Automated Identification of Relative Social Status

Katherine Metcalf METCALKA@INDIANA.EDU
David Leake LEAKE@INDIANA.EDU
Computer Science, Indiana University, Bloomington, IN 47405 USA

Abstract

Understanding the social context is a key requirement for any dialogue system that is able to interact in situations with multiple users, and also improves a dialogue agent's ability to identify expectations of a single user. This paper focuses on methods for identifying one sub-aspect of the social context, the relative social power between interactors. Specifically, this paper examines how alternative candidate dialogue feature representations affect the ability of different classifiers to predict relative social power. The performance of the classifier and feature representation pairs was compared for the task of determining whether or not two speakers have a difference in social power, for two different transcript data sets. Different feature and classifier combinations performed best at detecting the presence, versus the absence, of a difference in social power.

1. Introduction

Currently, most dialogue systems focus on interacting with a single user for a given task or passively interacting with a group of people (e.g., Mairesse & Young, 2014; Kim et al., 2014). In order to move beyond systems that are targeted towards single users to systems that act as part of a goal oriented team, methods are needed to identify a situation's interactional context (e.g., Tambe, 1997; Blunt & Black, 2000). Part of the interactional context is awareness of the social context, which is necessary for an agent to demonstrate social cognition and to seamlessly integrate itself into an interaction (e.g., de Weerd et al., 2014). However, neither the interactional nor the social context is always immediately evident, especially when agents must rely on unstructured data outside of a laboratory setting. As both are complex and fluid constructs, it is convenient to break them apart and build recognition methods for each up from more simple components. An important aspect of the social context is the interactors' relative social status. As power dynamics change within a given interactional context and over the course of a conversation, it is important that cognitive systems be able to track those changes. This paper focuses specifically on (1) examining the usefulness of function words in predicting relative social power, (2) calculating a metric for relating the feature sets of two different speakers for use for training classifiers and for subsequent inputs, and (3) investigating the accuracy of particular pairs of classifiers and metrics at predicting relative social power. This approach is an initial step towards systems able to dynamically track changes in social power.

The ability to identify the authoritative figure within a given interaction would benefit dialogue systems that interact with groups of people, enabling such systems to respond appropriately given social relationships. For example, if an automated system is working as part of a project team or a military unit, then a relationship-aware recommendation system could determine to whom to direct recommendations and which interactors would be the most acceptable to interrupt or contradict, if needed. The ability to identify social power relations could benefit systems such as, for example, CALO (Tur et al., 2008). The CALO project has focused on building an agent that is able to learn, take orders, and explain its reasoning process in both single and multiuser settings. With the additional capability to infer the social context, such an agent could function as an interactive part of new teams by providing unprompted recommendations based on the current social and interactional contexts.

Conversational dominance is not the same thing as social dominance, but there is a natural link between dominating a conversation and being the dominant interactor (Pennebaker, 2011). Consequentially, previous work on assessing relative social status builds on the idea of conversational dominance, using interactions as the unit of analysis, and bases the relative social status assignment on the presence or absence of dominant/authoritative and deferential features within a given interaction (Fitzsimons & Kay, 2004). Relative social status is treated as manifestations of gestures or enactments of dominance/authority or deference, both of which depend not only on the social situation, but also on the topic and goal of the interaction and the intentions and goals of the individual interactors (Duranti, 1988). The assignment of dominance/authority or deference is variable according to changes in interactor roles and intentions, which may change throughout a single interaction or over the course of a series of interactions. This paper focuses on testing how well different feature representation and classifiers pairs perform at predicting relative social power. This approach allows the performance of the classifier and feature representation pairs at identifying relative social power in general to be examined.

The paper reports on experiments comparing the accuracy of two different classifiers (linear-SVM and random forest) at two different tasks. The first task was identifying whether or not there is a difference in social power between two interactants, and the second task, if there is a difference in social power, is to determine which interactant has more social power. Each classifier was tested with six different interactant relation metrics, each reflecting different possible approaches to relating feature instances of two interactors, a topic that, to our knowledge, has not previously been explicitly studied in the literature. In the experimental results, the random forest classifier performed with the highest accuracy, precision, and recall. No single feature representation or metric provided the best performance across all three tasks, and a number of the classifier and feature representation pairs performed equally well. The difference in the most accurate classifier and feature representation pair for each task suggests a difference in the type of information important to each task, and, thus, to differences in the tasks as well.

