Regression Model for Politeness Estimation Trained on Examples

Mikhail Alexandrov[1], Natalia Ponomareva[2], Xavier Blanco[1]

[1]Universidad Autonoma de Barcelona, Spain
[2]University of Wolverhampton , UK
Email: dyner1950@mail.ru, nata.ponomareva@wlv.ac.uk,
xavier.blanco@uab.es

Abstract. Automatic assessment of subjective characteristics of customers like politeness, satisfaction or competence could provide services companies with information needful for improving service quality. In this work, we construct a regression model for politeness estimation of customers, which uses a) set of linguistic indicators and b) manual estimations of expert(s). We apply the suggested methodology for processing dialogues of passengers with directory inquires of Barcelona railway station. All linguistic indicators are proved to be statistically significant with a level of confidence equal to 5%. The constructed model is tested on independent data set and demonstrates good concordance with expert opinions.

Key-expressions: politeness estimation, regression model

1. Introduction

1.1 Problem setting

Politeness, competence, satisfaction, etc. are very important characteristics of customers whose analysis might help a service company to evaluate the needs of its clients and to improve the service quality. It should be said that automatic evaluation of mentioned personal characteristics is quite difficult, especially if we deal with short texts. For this reason, existing computer tools locate linguistic indicators (LIs) related to these characteristics in a text without giving any numerical estimation. In this paper, we show how to use LIs for constructing the simplest numerical model for politeness estimation. Our approach exploits the following steps: choice of LIs, their location in a text, construction of the regression model, checking model validity and calculation of model accuracy.

Linguistic patterns corresponded to LIs of politeness are revealed by means of NooJ. This tool allows the detection and summation of occurrences of given

lexical-syntactic patterns in texts [5]. In this paper, we neither discuss why we choose this set of LIs no aim to compare the model accuracy for different sets. In general case one should take into account all possible LIs, evaluate their contribution into the regression model and then eliminate insignificant indicators.

The experimental data are dialogues of passengers with railway directory inquiries of Barcelona station. The language of dialogues is Spanish. Obviously, the indicators of politeness are specific for each language and what we consider to be good indicators for Spanish might not be appropriate for other languages. But our idea is only to demonstrate the approach.

The paper is organized as follows. Section 2 we propose a set of indicators that could relate to a level of politeness. Section 3 describes our approach of empirical formula construction. Section 4 shows the experimental results. Conclusions and future work are drawn in Section 5.

1.2 Related Works

There are many publications where politeness is studied as an element of written and oral speech. The panorama of recent research in the area is presented in [4]. There are works devoted to software, which detect polite (impolite) expressions in a set of dialogues. The typical researches in this area are presented in [1,2]. Nevertheless we did not meet publications where quantitative estimations of politeness were specially studied.

In paper [3], we describe the general approach for constructing empirical formulae for formal estimation of various personal characteristics. As an example, politeness is considered and evaluated. This paper focuses on a linear regression model as the simplest case of polynomial models.

2. Indicators of politeness

We propose the following LIs for evaluation of a level of politeness of customers:

(1) first greeting (***G***);
(2) polite words (***W***);
(3) polite grammar forms (***V***).

In this work, we do not consider any indicators of impoliteness due to a lack of impolite examples appeared in our dialogue collection. It can be explained by the fact that a passenger needs the information and has no wish to be rude.

As an example of polite words such well-known expressions as "please" (por favor), "thank you" (gracias), "excuse me" (perdon), etc. can be mentioned. We also include a polite form of “you” (usted) inside this category. In Spanish it is

normal to omit personal pronouns; therefore, the use of these pronouns expresses a special respect to an interlocutor.

In Spanish, subjunctive and conditional verb forms are used to express a higher level of respect and politeness. This peculiarity of Spanish can also be found in English although the concordance is not complete. The examples of exact correspondence between Spanish and English polite verb forms might be: "I would like …" – "Me gustaría …" or "Could you …" – "Me podría …". However, in some cases English people do not utilize polite verbs whereas it is quite normal for Spanish. For example, Spanish polite variant of a phrase "How much does it cost?" can be formulated as "¿Cuánto me costaría?" where a verb in conditional form is used.

A special attention must be paid to the indicator "first greeting". It is characterized by presence or absence of a polite salutation in a dialogue. It is a binary indicator that takes a value 0 when a first greeting is impolite and 1, otherwise. Politeness of a first greeting is determined by two other indicators, namely, polite words and polite grammar forms. If a first greeting contains either the indicator $\boldsymbol{W}$ or $\boldsymbol{V}$ it is supposed to be polite and the indicator $\boldsymbol{G}$ takes a value 1. We consider the first greeting as a particular indicator because, in our opinion, it is a decisive factor of the level of politeness.

