Automated Content Analysis of Discussion Transcripts

Vitomir Kovanović v.kovanovic@ed.ac.uk

Dragan Gašević dgasevic@acm.org

School of Informatics,
University of Edinburgh
Edinburgh, United Kingdom
v.kovanovic@ed.ac.uk

31 Aug 2015,
University of Edinburgh,
United Kingdom

Asynchronous online discussions - *"gold mine of information"* (Henri, 1992)

- They are frequently used for all types of education delivery,
- Their use produced large amount of data about learning processes,
- Their use is well supported by the social-constructivist pedagogies.

Asynchronous online discussions - issues and challenges

- Produced data is used mainly for research after the courses are over,
- Content analysis techniques are complex and time consuming,
- Content analysis had *almost no impact* on educational practice (Donnelly and Gardner, 2011),
- There is a need for more proactive use of the data through automation:
  - Few attempts for automated content analysis,
  - Focus mostly on surface level characteristics, and
  - Not based on well established theories of education.

Overall idea

Overall idea

To examine how we can use text mining for automation of content analysis of discussion transcripts.

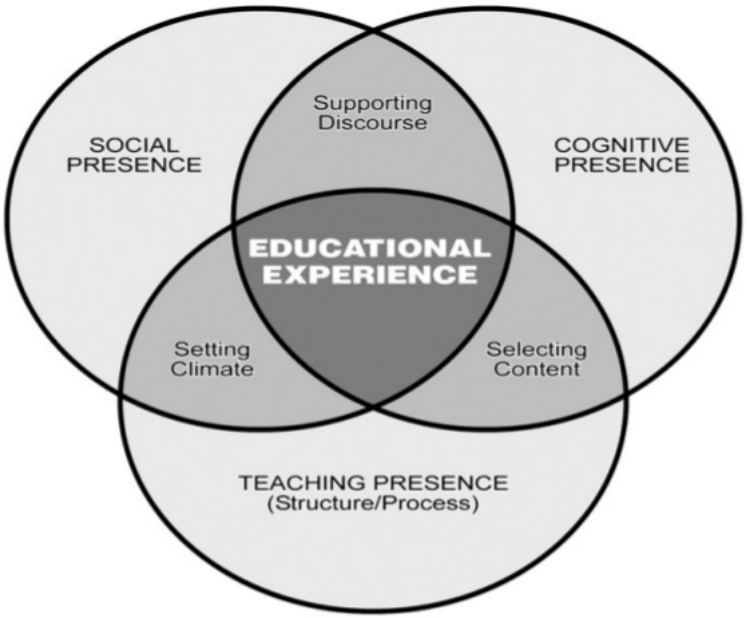
More specifically,

We looked at the automation of content analysis of **cognitive presence**, one of the three main components of Community of Inquiry framework.

Community of Inquiry (CoI) model

Community of Inquiry model (Garrison, Anderson, and Archer, 1999)

Conceptual framework outlying important constructs that define worthwhile educational experience in distance education setting.



Three presences:

- **Social presence**: relationships and social climate in a community.
- **Cognitive presence**: phases of cognitive engagement and knowledge construction.
- **Teaching presence**: instructional role during social learning.

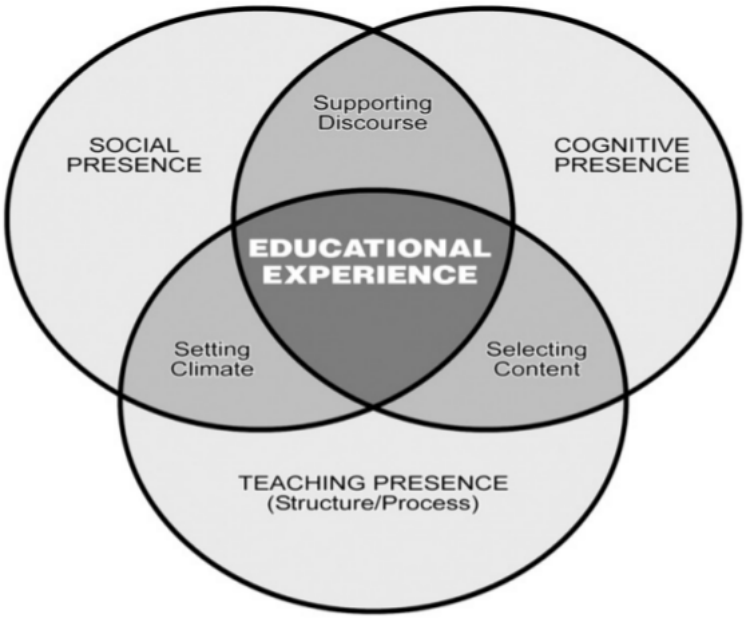
CoI model is:

- Extensively researched and validated.
- Adopts Content Analysis for assessment of presences.