1.1 Definitions

This paper is not focused on debating what is meant by social status versus social power versus dominance. Instead it takes a stance that is in line with Danescu-Niculescu and colleagues (2012; Danescu-Niculescu et al., 2013), Kacewicz and colleagues (2014), and Duranti (1988), viewing language as being linked to numerous social psychological events (Fitzsimmons & Kay, 2004), which allows it to inform on the roles and relative statuses that are embedded within a group's interactions (Danescu-Niculescu et al., 2012). Relative social power can come from formal, externally imposed roles, such as judge versus lawyer, from the respect one individual has for another, such as between colleagues, or from one interactor being able to deny another interactor something he or she needs (Danescu-Niculescu et al., 2012; Willer, 2006; Giles, 2008). From this perspective social power is fluid and dependent on both time and the goals of an interaction while still incorporating more long-lasting social statuses. For this reason, this paper focuses on the social power between individual speakers relative to a single interaction, such as meeting someone for the first time, instead of across the entire corpus of interactions.

2. Related Work

Until recently most of the work on social status focused on social class and was conducted within the context of gender and conversational dominance by looking at variations in language use between classes of people (Ross, 1954; Labov, 1966; Lakoff, 1975; Macaulay, 1977; Romaine, 2002; Rowe et al., 2007; Bramsen et al., 2011). Work on identifying power differences within a conversation is now carried out at the interaction level when identifying which speaker had the most influence (Rienks et al., 2006; Biren et al., 2012; Nguyen et al., 2014; Prabhakaran & Rambow, 2013), identifying leadership roles (Huffaker, 2010), and disambiguating social power (Danescu-Niculescu et al., 2012; Danescu-Niculescu et al., 2013; Prabhakaran & Rambow, 2013). Popular interactional domains include work meetings, political debates, and Supreme Court hearings, with a significant focus on online discussions.

Social status and social class identification have been approached by examining individuals' language features, grouping the individuals based on known characteristics, and then comparing the presence of language features across the groupings. The meaningful features identified with these approaches range from linguistically based function and content words (Pennebaker, 2011) to dialogue and conversation based turn taking (including number of turns) and turn length (Zimmerman & West, 1975; West & Garcia, 1988; Itakura, 2001), turn type (Mazur & Cataldo, 1989; Rienks et al., 2006; Biren et al., 2012; Nguyen et al., 2014), speech act features, degree of politeness (Locher, 2004), and language coordination (Gonzales et al., 2009; Danescu-Niculescu et al., 2011; Danescu-Niculescu et al., 2012; Danescu-Niculescu et al., 2013; Moon et al., 2014; Guo et al., 2015). To our knowledge, the productivity of function word linguistic features alone as a predictor of relative social status has not yet been examined. By examining differences in function word usage, the approach developed in this paper models participants directly instead of modeling the interaction and how participants fit into the interaction.

3. Representing Feature Sets and Feature Set Differences

We describe the representation of the input to the classifiers in two parts. As will be discussed in section 3.1, function-word-count feature-vectors were used to represent the feature sets of individual speakers. As will be discussed in section 3.2, the differences in feature production between two speakers were represented using various component-wise difference functions. Representing differences in feature production between two speakers was a key focus for this paper, as the way the differences are represented needs to have some relation to the important differences between the two speakers and what is means to have differences in speech production.

3.1 Linguistic Features

Individual speakers were represented with vectors whose components were the counts of thirteen different linguistic features: articles, conjuncts, quantifiers, auxiliary verbs, prepositions, first-person pronouns, singular first-person pronouns, plural first-person pronouns, second-person pronouns, third-person pronouns, singular pronouns, plural pronouns, and the total number of pronouns. Each linguistic feature represented the number of times speaker produced a given category of function word. The linguistic features were chosen based on work by Pennebaker and his colleagues (Chung & Pennebaker, 2007; Pennebaker 2011; Kacewicz et al., 2014;), which identified function words as meaningful features for social power differentiation.

Table 1. The feature comparison metrics used to compare the individual feature instances between two speakers.

Type	Description
Canberra	The Canberra distance between the two speaker vectors.
Euclidean Feature Vector	The Euclidean distance between the two feature vectors.
Pos-Neg Euclidean Feature Pair	The Euclidean distance between each of the two speakers' feature pairs. The Euclidean distance is made negative in cases where speaker two had a greater number of feature instances.
Euclidean Feature Pair	The Euclidean distance between each of the two speakers' feature pairs.
Delta	The absolute distance between each of the two speakers' feature pairs.
GTLT	Assembles a vector with one component per feature, each component representing a comparison between that feature’s frequency for the two speakers. If speaker one has more instances of a given feature, the value is 1; otherwise the value is 0.

