Women Like Backchannel, But Men Finish Earlier

Pattern Based Language Modeling of Conversations Reveals Gender and Social Distance Differences

Michal Ptaszynski Fumito Masui Dai Hasegawa
Kitami Institute of Technology Aoyama Gakuin University

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.

SENTENCE PATTERNS

Sentence patterns = ordered non-repeated combinations of sentence elements.[2]

for $1 \le k \le n$, there is $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ all possible k-long patterns, and

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

Normalized pattern weight

$$w_j = \left(\frac{O_{pos}}{O_{pos} + O_{neg}} - 0.5 \right) * 2$$

Score for one sentence

$$score = \sum w_j, (1 \ge w_j \ge -1)$$

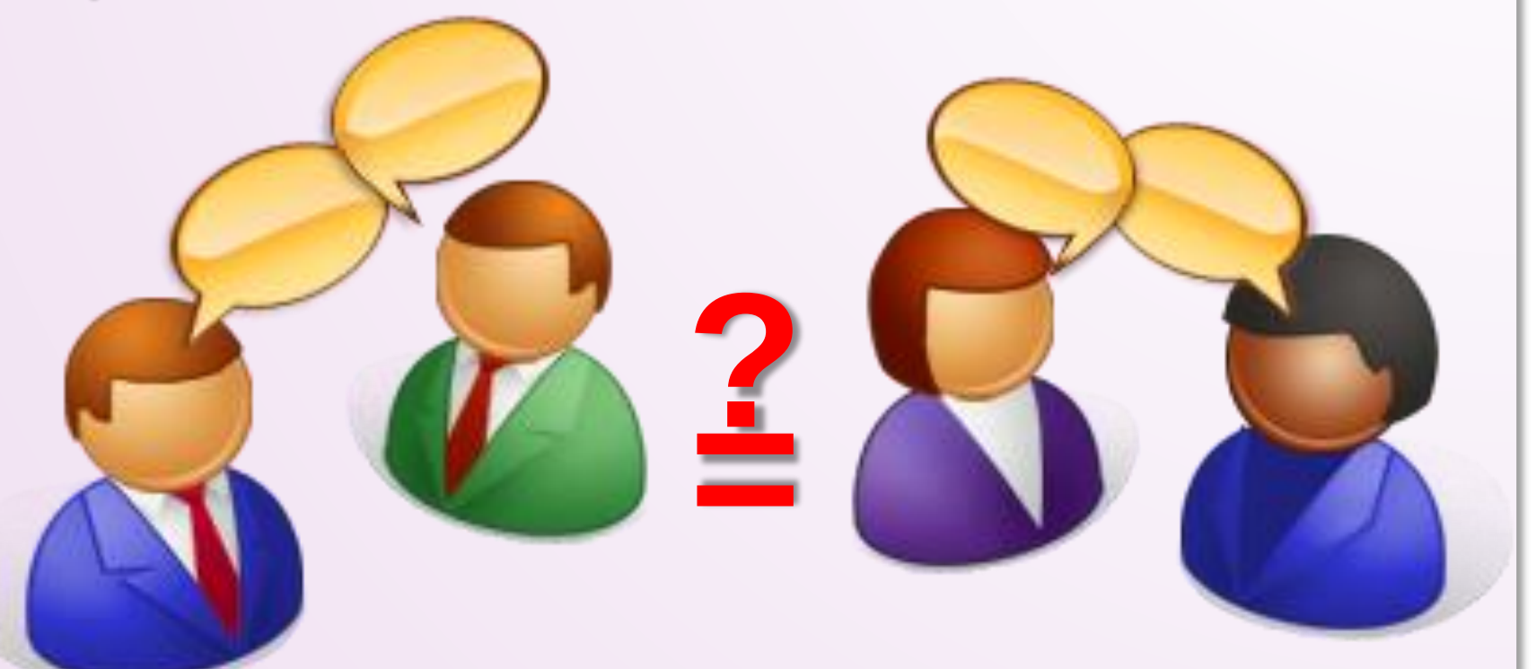
CORPORA COMPARISON METHOD

1. Compare results of automatic classification of conversations with opposite features. (10-fold cross validation, Precision, Recall, F-score)

