Women Like Backchannel, But Men Finish Earlier: Pattern Based Language Modeling of Conversations Reveals Gender and Social Distance Differences

Michal Ptaszynski†, Dai Hasegawa‡, Fumito Masui†

†Department of Computer Science, Kitami Institute of Technology
{ptaszynski,f-masui}@cs.kitami-it.ac.jp

‡Department of Integrated Information Technology, Aoyama Gakuin University
hasegawa@it.aoyama.ac.jp

Abstract. We propose a method for the support of conversation analysis research. In the method groups of conversations are compared with the use of language modeling and machine learning techniques. We compared conversations between people of different age, sex, and social status from a corpus containing over 1,600 minutes of conversations. On groups of conversations differing in one feature (e.g., male vs female interlocutors, or first meeting vs small talk among friends) we performed a text classification experiment with the use of a novel pattern-based language modeling method. This allows verifying the influence of each feature. Moreover, cross-referencing different features allows measuring how much each feature is influential in the context of other features.

1 Introduction

Comparative studies of differences in communication strategies have been researched in different subfields of linguistics [1–5]. Such differences can be viewed from either qualitative or quantitative perspective. The former, often found in comparative and sociolinguistics, focuses on thorough analysis of a small number of the most vivid differences (vocabulary, etc.). On the other hand, quantitative perspective, found in corpus and computational linguistics (CL), provides statistics of which words appear more often in which corpora. Unfortunately, such studies are usually based on words or n-grams (bigrams, trigrams), while actual patterns in language are usually more sophisticated. Finally, CL methods also provide scores representing the performance of machine learning (ML) classifiers trained and tested on selected datasets. Such results could be interpreted as a ratio of differences between the compared corpora, and thus could be of great use in linguistic studies. Unfortunately, ML classifiers require carefully selected training samples.

We propose a method dealing with all of the above drawbacks. It provides quantitative numerical values interpretable as a rate of difference between corpora, and information on which patterns in particular are used more frequently in which corpus. Moreover, it allows qualitative analysis with the use of more sophisticated patterns. To achieve our goal we applied a novel pattern-based language modeling method proposed by Ptaszynski et al. [7], which we further extended and applied in the task of comparing corpora of conversations between people of different age, sex and social status.

The rest of the paper is organized as follows. In section 2 we present the methodology employed in this research. We describe the system applying the language modeling method and explain how we apply it to corpus comparison. Section 3 gives an overview of the corpus and the specific samples used in experiments. Section 4 shows the results of experiments and discussion. Finally the paper is concluded in section 5.

2 Methodology

SPEC or **S**entence **P**attern **E**xtraction ar**C**hitecturte is a system created by Ptaszynski et al. (2011,2014) [7, 8] on the assumption that frequent patterns in language consist of combinations of sentence elements. The system automatically extracts frequent sentence patterns distinguishable for a given corpus (set of sentences). Firstly, the system generates all ordered non-repeated combinations from the elements of a sentence. In every n-element sentence there is k-number of combination groups, such that $1 \leq k \leq n$. The number of combinations generated for one k-element group of combinations is equal to binomial coefficient. The system creates combinations for all values of k in range $\{1, ..., n\}$. The sum of all initially generated combinations is calculated like in eq. 1.

$$\sum_{k=1}^{n} \binom{n}{k} = \frac{n!}{1!(n-1)!} + \frac{n!}{2!(n-2)!} + ... + \frac{n!}{n!(n-n)!} = 2^n - 1 \quad (1)$$

Next, all non-subsequent elements are separated with an asterisk ("*"). From all patterns generated this way SPEC retains only those which occur frequently (occurrence $O > 1$). To apply the method to binary classification task we extended it and used O to calculate normalized weight w_j of patterns according to equation 2. The score of one sentence is calculated as a sum of weights of patterns found in the sentence, like in eq. 3.

$$w_j = \left(\frac{O_{pos}}{O_{pos} + O_{neg}} - 0.5\right) * 2 \quad (2) \qquad score = \sum w_j, (1 \geq w_j \geq -1) \quad (3)$$

If the initial collection of sentences was biased toward one of the sides (e.g., more sentences of one kind, or the sentences were longer, etc.), there will be more patterns of a certain sort. Thus to avoid bias in the results, instead of applying a rule of thumb, threshold is automatically optimized. The above settings are automatically verified in the process of evaluation (10-fold cross validation) to choose the best model. The metrics used in evaluation are standard Precision (P), Recall (R) and balanced F-score (F).