Community of Inquiry (CoI) model

Community of Inquiry model (Garrison, Anderson, and Archer, 1999)

Conceptual framework outlying important constructs that define worthwhile educational experience in distance education setting.



Three presences:

- **Social presence**: relationships and social climate in a community.
- **Cognitive presence**: phases of cognitive engagement and knowledge construction.
- **Teaching presence**: instructional role during social learning.

CoI model is:

- Extensively researched and validated.
- Adopts Content Analysis for assessment of presences.

Cognitive presence

Cognitive Presence

"an extent to which the participants in any particular configuration of a community of inquiry are able to construct meaning through sustained communication." (Garrison, Anderson, and Archer, 1999, p .89)

Four phases of cognitive presence:

1. **Triggering event**: Some issue, dilemma or problem is identified.
2. **Exploration**: Students move between private world of reflection and shared world of social knowledge construction.
3. **Integration**: Students filter irrelevant information and synthesize new knowledge.
4. **Resolution**: Students analyze practical applicability, test different hypotheses, and start a new learning cycle.

Cognitive presence coding scheme

- Use of whole message as unit of analysis,
- Look for particular indicators of different sociocognitive processes,
- Requires expertise with coding instrument and domain knowledge.

Table 2. Exploration

Descriptor	Indicators	Sociocognitive Processes
Inquisitive	Divergence—within the online community	Unsubstantiated contradiction of previous ideas
	Divergence—within a single message	Many different ideas/themes presented in one message
	Information exchange	Personal narratives/descriptions/facts (not used as evidence to support a conclusion)
	Suggestions for consideration	Author explicitly characterizes message as exploration—e.g., "Does that seem about right?" or "Am I way off the mark?"
	Brainstorming	Adds to established points but does not systematically defend/justify/develop addition
	Leaps to conclusions	Offers unsupported opinions

Example: One reason I think it is seldom used is that it is too complicated to get cooperation. Another may be the mind-sets of those in charge to change practices.

Community of Inquiry (CoI) model

Issues and challenges:

- Very labor intensive,
- Crude coding scheme,
- Requires experienced coders,
- Can't be used for real-time monitoring,
- Not explaining reasons behind observed levels of presences, and
- Not providing suggestions and guidelines for instructors to direct their pedagogical decisions.

Data set

- Six offerings of graduate level course in software engineering.
- Total of 1747 messages, 81 students,
- Manually coded by two coders ($\texttt{agreement} = 98.1\%$, $\texttt{Cohen's}\ \kappa = 0.974$),

ID	Phase	Messages	(%)
0	Other	140	8.01%
1	Triggering Event	308	17.63%
2	Exploration	684	39.17%
3	Integration	508	29.08%
4	Resolution	107	6.12%
	All phases	1747	100%

Number of Messages in Different Phases of Cognitive Presence

Feature extraction

- Unigrams, Bigrams and Trigrams,
- Part-of-Speech Bigrams and Trigrams,
- Backoff Bigrams and Trigrams:

Example: *"John is working."*

Bigrams:

- john is,
- is working.