A reader familiar with Spanish might be surprised by the fact of absence, among the politeness indicators, the indicator, which would characterize a form of treatment: whether he/she utilizes a polite form of "you" or an unceremonious one. In English there is no difference between these two forms whereas in Spanish this difference exists. In this work, we do not take into account distinct forms of personal treatment, because nowadays it mostly refers to the age of a person and not to his/her level of politeness.

3. Regression model

There are different ways to calculate numerical values of selected indicators. It can be either a frequency of indicators occurred in a dialogue or just a binary value reflecting occurrence/absence of indicator in a dialogue. In our work, we make following assumptions to calculate numerical values of politeness indicators:

(i) Level of politeness is defined by a density of politeness indicators in a dialogue. A word 'density' refers to an indicator frequency normalized by a dialogue's length. The dialogue's length here is a number of customer phrases.

(ii) Level of politeness depends on the indicator density non-linearly: the contribution of each new polite word or verb form decreases with the growth of corresponding indicator density. It leads to the necessity of using any suppressed functions as, for example, logarithm or square root one.

Taking into account the aforesaid, numerical values of the introduced politeness indicators can be represented in a following way:

$$\boldsymbol{G} = \{0, 1\}, \ \boldsymbol{W} = Log_2(1+N_W/L), \quad \boldsymbol{V} = Log_2(1+N_V/L), \tag{1}$$

where N_W, N_V are a number of polite words and polite grammar forms respectively and L is a number of passenger's phrases.

It is evident that:

a) $\boldsymbol{W} = \boldsymbol{V} = 0$, if polite words and polite grammar forms do not occur;

b) $\boldsymbol{W} = \boldsymbol{V} = 1$, if polite words and polite grammar forms occur in every phrase.

Thus, these relations approximate minimum and maximum values of politeness indicators $\boldsymbol{W}$ and $\boldsymbol{V}$.

Table 1 shows an example of a dialogue (the records are translated from Spanish into English). Here **US** stands for a user and **DI** for a directory inquiry service. This example concerns the train departure from Barcelona to Zaragoza.

Table 1. Example of a real dialogue between passengers and directory inquires

US: ***Good evening. Could you*** tell me the schedule of trains to Zaragoza for tomorrow? **DI:** For tomorrow morning? **US:** Yes **DI:** There is one train at 7-30 and another at 8-30 **US:** And later? **DI:** At 10-30 **US:** And till the noon? **DI:** At 12 **US:** ***Could you*** tell me the schedule till 4 p.m. more or less? **DI:** At 1-00 and at 3-30 **US:** 1-00 and 3-30 **DI:** hmm, hmm <SIMULTANEOUSLY> **US:** And the next one? **DI:** I will see, one moment. The next train leaves at 5-30	**US:** 5-30 **DI:** hmm, hmm < SIMULTANEOUSLY > **US:** Well, and how much time does it take to arrive? **DI:** 3 hours and a half **US:** For all of them? **DI:** Yes **US:** Well, ***could you*** tell me the price? **DI:** 3800 pesetas for a seat in the second class **US:** Well, and what about a return ticket? **DI:** The return ticket has a 20% of discount **US:** Well, so, it is a little bit more than 6 thousands, no? **DI:** Yes **US:** Well, ***thank you very much*** **DI:** Don't mention it, good bye

Table 2 shows the results of parameterization of this dialogue and its manual estimation by an expert. Here the number of polite words is equal to 2 because the passenger uses a polite form of a particular pronoun "you" that is impossible to express in English translation.

Table 2.Parameterized dialogue

Parameter	Value
First greeting $\boldsymbol{G}$	Yes
Number of polite words $\boldsymbol{N_W}$	2
Number of polite grammar forms $\boldsymbol{N_V}$	2
Indicator $\boldsymbol{G}$	1
Indicator $\boldsymbol{W}$	0.13
Indicator $\boldsymbol{V}$	0.13

We consider the following model for politeness estimation:

$$F(\boldsymbol{G,W,V}) = A_0 + A_1\boldsymbol{G} + A_2\boldsymbol{W} + A_3\boldsymbol{V}, \quad (2)$$

where A_0, A_1, A_2, A_3 are undefined coefficients.