Although Pennebaker and his colleagues identified a number of features, the only features used in this experiment were function words. This decision was based on the relative stability of such linguistic features and the ease with which the features could be extracted. The criterion of stability of feature production aimed at using features that are less likely to change over time, as the language continues to evolve, and are, therefore, relatively more stable, which excluded the use of any type of morpheme or lexicon-based analysis as specific word meanings and usage vary with time. This meant that any feature requiring a dictionary or lexicon lookup was not used, such as emotion or cognitive words. Ease of feature extraction was seen to be important because of the eventual goal of online relative power identification. If a feature is too slow to be identified and extracted, then the online performance of the system can decrease and the system may not me able to keep up with the situation at hand.

3.2 Feature Vector Differences

Six methods for calculating feature differences were tested for how productively they associated the linguistic feature instances produced by one speaker with the speaker instances produced by another speaker (Table 1). The goal was to find the best way to calculate the distance between features produced by two different speakers. For each feature representation, the difference or distance is always calculated relative to one of the speakers, designated as speaker X and the other speaker was designated as speaker Y. This means that the question answered by the classifiers is, “Does speaker X have a more social power than speaker Y?” The designation of speaker X versus speaker Y is arbitrary and merely determines which speaker's value fills the x versus y arguments of the distance metrics. For this study, each speaker pair was evaluated twice so that each speaker was assigned to both X and Y. This was done to make sure that the assignment could be done arbitrarily without impacting the results of the classifiers. It was found that the speakers from a given pair could be arbitrarily assigned to either speaker X or speaker Y. without affecting the performance of the classifiers.

The difference calculation methods fall into one of two categories, representing the difference in the production of each feature or the difference in the overall feature production.

Table 2. Number of interactions, speakers, and utterances in each of the corpora.

	CSC	ICSI	Total
Interactions	325	60	385
Speakers	204	75	279
Utterances	886433	68759	955192

The overall feature production values were calculated using distance metrics that compressed the feature sets into the single value of the distance between the two feature vectors, in contrast to the feature level metric, which maintains the 14 distinct linguistic features.

To compare the feature sets, a distance value was calculated using the Canberra Distance formula (Lance & Williams, 1967), where speaker X is p and speaker Y is q, and using the Euclidean distance, where speaker X is (x_1, y_1, z_1, …) and speaker Y is (x_2, y_2, z_2, …). The Canberra Distance Formula finds the distance between two vectors, p and q, in a *n*-dimensional real-vector space[1]. For the Euclidean distance calculation, the linguistic features were treated as forming a 14-dimensional space. The four feature-level vector representations calculated the difference or distance between the individual feature points and represents, (1) the absolute difference between each feature value ([$s_1f_1 - s_2f_1$, … , $s_1f_{14} - s_2f_{14}$]); (2) whether or not speaker X has more instances of a feature, where 1 indicates that speaker X has more instances and 0 indicates that speaker X does not have more instances; (3) the Euclidean distance between speaker X and speaker Y feature instances, where the x and y for a given feature are both that feature's value; and (4) the Euclidean distance between speaker X and speaker two's feature instances, but with a negative value is assigned for those feature pairs where speaker Y has more instances than speaker X.

4. Data Sets

The experiments were conducted using two data sets, the ICSI Meeting transcripts (Janin et al., 2004) and the Cornell Supreme Court Trial transcripts (Danescu-Niculescu-Mizil et al., 2012). Both data sets contain the transcripts of face-to-face interactions within which the relative social power is in largely contingent upon the roles played by each individual. The adherence to topic and goal oriented power relations make these appropriate corpora with which to test the

[1]
$$d(p,q) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}.$$

feasibility of the different feature set representations. Differences in social power were assigned according to each speaker's official rank, for example, professor versus undergraduate. Together the corpora consist of 204 interactions (i.e. trials or meetings), 325 different speakers, and 10 different social positions (i.e. judge, lawyer, Professor, PhD student, and undergrad), which allows for a varied distribution of feature instances and power relations by providing examples of interactions that occur within and outside of static power structures that are the primary source of power differences (Table 2). One potential issue with using face-to-face interactions is that the speakers were able to enact and communicate relative social power with non-linguistic features, such as gesture and bodily positioning. This means that the productivity of a given linguistic feature might be contingent upon the co-production of a non-linguistic feature. Consequently, productive feature sets from these data sets might not transfer well to other domains where interactions are not face-to-face.

5. Evaluation

Descriptive statistics were calculated to assess how well each of the feature set representations correlated with the social power class assigned to each speaker pair. A goal of this analysis was to identify a subset of speaker pairs such that the difference in feature production was not significantly correlated with the social power class assigned to the pair. The identification of this subset was expected to give an indication of the extent to which feature production needs to differ between two speakers in order to indicate a difference in social power. The analysis was conducted on the two data sets, comparing the relationship between each of the six feature set types and between the features and their associated classification category. The analysis was done both within and across the datasets. For this analysis, the power relation classifications were treated as the independent variable and the features of the feature instances were treated as the dependent variables.