2.1 Corpora Comparison with SPEC

SPEC, as described in section 2, provides both qualitative information (specific patterns) and quantitative information (pattern weights w_j, and $score$ for each input sentence). Therefore we applied it in comparison of corpora.

One kind of the information provided by SPEC is the result of automatic classification of two provided collections of sentences of opposite characteristics. When these are exactly the same, Precision will be 0 for threshold $t > 0$ and 0.5 for $t \leq 0$. Recall will be 0 for $t > 0$ and 1 for $t \leq 0$. Any result different to the above will mean that the two corpora are different. Thus we can consider the result of the classification as a rate of similarity between two corpora.

Furthermore, patterns generated in the process could appear either uniquely on one of the sides (e.g. uniquely positive) or in both (ambiguous). Ambiguous patterns could appear more frequently on one of the sides (weight biased toward 1 or -1). Thus the weights can be interpreted as a rate of probability a pattern will appear in a corpus.

Finally, analysis of patterns characteristic only for one side and the sentences in which they appear could provide interesting linguistic discoveries. Since the patterns extracted automatically represent all probable frequent patterns hidden in the two compared corpora, then assuming the corpora cover a representative sample of the compared feature, the patterns already known to linguistics should also be included in the weighted pattern list. Moreover, we can expect new patterns unknown before. Some of them could be data-dependent. However, 10-fold cross validation filters out only those patterns which were useful across all tests.

3 Datasets for Experiment

In the experiment we used the BTSJ (Basic Transcription System for Japanese) corpus [9]. It contains 99 conversations (covering 1,604 minutes) between people of different age, sex, social distance and status. The conversations were performed either by friends or people who first met. The conversations are either small talks or on a specific topic. They are between men, women or mixed, students or adults. Each of those features of conversations can be considered as opposites. We extracted subsets for which only one feature would differ. Comparison of the subsets should provide patterns characteristic for the one differing feature. We extracted 24 small talks between female students, half of which by friends another half by people who first met, and 12 sets with similar conditions for male students. We compared how much the way of talking differs for female and male students when they talk to their friends or to unrelated peers. The summary of conversation sets used in the experiment is reported in Table 1.

4 Experiment Results and Discussion

General Observations
When the average number of sentences in conversations is compared, on "first met" male interlocutors exchanged more information than with friends. Females on the contrary, small talks between friends were about twice as long. Males used longer sentences and exchanged turns less often than females, which used backchannel more often. Although these findings need to be interpreted within the closed data, they seem to support other findings [1, 2], which suggest that for males it is important to convey specific information rather than keep up the conversation as it is for females.

Feature Differences
Comparison of the results achieved by the classifiers shows that higher F-scores were achieved for females rather than males, which means that the compared conversations were easier to distinguish. This suggests that women talk more differently than men to a person they just met than to friends. In particular, the results achieved by the classifier for male conversations were F=.79 with P=.74 and R=.85, while for women the highest F-score reached .85 with P=.79 and R=.96. Comparison of the results for male and female conversations are represented in Figures 2a and 2b.

Discussion
Next, we analyzed specific patterns characteristic to each of the compared sides. We noticed that there were the same patterns for both male and female students in similar situations. For example, the pattern *nanka*na* appears in friend conversations for both sexes. Example sentences with this pattern are given below. The first two examples are for female students. The latter two are for male students.

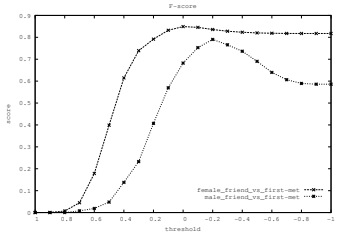
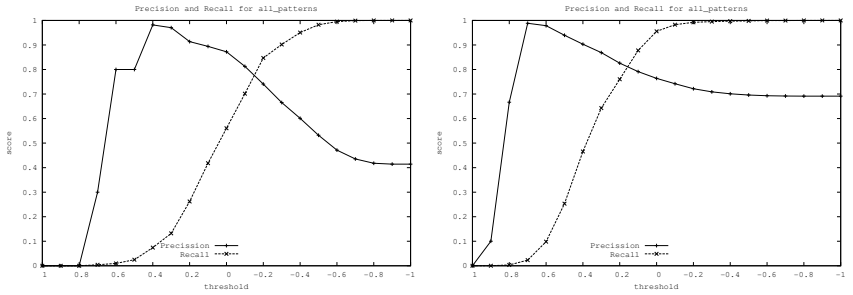


Fig. 1: F-scores for both datasets.