Let $\boldsymbol{N}$ be a number of dialogues. We have the following system of linear equations:

$$A_0 + A_1\boldsymbol{G}_i + A_2\boldsymbol{W}_i + A_3\boldsymbol{V}_i = \boldsymbol{E}_i \qquad \boldsymbol{i}=1,\ldots,\boldsymbol{N}, \quad (3)$$

where $\boldsymbol{G}_i$, $\boldsymbol{W}_i$, $\boldsymbol{V}_i$ are numerical values of the politeness indicators and $\boldsymbol{E}_i$ is a manual estimation of the level of politeness for a dialogue $\boldsymbol{i}$. Having constructed this model we need to evaluate the significance of its coefficients and to filter the insignificant ones.

4. Experiments

The corpus we used in our experiments are dialogues of passengers with railway directory inquiries of Barcelona station. The main characteristics of this corpus are presented in Table 3.

An example of data used in the experiments is presented in Table 4. Numerical values of the politeness indicators $\boldsymbol{G}$, $\boldsymbol{W}$, $\boldsymbol{V}$ are calculated using (1). Manual estimation is done in the framework of scale [0,1] with a step 0.25.

We used 15 dialogues for determination of model coefficients (2) and the rest 15 dialogues for checking precision of the constructed formula. Having solved the linear system (3) we obtained the following preliminary regression model:

$$F(\boldsymbol{G}, \boldsymbol{W}, \boldsymbol{V}) = -0.04 + 0.22\boldsymbol{G} + 3.72\boldsymbol{W} + 3.13\boldsymbol{V} \quad (4)$$

Table 3. Corpus characteristics

Characteristic	Value
Number of dialogues	30
Language	Spanish
Minimum dialogue's length	7
Minimum dialogue's length	62
Average dialogue's length	22.57
Average value of the indicator ***G*** per dialogue	0.87
Average number of polite words per dialogue	1.10
Average number of polite grammar forms per dialogue	1.73

Table 4. Example of data used in the experiments

G	***W***	***V***	**Manual estimation**
1	0.134	0.194	1
0	0.111	0.057	0.75
1	0.000	0.074	0.25
1	0.000	0.031	0
1	0.000	0.118	0.75
1	0.043	0.043	0.5
1	0.000	0.000	0.25
1	0.043	0.083	0.5
0	0.000	0.074	0
1	0.134	0.069	1

Global test (*F*-test) showed the statistical significance of a regression model with respect to all its variables. Individual test (*t*-test) for each variable showed that the intercept (first member) should be eliminated, but all other variables (regression coefficients) proved to be significant. Testing hypothesis was conducted with the confidence level of 5%. After recalculation we obtained the regression model:

$$F(\boldsymbol{G}, \boldsymbol{W}, \boldsymbol{V}) = 0.3\boldsymbol{G} + 3.7\boldsymbol{W} + 3.2\boldsymbol{V} \qquad (5)$$

Coefficient of determination of this model is equal to 80%. It means that the selected linguistic indicators cover 80% of variation in dialogue estimations. The testing procedure with 15 additional dialogues gives the relative mean square root error equal to 26%, which is comparative with the step of the manual estimation.

It can be observed that all indicators of politeness have positive coefficients. If a passenger does not use any politeness indicator then his level of politeness is 0, and if he says, at least, the first greeting his politeness level gets a positive value. These observations informally demonstrate a validity of the obtained model (5).

5. Conclusions

In this paper, we consider linguistic indicators of politeness, which can be used for formal evaluation of the level of politeness in dialogues. We show how to construct the simplest regression model based on these indicators.

The experiments confirm the statistical significance of all suggested indicators. The precision of the constructed model is comparative with a step of the manual estimation of dialogues, which is obtained on control data set.

In future, we intend to consider more complex indicators of politeness. We also plan to construct non-linear statistical models.

Bibliography

1. Alexandris, C, Fotinea, S.E.: Discourse particles: Indicators of positive and non-positive politeness in the discourse structure of dialog systems for modern greek. Intern. J. for Language Data Processing "Sprache Datenverarbeitung", **1-2** (2004), 19-29
2. Ardissono, L., Boella, C, Lesmo, L.: Indirect speech acts and politeness: A computational approach. In: Proceedings of the 17th Cognitive Science Conference. (1995), 113-117
3. Alexandrov, M., Blanco, X., Ponomareva, N., Rosso, P: Constructing Empirical Models for Automatic Dialog Parameterization. In: Proceedings of the TSD-07 (2007). Springer, LNCS, 4629: 455-462
4. Briz, A., et. Al (eds):Cortesia y conversacion: de lo escrito a lo oral. Valencia/Estocolmo: Universidad de Valencia y Programa EDICE, (2008), ISBN: 978-91-974521-3-7
5. NooJ description: http://www.nooj4nlp.net