5.1 Correlation of Feature Set Representations with Classifications

A Mann-Whitney-Wilcoxon Rank Sum test and a One-Way ANOVA were used to analyze the relationship between the social power classification and each of the feature values used to measure the distance between the two speakers. For each distance metric, both those distance measures between the individual speaker vectors and those between each of the features in the feature vectors, the feature vectors were tested to see how well they correlated with the social power classification assigned to each speaker pair. All of the feature-set level representations were significantly correlated with the associated classification ($p < 0.001$), and all of the feature-by-feature representations had at least 50% of the feature types significantly correlated with the associated classification ($p < 0.001$). This finding was particularly interesting as Pennebaker and colleagues (Chung & Pennebaker, 2007; Pennebaker, 2011; Kacewicz et al., 2014) identified first-person singular pronouns as the feature word most highly correlated with greater social power, whereas this study did not find first-person singular pronouns to be more highly correlated with the social power assignment than any other feature. The differences in correlations could point to a difference in social power related features when produced through a written versus spoken medium; namely, given the lack of other clues, language that emphasizes social power differences is more prominent in written versus spoken language as there are no accompanying physical traits. It would be interesting to see if this difference holds between in-person spoken interactions and spoken interactions that are conducted at a distance, such as when making a telephone call.

5.2 Level of Feature Differences Needed to Indicate Power Differences

We anticipated that those speaker pairs whose difference in feature production deviated the most from the mean would be less likely to exhibit the same, strong distributional correlation with the classifications and would be less distinct from one another. For this analysis, the feature production values for each of the social power classes were treated as a population; therefore, the aim was to identify those speaker pairs whose feature production deviated the most from the mean of the population. The question to address was, "To what extent do feature instances have to differ between two speakers for the features to indicate a difference in power?" Therefore, those speaker pairs who had different relative social power were treated as one class, regardless of which speaker had more power.

Interestingly, our tests did not identify a set of cases for which the features were not distinctly correlated with the difference in power and were not statistically distinct from one another. Such a set of cases was needed in order to examine the extent to which feature production needs to differ so as to indicate a difference in social power. These results suggest a categorical use of power features instead of continuous, however this interpretation seems counter intuitive. For that reason, this is an area that requires a more targeted examination.

6. Classifiers

The classification task had two components, identifying whether or not there is a difference in social power and, if there was a difference, identifying who had more power. The task of identifying who had more relative social power was broken into two sub-classification tasks, (1) whether or not speaker X had more power and (2) whether or not speaker Y had more power; this decomposition was used so that all of the classification tasks could be binary classifications. The productivity of each of the feature sets was compared across two different classifiers: random forest and linear support vector machine. The classifiers were trained and tested on each of the six feature representations with one classifier for each binary classification: (1a) speaker X has more power, (1b) speaker Y has more power, and (2) speakers X and Y have the same amount of power. In total, 36 binary classifiers were built using Scikit-Learn (Pedregosa et al., 2011).

Each of the classifier models was validated with a 10-fold cross-validation procedure and was trained and tested on the combined ICSI Meeting (Janin et al., 2004) and Cornell Supreme Court (Danescu-Niculescu-Mizil et al., 2012) corpus. The support vector machine was built using Scikit-Learn with a linear kernel and a c-value of one. The random forest classifier was built using Scikit-Learn with ten estimators. Each of the models was evaluated according to accuracy, precision, recall, and F1-score.

7. Results

No single classifier and feature set representation performed best at all of the classification tasks (Table 3; Table 4).

7.1 Classifier One: Speaker X

For the task of classifying speaker X as having more power, three of the classifier and feature representation pairs performed equally well when rounded out to the 8th decimal place: (1) Random Forest Classifier with the Delta feature set (*accuracy* = 0.99; *precision* = 0.99; *recall* = 0.99; *f-score* = 0.99); (2) Random Forest Classifier with the GTLT feature set (*accuracy* = 0.99; *precision* = 0.99; *recall* = 0.99; *f-score* = 0.99); and (3) SVM with the GTLT feature set (*accuracy* = 0.99; *precision* = 0.99; *recall* = 0.99; *f-score* = 0.99). The equivalent performance of the Random Forest classifier with the two feature representation pairs is not surprising, as the two feature representations encode the same relationship. The performances of the GTLT feature set can be explained in the same way. It is interesting to note that Random Forest with the Pos-Neg Euclidean Feature Pair (*accuracy* = 0.99; *precision* = 0.99; *recall* = 0.99; *f-score* = 0.99) did not perform significantly worse than the three top performing classifier and feature representation pairs. The similar performance may be due to the similar way in which it represents the relation between the feature sets.