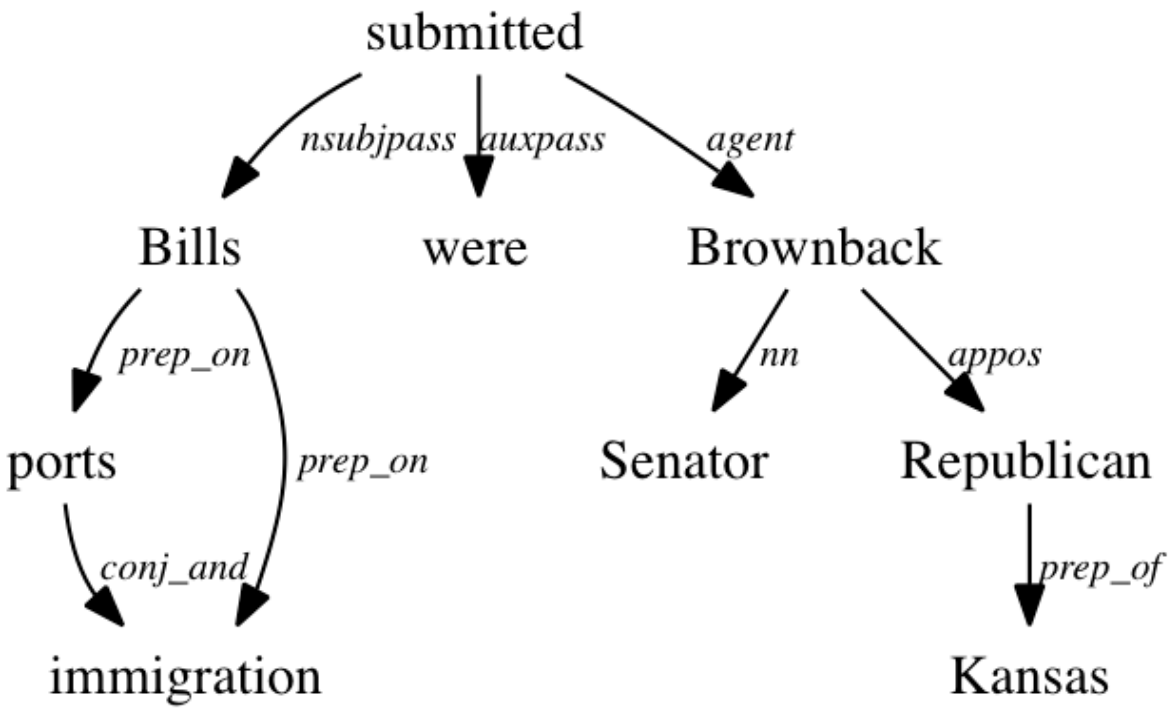
Backoff Bigrams:

- john ⟨verb⟩,
- ⟨noun⟩ is,
- is ⟨verb⟩
- ⟨verb⟩ working.

Feature extraction

- Dependency triplets: ⟨rel, head, modifier⟩

  Example: *"Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas."*



⟨nsubjpass, submitted, Bills⟩
⟨auxpass, submitted, were⟩
⟨agent, submitted, Brownback⟩
⟨nn, Brownback, Senator⟩
⟨appos, Brownback, Republican⟩
⟨prep_of, Republican, Kansas⟩
⟨prep_on, Bills, ports⟩
⟨conj_and, ports, immigration⟩
⟨prep_on, Bills, immigration⟩

Feature extraction

- Backoff dependency triplets:

  Example: *"Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas."*

  Dependency triplet:

  - $\langle$conj_and, ports, immigration$\rangle$

  Backoff dependency triplets:

  - $\langle$conj_and, $\langle$noun$\rangle$, immigration$\rangle$
  - $\langle$conj_and, ports, $\langle$noun$\rangle\rangle$
  - $\langle$conj_and, $\langle$noun$\rangle$, $\langle$noun$\rangle\rangle$

Additional features

- Number of named entities in the message
  Brainstorming should involve more concepts than posing a question,
- Is message first in the discussion?
  Posing questions is more likely to be initiating discussions,
- Is message a reply to the first message in the discussion?

Table 2. Exploration

Descriptor	Indicators	Sociocognitive Processes
Inquisitive	Divergence—within the online community	Unsubstantiated contradiction of previous ideas
	Divergence—within a single message	Many different ideas/themes presented in one message
	Information exchange	Personal narratives/descriptions/facts (not used as evidence to support a conclusion)
	Suggestions for consideration	Author explicitly characterizes message as exploration—e.g., "Does that seem about right?" or "Am I way off the mark?"
	Brainstorming	Adds to established points but does not systematically defend/justify/develop addition
	Leaps to conclusions	Offers unsupported opinions

Example: One reason I think it is seldom used is that it is too complicated to get cooperation. Another may be the mind-sets of those in charge to change practices.

Classification

Classifier:

- SVM classifier with RBF kernel.
- Accuracy and kernel parameter tuning evaluated using nested 5-fold cross-validation.
- Only features with support of 10 or more,
- Accuracy evaluated using 10 fold cross-validation,
- Comparison of models using McNemar's test.