A) If two corpora are the same,
- below threshold P,R and F-score = 0
- above threshold P=0.5, R=1, F=0.67

B) If two corpora have no similarities (none of the patterns extracted from one corpus appears in the other), P, R, F = 1

C) A Classification result in a range {A)… B)} is a rate of similarity between the two compared corpora



2. Weights of patterns can be interpreted as a probability rate of how often a pattern appears in the corpora

A) 1 or -1: pattern is characteristic to one of the two sides

B) 0: pattern is not characteristic to any side

C) Other (1>w>0,0>w>-1): pattern is biased toward one of the sides.

◆ A)~C) Applicable in corpus linguistic studies.

◆ Analyzing A) with corresponding sentences could provide interesting linguistic discoveries.

If the corpora cover a representative sample of the compared feature,
A) will contain the patterns already known to linguists.
Moreover, new patterns unknown before can be expected.
Some of them will be data-dependent. However, filtering through a 10-fold cross validation will retain only most useful patterns across all tests.

BTSJ CORPUS

The BTS (Basic Transcription System) for Japanese corpus [3] contains 99 conversation transcripts (1,604 minutes of talking) between:

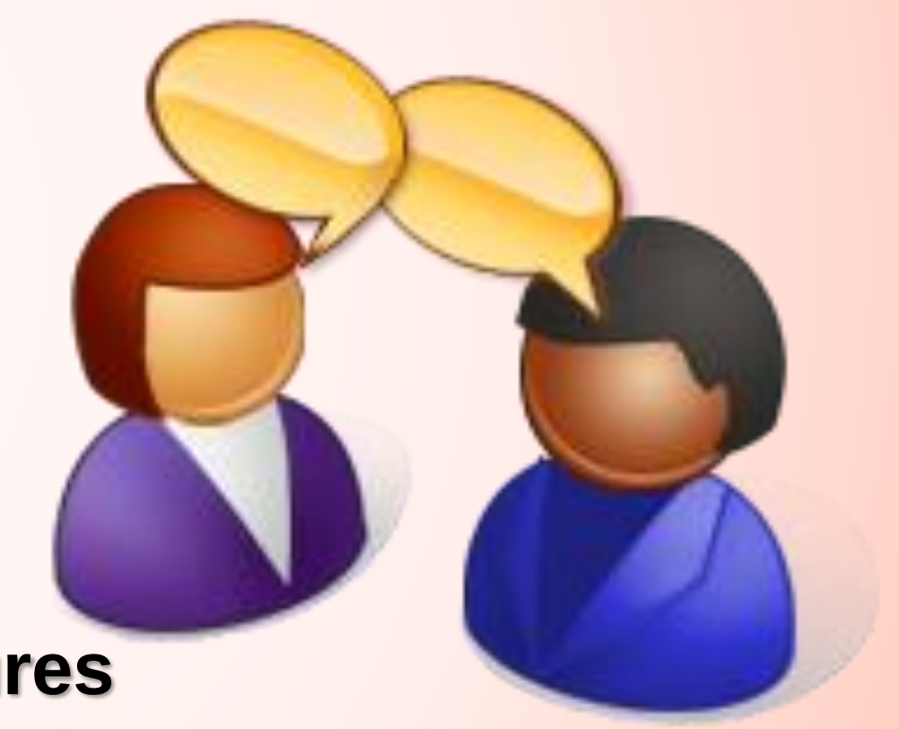
A) native speakers (used in this research), or a native speaker and a language learner

B) friends or people who first met

C) small talk, or specific topic

D) men only, women only, or mixed

E) students or adults



◆ A) ~ E) separate dimensions with opposite features

◆ Extract conversation subsets for which only one feature differs

◆ Comparing such subsets should provide sentence patterns characteristic for the differing feature.

DATASETS FROM BTSJ CORPUS

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

EXPERIMENT AND DISCUSSION

GENERAL OBSERVATIONS

First meeting vs. with friends

1. Men talk more on first meeting than with friends
2. Women talk 2-times longer with friends than on first meeting

Men
- use longer sentences
- exchange turns less often

Women
- use backchannel more often.