Table 4. Performance of the support vector classifier with each of the feature representations.

Speaker One Has More Power	precision	recall	accuracy	f1
Canberra	0.963	0.826	0.908	0.889
Euclidean Feature Vector	0.038	0.003	0.554	0.005
Pos-Neg Euclidean Feature Pair	0.987	0.897	0.949	0.940
Euclidean Feature Pair	0.037	0.003	0.554	0.006
Delta	0.987	0.897	0.949	0.940
GTLT	0.999	0.999	0.999	0.999
Speaker Two Has More Power				
Canberra	0.964	0.826	0.908	0.889
Euclidean Feature Vector	0.038	0.002	0.554	0.004
Pos-Neg Euclidean Feature Pair	0.988	0.898	0.949	0.940
Euclidean Feature Pair	0.129	0.002	0.553	0.005
Delta	0.988	0.898	0.949	0.940
GTLT	0.804	0.999	0.890	0.891
Equivalent Power				
Canberra	0	0	0.890	0
Euclidean Feature Vector	0.880	0.937	0.979	0.903
Pos-Neg Euclidean Feature Pair	0	0	0.890	0
Euclidean Feature Pair	0.861	0.939	0.976	0.894
Delta	0	0	0.890	0
GTLT	0	0	0.890	0

Table 3. Performance of the random forest classifier with each of the feature representations.

Speaker One Has More Power	precision	recall	accuracy	f1
Canberra	0.987	0.983	0.987	0.984
Euclidean Feature Vector	0.104	0.091	0.199	0.096
Pos-Neg Euclidean Feature Pair	0.999	0.999	0.999	0.999
Euclidean Feature Pair	0.097	0.076	0.196	0.0851
Delta	0.999	0.999	0.999	0.999
GTLT	0.999	0.999	0.999	0.999

Speaker Two Has More Power				
Canberra	0.988	0.985	0.987	0.986
Euclidean Feature Vector	0.097	0.088	0.197	0.095
Pos-Neg Euclidean Feature Pair	0.999	0.999	0.999	0.999
Euclidean Feature Pair	0.098	0.077	0.194	0.079
Delta	0.999	0.999	0.999	0.999
GTLT	0.809	0.996	0.893	0.893
Equivalent Power				
Canberra	0.981	0.980	0.996	0.982
Euclidean Feature Vector	0.991	0.975	0.997	0.983
Pos-Neg Euclidean Feature Pair	1	0.9890	0.999	0.997
Euclidean Feature Pair	1	1	1	1
Delta	1	0.994	0.999	0.996
GTLT	0.253	0.043	0.894	0.070

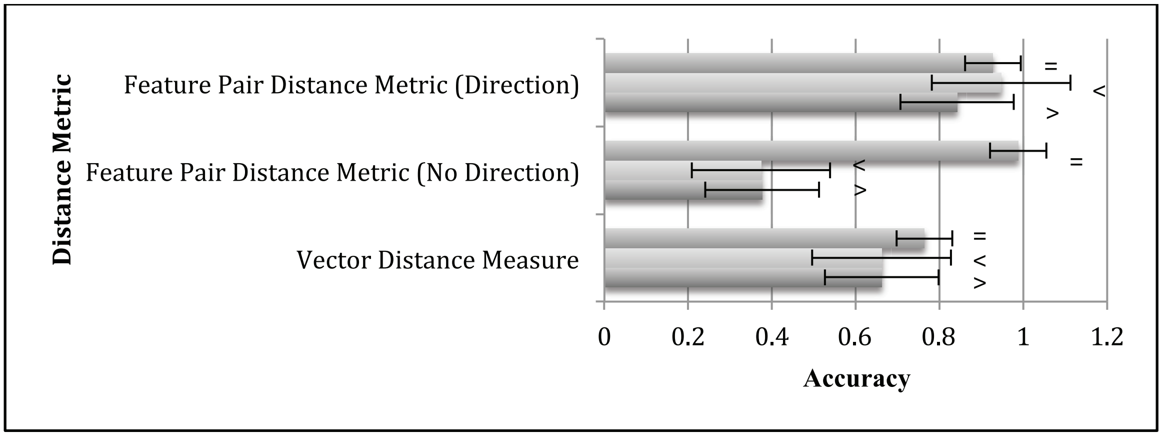

Figure 1. Average accuracy across the classifiers using the three types of distance metrics where “=” indicates that speakers one and two had the same power, “<” indicates that speaker two had more power, and “>” indicates that speaker one had more power.