Implementation:

- Implemented in Java,
- Feature extraction using Stanford CoreNLP[1] toolkit,
  - Tokenization, Part-of-Speech, and Dependency parsing modules
- Classification using Weka (Witten, Frank, and Hall, 2011) and LibSVM (Chang and Lin, 2011), and
- Statistical comparison using Java Statistical Classes (JSC)[2]

[1] http://nlp.stanford.edu/software/corenlp.shtml
[2] http://www.jsc.nildram.co.uk/index.htm

Results

- We achieved Cohen's κ of 0.42 for our classification problem.
- Better then the existing Neural Network system (Cohen's κ=0.31).
- Unigram baseline model achieved Cohen's κ of 0.33.

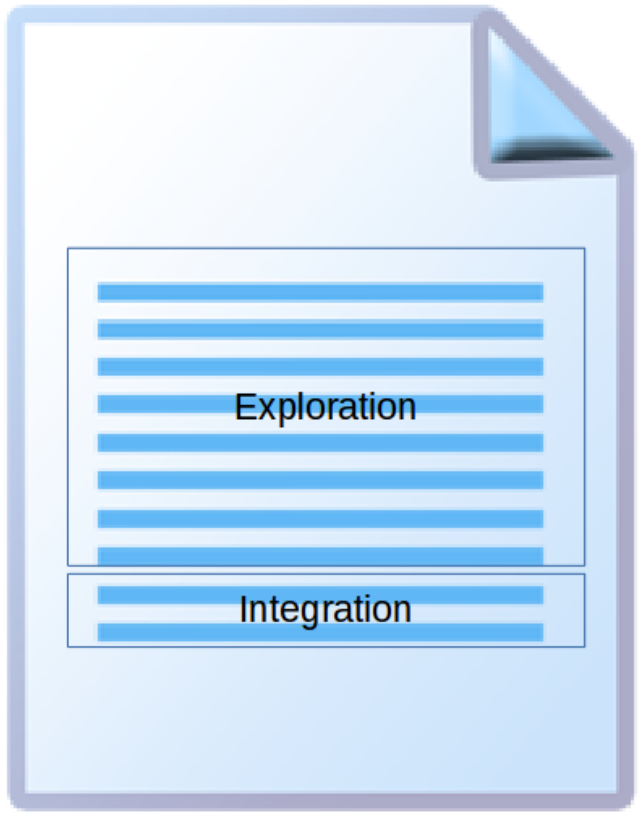
Error analysis:

	Predicted				
Actual	Other	Trigg.	Expl.	Integ.	Resol.
Other	**17**	04	05	02	00
Triggering	01	**42**	$\Rightarrow^{1}$ 14	03	01
Exploration	02	09	**98**	24	04
Integration	01	03	38	$\Leftarrow^{1,2}$ **56**	04
Resolution	00	00	03	15	$\Leftarrow^{2}$ **03**

