Gossip is more than just story telling.

Topic modeling and quantitative analysis on a spontaneous speech corpus

Boróka Pápay, Bálint György Kubik, Júlia Galántai

Presenter: Júlia Galántai

MTA TK "Lendület" Research Center for Educational and Network Studies (RECENS) Hungarian Academy of Sciences, Centre for Social Sciences, Budapest, Hungary

http://recens.tk.mta.hu

Text2Story 2018

First Workshop on Narrative Extraction From Text

26th March, 2018 - Grenoble, France

Acronym: EVILTONGUE

Project number 648693

European Research Council

Established by the European Commission

European Commission

Horizon 2020

European Union funding

for Research & Innovation

Introduction

- Quantitative approach to identify gossip in a spontaneous speech corpus with LDA topic modeling and quantitative analysis.
- Test literature assumptions about gossip on a spontaneous human speech corpus.
- Tools: text preprocessing, topic modelling, quantitative text characteristics
- Relationships between topic memberships, manually annotated features with emphasis on gossip
- distinguish gossiping and storytelling topics by dividing gossip and non-gossip texts.

Gossip and its roles in society

- Gossip is one of the most widespread human activities like enhancing human cooperation, establishing social order, information sharing, norm enhancing or stress reduction.
- Two thirds of human conversations are about social topics that can be labeled as gossip. Gossip is the core of social relations and society itself [Dunbar 2004]
- Language caused a significant increase in communication in groups and in information exchange. It also allows us to get information about what happens in a social group, while gossip plays an important role in the sustaining of human cooperation [Dunbar 2004]
- Gossip also often transmits reputational information about individuals, establishing social order. [Feinberg 2014], [Hess 2006], [Novak2005].

Data and database 1

- For our analysis we used a unique corpus of Hungarian language which consists of approximately 550 hours of spontaneous speech.
- The documents are transcripts of organic human dialogues, separated by natural silence no longer than 2 seconds.
- The high-quality audio recordings were recorded during a Hungarian entertainment programme covering a period of 8 days.
- The recordings were obtained using personal microphones of eight participants of a gameshow covering the whole interval of their wake times.
- Manual annotation, tagged those parts of the text where the speakers were talking about a person who was a participant or former participant of the gameshow, but was not present. These parts of the text are gossip dialogues.

Data and database 2

- When the speakers were mentioning multiple participants who were not present at the dialogue all mentioned participants were marked individually as a gossip target.
- Those dialogues were not tagged as gossip where the person whom the speakers were talking about was not present **but** was not a participant or former participant of the gameshow (like acquaintances, family relatives, and so on). These discourses were mentioned as storytelling later on in our topic model.

Annotation of the corpora 1

- Manual annotation process: annotation codes to mark the speech about a third person who is not present during the conversation (gossip).
- The sender and the receiver could be identified by annotators - names and target of gossip.
- Name tags - turn-taking and simultaneous speaking.
- The time interval of speech was tagged by using timestamps
- Indicating silent participants via annotation: to measure the number of persons present while gossiping.
- Codes signing incomprehensible, unidentifiable speech.

Annotation of the corpora 2

- Conversation behavior when gossiping - verbal and non-verbal emotions of speakers.
- Annotation marks non-verbal signs of emotions during the conversations were: laughter, crying, sighing, etc.
- Verbal emotions semantically with emotion and sentiment analysis [Szabó15] [Szabó16]
- Work quality of the annotators: same text files to compare them by means of matching annotation tags, name tags, and timestamp usage - comparing their work to each other and by using a reference annotator.
- Text similarity - cosine similarity and Levenshtein distance.

Research Directions

1. What were the participants of the gameshow talking about? Majority of their speech is about other people.

2. Gossip is confidential, during gossip, less people are present. Gossiping entails confidential topics among people close to each other [Shimanoff85].

3. Other people can be outside of the closed environment or can be fellow players.

4. Gossip is usually among a few individuals [DiFonzo07], with less speakers than in non-gossip situations.

5. Close acquaintances or friends speak longer. The level of confidentiality required for gossiping takes longer time to form. Segments that contain gossip are longer.

Text preprocessing

- The number of unique terms were reduced using lemmatization with stopword dictionary, and frequency-based filtering.
- Preprocessed data an input for topic modeling with Latent Dirichlet Allocation (LDA).
- Agglutinative language - Hungarian requires lemmatization, large number of words with similar meanings. Magyarlánc software to lemmatize for morphological analysis and part-of-speech tagging [Zsibrita13].
- Stopword dictionary: using Magyarlánc. Excluding: adverbs, auxiliary verbs apart from verbal adverbs, adpositions, auxiliary verbs, interjections, particles, determiners, coordinating and subordinating conjunctions.