7.2 Classifier Two: Speaker Y

For the task of classifying speaker Y as having more power, the Random Forest Classifier with the Delta feature set (*accuracy* = 0.99; *precision* = 0.99; *recall* = 0.99; *f-score* = 0.99) performed the best. The precision of the Random Forest with Delta set differed from the Random Forest with Pos-Neg Euclidean Feature Pair set by less than 0.001. We believe that the lack of a significant difference between the two feature representations is because the two metrics represent approximately the same type of relationship between the two feature sets; the two methods for calculating the differences differ only slightly (ie. ($spkr_1feat_i$ – $spkr_2feat_i$) versus $((spkr_1feat_i - spkr_2feat_i)^2)^{1/2}$).

7.3 Classifier Three: Equivalent Power

For the task of classifying speakers X and Y as having the same power, the Random Forest classifier with the Euclidean Feature Vector metric (*accuracy* = 1; *precision* = 1; *recall* = 1; *f-score* = 1) had the best performance. However, the Random Forest classifier and Pos-Neg metric (*accuracy* = 0.99; *precision* = 0.99; *recall* = 0.99; *f-score* = 0.99) did not perform significantly worse than the Random Forest with the Euclidean metric.

7.4 Results Summary

Although the majority of the classifier and feature pair metrics performed well (*accuracy* > *0.89*), four of the classifier and feature metric pairs performed with below 20% accuracy and four others performed with below 60% accuracy. The Euclidean Feature and Vector distance metrics did not perform well at tasks one or two. We hypothesize that this is because the two measures did not represent the direction of the difference in feature production (who had more power), which has been considered a key aspect in identifying differences between speakers (Pennebaker et al., 2007). However, the Euclidean Vector distance metric had the highest performance for the SVM classifier (*accuracy* = 0.98; *precision* = 0.90; *recall* = 0.88; *f-score* = 0.94) and the Euclidean Pair distance metric had the best performance for the Random Forest classifier for task of determining whether or not the speakers have the same social power. We hypothesize that this is because the direction of the difference in not meaningful, only the size. Therefore, removing the direction of the difference could reduce a potentially confounding feature. These results indicate that when identifying a difference in social power, representing the magnitude of the difference is the meaningful, whereas when determining who has the more social power, the direction of the difference is most meaningful.

8. Discussion

The results of the classifier and representation pairs indicate that certain feature representations better capture differences in social power and that differences in social power can be reliably classified (Figure 1). Additionally, the suitability of these feature representations for achieving high performance at this task suggests that they might also be useful in identifying other aspects of the social context, such as quality of relation and relationship type.

The performance of the classifier and distance metric pairs suggests that, in general, a Random Forest classifier performs better, and that its performance is maximized by the Pos-Neg Euclidean metric. The majority of the best performing feature representations for classification task one contain information about the differences in production for each feature between the two speakers. The top performing feature representations for classification task one represent both the size of the difference in the production of each feature and the direction of the difference (negative when speaker two has more instances), whereas as the best performing feature representation for classification task two represented the difference between the feature sets.

The differences in feature performance suggest that in cases where there is a power difference, representing the feature differences on more than one dimension and with direction produces higher performance; size and direction of feature differences are important. In detecting cases where there is not a power difference, a more general, single measure of the difference produces the highest performance, which implies that the size of the a general difference in feature production is important, but not the direction. Additionally, the difference in best performing feature representation between classifiers one and two and classifier three indicates that a subset of the linguistic features might be more meaningful than others when predicting who has more power, which could be why metrics that represented individual feature difference performed better at this task.

8.1 Future Directions

This paper examines the ability of different productive feature representations and classifiers to predict relative social power, using data from an entire transcript of an interaction. However, tracking changes in relative social power within a single interaction is an important task for a situated, interactive dialogue agent, so is a promising direction for future research.

Identifying differences in relative social power is one aspect of interactional context, which is itself just one aspect of the social context. These results suggest that relative social power can be reliably used as feature in relationship classification and in interpreting the social and interactional context, as a result, the next step is to examine how well these classifier and feature representation pairs perform when tracking changes in relative social power within an interaction, investigate generalizing the classifier and feature representation pairs to other aspects of relationship classification, such as relation quality and type, and to develop a high-performance relationship classifier that can be implemented as part of a dialogue system.