Small talk conversations		No. of samples	Avg. sent. length	Avg. sentences per conversations
Female-student	first met	12	12.7	288.9
	friends	12	9.3	550.0
Male-student	first met	6	12.4	326.5
	friends	6	14.5	245.3

Table 1: Summary of the conversation sets.



(a) Male students conversation dataset. (b) Female students conversation dataset.

Fig. 2: Precision and Recall with Break-Even Point for both conversation datasets.

Ex.1: *Nanka... bannō nabe mitai na yatsu.* (Something like a... universal cooking pot.)
Ex.2: *Nanka gakugaku, mitai na.* (Something, like a sound of knocking.)
There were also similar patterns for both sexes under the "first met" condition, such as the pattern *so*desu* in the examples below.
Ex.3: *Aaa, sō nan desu ka* (Oh, so that's the case.)
Such patterns could be characteristic for social distance rather than sex.
There were also patterns specific for a particular sex. Self referential *ore* for boys and *atashi* for girls (both meaning "I/me") are good examples.
Ex.4: *Ore 1-kai mo nai kara ne.* (I[masculine] haven't [done it] even once, you know.)
Ex.5: *Nanka atashi, tento tte sugoi suki.* (Oh, I[feminine] just love tents so much.)
There were also patterns characteristic for specific social distance. For example, a pattern *sō sō sō!* ("yes, yes that's right!") does not contain any distance-specific vocabulary (like in the case of gender-related *ore* vs. *atashi*). However, in practice although the pattern is often used in friends' conversations, it does not appear at all in first-met conversations. On the other hand a pattern similar in meaning *hai hai hai* ("yes, yes, yes") is used in first-met conversations, but does not appear in friend-friend conversations. We also looked at conversation topics similarly to previous research [1, 2]. For friend-students (both sexes), the topic of "an exam" was equally frequent. However, a topic of "a marriage" appeared only in female student conversations, similarly to "food" and "alcohol". On the other hand "newspapers" were the male-specific topic.

5 Conclusions and Future Work

We studied differences of how people talk by comparing frequent patterns appearing in conversations. We found out that male interlocutors used longer sentences and exchanged turns less often than females. When it comes to differences of talking to friends and newly met people, for females they were much grater than for males. Some patterns appeared for both males and females, which suggests they could be typical for linguistically expressed social distance. Some patterns were specific for a particular sex (like *ore* [masc.] and *atashi* [femin.]), while others although could be used in any context, in practice were used by only one side (*so so so* (friend-friend) vs. *hai hai hai* (first-met)).

- *Could you think of some specific patterns you use/recognize in your everyday conversations?*
- *Could some of those patterns be specific to a wider group (men, women, linguists)?*
- *Could we talk about slang-patterns similarly we talk about slang words?*

References

1. Adelaide Haas. 1979. Male and female spoken language differences: Stereotypes and evidence. *Psychological Bulletin*, Vol. 86, No. 3 (1979), pp. 616-626.
2. Lynette Hirschman. 1994. Female?male differences in conversational interaction. *Language in Society*, 23, pp 427-442.
3. Cameron, D. 1998. Gender, language, and discourse: A review essay. Signs: *Journal of Women in Culture and Society*, 23, pp. 995-973.
4. Eckert, P., & McConnell-Ginet, S. 2003. *Language and gender*. New York: Cambridge University Press.
5. Holmes, J., & Meyerhoff, M. (Eds.) 2003. *The handbook of language and gender*. Malden, MA: Blackwell.
6. Kaori Sasai. 2006. The Structure of Modern Japanese Exclamatory Sentences: On the Structure of the *Nanto*-Type Sentence. *Studies in the Japanese Language*, Vol, 2, No. 1, pp. 16-31.
7. Michal Ptaszynski, Rafal Rzepka, Kenji Araki and Yoshio Momouchi. 2011. Language combinatorics: A sentence pattern extraction architecture based on combinatorial explosion. *International Journal of Computational Linguistics (IJCL)*, Vol. 2, Issue 1, pp. 24-36.
8. Michal Ptaszynski, Fumito Masui, Rafal Rzepka, Kenji Araki. 2014. First Glance on Pattern-based Language Modeling. *Language Acquisition and Understanding Research Group (LAU)*, Technical Reports, Summer 2014, Sapporo, August 09, 2014.
9. Mayumi Usami (Ed.). 2007. *BTS ni yoru nihongo hanashikotoba kōpasu 1 (hatsu taimen, yuujin; zatsudan, tōron, sasoi)* [Conversation corpus of spoken Japanese using the Basic Transcription System (first meeting, friend's conversation, small talk, discussion, invitation)] (In Japanese), Tokyo University of Foreign Studies, Tokyo, Japan, 2007.