? → For man it could be important to convey information (goal oriented) rather than keep the conversation going (state oriented).

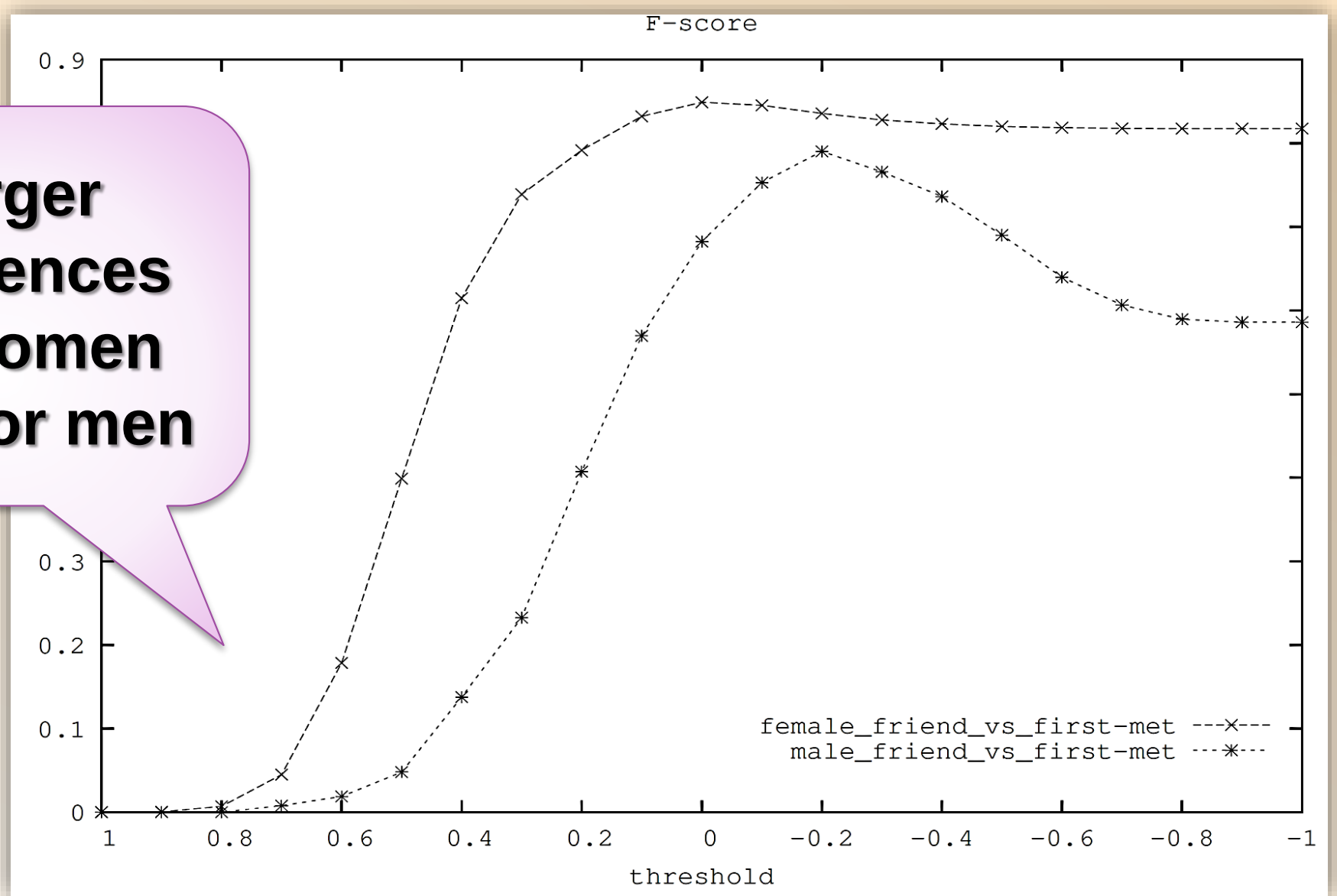
FEATURE DIFFERENCES

↓ Higher classification F-scores were achieved for women rather than men

↓ Higher F-score = the compared conversation sets were easier to distinguish

↓ Comparing to men, women talk more differently to a person they just met than to friends.