9. Conclusion

ICSI meeting and Supreme Court trial transcripts were used to test the ability of different classifier and feature representation pairs to accurately predict differences in social power. Classifier and feature representation pairs were identified that are able to classify social power with a high level of accuracy. At the first classification task, determining whether there is a power difference, those feature representations that contained information about both the direction and size of the each of the feature instances performed the best. For the second classification task, determining whether there is no power difference, feature representations that contained information about only the size of the difference performed the best. This difference in which feature representations have the highest performance points to differences in the two classification tasks and the manifestations of whether a person has more power in terms of linguistic feature sets. Overall, the high accuracy of the classifier and feature representation pairs suggests that relative social status is a feature that could be reliably identified and used by higher order social classification and dialogue systems.

References

Berry, P. M., Gervasio, M., Peintner, B. & Yorke-Smith, N. (2011). PTIME: Personalized assistance for calendaring. *ACM Transactions on Intelligent Systems and Technology*, 2, 40:1-40:22.

Biran, O., Rosenthal, S., Andreas, J. McKoewn, K. & Rambow, O. (2012). Detecting influencers in written online conversations. In *Proceedings of the workshop on language in social media (LSM)*.

Bunt, H. & Black, B. (2000). The ABC of computational pragmatics. In H. C. Bunt and W. Black (Eds.), *Computational pragmatics: Abduction, belief and context* (pp. 25-46). Amsterdam: John Benjamins.

Bramsen, P., Escobar-Molano, M., Pang, B. & Kleinberg, J. (2011). Extracting social power relationships from natural language. In *The 49th Annual Meeting of the ACL*, 773-782.

CALO: Cognitive Assistant that Learns and Organizes. (n.d.). Retrieved February 24, 2015, from http://www.ai.sri.com/project/CALO

Danescu-Niculescu-Mizil, C., Pang, B., Lee, L., & Kleinberg J. (2012). Echoes of power: Language effects and power differences in social interaction. Proceedings of WWW 2012.

de Weerd, H., Verbrugge, R. and Verheij, B. (2014) Theory of mind in the Mod game: An agent-based model of strategic reasoning. In A. Herzig and E. Lorini, (Eds.), *Proceedings of the European Conference on Social Intelligence (ECSI-2014), CEUR Workshop Proceedings 1283*, (pp. 128-136).

Duranti, A. (1988). Ethnography of speaking: Toward a linguistics of the praxis. In F. Newmeyer (Ed.), *Linguistics: The Cambridge Survey*, vol. 4, pp. 210-228. Cambridge, UK: Cambridge University Press.

Fitzsimons, G.M. & Kay, A.C. (2004). Language and interpersonal cognition: Causal effects of variation in pronoun usage on perceptions of closeness. *Society for Personality and Social Psychology, Inc.*, 547-557

Giles, H. (2008). Communication Accommodation Theory. *Engaging theories in interpersonal communition: Multiple perspectives*, 161-173.

Gonzales, A.L., Hancock, J.T. & Pennebaker, J.W. (2009). Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 20, 1-17.

Guo, F., Blundell, C., Wallach, H. & Heller, K. (2015). The Bayesian Echo Chamber: Modeling social influence via linguistic accommodation. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*. San Diego, USA.

Huffaker, D. (2010). Dimensions of leadership and social influence in online communities. *Human Communication Research*, 36, 593-617.

Itakura, H. (2001). Describing conversational dominance. *Journal of Pragmatics*, 1859-1880.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg E., Stolke, A., & Wooters, C. (2004). ICSI Meeting Speech LDC2004S02. Web Download. Philadelphia: Linguistic Data Consortium, 2004.

Kacewicz, E., Pennebaker, J.W., Davis, M., Jeon, M. & Graesser, A. (2014). Pronoun use reflects standings in social hierarchies. *Journal Of Language and Social Psychology*, 125-143.

Kim, D., Henderson, M., Gasic, M., Tsiakoulis, P. and Young, S. (2014). The Use of Discriminative Belief Tracking in POMDP-based Dialogue Systems. *IEEE SLT 2014*, Lake Tahoe, NV.

Labov, W. (1966). *The social stratification of English in Ney York City*. Center for Applied Linguistics. Washinton, DC.

Lance, G. N. & Williams, W.T. (1967). Mixed-data classificatory programs I - Agglomerative systems. *Australian Computer* Journal, 1, 15–20.

Litman, D.J. & Silliman, S. (2004). ITSPOKE: An intelligent tutoring spoken dialogue system. *HTL-NAACL – Demonstrations 2004 Demonstration Papers at HTL-NAACL 2004*, 5-8.

Locher, M.A. (2004). *Power and politeness in action: Disagreements in oral communication - Language, Power, and Social Process*. Berlin: Mouton de Gruyter.