LDA

- The document-term matrix of unigram counts - text preprocessing - low number of unique lemmas, terms appearing in less than 5 documents and words present in more than 60% of texts were removed to discard overly rare and overly frequent unigrams.
- The final document-term matrix had 12.961 documents and 8.530 terms.
- Gensim version 3.2.0, a topic modeling library for Python 3, was used [Rehurek 2010].
- Randomly split corpus into train, test, and validation set with 50%, 25%, 25% of data.
- We decided to use 50 topics, a number providing coherent topics and still enabling qualitative assessment. The average semantic coherence metric as defined by Mimno et al [Mimno 2005] was -3.31.
- Model building, choosing the number of topics metrics logarithmic perplexity (measured on the test and validation set). Jaccard distances and Kullback-Leibler differences between consecutive training steps, coherence metric.

Results

- Our results include 50 topics that are present in our speech segments.
- For our first research direction, we were able to categorize these topics by their main theme. 24 of these topics included speech about everyday life, or so called ‘internal issues’ like kitchen and food, clothes, body care.

Wordclouds of the most important terms in the two topics with gossip. Words were translated from Hungarian to English. Word sizes are proportional to LDA weights. (Source: own visualization)

Results

- 10 of the topics seem to be about the entertainment show of which the speakers were part, selection process, duels, other organized games.
- In other 12 topics, they mostly told stories about other people, who are not part of the participants.

Wordclouds of the most important terms in the two topics with gossip. Words were translated from Hungarian to English. Word sizes are proportional to LDA weights. (Source: own visualization)

Results

- Third research direction: Two topics were distinctively about each other, called later ‘gossip topics’, two of the most coherent ones.
- Topics were categorized as “gossip topics”, if the number of gossip annotation tags provided by human transcribers were significantly high.

Wordclouds of the most important terms in the two topics with gossip. Words were translated from Hungarian to English. Word sizes are proportional to LDA weights. (Source: own visualization)

Results

- Topics with gossip contain the names of some of the participants with a high weight. These words most probably describe the actions of these individuals and feelings or actions associated with them.
- Topics containing gossip about each other are distinct from other non-gossip topics, including those that are about people from the outside.

Wordclouds of the most important terms in the two topics with gossip. Words were translated from Hungarian to English. Word sizes are proportional to LDA weights. (Source: own visualization)

Results

- Certain topics correlated with other characteristics and variables from segments.
- Segments containing gossip differ quantitatively from other, non-gossip segments.
- In contrary to our assumption, texts that contain gossip are not necessarily longer than their non-gossip counterparts.

Segment characteristics	internal issues	story	gossip1	gossip2
length (in rows)	-0,074	-0,054	-0,018	-0,076
gossip ratio	-0,064	0,027	0,202	0,235
people present at the conversation	0,081	non sign.	non sign.	-0,041
people speaking at the conversation	non sign.	-0,037	-0,012	-0,051
turn taking at conversation	non sign.	-0,037	non sign.	-0,047
ratio of "joyful" words	-0,016	-0,036	-0,058	-0,091
ratio of words associated with sadness	0,016	non sign.	non sign.	-0,037
ratio of words associated with anger	non sign.	-0,021	0,018	0,024
positive_ratio	-0,023	-0,045	-0,067	-0,092
negative_ratio	-0,02	-0,051	-0,058	-0,072
ratio of non-verbal annotation tags	-0,026	0,014	-0,046	-0,105
the ratio of laughter annotation tags	-0,042	non sign.	-0,073	-0,123
the ratio of crying annotation tags	0,033	0,078	non sign.	-0,048

Conclusions

- Less people were present during conversations, and in both topics the number of individual speakers is significantly lower.
- Gossip topics usually contain names, personal pronouns, simple verbs as ‘say’, ‘go’ or ‘think’ and verbs related to expressions of own emotions such as ‘feel’ and ‘understand’.
- More anger in both gossip topics, but less joy related words were present in them.
- Story topics in general contain less anger, harsher nonverbal emotions.
- The participants underused nonverbal communication during conversations containing gossip such as laughter or crying.
- Gossiping is different from storytelling and other social topics. In the only category the usage of non-verbal emotions is the ‘story’ category where participants mainly talk about their outside acquaintances.

Thank you for your attention!

papay.boroka@tk.mta.hu
kubikbalint@gmail.com
galantai.julia@tk.mta.hu