Confusion Matrix

Challenges

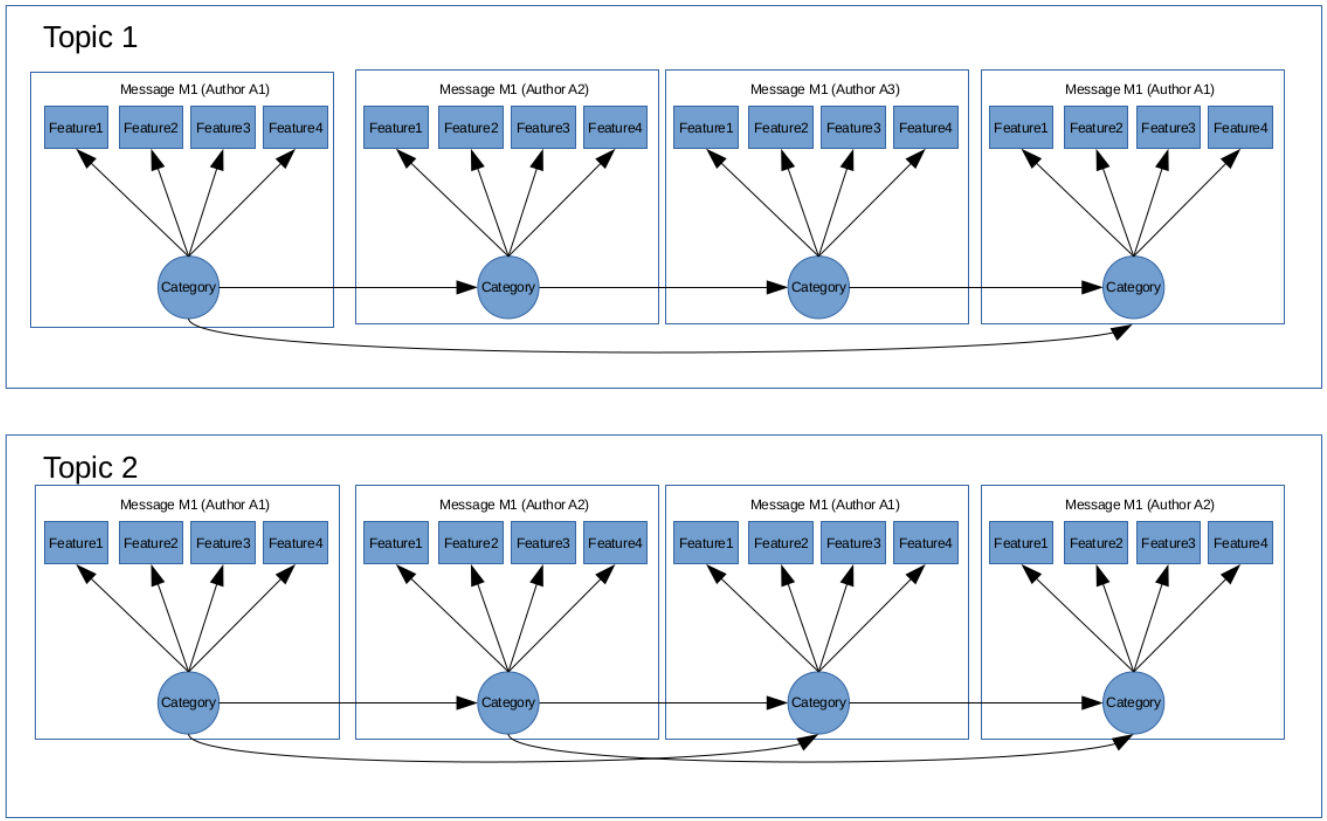
1. Effect of the large relative size of the exploration class,
2. Effect of the code-up rule for coding,
3. No relative importance of features, and
4. Context is not taken into the account.



Code-up rule for coding

In progress: making use of tread context

- Discussions (and students' learning) progresses from triggering to resolutions.
- Content of a message depends on the content of the previous messages.
- Content of a message depends on the learning progress of a given student.



Model for message classification

Approach: Hidden Markov models (HMMs) & Conditional random fields (CRFs)

- Hidden Markov Models:
  - HMMs used to models system states and their transitions in a variety of contexts.
  - Widely used, Bayesian Knowledge Tracing models based on HMMs.
  - Challenges with HMM:
    - Can this be modeled as HMM (2nd order HMMs?)
    - Dependency only on a single previous state,
    - One manifest variable for each state
- Conditional random fields:
  - Used for structured predictions (e.g., speech recognition)
    - For speech recognition, take into the account the classes of all letters in a word.
  - Widely used in natural language processing,
  - More flexible than HMMs,
  - Challenges with CRF:
    - Too many parameters to estimate with little data

Conclusions and future work

Summary:

- Promising path to explore,
- Use of backoff trigrams, plain and backoff dependency triplets, entity count and first message indicator seems useful,

Future work:

- Additional types of features which look at the context of previous messages (e.g., convergence vs. divergence),
- Moving away from SVM, explore other classification methods which are better at explanation
- Give associated probabilities for each classification,
- Give relative importance of different features.

Challenges:

- Challenges with message unit of analysis and surface-level features,
- Low frequency of resolution messages.

Thank you

Vitomir Kovanovic
v.kovanovic@ed.ac.uk

References I

Chang, Chih-Chung and Chih-Jen Lin (2011). "LIBSVM: A library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology* 2 (3), 27:1–27:27.

Donnelly, Roisin and John Gardner (2011). "Content analysis of computer conferencing transcripts". In: *Interactive Learning Environments* 19.4, pp. 303–315.

Garrison, D. Randy, Terry Anderson, and Walter Archer (1999). "Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education". In: *The Internet and Higher Education* 2.2–3, pp. 87–105.

Henri, France (1992). "Computer Conferencing and Content Analysis". en. In: *Collaborative Learning Through Computer Conferencing*, pp. 117–136.

Witten, Ian H., Eibe Frank, and Mark A. Hall (2011). *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*. 3rd ed. Morgan Kaufmann.