all clusters is 62%. For words encoded in Danescu-Niculescu-Mizil et al. (2013)'s model (e.g. pronouns, hedge words, etc.), the accuracy is 67%.

4 Politeness Classification Results

We quantitatively validate our approach by assessing the predictive power of the patterns that it discovers. Clusters constructed at training can be used to generate fine-grained, unigram-like features for the target task. Given a training item, its features are simply the cluster ids of its constituent words. For example, if the item is *Can you help me?*, its features are Can_0, you_0, $help_1$, me_3, $?_1$. For an unseen item, we predict the cluster id of each of its constituent words. Specifically, for each word in the request, we calculate the euclidean distance between the word's contextualized representation and each cluster for that word. The feature for that word is simply the id of the closest cluster to the word's representation.

We extract features using the proposed model for two subsets of words. In the first subset, we only consider words encoded within Danescu-Niculescu-Mizil et al. (2013)'s features (politeness words) with the goal of validating that the proposed approach captures existing politeness cues. Examples of those words are shown within Table 1. In the second subset, we consider all words in the vocabulary to assess the predictive power of discovered patterns. To be consistent with previous work, models are trained with the Wikipedia training set and are tested on both the Wikipedia test set (in-domain) and the Stack-Exchange test set (cross-domain). We use Pedregosa et al. (2011)'s implementation of Support Vector Machines (SVMs).

Unigrams: We keep track of unigrams (non-contextualized) in the training set as features.

Politeness Strategies (DNM): We use the features from Danescu-Niculescu-Mizil et al. (2013). These include 21 linguistic politeness markers originally studied by Brown & Levinson (1987). Those features

Model	Features	Feature count	In-domain	Cross-domain
SVM	unigrams	1910	78.6	67.0
SVM	Politeness Strategies (DNM)	21	78.8	60.3
SVM	Clustering (politeness words)	565	80.5	62.8
SVM	unigrams + Politeness Strategies (DNM)	1931	82.2	67.2
SVM	unigrams + Clustering (politeness words)	2475	82.9	66.7
SVM	Clustering (all words)	3252	84.1	63.1
SVM	Clustering (all words) + Politeness Strategies	3273	85.1	65.9
Fine-tuned BERT			**89.1**	**75.5**

Table 4: Accuracy on the test set for each model along with the features used to produce the results.

.

are described in Table 1.

Clustering Features (politeness words): These are features extracted using clusters of words encoded within Danescu-Niculescu-Mizil et al. (2013)'s features (e.g. *I*, *you*, *think*, *why*). Examples of those words are shown within Table 1. The number of features (clusters of those words) is 565.

Clustering Features (all words): These are features extracted using clusters that correspond to all words in the vocabulary. They include 3252 features.

Table 4 gives the accuracy of these feature-based models on the test set. In the in-domain settings, using Clustering Features (all words) as features helps in getting a 5.5% improvement when compared to a standard unigram model and a 1.9% improvement when compared with Danescu-Niculescu-Mizil et al. (2013)'s politeness strategies in addition to unigrams. The fact that combining unigrams and Clustering (politeness words) performed comparably with using unigrams and politeness strategies (82.2 vs. 82.9 in the in-domain setting and 67.2 vs. 66.7 in the cross-domain settings) indicates that our clustering model captures cues encoded in the politeness strategies. Clustering Features (all words) combined with Danescu-Niculescu-Mizil et al. (2013)'s politeness strategies gives the best feature-based results in the in-domain settings with an accuracy of 85.1%. Using unigrams with politeness strategies (Danescu-Niculescu-Mizil et al., 2013) achieves the best cross-domain accuracy among the feature-based models. This shows that contextualization has a downside in that it is encoding some specific aspects of the domain, which likely makes it less robust in the cross-domain context. Perhaps training on several domains and testing on a new domain might improve the cross-domain performance. Unsurprisingly, a fine-tuned BERT model achieves state-of-art results on this classification task with 89.1 in-domain accuracy and 75.5 cross-domain accuracy.

Our classification results indicate that pre-trained BERT embeddings can help in obtaining fine-grained word features which enhance the performance of in-domain politeness prediction in comparison with other standard feature-based models. The fact that features based on clusters performed significantly better in the in-domain settings than in the cross-domain settings could be due to the difference in nature and topics between the Wikipedia discussion and the Stack-exchange datasets. We reserve discovering politeness features that would generalize to more domains to future work.

5 Conclusion

In this work, we proposed a model for obtaining insight on how words are used given the context of their usage and the labels of the items they belong to. When applied to requests annotated for politeness, our model not only re-discovers existing linguistic markers discussed in the literature but also enriches them by discovering novel finer-grained patterns of using words in (im)polite language. We quantitatively validated the model by showing that fine-grained features improve over politeness-word baselines. This shows that examining the context of using words helps enriching the linguistic analyses of the task/phenomenon being studied. In the future, we plan to extend our model to obtain insight on tasks based on annotated text pairs (posts & replies). We also plan to use our model to inform sociolinguistic analyses on gender in order to get a better understanding of the relationship between language use and gender.

References

Aubakirova, M., & Bansal, M. (2016). Interpreting Neural Networks to Improve Politeness Comprehension. In *Empirical Methods in Natural Language Processing*.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender Identity and Lexical Variation in Social Media. *Journal of Sociolinguistics*, *18*, 135–160.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage* volume 4. Cambridge University Press.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of The 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Hyland, K. (2018). *Metadiscourse: Exploring Interaction in Writing*. Bloomsbury Publishing.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of the 14th international conference on World Wide Web* (pp. 342–351).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. In *International Conference on Learning Representations Workshop: Workshops Track*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Niu, T., & Bansal, M. (2018). Polite Dialogue Generation without Parallel Data. *Transactions of the Association for Computational Linguistics*, *6*, 373–389.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*, 2825–2830.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of The North American Chapter of the Association for Computational Linguistics*.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine learning*, *1*, 81–106.