Highest results
for men: F = 0.79, P = 0.74, R = 0.85
for women: F = 0.85, P = 0.79, R = 0.96



Larger differences for women than for men

DETAILED ANALYSIS

Extracted patterns

	women		men	
	freq.	example pattern	freq.	example pattern
friends	257	なん＊な	83	ん。
	251	わ	50	俺
	244	う＊よ	39	だね
	202	なんか＊な	35	なんだ
	162	なんか＊か	27	そうだよ
	160	かな	26	なんか＊な
	157	ん。	22	そうだよね
	152	んで	21	なー。
	149	みたい	18	だから＊う＊。
	140	でも＊、	15	そうそうそう
	122	みたいな	13	すか
	94	じゃん。	13	まじ
	92	んない	12	やっぱ＊な
	91	う＊よね	12	みたいな。
	51	ちゃん	12	やん
	51	んだっ	11	でしょ？
	51	たんだ	11	奴
	50	なんか＊た＊な	10	お前＊。
	50	あたし	10	だろうね
first-met	155	う＊です＊。	243	そ＊です＊。
	125	い＊です	199	ですね
	103	う＊んです	100	そうですね
	93	なんです	79	、＊んですか
	62	たんで	74	あぁ
	59	そうですね	69	なんです
	58	あ＊ですか＊。	55	あ＊んですか
	58	ー、＊です	49	ーん
	22	でも＊です＊。	44	ええ。
	19	あ＊ですよね	32	ないんで
	16	あ、そう＊ですか。	28	あ＊そう＊んですか
	16	あ、そうです	23	んですよ。
	16	、なるほど	18	結構
	15	なるほどね	17	一応
	12	よろしくお願いします	16	あるんで
	15		14	、はいはいはい

Example sentences

Example 1. ♀
なんか… 万能鍋見ないなやつ
Nanka… banno nabe mitai na yatsu.
(Something like a… universal cooking pot.)

Example 2. ♀
なんかすごい高性能なスキャンなーだとー
Nanka sugoi koseino na sukyana da to–
(Oh its like an amazingly high-performance scanner!)

Example 3. ♂
なんかがくがくみたいな。
Nanka gakugaku, mitai na.
(Something, like a sound of knocking.)

Example 4. ♂
インターネットとしては、なんか結構、不足なとこもある。
Intanetto to shite wa, nanka kekko, fusoku na toko mo aru.
(So when it comes to the Internet, it has like pretty a lot of deficiencies.)

Example 5. ♀ ♂
あぁぁ、そうなんですか
Aaa, so nan desu ka
(Oh, so that is the case [I understand now])

Example 6. ♂
俺一回もないからね。
Ore 1-kai mo nai kara ne.
(I[masculine] haven't [done it] even once, you know.)

Example 7. ♀
なんかあたし、テントってすごい好き。
Nanka atashi, tento tte sugoi suki. (Oh, I[feminine] just love tents so much.)

CONCLUSIONS

Investigated differences of how people talk, by comparing sentence patterns from conversations.

1. Sentence pattern = ordered combination of sentence tokens.
2. Automatically extracted frequent patterns from conversations.
3. Performed a text classification experiment using those patterns.
4. Used classification results to explain differences between conversations.

- Men use longer sentences and exchange turns less often than women.
- Difference between talking to strangers and friends is greater in women.
- Some patterns are typical for linguistically expressed social distance (first met はいはいはい vs. with friends:そうそうそう).
- There were also patterns specific for a particular sex (words like 俺/ore/ and あたし/atashi/)

In the future we will analyze other conversations and compare different kinds of corpora, not limited to conversations.

REFERENCES

[1] 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.

[2] 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.

[3] Mayumi Usami (Ed.). 2007. *BTS ni yoru nihongo hanashikotoba kopasu1 (hatsutaimen, yujin; zatsudan, toron, 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.

[4] Adelaide Haas. 1979. Male and female spoken language differences: Stereotypes and evidence. *Psychological Bulletin*, Vol. 86, No. 3, pp. 616-626.

[5] Lynette Hirschman. 1994. Female-male differences in conversational interaction. *Language in Society*, 23, pp 427-442.