Macaulay, R.K.S. (1977). *Language, social class, and education: A Glasgow study*. Edinburgh, UK: University of Edinburgh Press.

Madani, O. & Yu, J. (2011). Discovery of Numerous Specific Topics via Term Co-Occurrence Analysis. (569) California: SRI International.

Mairesse, F. & Young, S. (2014). Stochastic Language Generation in Dialogue Using Factored Language Models. *Computational Linguistics*, 40, 763-799.

Mazur, A. & Cataldo, M. (1989). Dominance and deference in conversation. *Journal of Social Biological Structures*, 87-99.

Moon, S., Potdar, S. & Martin, L. (2014). Identifying student leaders from MOOC discussion forums thorugh language influence. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 15-20). Doha, Qatar.

Myers, K., Kolojejchick, J., Angiolillo, C., Cummings, T., Garvey, T., Gervasio, M., Haines, W., Jones, C., Knittel, J., Morley, D., Ommert, W. & Potter, S. (2011). Learning by Demonstration Technology for Military Planning and Decision Making: A Deployment Story. In *Proceedings of the Conference on Innovative Applications of Artificial Intelligence (IAAI-11)*.

Nguyen, V.A., Boyd-Graber, J., Resnik, P., Cai, D.A., Midberry, J.E. & Wang, Y. (2014). Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, 95, 381-421.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Sciki-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 2825-2830.

Pennebaker, J. (2011). *The secret life of pronouns: What our words say about us*. New York, NY: Bloomsbury Press.

Prabhakaran, V. & Rambow, O. (2013). Written dialogue and social power: Manifestations of different types of power in dialogue behavior. *International Joint Conference on Natural Language Processing* (pp. 216-22). Nagoya, Japan.

Rayner, M., Hockey, B.A., Chatzichrisafis, N., Farrell, K. & Renders, J.M. (2005). A voice enabled procedure browser for the International Space Station. *ACLdemo 2005 Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, 29-32.

Refanidis, I., Alexiadis, A. & Yorke-Smith, N. (2011). Beyond Calendar Mashups: Intelligent Calendaring. *In Proceedings of ICAPS'11 System Demonstrations*. Freiburg, Germany.

Rienks, R., Zhang, D., Gatica-Perez, D. & Post, W. (2006). Detection and application of influence rankings in small group meetings. In *Proceedings of the international conference on multimodel interfaces, ICM 2006*.

Romaine, S. (2002). Language and social class. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences* (pp. 8308-8312). Oxford: Pergamon.

Ross, A.S.C. (1954). Linguistics class indicators in present-day English. *Neuphilologische Mitteilungen*, 171-185.

Rowe, R., Creamer, S., Hershkop, S. & Stolfo, S.J. (2007). Automated social hierarchy detection through email network analysis. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web Mining and Social Network Analysis, ACM*.

Schaul, T., Bayer, J., Wierstra, D., Yi, S., Felder, M., Sehnke, F., Ruckstie, T., &Schmidhuber, J. (2010). PyBrain. *Journal of Machine Learning Research*, 743-746.

Walker, M.A., Whittaker, S.J., Stent, A., Maloor, P., Moore, J., & Vasireddy, G. (2004) Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28, 811-840.

Tambe, M. (1997). Towards flexible teamwork. *Journal of Artificial Intelligence Research*, 7, 83–124.

Thye, S., Willer, D. & Markovsky, B. (1999). From status to power: New models at the intersection of two theories. *Social Forces*, 1471-1495.

Tur, G., Stolcke, A., Voss, L., Dowding, J., Favre, B., Fernandez, R., Frampton, M., Frandsen, M., Frederickson, C., Graciarena, M., Hakkani-Tur, D., Kintzing, D., Leveque, K., Mason, S., Niekrasz, J., Peters, S., Purver, M., Riedhammer, K., Shriberg, E., Tien, J., Vergyri, D. & Yang, F. (2008). The CALO meeting speech recognition and understanding system. *Spoken Language Technology Workshop: SLT 2008*, 69-72.

West, C. & Garcia, A. (1988). Conversational shift work: A study of topical transitions between men and women. *Social Problems*, 551-575.

Yorke-Smith, N., Saadati, S., Myers, K. & Morley, D. (2012). The Design of a Proactive Personal Agent for Task Management. *International Journal on Artificial Intelligence Tools*, 21, 90-119.

Zimmerman, D. H. & West, C. (1975). Sex roles, interpretations and silence in conversation. In B. Thorne & N. Henley (Eds.), *Language and sex: Difference and dominance* (pp. 105-129). Rowley, MA: Newbury House.

IT Press.